

INVESTIGATIONS INTO THE PATIENT VOICE
A MULTI-PERSPECTIVE ANALYSIS OF INFLAMMATION

by

SAMANTHA C PENDLETON

A thesis submitted to the University of Birmingham for the degree of
DOCTOR OF PHILOSOPHY

Institute of Cancer and Genomic Sciences

College of Medical and Dental Sciences

University of Birmingham

November 18, 2022

University of Birmingham Research Archive e-theses repository



This unpublished thesis/dissertation is under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence.

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:



Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

Unless otherwise stated, any material in this thesis/dissertation that is cited to a third-party source is not included in the terms of this licence. Please refer to the original source(s) for licencing conditions of any quotes, images or other material cited to a third party.

ABSTRACT

The patient is the expert of their medical journey and their experiences go largely unheard in clinical practice. Understanding the patient is important as bridging gaps in the medical domain enhances clinical knowledge, benefiting patient care in addition to improving quality of life. Valuable solutions to these problems lie at the intersection of Machine learning and sentiment analysis; through ontologies, semantic similarity, and clustering. In this thesis, I present challenges and solutions that explore patient quality of life pertaining to two inflammatory diseases: Uveitis and Inflammatory Bowel Disease, which are immune-mediated inflammatory diseases and often undifferentiated. This thesis explores how a patient's condition and inflammation influences their voice and quality of life via sentiment analysis, clustering, and semantic characterisations.

Methods

With guidance from domain experts and a foundation derived from clinical consensus documents, I created an application ontology, Ocular Immune-Mediated Inflammatory Diseases Ontology (OcIMIDo), which was enhanced with patient-preferred terms curated from online forum conversations, using a semi-automated statistical approach - with application of annotating term-frequency and sentiment analysis. Semantic similarity was explored using a preexisting embedding model derived from clinical letters to train other models consisting of patient-generated texts for systematic comparison of the clinician and patient voice. In a final experimental chapter, blood markers were clustered and analysed with their corresponding quantitative quality of life outcomes using patients in the UK Biobank with Inflammatory Bowel Disease.

Results

OcIMIDo is the first of its kind in ophthalmology and sentiment analysis revealed that first posts were more negative compared to replies. Systematic comparisons of embedding models revealed frequent misspellings from clinicians; use of abbreviations from patients; and patient priorities - models performed better when the clinical domain was extended with equivalent-sized, patient-generated data. Clusters unveiled insight into the presence of inflammatory stress and the relationship with happiness and the presence of a maternal smoking history with a Crohn's disease diagnosis.

Summary

Patient-preferred terms prove the patient voice provides meaningful text mining and fruitful sentiment analysis, revealing the role a forum plays on patients; semantic similarity highlighted potential novel disease associations and the patient lexicon; and clustering blood markers featured clusters presenting a relationship with sentiment.

In summary, this deeper knowledge of quality of life biomarkers through the patient voice can benefit the clinical domain and patient outcomes as understanding the patient can improve the clinical-patient relationship and communication standards: all benefiting the diagnosis process, developing treatment plans, and shortening these intensive time hauls in clinical practice.

Word count: **40,753**

DECLARATION

I declare that all work provided in this thesis is my, **Samantha C Pendleton**, own work and taken upon my own personal responsibility unless otherwise stated. The research was in fulfilment of a doctorate of philosophy in Clinical Informatics, under the supervision of **Georgios V Gkoutos**. I understand the consequences and the penalties in the results of plagiarism. I agree with my work being publicly available. I complied with the guidelines of the University of Birmingham.

The work in this thesis was funded by the **Medical Research Council** (MRC reference MR/S502431/1). Ethical clearance references were approved from the University of Birmingham Ethics Committee (ERN_20-0047 & ERN_21-1420) - see the University of Birmingham Ethics framework for further details. Various data obtained was publicly available, whilst specifically some data retrieved was from the other sources and will be stated accordingly in the thesis. Data from the UK Biobank received ethical approval from the research ethics committee (reference REC 11/NW/0382) and analyses conducted under UK Biobank application number 31224.

The computations described in this thesis were performed using the University of Birmingham's BlueBEAR HPC service, CaStLeS by Thompson, et al. (2019)¹. Providing a High Performance Computing service to the University's research community. See <http://www.birmingham.ac.uk/bear> for more details.

¹CaStLeS (Compute and Storage for the Life Sciences): a collection of compute and storage resources for supporting research at the University of Birmingham. 2019. Zenodo; Available: https://zenodo.org/record/3250616#.YP_kHS1Q1B1.

there is no single correct ontology for any domain

- Noy & McGuinness (2001)²

²Ontology Development 101: A Guide to Creating Your First Ontology. Stanford University; 2001. Available: https://protege.stanford.edu/publications/ontology_development/ontology101.pdf.

DEDICATION

to mum

ACKNOWLEDGEMENTS

This thesis provided opportunities to work with many experts across the University of Birmingham, the UK, the NHS, and even extended to other countries. Even though it is finished, it is not the end of this research. I owe gratitude to many people.

Firstly, George. George Gkoutos has been my supervisor for more than 3 years and I admire their work and dedication. George, I thank you for the support you provided, you have allowed me to grow into the ontologist, the philosopher, and data scientist I am. We were both a little stubborn regarding the genetic work, but I am extremely grateful that you allowed me to pursue my research ambitions to explore the patient voice. I wish to have the opportunity to work with you and your group in the future.

Secondly, two post-doctoral supervisors, Luke Slater and Andreas Karwath, both guided me throughout this process. Luke and Andreas taught me a paramount of information without which I could not have finished this work. Andreas, thank you for the thorough conversations concerning best Machine learning practices, I believe I have become a better communicator of my work. Luke, thank you for always showing your interest in ontologies and ontology accessories - that ambition was contagious and I aspire to dive deeper into NLP and ontologies in the future.

To both my second supervisor, Alastair Denniston, and mentor, Tasanee Braithwaite. Both clinical experts - including their colleagues - with whom I worked closely during the ocular inflammatory ontology project. Each provided the much-needed clinical interpretation to allow a fruitful thesis. Thank you Alastair for the support throughout my PhD. Thank you Tas for being a mentor: you provided much-needed clinical interpretation, we worked well as a team from two different domains, and created the first ontology of its kind in ophthalmology.

To all members of the Gkoutos group, I thank each and every one of you for the utmost support. You are all a foundation of this thesis through thought-provoking meetings and helpful guidance. I am grateful for all your patience and I hope we keep in touch.

To my family. Despite the hardships throughout the past few years, you were always on the other end of the phone, allowing me to express myself through the highs and lows - I miss you all deeply, and I hope we will be able to catch up on lost time in person. To my mum - the strongest woman I know - if you did not encourage me to pursue an independent life, my education, and a career; I would not be in this position. I owe you a lot.

To my friends, with a special mention to Frances & Tom:- my PhD was constantly reassured by your experiences. Knowing each owns' PhD experience opened my eyes and understood the process is similar yet different.

Finally, Sam. The night-owl, “prima donna” programmer, spent nights working on infrastructure for the UK’s COVID response. In contrast, I am the morning lark, so I want to thank you for not waking me at 4am. Sam, you are my backbone and kept me sane. You spent your evenings and weekends fixing my laptop when I trusted random code; you lent an ear for deep conversations when I wanted to express experiment plans; and you “begrudgingly” read through my thesis without holding back, despite my grammar issues. Without your support, guidance, and programming help, I would not have become the scientist I am - I greatly appreciate it. We are both on the next career expedition in our lives, so I look forward to that!

To all those involved, who empowered this work, thank you.

Special acknowledgements

I wish to thank the patient support charities, Ms Nikki Davies, on behalf of Olivia’s Vision <https://www.oliviasvision.org/>; Professor Stephen Foster, on behalf of The Ocular Immunology and Uveitis Foundation <https://uveitis.org/>; and Patient Platform Limited <https://patient.info/>. All kindly gave permission to access the public fora which were explored in this thesis. Additionally, the UK Biobank and members of the access team, including Mr Paul Flood, assisted greatly with understanding the data standards and mappings available. Finally, my funders, the Medical Research Council (MRC).

List of Abbreviations

W	Wilcoxon signed-rank test
mBR	<i>modified</i> BioReddit
$mBR + PI$	<i>modified</i> BioReddit + PatientINF
r	Pearson correlation coefficient
x^2	chi-squared
ANOVA	Analysis of variance
aRi	adjusted Rand index
BERT	Bidirectional Encoder Representations from Transformers
BFO	Basic Formal Ontology
BMI	body mass index
CB	ClinicalBERT
CB + PI	ClinicalBERT + PatientINF
CBC	complete blood count
CD	Crohn's disease
COID	Combined Ontology for Inflammatory Diseases
CRP	C-reactive protein
DOID	Disease Ontology
DR	dimension reduction
EHR	electronic health record
GE	gastroenteritis
GO	Gene Ontology
GP	general practitioner
HP	Human Phenotype Ontology
IBD	Inflammatory Bowel Disease
IBD-D	IBD patients diagnosed
IBD-D-FD	IBD patients diagnosed with future diagnosis
IBD-NYD	IBD patients not-yet diagnosed
IBS	irritable bowel syndrome
ICD	International Statistical Classification of Diseases
ICU	intensive care unit
IMID	immune-mediated inflammatory disease
IRI	Internationalized Resource Identifier
MIMIC-III	Medical Information Mart for Intensive Care III
ML	Machine learning

NCIT	National Cancer Institute Thesaurus
NER	named entity recognition
NHS	National Health Service
NLP	Natural language processing
NLR	neutrophil-to-lymphocyte ratio
OcIMIDo	Ocular Immune-Mediated Inflammatory Diseases Ontology
OR	odds ratio
ORDO	Orphanet Rare Disease Ontology
OV	Olivia's Vision
PATO	Phenotype and Trait Ontology
PCA	Principal component analysis
PCs	Principal components
PI	PatientINF
PIMP	permutation variable importance measure
PROMs	Patient Reported Outcome Measurements
QoL	quality of life
RCOphth	Royal College of Ophthalmology Uveitis Clinical Dataset
RO	Relation Ontology
ROC	receiver operating characteristic
SCDO	Sickle Cell Disease Ontology
SNOMED	Systemized Nomenclature of Medicine
SYMP	Symptom Ontology
t-SNE	t-distributed stochastic neighbor embedding
TF-IDF	term frequency-inverse document frequency
UBERON	Uberon Multi-species Anatomy Ontology
UC	ulcerative colitis
UKB	UK Biobank
UMLS	Unified Medical Language System
UVE	Uveitis Foundation
WBC	white blood cells
x-ref	database cross reference

Contents

	Page
1 Introduction	1
1.1 Symptoms vs. Signs	2
1.2 Quality of life	3
1.3 The Patient Voice	5
1.3.1 Online communities	11
1.4 Inflammation	16
1.4.1 Biomarkers	17
1.4.2 Immune-mediated diseases	19
1.4.2.1 Uveitis and ocular immune-mediated inflammatory diseases	20
1.4.2.2 Inflammatory bowel disease: autoimmune condition	22
1.5 Ontologies	27
1.5.1 Current biomedical ontologies	32
1.5.2 Medical terminologies	35
1.6 Machine learning	42
1.6.1 Unsupervised machine learning	44
1.6.2 Data preprocessing methods	47
1.6.3 Dimension reduction	48
1.6.4 Natural language processing	50
1.6.5 Semantic similarity and vectorisation	56
1.6.6 Machine learning clinical advances	63
1.7 Challenges and research questions	65
1.7.1 Lack of lay synonyms mappings to clinical terms	66
1.7.2 Lack of open embedding models for interpreting of the patient lexicon	69
1.7.3 Lack of understanding physiological markers adjacent to quality of life	70
1.8 Summary of contributions	72

2	The Ocular Immune-Mediated Inflammatory Diseases Ontology	74
2.1	Introduction	74
2.2	Data and methods	75
2.2.1	Building the ontology	75
2.2.2	Patient-preferred terms	77
2.2.2.1	Validation	78
2.2.2.2	Annotation	79
2.2.3	Sentiment analysis	79
2.3	Results	80
2.3.1	Description of the ontology	80
2.3.2	Enhancing ontology with lay terms	83
2.3.2.1	Validation	84
2.3.2.2	Annotation	85
2.3.3	Application into sentiment	86
2.4	Discussion	89
2.4.1	Implications	90
2.4.2	Limitations and future endeavours	91
2.4.3	Conclusions	94
3	The Patient Voice as a Clinical Narrative via Embeddings	96
3.1	Introduction	96
3.2	Data and methods	97
3.2.1	Development of an ontology for inflammation	97
3.2.2	Forum and creation of embeddings	99
3.2.2.1	ClinicalBERT	99
3.2.2.2	PatientINF	99
3.2.2.3	BioReddit	100
3.2.2.4	Design of two experiments	100
3.3	Results	102
3.3.1	Combined Ontology for Inflammatory Diseases summary	102
3.3.2	Existing embedding models and description of data	103
3.3.3	Experiment one: clinical and patient differences	104
3.3.4	Experiment two: patient and patient differences	118
3.4	Discussion	129

3.4.1	Implications	131
3.4.2	Limitations and future endeavours	133
3.4.3	Conclusions	134
4	A Biological Perspective into Inflammatory Bowel Disease and Sentiment	136
4.1	Introduction	136
4.2	Data and methods	138
4.2.1	Workflow	140
4.2.2	Features included	141
4.3	Results	143
4.3.1	Dimensionality reduction	144
4.3.2	Clustering	145
4.3.2.1	IBD patients diagnosed	145
4.3.2.2	IBD patients not-yet diagnosed	146
4.3.2.3	IBD patients with future diagnosis	146
4.3.3	Statistical analysis	147
4.3.3.1	IBD patients diagnosed	147
4.3.3.2	IBD patients not-yet diagnosed	150
4.3.3.3	IBD patients with future diagnosis	152
4.3.4	Supervised evaluation metrics	154
4.3.4.1	Permutation variable importance	154
4.3.4.2	Regression	154
4.4	Discussion	155
4.4.1	Implications	157
4.4.2	Limitations and future endeavours	159
4.4.3	Conclusions	161
5	Discussion	163
5.1	Implications	165
5.1.1	Patient-preferred terms	165
5.1.2	Sentiment analysis	167
5.2	Future endeavours and limitations	168
5.3	Conclusion and overall scientific impact	171
	List of Figures	173

List of Tables	175
List of Code	176
Bibliography	177
Appendices	189
A Supplementary material	190
B Ocular inflammatory diseases	192
C Embedding models	197
D Clustering IBD patients from the UK Biobank	202

1 INTRODUCTION

Globally, 3 out of 5 people suffer and die due to chronic inflammatory diseases [1]. Inflammation is the body’s first response to external threats and has a complex relationship with a wide range of diseases [2]. Inflammation is a fundamental part of a patient’s experience: inflammatory episodes can provide insight into the patient perspective, yet there is a need for understanding the patient voice [3–8].

Symptoms are disease manifestations, e.g. pain, that are experienced and expressed by a patient. Each symptom can be an important indicator for a diagnostic sign with the possible dismissal of a patient’s personal report of symptoms, because they are “subjective” [4], provokes misdiagnosis and negative patient-clinician experiences. While the patient and clinician can work together, they may have different priorities and judgements in the different stages of condition development [9]. Furthermore, there are communication gaps between the clinician and patient that are not limited to terminology [8].

The patient is the expert of their experience. Patients have valuable and deep knowledge of their conditions and the context of how these conditions affect their day-to-day lives. The relationship of inflammatory diseases, with phenotypes and biomarkers, adjacent to a patient’s emotional state and quality of life, can both allow a deeper understanding of the patient experience. Patient-generated responses, be that electronic questionnaire feedback, or manually-typed conversations, all can be explored yet such data is unstructured and lacking in research for secondary use.

Although a symptom/phenotype can be observed by the clinician, the patient has first-hand symptomatic experience [3, 4, 7]. In some cases this knowledge is inaccessible and clinical sources are scarce for secondary research.

1.1 Symptoms vs. Signs

The word “symptom” is derived from Greek and is primarily concerned with diseases [4]. On the other hand, “sign” is Latin and had no connection with disease [4]. Landre-Beauvais [10] (1813) and King [4] (1968) discussed that symptoms are changes in the body resulting from a disease and signs are symptoms that lead to inner knowledge. The descriptions of symptoms (e.g. headache), when understood by a clinician and merged with inner knowledge (e.g. imaging scans), become signs as they lead to the indication of something, e.g. a diagnosis [4, 10] (e.g. hypertension).

Signs may not be noticed directly or immediately by the clinician but reference present, past, or future “prognostic” events and are significant in medicine as medicine is largely based on signs [4, 10, 11]. Many investigations, such as blood tests can “point” to a cause of disease manifestation, observe current disease activity, or diagnosis a disease [4]. There are both “good” and “bad” signs, e.g. changes in blood tests can indicate remission.

Patients having the ability to assess their own symptoms has advantages in the medical domain, leading to improvements in patient care [4, 12, 13]. In some circumstances, patients have diagnosed their own condition. In 1968, King [4] stated that 20th century clinicians would immediately state that a symptom is subjective and the manifestation of a disease apparent only to the patient; whilst a sign is objective evidence and the manifestation of a disease that the physician perceives - this misunderstanding comes from the specialised knowledge clinicians have which a patient “does not” [4].

An early 20th century discovery of a new disease is an example of converting symptoms into signs [14]. A Norwegian mother noticed that her two children, who were both physically and mentally disabled, had a “peculiar and distressing” odour, associated with their urine (symptoms). The mother did not receive the help from doctors until she consulted Dr Asbjorn Foiling, who examined the urine, noting the green colour and unusual metabolites [14]. Dr Foiling decided to perform additional tests which identified the presence of phenylpyruvic acid in the urine (signs). With these further investigations

and knowledge, the symptom inference a new metabolic disease, phenylketonuria [14], and by Dr Foiling listening to the mother and conducting further investigations were the gaps between clinician and patient being overcome. Another case of self-diagnosis and self-assessment includes patient, Jill Viles, being able to diagnose their own condition as Emery–Dreifuss, a slow progressive chronic disorder. The patient received scepticism from clinicians but was able to present new discoveries in the biology and improved management for the disease [13], improving quality of life.

1.2 Quality of life

Patient quality of life (QoL) is important: it is valuable for patients to be included and have input on health-related decisions, leading to a better improved health and patient outcomes - both of which transform patient care [3, 7, 15–17]. Better QoL means a patient can be happier, more comfortable, and in some cases be able to resume normal life events in spite of their illness. Diseases and severity can affect QoL, specifically there is a relationship between QoL and chronic diseases: patients with multiple chronic diseases (multimorbidity) have a lower QoL [18].

Morbidity, comorbidity, and multimorbidity

Morbidity is the presence of a disease, (or symptoms of a disease), comorbidity is the presence of two morbidities, and multimorbidity is the presence of three or more morbidities [19].

In 2008, Uijen and van de Lisdonk [19] reported around 55% of patients aged 75 years or older have four or more chronic diseases. However in 2012, Barnett et al. [20] reported that around 40% UK and 50% US population have a long term condition; with 65% of those aged 65-84 years in the UK have multimorbidity. Barnett et al. [20] also reported that most people with multimorbidity are younger than 65 and those with multimorbidity are 3-7 times more likely to have a mental health disorder.

The relationship between QoL and a disease is important. In 1996 Porcelli et al. [21] performed a study involving patients with Inflammatory Bowel Disease (IBD) noting

that anxiety and depression scores were influenced over time by disease activity, and stated that many patients with IBD have the inability to explain their own emotions [21]. In 2001 another study conducted by de Rooy et al. [22] looked into both physical and psychosocial factors impacting a patient’s QoL: the most concerning responses were both physical (fatigue) and psychosocial (ability to achieve full potential) [22], but not relating to disease duration. Then in 2007, Fortin et al. [18] studied QoL and the relationship with multiple chronic diseases, stating that anatomical systems have a strong negative influence on QoL, such as gastrointestinal (digestive) conditions [18]. By 2006, Pizzi et al. [23] looked into disease severity of chronic comorbidities with IBD in respect to impact of QoL in patients. Authors reported predictors of QoL in IBD including heart disease comorbidities, age, and back/shoulder pain [23].

There are many domains covered in QoL and the QoL framework is used for assessing a patient’s life [24]. QoL can be interpreted differently, by both the clinician and the patient through professional judgements and patient experience [24, 25]. Felce [24] (1997) described 6 main categories of QoL with areas that can overlap with the other, but recent years means more developments and different medical domains have their own categories, such as Prem Senthil et al. [26] having a QoL item bank for an ocular inflammatory disorder, uveitis, that includes specifically “symptoms” as a category.

- physical (health, nutrition)
- material (privacy, food)
- social (relationships, support)
- productive (independence, job, education)
- emotional (mental health)
- civic well-being (privacy, responsibility)

A study in 2020 Prem Senthil et al. [26] looked into QoL and patients with anterior uveitis, an eye inflammatory disorder. The authors asked a series of questions provoking QoL domains to be covered then manually annotated concepts that patients mentioned on their own accord. Prem Senthil et al. [26] reported that patients described a variety of

symptoms due to their condition which affected their daily lives and tasks; for example, light sensitivity affecting their ability to use computers, drive, or read books. Patients also repeatedly described frustration due to prognostic uncertainty and associated discomfort, in addition to concerns of inflammation relapse. One of the main pieces of QoL aspects covered was medication: patients were concerned about the side effects of their long-term medications and additional medication-related inconveniences, e.g. ability to use eye drops affecting activity limitations [26].

Each patient experience is different and so QoL can be subjective. For example a decline in vision may be more important to those who value driving their car and so these QoL domains are perceived differently. Moreover, a person could be depressed but have a happier outlook on their physical health [24]. There are correlations between a patient’s health and happiness via life satisfaction: in the UK 38% of people who reported poor health had high levels of life satisfaction, and 20% of those who reported good health had low life satisfaction [27, 28]. Whilst an objective (e.g. disease) can affect the patient’s life, QoL is very subjective in that the patient has their own priorities so importance varies with each patient. To have a deeper understanding of the patient’s perspective, we must listen to the patient.

1.3 The Patient Voice

The patient experience is an untapped resource in research, being difficult to access or sparsely available, which their voice can provide insight into their medical journey [3, 6–8, 12, 15, 24, 29–32]. There is growing recognition in the last century of the patient voice, that in the literature it has been raised the importance of integrating the digital world into research and clinical practice [3, 7, 8, 11, 15, 29–42]. Specifically in 2000, Ferguson [6] brought the attention of the emerging world of the electronic patient “e-patient” and the digitising of medical practice. The concept of the e-patient plus the e-patient’s insight and knowledge into niche conditions can improve clinical practice and patient outcomes, better patient pathways, and treatment plans [6, 8, 29, 42].



Figure 1.1: “Patiently waiting for advice on health, last century” by Ferguson [6].

In 2013 deBronkart [17], dubbed as “Dave the e-patient”, described their journey of cancer treatments and the importance of their voice being heard. Dave was not aware they had cancer but only by going to a doctor and reporting fatigue, they had an x-ray showing a shadow in their lung. The radiology team quickly suggested a probable cause, but Dave’s doctors did not make a plan until further tests, including a biopsy, confirm the condition [17]. Dave deBronkart [17] reported their treatment plan was “unique” and because their doctor explained “negative” side effects, Dave immediately sought advice and help from online patient communities: three out of four patients with the same cancer never heard of the treatment but Dave stated the importance being able to choose treatment plans and communicating with other patients [17]. In conclusion, deBronkart [17] stated they were happier: the treatment was successful and so Dave the e-patient was able to continue on with life experiencing things which they could have missed.

Although Dave the e-patient and other patients may have thorough conversations with their doctor/clinician, the terminology gap between patients and clinicians is an obstacle in clinical practice [12, 43]. This gap has limited patient participation both in research studies and clinical phenotyping [44]. A clinician is trained for years to become an expert, yet patient self-phenotyping¹ has, in numerous cases, led to their own diagnosis [12, 13].

¹activity of assessing one’s own phenotypes.

A phenotype is an observable trait [45, 46]; in humans, it could be eye colour. In the context of this thesis, phenotypes are abnormal traits. Phenotypes change over time due to disease activity, manifestations possibly due to genetic sex [7, 45, 47], or external factors (e.g. environmental). Phenotypes often go unnoticed by a medical professional [11] which may be due to short appointments in general practitioner (GP) sessions or limited face-to-face meetings in hospitals. In the UK, the average GP consultation is 9.2 minutes (with GPs offering 10 minutes maximum unless a patient requests otherwise) and includes on average a discussion of 2 health-related problems [48] - these are shorter times on average compared to other economically-advanced nations. Patients with long-term conditions account for around 50% of all GP appointments [48].

The clinician and patient relationship

Clinicians need a good relationship with their patients in order to better understand phenotypes: symptomatic descriptions, communicate important information, and make treatment-related decisions [7, 8, 49]. Clinicians make decisions based on their professional judgements [24, 25] whether or not particular phenotypes are present. Mutual understanding and a positive relationship are important for both clinician and patient. An early diagnosis means early medical intervention, which can stop diseases from serious development. Though in some cases, a patient may prefer their own autonomy in treatment decisions and feel excluded if a clinician does not communicate options [7, 15, 16, 31, 50]. Patients want their concerns to be listened to [29, 43, 50]. Some patients have their own priorities or preferences that could involve the type of investigation for diagnosis, or disease activity and progression activity tests [51], in that there is the need for a shift in the way a patient is diagnosed to involve less invasive and uncomfortable procedures [51, 52].

Positive clinician-patient relationships benefit the health of the patient, being studied that patients' biological markers, such as blood pressure, reflect a healthier outcome with a positive relationship [7]. On the other hand, poor relationships can lead to dissatisfaction, and when patients are dissatisfied, they often seek support elsewhere [16, 50].

The clinician-patient relationship is particularly important for patients with chronic, life-long illnesses [7, 50, 51]. For these patients, frequent visits are common for the management of their condition: health updates, routine tests, disease activity investigations, and the discussions of treatments (e.g. any potential side effects). Some patients may feel that clinicians do not listen or believe them: clinicians can state blood test results have improved so the patient “must” feel better, however, the patient is the judge of their own feelings, symptoms, and QoL [3, 15, 24, 31, 32]. An example of this was discussed in 2021 by Beach [9], who interviewed patients who were told their cancer was stable, alongside their doctor’s opinion. Around 39% of patients were hesitant that stable meant “good news”, with a further 30% lacking enthusiasm or were non-verbal regarding the news, yet doctors were positive that stable was good news. This article is an important insight into a patient’s “disappointment” and how clinicians/patients manage news differently: revealing possible tension in the relationship [9].

Furthermore in a 2021 study with McKay et al. [51], patients with chronic liver disease were investigated whether or not they understood their progress and results, depending if they had a non-invasive liver test. The study concluded with stating that patients expressed both positive experience and preferences for non-invasive tests, with those having a better experience also an increased understanding of their results [51].

The clinician’s knowledge alone may not be enough to understand the patient [53] as studies have shown that preconceptions of some diseases do not align with patient experience. Clinicians may not understand patient-reported symptoms believing they are mistakenly attributed; or may not diagnose a condition as they believe a particular symptom should be present [53] (patients may report symptoms “differently”). Clinicians themselves have been “let down” by their own profession when experiencing the condition they specialise in [53], for example, a clinician specialising in rheumatology when diagnosed with rheumatoid arthritis discovered differences in their domain knowledge [53].

Patients themselves may not understand the realities of their condition [8, 9, 38, 51, 53], this could be due to inadequate communication or the patient lacking medical certainty

of important aspects of their condition, such as the effectiveness of medical interventions. There have been studies showing the importance of specialists being involved in patient care [16, 17] as they provide extensive knowledge and communication skills, compared to GPs. However, there is a lacking number of medical professionals to meet the needs of the patient [54]. A survey conducted in 1987 by Buller and Buller [16] asked patients about their satisfaction in meetings with clinicians, stating that most patients perceived their condition as “moderately serious” and patients were satisfied when communicating with more specialised professionals [16].

As medical knowledge advances, clinicians have changed the way they treat their patients: shifting to working with patients, rather than doing things “to” patients [7, 34]. Clinicians are making decisions and communicating better with patients because they are aware that patients have their own preferences, including their own health priorities or treatment plans [3, 7, 9, 17, 30, 32]. For example, by explaining and talking through possible treatment plans with the patient [7, 17], patients are taking a more active role in decisions: they value doctors’ opinions and by their doctor taking a co-operative role, patients are more receptive [7, 50]. Some patients continue to rely on their doctor’s knowledge, allowing doctors to lead treatment plans [8]. Some clinicians are listening to patients for longer [34] and using natural language (layman) to better communicate [33]. All benefit and improve the clinician-patient relationship.

Communication issues

Patients are familiar with lay terminology, whereas a clinician uses clinical terms. We can see the use of clinical terms in: clinical letters, electronic health records (EHR), and consensus documents [30, 55–57]. However, patients can have difficulty understanding clinical terms: medical jargon [11, 16, 31, 34, 38, 39]. Patients can develop their own vocabulary primarily consisting of layman or a mixture of lay and clinical [12]. Conversely, clinicians may be unfamiliar with patient-preferred terms and the gap between the clinician and patients has been widely acknowledged [11, 30, 31, 34, 39, 58].

In 2018, Staffieri et al. [59] conducted a questionnaire through social media advertisements to observe how non-medical individuals described leukocoria through photographs, a common sign for many eye diseases seen through a white reflection. The authors curated search terms of the participants and reported that individuals may have landed on the incorrect sites with wrong information due to the wrong use of search terms. However, the authors did also suggest that the study was at bias with the “Question-Behaviour Effect” that individuals may not have noticed anything wrong but the nature of the questionnaire prompted them to seek out information. Staffieri et al. [59] concluded that participants did not use clinical terms with the exception of “pupil”.

Patient satisfaction can rely on jargon used, control in medical decisions, and social norms to be happy with clinical care [11, 16, 31, 34, 38, 39, 43] - patients can feel overwhelmed from a lack of understanding. Although clinical letters or clinical notes are written and maintained by clinicians, they often lack deep information about the patient (e.g. their description of symptoms or emotional well-being) and rather they are the clinician’s professional judgement [25]. Furthermore, they can be full of verbose language, spelling errors, and duplicates in phrases (e.g. “copy and paste”) [25, 60, 61].

A study as early as 1968 by analysed 800 patient-doctor interviews and reported that 76% (mothers of) patients were satisfied with visits, however noted the main communication barriers included clinician failure to listen to concerns and the use of medical jargon [43]. Then in 2009, research looked at patient understanding of jargon: majority of patients understood clinical terms but used them in the wrong context: when participants were corrected, they had a negative response [39]. In another study on patient understanding of medical terminology, Pieterse et al. [58] (2013) stated only few participants understood all (cancer-related) terms, but most participants had understanding of majority of terms.

Patients may not mention specific details during the medical encounters for a variety of reasons, such as forgetting [25], thinking particular details are not relevant², or embarrassment. There are also social-economic [47] issues, in that patients may be concerned about their

²e.g. not mentioning abdominal pain specifically in the lower right (a sign of appendicitis).

safety when reporting their identity, even to their clinician. For example, patients with a menstrual cycle may be embarrassed to express information and can face being stigmatised due to the limited medical knowledge as many conditions are poorly understood [47]; for example, there is little known about endometriosis and can only be diagnosed via an invasive surgical procedure. Patients may face discrimination based on their age, sex, or race [47]. Some previous clinical studies had recommended “hospitals for violent white patients and prison for violent black patients” and care-management plans for patients were found to exclude black patients [47].

Women are faced with discrimination and judged more harshly on their weight and appearance compared to men [27], and children as young as 3 are aware of the stigma of weight [62]. A clinician with a negative bias toward heavy patients can lead to falsely attributing irrelevant phenotypes to weight, at the detriment of the patient leading to influence on their actions. Ul-Haq et al. [27] discuss the relationship between weight, self-reported health, and happiness in patients. In their study, of those who were highly obese were more likely to feel unhappy but the associations differed by sex - amongst women there was a significant association between unhappiness and all levels of obesity.

The degree to which patients understand the terminology used in medical encounters has been evaluated by various methods and across different disciplines. A study in 1983 by Spiro and Heidrich [38] discussed how there was “significant” confusion amongst patients in clinical settings irrespective of their age and background, although a positive correlation of education and knowledge was acknowledged. As mentioned, a poor clinician-patient relationship can lead to dissatisfaction and dissatisfied patients often seek support elsewhere [16, 29, 50], such as online communities.

1.3.1 Online communities

Online communities are a powerful source of the patient voice and over the decade patients are emerging online and exploring the internet for health-related advice [6, 17, 29, 41, 42, 50, 54, 63–68]. These patients are using social media or patient communities to express

their experiences, gain support, and seek advice/information [32, 69]. Having this online presence is easy: it is effortless to access, convenience of time and location, and patients can retain anonymity [3, 37, 50, 70]. Bowes et al. [50], Pletneva et al. [66]. It has been mentioned in the literature the importance of online patient communities and how they are valuable for research as they consist of patients with niche conditions and “impressive” knowledge into day-to-day problems [6] that a clinician may not be aware of.

Some patients may have difficulties using online sites [69], such as age or conditions (chronic) and so their carers may do it on their behalf. On the other hand, some patients used the internet for advice due to not wanting to bother their GP: time constraints or GPs showing a lack of engagement. However patients continued to value their GP’s advice if they were open to discussion and the online presence was not a replacement to their traditional health services [41, 50, 70]. Furthermore, studies have shown that some patients go online first before consulting with their doctor [63]. As mentioned, carers may use the internet for advice for the patient. A questionnaire by Pehora et al. [71] looked into how caregivers (of children) used the internet to seek advice and if these websites were reliable or safe for use. Of their results, 98% of participants use the internet to seek information about their child’s health, but only 20% used medical-reliable sites and only 50% checked the information with a clinician. The authors summarised by stating that caregivers will continue to use the internet in future for medical advice [71].

Issues with online text

Online patient-generated content: social media posts, blogs, and forum conversations, all add a new dimension to the data landscape and represent a rich, underutilised resource for advancing patient care [3, 54]. All types of patient-generated content consist of patient exchanges [67] and impressive niche information [6]. However, online text is relatively inaccessible and generally complicated to extract [55, 72].

Patient-generated data is unstructured and extraction of the patient voice can include

retrieving patient-preferred terms (layman). Online, unstructured, text provides a candid insight into the life of a patient, which is not otherwise captured and can involve discussions of personal medical experiences, e.g. medical therapies (including side effects) [3, 6, 11, 12, 15, 17, 30–32, 69]. The internet has become a popular source for information retrieval, approaches targeting patients [29, 41, 63, 64, 66], as medical advice is becoming more freely available and easily accessible [3, 29, 37, 59, 69–71].

The advice provided from these easily-accessible sites may be poor in quality: consisting of communication at a substandard level, biased writing, and inaccuracies or misinformation [6, 35–37]. Despite the UK government creating NHS Choices (<https://www.nhs.uk/>) for health advice [50], there still exists other online health sources that need to be improved with the correct and appropriate information to better help patients [6, 36, 37]. Misinformation can be persuasive and can lead to relationship changes between the patient and clinician [40, 41, 50], this can change the way clinicians treat their patients [6, 29, 34, 42]. For example, some clinicians feel devalued, anxious or are uncomfortable with the internet and online medical advice as some patients ask for specific treatments or believe incorrect information concerning their welfare [6, 29, 50]. On the other hand, this shift to online health is embraced by some clinicians and combined with traditional consultations: the two can complement each other for more informed discussions [42, 50]. In fact, patients state online communities are not a replacement for traditional doctor visits [70] and rather an opportunity to present an active role [50]. Highlighting a future where online health must be integrated into clinical practice.

Using these support forums provide patients with the ability to be more independent and make decisions based on what other patients (with the same condition) experience - it is a place for patients to meet, talk, support, and empower each other. Patients learn valuable information about coping with symptoms or medication side-effects [69]. Carers benefit too: carers can seek help and advice regarding legal support [68]. Patients are spending more time defining their own health plans, and are more willing to adopt positive changes when they have the knowledge and ability to manage their own health

[8, 15, 17, 29, 31, 41, 66, 73, 74]. A study in 2008 by Iverson et al. [41] stated that 50% of participants use the internet to seek health-related information and change their opinion based on that information, such as dietary changes.

These online communities and health-related sites give us insight into patients' condition management due to advice they take: taking advice from others (e.g. other patients) not included in their health management (e.g. healthcare providers) can explain health-related issues [17], e.g. a clinician understanding why a patient is avoiding a particular medication because they read online about specific side effects (perhaps that other patients experienced); or explanations for morbidities and QoL. Online sources are mitigating the National Health Service (NHS) shortage crisis of health services and professionals [75]. Online communities and resources are adaptable to new information and easily accessible for some patients who are unable to access adequate care [75].

Patient Reported Outcome Measurements

Patient communities are a perspective into Patient Reported Outcome Measurements (PROMs) [6, 17], which commonly appear in the form of electronic-based questionnaires, but can also be retrieved from focus groups [3, 73, 76]. PROMs, being another form of patient-generated information, are important as they also represent patient experiences without clinical interpretation or intervention [73, 76] to better understand a patient's medical journey, quality of care, and how to better medical environments.

All of these types of invaluable patient voice resources can yield important qualitative data, which leads to the patient being heard [3, 32, 76]. PROMs from online communities gives us insight into patient conditions and management, the advice from others, and QoL [15, 31, 73], addressing gaps in clinician knowledge.

One study of QoL via online health forums presented an alternative analysis based on auto-questionnaires [67]. However, these questionnaires are biased as they focused on topics the clinician is already aware of and so may not capture information unprompted

from online forums, highlighting new topics arising in PROMs should not be overlooked. These rare occurrence is an interesting result itself since it suggests that these “irrelevant” topics are important to the patient: they play a role, minor or not, in the actualities of a patient’s life. However, accessing and extracting text from online communities is difficult as text in online forums are largely unstructured: unlabelled and noisy. Responding to these challenges involves calling for adapted preprocessing and query methods [42, 67, 77].

The patient voice is vital in understanding the sort of terms they use and QoL [6, 17]. Online communities provide an environment for people to discuss their conditions with other patients suffering from a similar condition and these patients can be considered “allies” for clinicians by contributing to the medical domain with their particular information.

Ferguson [6] mentioned the importance to remember how patient communities may not be accurate due to lack of clinician contribution, yet a patient’s own knowledge of their condition is impressive and these online communities can appreciate both clinical and patient opinion. Ferguson [6] discussed with many clinicians regarding this “new type of patient” who use the internet to research their condition. One response from clinicians was that they are uncomfortable with the internet as some patients ask for particular treatments, or gained incorrect information.

On the other hand, a clinician stated that online patient services is where health care is heading and how a century ago clinicians had doubts about telephone appointments; and another clinician specifically stated the internet was “the greatest untapped resource in health care is the consumer”. Finally, it was brought up that years ago clinicians felt that parents had no business being involved in their child’s medical care, but it soon became apparent that well-informed parents can take much better care of their child. Moreover, the internet is providing better encouragement and support for adults, and clinicians are observing their patients being more helpful and responsible.

1.4 Inflammation

Inflammation is a complex, immediate body response to cell injury [2] with thousands of diseases having inflammatory properties [5]. The term “inflammation” is derived from the Latin “inflammare” - “to set on fire” [78]. Healthcare has been largely shaped by organ-based anatomy since 150-200 A.D. by the Greek physician and philosopher Galen [79]. Many inflammatory disorders have the suffix “-itis”, meaning “inflammation of”, such as appendicitis, which is inflammation of the appendix (organ). Though this does not apply to all inflammatory diseases, such as asthma or Crohn’s disease (a type of IBD).

Inflammation can be broadly separated into two types, acute and chronic [1, 2, 80–83]. There are five fundamental clinical signs of both acute and chronic inflammation: heat, redness, swelling, pain, and loss of function [2, 78] - these are usually described by the patient, but not limited to observation by the clinician. The first four were derived from Roman encyclopaedist, Cornelius Celsus (25 B.C. - 50 A.D.) [78], which was then expanded to include loss of function by Physician Rudolf Virchow (1821-1902). Acute inflammation is the instantaneous response of the immune system to cell injuries lasting no more than six weeks [2] - this tissue damage can be a result of microbial invasion (infection), toxins, or injury (trauma) [1, 2, 80]. Some acute inflammatory disorders, such as acute pancreatitis, can be a result of gene mutation [2].

Prolonged acute inflammation or continuous tissue stress can lead to chronic inflammation [81–83]. Chronic inflammation is more persistent and in some cases, it can be permanent. Chronic inflammation can occur due to the extent of the trauma or the inability to repair the damaged tissues [1, 83]. Other factors can promote chronic inflammation, including natural ageing, environmental, and lifestyle: smoking or obesity (e.g. high body mass index (BMI)) [1, 84, 85]. Chronic inflammation can progress silently, and without treatment/prevention, can lead to the development of chronic diseases such as IBD, Alzheimer’s disease, diabetes and cancer [1, 81, 84] - chronic inflammatory diseases are the most significant causes of mortality in the world [1].

Treatments for inflammation vary by the type and causes, e.g. therapeutic interventions or lifestyle changes. In acute inflammation, bacteria can be fought with antibiotics. As chronic inflammation can be life-long, treatments can alleviate inflammation [1], despite not curing them. Patients with a chronic inflammatory condition may undergo intensive treatments, such as medical or surgical therapies, but like all chronic conditions, patients can have acute inflammation re-emergence of the condition, known as recurrent (flare-up).

1.4.1 Biomarkers

Biological markers (often shortened to “biomarkers”) refer to clinical signs (or phenotypes) [86]. Biomarkers are important in clinical practice for diagnosis, observation of disease activity, knowledge of treatments, and understanding physiology [86]. In clinical practice, the use of biomarkers is commonplace and the presence of biomarkers is accepted without question [86] whilst research on phenotypes is broad [45]. Some examples of biomarkers include initial measurements: age, BMI, and blood pressure; and more complex tests: blood test of complete blood count (CBC) or biopsies (tissue samples) [86]. Biological (e.g. age, genetic sex) and external factors (e.g. lifestyle behaviours) can have a large effect on biomarkers [87].

Each biomarker plays a different role in a disease and our knowledge of biomarkers has stemmed from various sources, including case studies, literature, and clinical trials: biomarkers predicting a clinical outcome [86]. Biomarkers are not clinical outcomes and rather biomarkers can indicate outcomes in prediction studies, e.g. biomarker of BMI with endpoint as survival [86], higher BMI can mean higher mortality.

Clinical outcomes

Many clinical trials aim to predict a clinical outcome. In some prediction studies, biomarkers have been used as clinical outcomes. However, it is important to note they are not replacements - biomarkers can be used as a “surrogate” endpoint if there is strong evidence contributing to the outcome (e.g. strong correlation) [86].

Where inflammation is present, the production of leukocytes in the blood is triggered in order to repair the affected tissues [83, 84, 88, 89]. Leukocytes are known as white blood

cells (WBC), and they consist of basophils, eosinophils, lymphocytes, monocytes, and neutrophils [78, 88, 90, 91] - WBC and neutrophils have a strong correlation [88]. Welsh et al. [88] (2018) looked at WBC and leukocytes (e.g. lymphocytes and neutrophils) associations with inflammation and cardiovascular diseases: they reported men with higher neutrophil counts were at a higher risk of fatal and nonfatal cardiovascular deaths. Additionally, patients with higher WBC were older, higher systolic blood pressure, and had a higher BMI, plus more likely to be a smoker [88].

There are various blood biomarkers that can determine acute or chronic inflammation [2]. In acute inflammation, WBCs, predominantly neutrophils plus monocytes, clear out infected and damaged tissues to initiate repair [1, 2, 78, 89, 90], see Figure 1.2. In clinical practice, WBCs are used to measure acute inflammation, immune response (presence of infection), and can also reflect poor health plus the risk of death [88], although neutrophils initiate repair, they can also cause tissue damage [90].

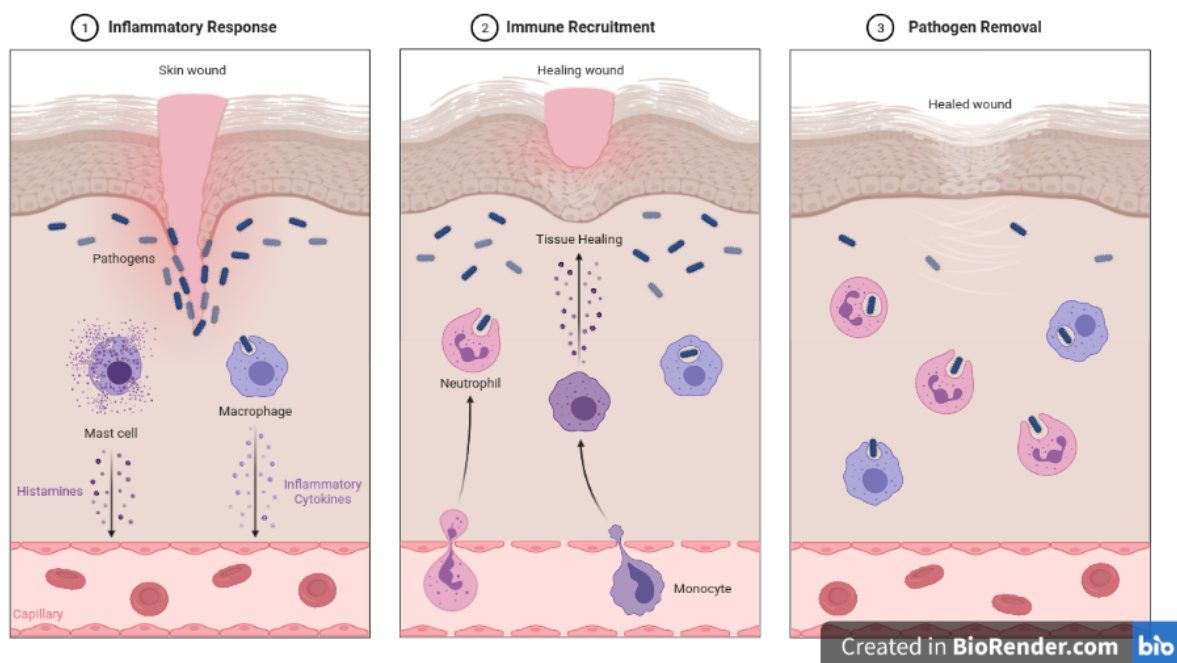


Figure 1.2: Acute inflammatory response in anatomy. In the diagram, (1) the inflammation is caused by a wound in the skin, (2) neutrophils and monocytes emerge and initiate tissue repair, and finally (3) the wound has healed. Adapted from BioRender template [92].

The transition to chronic inflammation includes WBCs in acute inflammation changing: the neutrophils are replaced predominantly by lymphocytes and macrophages (such as basophil or eosinophil) [1, 78] - the presence of lymphocytes in the blood acts as the main biomarker for diagnosis of chronic inflammation. The neutrophil-to-lymphocyte ratio (NLR) indicates inflammatory stress in the body [93].

Other blood biomarkers indicate the presence of inflammation including the increased production of uric acid (urate) crystals, circulating platelet aggregation, elevated C-reactive protein (CRP) [1, 80], and a decrease in protein biochemistry albumin [94]. Other blood biomarkers, such as creatinine, rheumatoid factor, and CBC (e.g. platelet count) can rule out other diagnoses [1]. An early study in 1995 by Beattie et al. [95], explored potential IBD biomarkers and concluded that elevated CRP levels were the best biological parameter to differentiate those with IBD and without IBD. Then in 2004, Vermeire et al. [96] discussed that increased CRP in patients with IBD indicated severe or uncontrollable inflammation in the gut and highlighted that one of the best blood biomarkers for diagnosing IBD was CRP.

1.4.2 Immune-mediated diseases

Immune-mediated diseases are distinguished by the body's immune system attacking itself [97] via over or under-reacting. The immune system's main goal is protecting against infections, viruses, disease pathogens, and tumour cells. Some immune-mediated diseases are characterised by excessive activation of certain WBCs called mast cells and basophils by a type of antibody known as IgE, resulting in an extreme inflammatory response, immune-mediated inflammatory disease (IMID). Autoimmune diseases are a subset of IMID [97]. The reason why the immune system starts to attack itself is unknown, but research has shown that there are some possible contributing factors, for example, genes (genes: sex, ethnicity, and heredity) and phenotypes (environment: allergen, pollution, or diet) [97]. These factors do not contribute in the same way [97]. For example, some diseases are caused by genetics more or less than lifestyle factors. Some IMIDs can be

caused by infections (e.g. viral HIV) or allergies (e.g. rhinitis) and are not specifically immune-related [97], though eventually affects the immune system.

1.4.2.1 *Uveitis and ocular immune-mediated inflammatory diseases*

Uveitis is inflammation of the uvea [26] and is an ocular IMID. The anatomy of the eye is presented in Figure 1.3 illustrating the uvea (commonly known as the uveal tract) consists of the iris, ciliary muscle/body, choroid and is a part of the middle layer of the eye [26]. Some other notable ocular IMIDs include scleritis³, optic neuritis⁴, blepharitis⁵, and well-known conjunctivitis⁶.

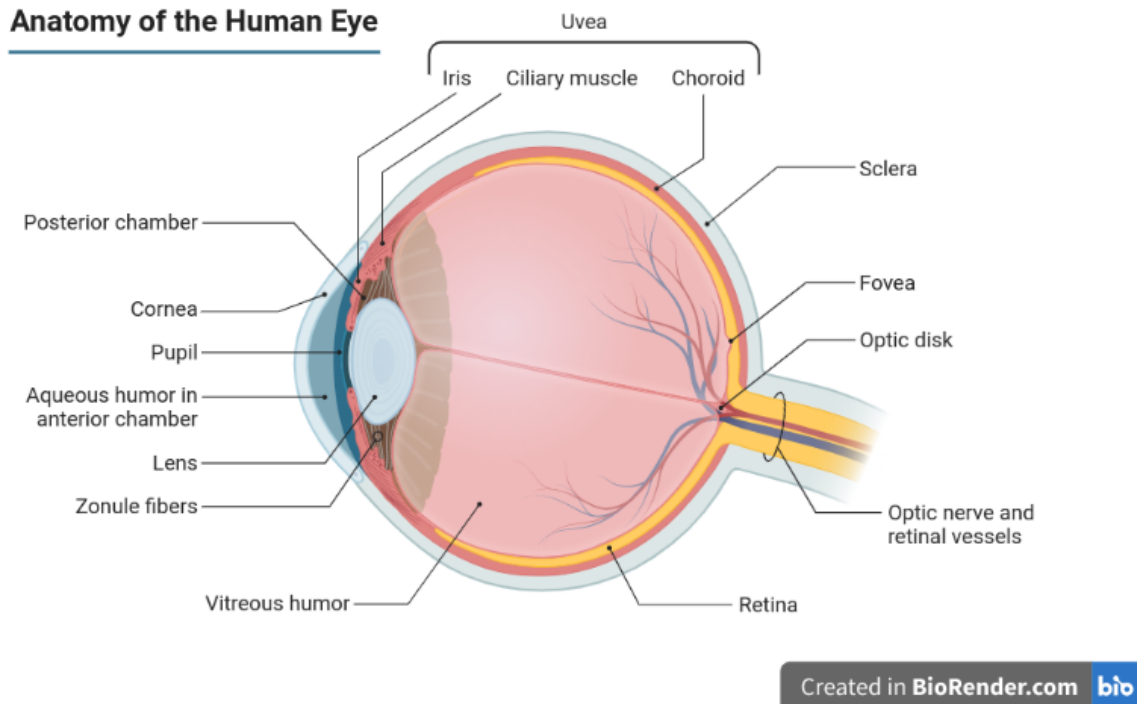


Figure 1.3: The anatomy of the human eye, highlighting the uvea, the middle layer of the eye, which consists of the iris, ciliary muscle, and the choroid. Uveitis can occur in the anterior, posterior, or both chambers.

Adapted from BioRender template [92].

³occurs in the sclera, the white layer of the eye.

⁴occurs in the optic nerve, located at the back of the eye

⁵occurs in eyelids.

⁶occurs in the conjunctiva, which is the connective tissue between eyelid and eyeball.

Ocular IMIDs and phenotypes are important as they are life-threatening. Uveitis is one of the top causes of blindness [3, 26] which adversely impacts the QoL as loss of sight is one of the most feared impacts of eye inflammatory disorders [3]. Ocular IMIDs can be acute, chronic, or recurrent (as in they can acute inflammatory episodes) and are present throughout the life course [98]. Additionally, these disorders have associations with multi-organ immune-mediated inflammatory diseases of infectious and autoimmune aetiology [26, 99, 100]. Uveitis can occur in people who currently have an autoimmune disease, as the immune system starts to become over-active in the visual system; but also can be caused by infections [26, 101], such as toxoplasmosis⁷. Other ocular IMIDs, such as blepharitis and conjunctivitis, are most commonly caused by allergies. Uveitis and other ocular IMIDs are not thought to be hereditary, however, the gene HLA-B27 is linked with uveitis and is found in patients with some autoimmune conditions, including IBD [102]. Studies estimate between 30-60% of patients with uveitis are “undifferentiated”: meaning no identifiable cause [26, 101].

Uveitis, and other ocular IMIDs, can cause health complications, including cataracts (cloudy vision). There are various subtypes of uveitis [26, 103]. Anterior uveitis, the most common, occurs in the front part of the uveal tract (in the anterior chamber), with subtype iridocyclitis, causing redness of the eye [26, 104]. Uveitis in the middle uveal tract of the eye is known as intermediate uveitis, with subtype pars planitis, which can cause blurred vision [26, 105]. Uveitis at the back of the uveal tract (in the posterior chamber) is posterior uveitis, which can cause more severe vision problems [26, 105], such as “floaters”. Uveitis which occurs in both anterior and posterior chambers is known as panuveitis [26].

The treatments for uveitis and ocular IMIDs depend on where inflammation is occurring and the severity. Many conditions require systemic immunosuppressive or immunomodulatory therapies [26, 98, 106]. Some treatments include corticosteroids (steroids) in the form of oral, or topical (eye drops or ointments) medicines. If the inflammation is caused by

⁷a disease from a parasitic infection: *Toxoplasma gondii*, a very common parasite.

an infection, then antibiotics (bacterial), antiviral (virus), or antifungal (fungal) can be provided [26, 106]. Just as conditions can cause symptoms, medical therapies can cause side effects, such as headaches.

1.4.2.2 *Inflammatory bowel disease: autoimmune condition*

Inflammatory Bowel Disease (IBD) is characterised as a chronic, non-infectious, autoimmune, digestive, inflammatory disease [107–110], divided into two main conditions: Crohn’s disease (CD) and ulcerative colitis (UC), but often with exception of recurrent gastroenteritis (GE). IBD involves inflammation and subsequent damage to the gastrointestinal tract [107], leading to many complications including issues with the liver or gallbladder issues, and cancers [102, 110–112]. IBD is commonly diagnosed between the ages of 15 and 40; UC is more common in men and primarily damages the surface of the colon (large intestine); on the other hand CD is more common in women and can damage any part - from mouth to anus - of the digestive system [107, 113].

The aetiology of IBD is presented in Figure 1.4 in that IBD occurs in the digestive tract: in normal anatomy, the intestines have a mucus layer that prevents bacteria from causing infections. However in a person with IBD, the immune system starts to attack itself and there is loss of the mucus layer so bacteria permeability is increased - leading to inflammation of the gastrointestinal tract and infections caused by *Clostridium difficile*, resulting in diarrhoea and UC. Presence of bacteria and genetics can both contribute to the trigger of an abnormal immune response, but being autoimmune, there is no known reason as to why the immune system starts to attack itself [97].

Heredity, bacteria, diet, smoking, and antibiotic use can all contribute to triggering the abnormal immune response, leading to IBD [107, 108, 114]. Other factors have been known to heavily outweigh the genetic factors in IBD [109, 113, 114], such as the higher consumption of meat [113] and those who smoke are twice as likely to get CD than non-smokers. In 2009, a group looked into the relationships between smoking and IBD diagnosis: majority of patients who smoked had CD and children acquired IBD earlier

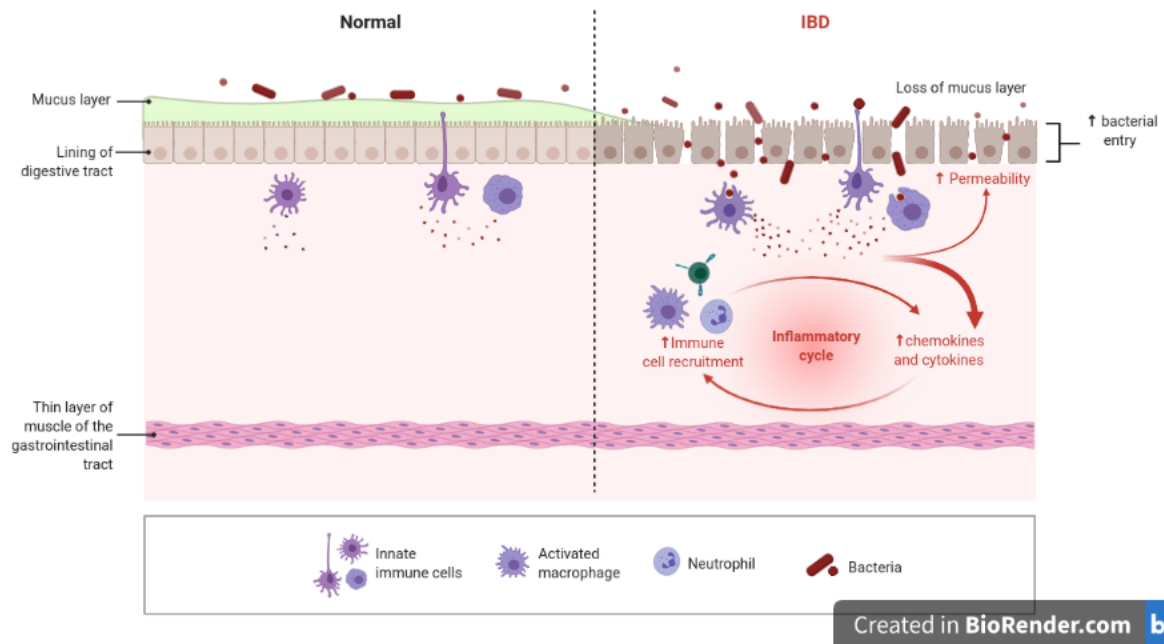


Figure 1.4: Immune response in Inflammatory Bowel Disease (IBD). In IBD, the immune system starts to attack itself leading to loss of the intestine’s protective mucus layer, leading to increased bacteria permeability and inflammation of the gastrointestinal tract. Adapted from BioRender template [92].

in life compared to their parents, suggesting genetic anticipation and the influences of environmental factors in families [115].

Studies have documented that a patient’s QoL is severely affected by IBD specifically those with IBD have poorer QoL compared to the general population [22, 23, 65, 108, 116–118]. We lack a deep understanding of how life is more challenging [22, 23, 108, 117] yet it is researched that IBD patients are more likely to report increased levels of depression with increase disease activity [21, 23, 108, 119], specifically in those with UC.

IBD is chronic but also recurring: flare-up episodes of inflammation [108], which are eased with dietary changes or medication: in 2021, Kikut et al. [114] highlighted the importance of dietitian being involved in the care of patients with IBD. These therapeutic interventions need to be modified with each separate case of IBD in a patient, as each patient is different [52]. Many patients follow elimination diets after remission due to fears of another flare-up, and their new relationship with food affects their QoL [65, 114, 116]. Although lifestyle changes can relieve inflammation/symptoms and is

a key component of care, some dietary changes can cause deficiencies, over-consumption, or impaired absorption [113, 114]. Zallot et al. [116] (2013) conducted a questionnaire, focused on the relationship or impact IBD and diet has on QoL: authors reported 16% of patients believed that diet caused their diagnosis, 58% believed that diet can cause a relapse, and 67% said they avoided certain foods to prevent a relapse (with 25% having a normal diet outside disease activity) and specifically 20% stated their diet post diagnosis affected their social life [116]. The authors concluded their report stating that patients' QoL was affected, specifically social life: participants avoided certain foods, outdoor dining, not eating the same foods as family members; as they lived in fear of a relapse [116]. In 2020 Bischoff et al. [113] released guidelines for clinicians with advice for providing the best dietary care to patients with IBD, specifically mentioning that patients can benefit from an active role. One such suggestion included observations into energy intake and malnutrition levels [113] as low iron can result in anaemia (common in patients with IBD).

There are numerous common symptoms of IBD, including abdominal pain, swelling, high temperature, recurring or bloody diarrhoea [107, 108, 117, 120] - all "fundamental signs of inflammation" [2], with the exception of the loss of function, though patients do report the loss of bowel control [108]. Bloody diarrhoea is more common in patients with UC [107]. Patients also report the burden of fatigue [120].

The European Federation of Crohn's and Colitis Associations patient survey [117] in 2007 reported that studies fail to capture the true experiences of IBD specifically symptoms, medications, and QoL. Ghosh and Mitchell [117] reported that approximately 75% patients were satisfied with the response from their current medical therapies, more than 86% stating they would rather try a new type of drug than endure surgical therapies, and 75% of patients mentioned their symptoms affected the ability to enjoy leisure activities [117]. However, only 39% of patients said their doctors discussed new medical therapies, and nearly half reported that their doctor does not ask about their QoL [117].

Different surgical interventions treat the different types of IBD since they affect different anatomical sections [107, 113]. In patients with UC surgery to remove parts of the colon

(colectomy) - or the whole colon in place of an opening in the body (ileostomy) - may be performed; in patients with CD surgery is less intense and usually fixes are done to the small intestine.

Investigations and surgical therapies^a

- colonoscopy - visual test to check inside bowels via the anus,
- endoscopy - visual test to check small intestines via the mouth,
- colectomy - surgery to remove part of the colon,
- ileostomy - small intestine is made into a stoma (opening into the body) through the abdomen.

^aNote: -scopy is the suffix for observation investigations and -tomy is the suffix for surgical removal.

Although there are different types of IBD, in 1999 Kim et al. [119] conducted a questionnaire for patients to investigate the “worse type” of IBD. Participants scored their QoL higher at the beginning of their disease manifestations. Kim et al. [119] concluded that investigations are needed to observe the relationship of QoL with demographic parameters; such as age, sex, and duration of disease.

It can be difficult to distinguish the type of IBD in a patient without several diagnostic methods. It is important to note that while there is no single test that can definitively diagnose IBD but multiple tests can evaluate and lead to a diagnosis. Stool samples are one initial test, which are analysed for blood or the presence of bacteria: indicating IBD, and more likely UC [52, 94].

Blood tests are also one of the first investigations conducted for the diagnosis development of patients [52]. A complete blood count (CBC) is done to test for inflammatory biomarkers but also to rule out other diagnoses, such as changes in platelet count [1], creatinine, and rheumatoid factor. Blood biomarkers, such as urate and CRP are present in acute inflammation and can indicate IBD [1, 52, 94–96, 111], in addition to WBC leukocytes, platelets, albumin (these are not as extensively studied as CRP), and red blood cells (for signs of anaemia), which is common in IBD patients as IBD can lead to bleeding. As IBD can lead to liver issues, both bilirubin and albumin abnormalities can indicate problems [111]. CRP is the most-telling blood biomarker for IBD, and levels at 10-40mg/L

are indicative of inflammation [94] with abnormally high $>45\text{mg/L}$ indicates severe or ongoing inflammation in the gut [96] and the possible need for surgical intervention. CRP at extremely high levels, $>200\text{mg/L}$ are severe [94].

CRP sharply decreases after inflammation is resolved and is correlated with disease activity in IBD, specifically CD [94] - plus helps distinguish from irritable bowel syndrome (IBS). IBS can be misdiagnosed as IBD due to very similar symptoms [121]. In 2017, a group looked at stool samples from patients with IBD to observe similarities with IBS symptoms [121]. In their study, authors reported that a quarter of patients displayed symptoms of IBS and found signs of mucosal inflammation; additionally patients with IBD who reported symptoms overlapping with IBS had a lower QoL [121].

Irritable bowel syndrome (IBS)

IBS is a similar condition to IBD in that both occur in the gut. However, they are not the same. IBD can be misdiagnosed as IBS as symptoms are similar: bloating (similar to swelling), diarrhoea, and abdominal pain.

After initial tests are done and IBD is a possible concern (with other diagnoses ruled out), then further tests are performed for the definite diagnosis of IBD and the subtype. These investigations include antibody tests or biopsies [52]. Biopsies are body samples from a colonoscopy or endoscopy and these assessments are the most accurate in diagnosing IBD as well as observing disease activity. Both stool samples and biopsies are disliked by the patient due to being uncomfortable and invasive, whilst blood tests are a popular alternative as they are less invasive [51, 52]. More research into additional blood biomarkers could avoid expensive medical tests, invasive procedures, and improve the diagnosis experience.

A study in 2010 demonstrated using computational techniques and the possibilities of blood metabolites (blood biochemistry) including measurements of lymphocytes and urine measurements as a diagnostic tool in active and quiescent IBD [122]; and in 2013 Zhang et al. [123] indicated metabolites are possibly non-invasive diagnostic tool for UC. Furthermore, in 2018 Probert et al. [52] investigated the use of blood markers as

alternatives for endoscopic assessments of UC, however their study consisted of a small cohort ($n=40$). Probert et al. [52] reported that metabolites provided potential to provide possibly alternatives to endoscopic assessments [52].

All of these investigations are not only for diagnosis but also conducted throughout a patient's life as clinicians may want to assess and monitor a patient's IBD disease activity, e.g. investigate a flare-up or new symptoms. Whilst some patients can lead relatively normal lives, the complications of IBD can be debilitating and QoL continues to be affected after the inflammatory episode [21–23, 108, 117–119]. A patient can experience physical pain, but also mental and emotional consequences [21–23, 108, 117, 118], such as employment issues, diet affecting social relationships, and constant fatigue [22, 23, 65, 116, 117]. All of these issues are not always apparent. PROMs or questionnaires focuses on complications or IBD symptoms, yet not about how they can impact well-being [65, 108, 116, 117].

1.5 Ontologies

Ontologies represent a domain of knowledge, a *thing*, providing domain vocabulary [124–128] in a condensed, structured, and organised manner: assembled as a network of information. The idea of categorisation dates back to 400 B.C. when philosopher Aristotle, expressed the importance of classification [127], introducing the notion of “species” for categorisation. Greek philosopher Porphyry (dated around 300 A.D.) expanded upon Aristotle's “Categories” to draw what might qualify as the first semantic network [127]. The computational ontology was explored only recently in 20th century due to advancements of computers [127]. Throughout this thesis, “ontology” refers to computational ontology.

Modern ontologies are inspired by the “semantic web” from the computing domain [129, 130]. Ontologies allow for enhanced data integration, text mining efforts, annotation, semantic similarity, and unifying knowledge across domains [125, 131]. Semantic similarity is described as assessing the relatedness between two entities by the similarity in meaning [132], and will be described in more detail in section 1.6.5 (page 56).

The **three fundamental concepts** of ontologies: **classes**, **relationships**, and **metadata** [125, 130]. A working example of these concepts are illustrated in Figure 1.5.

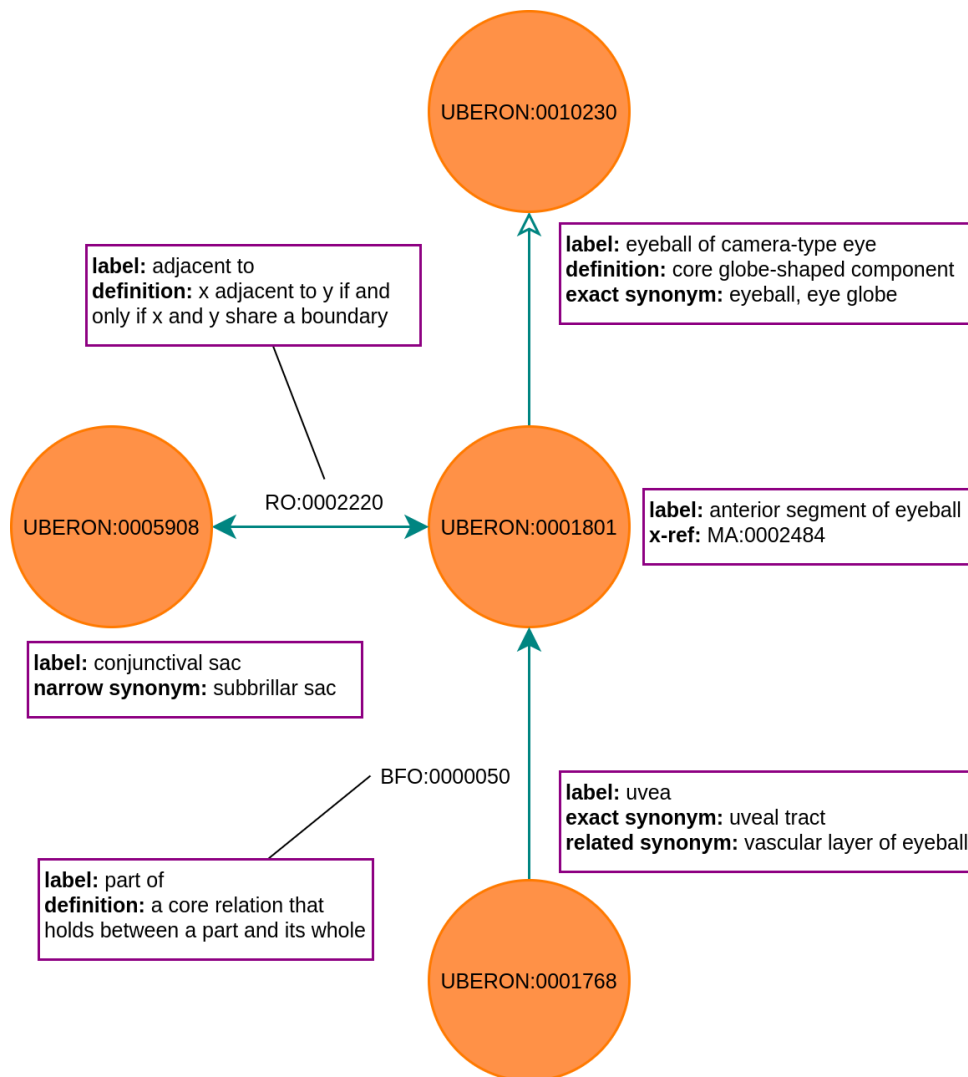


Figure 1.5: Part of the cross-species anatomy ontology showing various concepts of an ontology. **Classes** are circles with IRI labelled in them. Undefined **relationships** are blank arrows whilst filled arrows are defined axioms. Boxes are the **metadata** for each concept, which includes the label, definition, x-ref, and types of synonyms.

Classes represent an entity in the domain [125], such as **uvea** in anatomy. Classes are organised in a hierarchy: superclasses can be further defined by subclasses [127]. Each class is uniquely identified with an Internationalized Resource Identifier (IRI): a web link that allows more characters (e.g. unicode) and for ontologies to point to where the definition of the class is stored. The IRI is followed by a prefix string (usually an abbreviation of ontology) and a set of digits [124, 125] - IRI builds on top of URI.

Relationships link **classes** together [125]. Ontology classes with defined relationships provide a deeper meaning to an ontology [125, 127]. **Subclasses** have natural **relationships** linking them together, however we can also explicitly define relationships. A commonly used relationship is “**part of**” - such as **choroid part of uvea** (linking back to Figure 1.3, page 20). **Axioms** are **logical relationships** that are true without question. Axioms provide a deeper background knowledge in our domain [125, 127].

An example from Greek mathematician, Euclid of Alexandria (dated around 300 B.C.), “*things which are equal to the same thing are equal to each other*” [133], such as **axioms** automatically inferred via: **class A subclass of class B** and **class B subclass of class C**, then we can logically infer that **class A subclass of class C**. Defined relationships have an IRI, both relationships and axioms are machine-readable, and they facilitate semantic interoperability. One such example of an **axiom** stated in an ontology can be observed in Listing 1.1 which represents uveitis in Extensible Markup Language (XML): in a condensed structure we can learn that **uveitis is an abnormal increased inflammatory response occurring in the uvea**.

```

1 uveitis
2 has part some (
3     increased amount and
4     characteristic of some (
5         inflammatory response and
6         occurs in some uvea ) and
7     has modifier some abnormal )

```

Listing 1.1: Example of an axiom in a biomedical ontology representing uveitis.

The third and final concept is **metadata**, or better known as ontology **annotations**. **Metadata** in ontologies are important for users to understand and have the ability to use an ontology [125]. There are various types of annotations: despite defining a class or relationship, for it to have meaning we need to annotate this concept with a **label**. **Annotating** the **class** (IRI EXAMPLE:002) with the **label** “uvea” gives the class conceptual meaning. **Labels** in an ontology are from a formal, standard set of vocabulary which describes the domain [124–126, 128]. Concepts can also be annotated with a **definition**

to further describe the concept, this is important for users to grasp what it is and the purpose of use [125]. Another common annotation is a **database cross reference** (**x-ref**), which is an IRI to another ontology: promoting the semantic web allowing the linking knowledge and classes across domains [125]. Ontologies having **x-refs** provide robust and sophisticated terminology standards enhancing the compatibility of domains [131]. Another vital annotation are **synonyms**: terms which could be used instead of the **label**, however less formal [124, 125]. Synonyms in ontologies can be described as **exact** (precise alternatives), **broad** (more general), **narrow** (more specific), and **related** (associated) [30].

There are numerous types of ontologies, “upper ontologies” (top-level or foundation) are aimed at standardising the design of ontologies [127]. Two other ontological frameworks are domain ontologies, which cover a large domain (e.g. all diseases) [127, 129] and application ontologies, which are specific to a domain or applications (e.g. a specific disease) [129, 134]. When domain ontologies are first developed the concepts and axioms are detailed. These initial designs derive from large meetings, conferences, or workshops with ontological and domain experts contributing [124]. These methods of ontological design are not efficient in eliciting feedback due disagreements of a standardised vocabulary or ontological structure [124]: many classes lack information (e.g. metadata: definitions or synonyms). This includes future developments with new additions of expansions. Despite continuous development of ontologies, their terminology should be standardised across superclasses [128] to enable consistent and coherent ontologies for integrative analysis.

Terminology

Ontologies are designed by domain experts: such as clinicians or researchers. Class labels are terms that experts deem to be the most formal for medical orientation applications: a clinical term. Clinical terms are standardised and consistent in the domain’s coherent vocabulary and attempt to avoid confusion or repetition for annotation tasks [5, 124–126, 128, 131, 135]. However, clinicians and researchers shape their vocabulary and terminology to fit their organisation or environment [124]. Synonyms are other terms that are used

in the domain, in the example visualised in Figure 1.5 (page 28), the **class** “eyeball of camera-type eye” has an **exact synonym** “eyeball”, telling us that domain experts commonly use “eyeball” but “eyeball of camera-type eye” was the best-defined **label** for this concept in their domain. Having these clinical terms and synonyms, plus the structure and axioms in ontologies, make them powerful semantic networks for computational analysis and contribute to the semantic web [55, 124, 125, 127, 131, 135–137]. Ontologies have been used for semantic similarity objectives to predict the diagnosis of disease based on phenotypic similarity [125, 136, 137]. Using ontologies for text-based tasks have previously identified associations: a particular study Zhou et al. [138] presented diseases based on phenotypes, though for these tasks to be successful the ontology concepts need to be a full coverage of the domain and be a standard set of vocabulary [125].

Ontologies have previously demonstrated success to address the challenges presented by unstructured text [125, 130]. A lack of standardised vocabulary means computational analysis is delayed as computers may not recognise terms without explicit definitions or relationships [5]. However resources of lay vocabulary is scarce [55], in fact, mining these sort of data sources can expand and improve medical domains by building upon this lay vocabulary to connect them to their clinical counterparts [139]. In an example from 2016, Eholié et al. [55] and fellow research collaborators acknowledged gaps in clinical and patient vocabulary, especially in other languages. Eholié et al. [55] aligned vocabulary produced from forums and Facebook groups about breast cancer, with standard medical terminologies in the form of an ontology, to represent what they found. A total of 64 expert concepts with 173 relations were built: 89% of terms were automatically matched with concepts from biomedical ontologies and 11% had to be done manually [55].

Vasilevsky et al. [12] (2018) depicted the importance of plain language (lay) in biomedical ontologies and acknowledging patients’ role in diagnosis. They also outlined the development of explicit and implicit knowledge encapsulated as self-phenotyping information [12, 13]. Patients develop their own vocabulary, which can be in lay, or a mixture of lay and clinical [12]. Vasilevsky et al. [12] acknowledge the importance of ontologies for phenotyping,

however explain that they are “fuzzy” when matching lay and clinical terms from the lack of knowledge of the patient language. Vasilevsky et al. [12] explain how current biomedical ontologies have deep phenotyping design, whilst useful, is also time-consuming and lacking in lay terms. Finally, the authors delineate the importance of layperson knowledge and expressions openly on the web especially for developing phenotyping profiles of rare diseases [12].

1.5.1 Current biomedical ontologies

Gene Ontology

In biological science, biomedical ontologies started to emerge, such that in 1998 the development and release of one of the most well-known and used ontologies, the Gene Ontology (GO) [124, 125, 140]. GO was proposed as a structured system to enhance the availability and usability of data gathered on several different organisms [124, 140], which includes three main superclasses: biological process, molecular function, and cellular components [140]. GO was used for genetic data throughout the early 2000s and continues to be used for bioinformatics process of large genetic datasets, such as sequences through annotation via their standardised domain vocabulary [125, 127]. As of 2021, GO has 50,653 classes and has become a standard of the display of data [124]. Whilst GO is a very-well known biomedical ontology for genetic analysis and includes inflammatory biological processes, it lacks the ability to annotate textual patient data.

Human Phenotype Ontology

The Human Phenotype Ontology (HP) was developed in 2008 by Robinson et al. [5] and consisted of 8000 classes. HP was created to enable deep phenotyping and designed to overcome the issues of analysis of phenotypic data of human diseases [5]. HP was curated from literature and has expanded over time to include other rare diseases and further disease characterisations [11]. As of 2021, HP has 19,989 classes.

A study in 2016 by Vasilevsky et al. [30] aimed to make HP more accessible to patients by expanding it with layperson terms. They manually reviewed HP, and then manually reviewed public sites and other biomedical ontologies. The authors acknowledged that despite some existing lay terms in other biomedical ontologies and medical terminologies, they are generally broad. Vasilevsky et al. [30] marked 6,240 synonyms or primary labels as layperson - they had issues as to whether certain layperson terms were useful, such as the layperson term “colourblindness” can describe multiple clinical terms. The authors reported issues that many HP classes were already layperson: showing an issue at the lack of standardised and consistent vocabulary. An issue is that when they added a layperson term for a class, they also added that for the subclasses, but axioms would automatically infer this. Vasilevsky et al. [30] concluded their study by discussing that further validation is needed as to whether these layperson terms reflect the patient community, to do this they suggest a crowd-sourcing approach, which has been previously discussed by the Staffieri et al. [59] study in that crowd-sourcing is time-consuming. Despite the expansion of HP with lay terms, Köhler et al. [11] pointed out that patients hinder the use of HP as they do not use medical terms when searching for information but highlight the benefits of patients being able to use HP: patients may gain better understanding of symptoms, diseases, and use medical terminology.

Disease Ontology

The Disease Ontology (DOID) was developed in 2012 by Schriml et al. [141] and had 8043 diseases. DOID semantically integrates diseases through standardised vocabulary from medical terminologies and x-ref [141]. As of 2021, DOID has 17,563 classes.

Uberon Multi-species Anatomy Ontology

The Uberon Multi-species Anatomy Ontology (UBERON) was developed in 2012 by Mungall et al. [142] and consisted of 6,500 classes. UBERON integrates other ontologies [142], such as GO [140]. UBERON not only contains information of human anatomy, but

also anatomical structures of other species allowing multi-analysis of various species with a single ontology [142]. As of 2021, UBERON has 21,912 classes.

Orphanet Rare Disease Ontology

The Orphanet Rare Disease Ontology (ORDO) was developed in 2014 by Vasant et al. [143] with 11,699 classes of rare diseases, designed by phenome type [143]. ORDO is based on the Orphanet database, which is a resource consisting and improving of knowledge on rare diseases [144], validated by experts in the domain. ORDO is Orphanet in ontological form [143]. As of 2021, ORDO has 14,886 classes.

Phenotype and Trait Ontology

The Phenotype and Trait Ontology (PATO) was developed in 2018 by Gkoutos et al. [134]. The main aim of PATO was overcoming the issues of capturing machine-readable phenotypic information for computational analysis [134]. Phenotypic information is in the form of natural language and are in biomedical ontologies, and so PATO is available for ontology design patterns and allows mapping between phenotype ontologies [134]. As of 2021, PATO has 3664 classes.

Repositories

By the late 2000s, there was sufficient interest and activity in biomedical ontologies that repositories were developed in order for ontologies to have a platform to be explored [124, 125]. The Ontology Lookup Service (<https://www.ebi.ac.uk/ols/index>) is one of the first online repositories, released in 2006 for biomedical ontologies [145]. When the Ontology Lookup Service was released it contained 42 ontologies and approximately 135,000 terms [145], whilst as of 2021 it contains >250 ontologies with >6 million classes. The Open Biomedical Ontologies Foundry (<http://obofoundry.org/>) was developed in 2007 [146]. When the Foundry was released, approximately 16 biomedical ontologies were available, including GO and DOID. To have an ontology available on the Foundry, the

group Smith et al. [146] recommends first submitting to their forums where other expertise can first analyse and then agree if the ontology meets the Foundry’s principles.

BioPortal (<https://bioportal.bioontology.org/>) [147] was developed soon later in 2009 and is one of the most-used and largest biomedical ontology repositories [125, 148]. When BioPortal was first released in 2008 they had only 72 ontologies and approximately 300,000 classes [147, 148], whilst as of 2021 it contains >900 ontologies with a total of >13 million classes. Some of the advantages of BioPortal is the ease of uploading an ontology, including allowing different formats and allowing users to comment on ontologies [147]. In comparison to the somewhat stricter approach from the Foundry [146]. BioPortal has the unfortunate case of high reuse of classes [149], but this can be fixed with x-ref or importing from IRI directly. Finally, one most recent repository is Aber-OWL (<http://aber-owl.net/>) developed in 2015 [150], intended for a deeper semantic overview of ontologies via queries.

1.5.2 Medical terminologies

Medical terminologies are systems consisting of clinical terms in order to create a consistent and standardised set of terms in clinical settings [30, 33, 55–57, 134, 151]. These terminologies may also be referred to as “classification”, “thesaurus”, and “vocabulary” [56, 151].

Like ontologies, these terminologies are structured in a hierarchical manner and each term has a code that is “significant” or “non-significant” [56]. Significant codes are structured (mnemonic), e.g. **d**igestive system **d00** with subclass mouth **d01**. Non-significant systems are randomised codes so when the structure changes there will be no consequences and new terms can be added easily [56], e.g. digestive system **341**, and mouth **711**. The codes, or vocabulary, should be consistent in medical terminologies [151] and safe, but also clear to avoid misunderstandings. Kubben et al. [151] bring the attention that complications can arise if no agreement is set (or description) to what a code means, for example D_1 =adolescent and D_2 =age in years can cause confusion as no description for what years D_1 are set.

The first modern medical classification that can be considered a true ontology of diseases (nosology) was developed by Carl Linnaeus (1707–1778) [152]. Linnaeus was known as the “father of modern taxonomy” and created biological plus logical taxonomic organisation. These broad categories are “binomial” concepts in that all Kingdom names are standardised as Latin so these should be used when referring and classifying a concept [152] e.g. *Homo sapiens* for human, *Felis catus* for cat, and *Strix aluco* for tawny owl. Although Linnaeus’ classification contains some errors, such as the notion that leprosy can be caught by eating herring worms, his classification laid the foundation for work that eventually led to the first edition of the International Statistical Classification of Diseases, in 1893 [130].

International Statistical Classification of Diseases

The International Statistical Classification of Diseases (ICD) was introduced in 1893 and was known as the International List of Causes of Death [130, 153]. Since 1948, ICD has been maintained by World Health Organization (WHO) [153], which the sixth version, ICD-6 was released - this version was the first to include morbidity [153]. Over time, ICD has been modified to adjust to new findings, research, and advanced in healthcare to become a standard for medicine [130, 153]. These modifications have been published as a series of revisions. ICD over the years has been organised by “chapters” according to the anatomical system [56, 153], such that all digestive-related diseases are in the same group despite phenotypic features.

In 1975 ICD-9 was released, and even today many organisations continue to use them. ICD-9 was expanded to ICD-9-CM, which included specific clinical modifications (additional diagnostics and therapeutic procedures) [56, 153]. There are 38,000 ICD-9 codes and they are hierarchical, such that diseases of the circulatory system all have a code of 390–459, e.g. rheumatic chorea 392.

By 1990, ICD-10 was released, with more than 55,000 codes. ICD-10 codes are also hierarchical but also mnemonic, such that diseases of the circulatory system all begin

with I followed by a number, e.g. rheumatic chorea I02 [56, 153]. ICD-10 is the most recent version of ICD which includes additional concepts regarding health status V and external causes of injury and poisoning E [56, 153]. World Health Organization (WHO) [153] report ICD-10 has been cited by more than 20,000 scientific articles, used by more than 150 countries, and translated into more than 40 languages. ICD-10 also has a CM version, which is more in-depth, such as acute appendicitis in ICD-10 is K35 and only expands to an additional number to K35.8. However acute appendicitis in ICD-10-CM is K35 yet can expand very detailed to K35.891 as “Other acute appendicitis without perforation, with gangrene”.

In January 2022, ICD-11 will be fully available [153], which includes additional clinical needs - with the main aim of migrating ICD from a statistical framework to a clinical classification for statistical use [153]. ICD-11 aims to bridge gaps in mental health to include “Mental, behavioural or neurodevelopmental disorders” or “Factors influencing health status or contact with health services” [153].

ICD is one of the most known medical terminologies used in medicine [56, 153] due to the major advancements since the initial development, to include mortality and cause-of-death concepts [130, 153].

Read codes

Read codes were developed by general practitioner (GP) James Read and implemented by computational expert, Tim Benson [154, 155], with the first demonstration and publication in 1986 [154]. Dr Read’s idea for Read codes stems from multiple sources. Initially, it was motivated by Dr John Preece, who in the 1960s was the first to use a computer in a GP consulting room, in the UK [155]. Secondly, Dr John Perry wrote a system, Oxford Medical Information System (OXMIS), aimed for GPs based on ICD-8. However, after Dr Perry’s death, OXMIS was no longer being used [155]. In the 1980s, Tim Benson developed a single-user GP system. Whilst working with Dr Read, they developed the Read codes [155], which originally was named, Abies Medical Dictionary. With the aim

of automated description of clinical and administrative data [56, 154]. Read codes were based on ICD-9 however only concepts seen in GPs were extracted [154, 155] as Read codes were developed specifically for GPs, by a GP.

Over the years, there have been multiple versions of Read codes, with the most recent being version 3. By 1992, Read codes became NHS Clinical Terms [56, 155]. As of 2021, there are 140,065 concepts of NHS Read codes version 3. Read codes are ordered hierarchically, with version 3 onward each concept has a unique code consisting of characters and numbers [56, 154], an example is “Disorder of appendix” Xa9C3 and a subclass appendicitis Xa9C4.

Systemized Nomenclature of Medicine

The Systemized Nomenclature of Medicine (SNOMED) was published in 1975 by the College of American Pathologists (CAP), originally named Systematized Nomenclature of Pathology (SNOP) [56]. SNOMED aims to provide terms for a broad range of clinical domains and intended to be incorporated into patient records [56, 57]. Like ICD, SNOMED has changed over the years, with SNOMED-II, SNOMED-International, and SNOMED-RT (Reference Terminology) [56, 57], and then in 1998 SNOMED-CT (Clinical Terms) was developed by merging SNOMED-RT with Read codes (version 3).

Since 2007, SNOMED has been managed by International Health Terminology Standards Development Organisation (IHTSDO) [156] and as of 2021 there are 358,483 concepts. Kubben et al. [151] discussed that whilst SNOMED-CT is set to be an ontology in the sense it will have synonyms and other metadata, the ontological standard is not yet completed. SNOMED-CT has non-significant codes: they are randomised and concepts are organised into various categories, e.g. “procedure”, or “disease and diagnosis”, which almost all of ICD-9 fits into [56]. SNOMED-CT “Disorder of respiratory system” is 50043002, whilst asthma is 195967001.

Mappings of medical terminologies

Medical terminologies were originally designed for administrative purposes. Clinicians benefit from using a common language which is consistent across their domain and environment, e.g. EHR systems or clinical letters [56, 134, 139, 151]. However with the semantic web and the importance of research, medical terminologies have been used more frequently with researchers choosing a terminology system based on the sort of coded data being handled, or by evaluating the existing terminologies [56, 151]. Medical terminologies and ontologies can be used for semantic analysis of expert-derived texts [55], such as literature, clinical letters, or EHR.

Despite the co-existence of ontologies and medical terminologies, they are different. Some similarities include both consisting of clinical terms and having concepts that are structured in condensed, organised systems. However medical terminologies are not logically annotated: ontologies have extensive metadata [56, 130, 134]. Many biomedical ontologies, such as HP [5], DOID [141], and UBERON [142] all aim to map to GO [140] allowing the connection of gene analysis through phenotypes, diseases, and anatomy. However not being up-to-date or at a consistent standard causes limitations.

Clinical research can use ontologies to benefit clinical practices, and although some biomedical ontologies are medical terminologies enhanced with deeper representations and axioms, they may fall of expectations when dealing with unstructured text which could be grammatically complex [157], e.g. clinical letters or EHRs. Medical terminologies are dynamic and often updated or republished due to continuous advancements in clinical care and practice [158]. Medical terminologies do not yet have the ability to represent complex relationships. Moreover, one of the fundamental issues is the lack of ability to utilise these systems, there is a lack of tools available for direct manipulation of an ontology.

A lack of standardised vocabulary across these medical terminologies, plus biomedical ontologies, means computational analysis is delayed, an example by Robinson et al. [5] describes “generalized amyotrophy”, “generalized muscle atrophy”, and “muscular atrophy, generalized” are different diseases yet a computer may not recognise this. Ceusters

et al. [33] highlight the grammatical construction preferences, in that SNOMED some terms do not mean what is expected and that clinicians use different variations for the same clinical meaning. Despite ICD being one of the most known medical terminologies and having codes for rare diseases [56, 153], ICD has been noted by Schieppati et al. [159] that it is not useful in regards to rare diseases and the works by Palestine et al. [160] has noted that ICD-10 lacks reproducibility in data analysis due to imprecision in coding, and lack of descriptions for some codes meaning users can interpret uses of ICD-10 differently.

Tim Benson himself stated many issues with medical terminologies, including their own Read codes. Benson [155] discussed how GPs can enter the wrong code into the system due to mistakes, synonyms, or abbreviation issues. Benson also mentioned that although Read codes are useful for GPs, they have proven unsuitable for hospitals due to the “simple” hierarchical structure and hospital doctors did not understand the Read code information retrieval. Benson also brings attention to the issue of inaccuracies, providing the example of tuberculosis meningitis, being both a type of tuberculosis but also a type of inflammatory disease, having a code in two separate chapters: two separate codes creates code redundancy. Finally, the levels to Read codes and other terminologies are mentioned as some conditions will need further in-depth detail [155].

SNOMED has complex concepts, which tend to have higher error rates that Zheng et al. [161] state causes difficulties. de Keizer et al. [56] discuss that it is questionable whether SNOMED can generate sensible compositions. Furthermore Ceusters et al. [33] mentions the linguistics issues: ambiguous meanings to some concepts “Implantation of mammary artery into ventricle (procedure)” 55230008, which clinicians may use different variations for the same clinical meaning. Ceusters et al. [33] also discuss that SNOMED has a mismatch of concepts “Urinary bladder excision (procedure)” 108034003 does not mean “excision of the urinary bladder”, but rather “excision from”.

Whilst each medical terminology system has their own advantages, different organisations use terminologies differently but also specifically for their data resources. It is said that United States medical environments are better adjusted using ICD-10 compared to ICD-9

or SNOMED-CT as ICD-10 is more clinically precise [162], however this can complicate organisation terminology issues. On the other hand, in the UK, medical environments are changing over to SNOMED-CT, the implementation started in 2018 for GPs to change from NHS Read codes to SNOMED-CT. As of 2021 the England's health department state healthcare providers should all be using SNOMED-CT as mentioned by International Health Terminology Standards Development Organisation (IHTSDO) [156]⁸. However, de Lusignan et al. [163] discusses concerns with SNOMED-CT in that being large lies issues when searching as multiple categories (superclasses) make it difficult. Chen et al. [47] bring to light that diagnosis codes in hospitals are not always 100% true in that many codes are for billing purposes and some hospitals over or under-report patient diagnostic codes and so some diseases in datasets or EHR could be over-represented.

All of these issues with medical terminologies present potential problems for the semantic web [33, 163]. These systems can have disagreements in their preferred formal terms, and so text-based efforts are affected due to conflicting linguistic designs [33]. Another the major issues is that the advances and changes in healthcare is inevitable and so how terminologies adjust to these changes will depend on their own schema [158, 163]: how can these comprehensive systems stay updated?

The Unified Medical Language System (UMLS) was developed in 2004 by Bodenreider [164] as a mapping system for various medical terminologies and ontologies. The main aim for UMLS was for clinicians and medical records allowing a standardised vocabulary used across terminologies [164]. The UMLS consisted of more than 900,000 concepts from 60 different sources [164]. As of 2021, to access the UMLS you need to sign-up to their license agreement. The UMLS may have mapping issues for medical terminologies as stated from Benson [155] and Ceusters et al. [33] that some concepts in medical terminologies may need multiple codes, mean the same, or actually not be as expected. UMLS has also been said to have flaws due to being a giant thesaurus and covering too broad of topics and perhaps not including more specific detailed information [33]; additionally de Keizer et al.

⁸read more here: <https://www.snomed.org/our-customers/member/united-kingdom>.

[56] report that UMLS does not support the composition of new concepts. Furthermore the UMLS lacks in informal or lay terms [165], which highlights also the limitation in other biomedical ontologies and medical terminologies.

The terminology gap between patients and clinicians is an obstacle in clinical practice [12], and the gap has limited patient participation both in research studies and in clinical phenotyping [44]. A study in 2016, by Park et al. [44] looked into understanding the way patients use medical terms by evaluating the coverage of lay terms in the UMLS. They extracted two social media corpora of diabetes-related posts: blogs that had a tag with “diabetes” and Q&As in a diabetes forum. Park et al. [44] identified UMLS concepts in the datasets and looked at which medical terminology was most present. The authors stated that SNOMED-CT was the dominant vocabulary ranging from 84.9% (blogs) and 95.9% (Q&A). To conclude, Park et al. [44] brought to attention that although the datasets varied in size, they exhibited similar conceptual coverage and even though the medical term coverage was low, they show that patients are using medical terms [44].

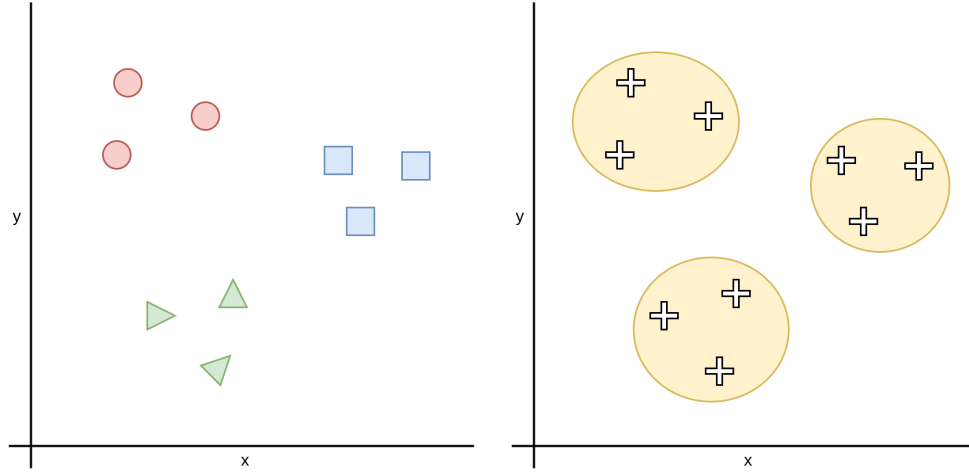
1.6 Machine learning

Mathematical algorithms determine the ability of Machine learning (ML) [127] and can be understood through algebra, logical deduction, and probability. This branch of science is related to the idea that we mimic a brain: a lot of computational power is needed to produce solutions [127] and despite growth during the 20th century, the connection of computers to the brain has been said to derive from 400 B.C. philosophers, such as Aristotle, who commented that the mind is in some ways like a machine [127].

ML aims to find complex patterns in data, potentially in order to replace the manual work of a human [166, 167]. ML plays a major role in our lives today from search engine optimisations, to advances in medical care. ML is computationally expensive because the algorithms are complex and often have recursive or exhaustive steps to develop models: learning from the data and changing parameters [166, 167]. Despite the valuable use of ML techniques in clinical practice, there have been many issues with the methods of reporting

the analysis and results [168] - lack of descriptions and a need for better explanations as research can be under or over-estimated.

ML can be broadly categorised into supervised and unsupervised learning, Figure 1.6 illustrates the two variations. The majority of ML related work in this thesis relies on unsupervised learning.



(a) An example of supervised ML. Samples are datapoints and their shape is based on their endpoint. (b) An example of unsupervised ML. Samples are datapoints and they are assigned to a cluster via distances.

Figure 1.6: Machine learning supervised and unsupervised visualised in plots.

Supervised machine learning

Supervised ML relies on a dataset where each sample is labelled with an endpoint (or target, e.g. clinical outcome survival) [151, 166]. These endpoints could be categorical (classification problem) or continuous (regression) [166, 167, 169].

Supervised algorithms work via splitting the input data into *training* (with endpoints) and *testing* (hidden endpoints). The training data consists of majority of the data and used to build a model. This model is then used to predict the “unseen” endpoints in the test set [166, 167]. Subfigure 1.6a shows an example of plotting samples shaped by the endpoint - models will predict the shape of new samples based on what the model learnt. A major issue in supervised learning is overfitting. Overfitting is a model performing well on the training set yet performs poorly unseen data [166]. One potential reason for

overfitting is the inclusion of too many features, making the model very specific to the observed data and being too complex [166].

A type of supervised algorithm is decision trees (and random forests which are multiple trees) [151, 166], essentially using endpoints to build a structure. Decision trees can be used as a form of feature importance in a model [166]. Trees can avoid the overfitting issue or no distribution techniques (e.g. no normalisation) [151], however, can struggle with smaller sized data. Another type is the k-nearest neighbour classifier, known as a lazy algorithm [166] as it memorises the training set rather than building a logical model.

Evaluation of prediction

There are numerous evaluation techniques, including cross validation: the splitting into training and test sets is done over multiple iterations to create a model that should be able to validate overall precision. Cross validation can overcome the overfitting problem. Moreover, a receiver operating characteristic (ROC) [151] is a performance measurement for the supervised task: representing capabilities of distinguishing the endpoint.

1.6.1 Unsupervised machine learning

Unsupervised ML, or “clustering”, aim to find complex structures and underlying patterns in the data by learning natural distances between samples [166, 170, 171] in a way that samples most “similar” are placed in the same group and so revealing what constitutes a group, an example can be seen in Subfigure 1.6b (page 43). Not only are there different algorithms, but there are also various methods to get the optimal number of clusters K and all have their own parameters to set [170].

One challenge of clustering is that there is no effective evaluation method to validate if algorithms are working as intended or expected [172], therefore a team may use multiple algorithms in parallel. Clustering is best interpreted with expert opinion [173], for example clinical expert can best characterise clusters as they have the domain knowledge to understand.

Algorithms

K-means

The most well-known clustering algorithm is K-means, which is essentially assigning a sample to a cluster iteratively [151]. K-means is a centroid-based clustering algorithm. Centroid-based clustering algorithms are based on finding the best central vectors and then assigning samples to whichever centre vector is closest, such that the squared distances from the cluster are minimised [166, 167, 170], these central vectors greatly affect the final clusters. K-means is computationally fast (the fastest algorithm) and very understandable, however, identifying the number of clusters (K) and their central vectors is difficult as they are dependent on the initial structure of the data [151, 167]. Additionally, K-means performs best on low-dimensional datasets [172]. Another algorithm that performs well is K-means++ which is essentially K-means and Hierarchical combined, it overcomes K-means' centroid drawback of randomisation [166].

Hierarchical

Hierarchical is partitional clustering: using the complete dataset and separating samples into groups continuously (divisive), or reverse which is starting with single samples and continuously aggregating them into clusters (agglomerative) [166, 167, 170], via a criteria. Hierarchical does not need to determine the number of clusters K . A limitation with hierarchical clustering is that if a small error occurs at the beginning stage when splitting or joining samples, then they cannot be fixed at a later stage [167], in addition, Hierarchical performance is much more sensitive if there are potential outliers [151].

DBSCAN

A type of density-based clustering is DBSCAN, which uses the distances between samples to group them. DBSCAN uses a grouping criteria and can consider some samples as outliers (noisy data) if they do not meet the criteria [166, 170]. An issue with DBSCAN

is finding the cut-off criteria and parameter fine-tuning: these parameters are problematic if the data is heavily dense [166]. Additionally, DBSCAN is affected heavily by the curse of dimensionality [166, 172]. DBSCAN can be tricky in datasets as the aim of the study needs to be understood: in data which noise can be excluded can be conducted, however in data which noise can be interesting needs to remain [166]. DBSCAN is useful in clustering data of arbitrary shapes [166], as seen in Appendix Figure A.1 (page 191).

Spectral

A final type of clustering is graph-based, and one example is spectral. Spectral clustering is quite modern and outperforms other types of clustering using graph-based similarity [166]. Spectral clustering uses eigenvectors of sample and matrix distances to derive cluster relationships [166].

Evaluation of clusters

With all of these clustering algorithms, there are methods to choose the optimal number of K [166, 174]. When running multiple algorithms, which is best practice [172], with multiple optimal clusters methods, you can evaluate through visualisation, clustering algorithm agreement scores, and the model which provided the highest gap score - the adjusted Rand index (aRi) score compares each clustering method, to observe statistical agreement of cluster assignments [175].

The elbow, silhouette coefficient, and gap statistic all compare K distortion [166], see an example in Figure 1.7. In the elbow plot (Subfigure 1.7a), you choose K by the most angle; in both silhouette and gap statistic (Subfigure 1.7b) K is the highest returned value [174]. Moreover, the silhouette coefficient can also be used as an evaluation metric [166, 174] as after clustering, a visualisation of silhouettes can be used to observe cluster size fluctuations: tightness and separation [174]. It is important to note that not all methods work with all clustering algorithms and so it can be difficult to choose K with different algorithms without bias [166].

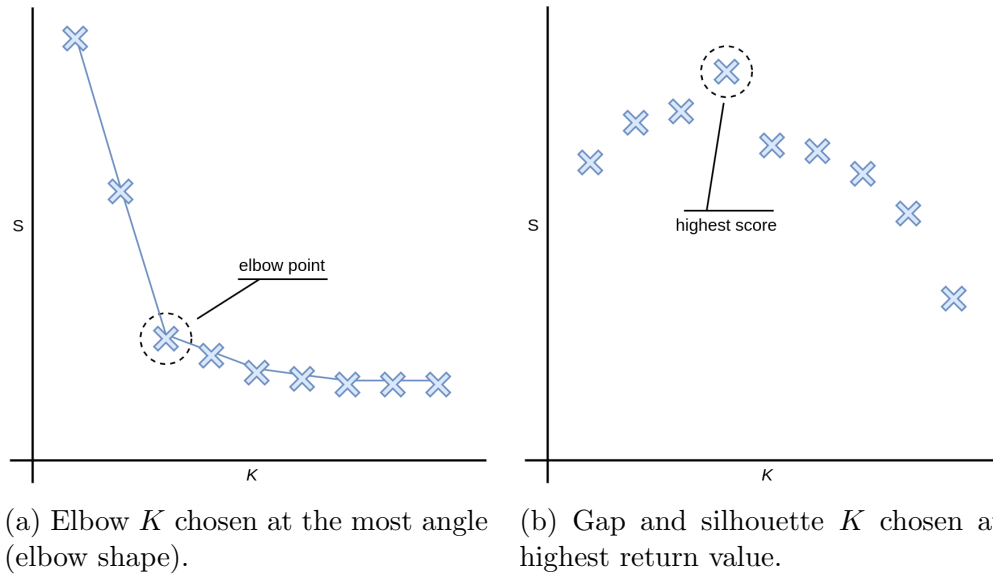


Figure 1.7: Visual representations of methods to gain the optimal number of clusters K .
Note: S =score.

1.6.2 Data preprocessing methods

Notation

- Matrix M - otherwise known as table, spreadsheet
- Sample m - otherwise known as row, observation
- Feature n - otherwise known as column, variable

For brevity this section will make reference to the notation above.

In an ML study, there are various methods in order to prepare data (M) [166]. One initial step is sorting categorical features, e.g. correcting into ordinal (ordered) such as BMI scale [166]. In both supervised and unsupervised ML, features in M should be comparable and scaled into the same format for optimal results [166]. Two ways to scale data include normalisation (bound between two values, e.g. 0 and 1) and standardisation (mean 0 and standard deviation 1) [166]. Standardisation is best for a Principal component analysis (see section 1.6.3, page 49) and still retains outlier information [166]. An example of normalisation and standardisation scaling can be observed in Appendix Table A.1 (page 190).

Missing value handling

It is common in experiments to consist of missing data. Removing missing data from M can be either done via features (n) and/or samples (m). Removing too many features can risk losing valuable information [166, 176], sometimes a percentage of missing values can be used as a cut-off, e.g. 5%. Removing too many samples can affect the statistical power but avoids biased [176]. In some circumstances, imputation is an option, which is filling missing values based on other features [166, 176] (such as averages or most common value), however by doing this we can confuse the model.

Feature elimination

Feature elimination (unsupervised) or selection (supervised) removes features to avoid noise in M [166]. One initial method is dimension reduction (see section 1.6.3), another is correlation. Using correlation highlights features that are strongly positive/negatively dependant with another [166] to remove additional noise. Two commonly used methods includes, Pearson correlation coefficient and Spearman's rank correlation coefficient. Pearson should be conducted with raw features, e.g. systolic/diastolic blood pressure; Spearman works with rank-ordered features, e.g. pain scoring. Removing features which do not meet a variation threshold is another form of feature elimination [166]. The variance threshold removes those features since they do not contribute to the pattern of M [166]: these features have similar values in samples.

There are alternatives for supervised ML, such as the leave-one-out approach which tests for feature selection by running the model n =feature times and each without a feature [166], though the leave-one-out approach is only suggested in very small datasets.

1.6.3 Dimension reduction

A subtype of ML is dimension reduction (DR) [166]. One of the main purposes of DR is visualisation, compression, and feature elimination [166]. DR is a form of data compression: reducing an $M \times n$ (features) to D dimensions: $D < n$. The aim is to retain as much variation present in the data [171], whilst compressing it into a smaller set:

this smaller dataset avoids the large dimensionality problem and is more computationally efficient [166]. An M of $n > 16$ is considered highly dimensional [172], yet argued that $n > 10$ can also be highly dimensional. Even with DR, it can be difficult to visualise M if $D > 2$ as it may be inconvenient to interpret without a clear distinction of plot markers [177]. Which is why D s 1 and 2 are easier to visualise the data structure in a plot.

Using DR for feature elimination removes features that are not informative or do not contribute to the structure (noise) [166, 172] and so clustering can be more successful in finding patterns. If DR methods are applied to M with little or no uninformative features then it can degrade the predictive performance [166]. On the other hand, if DR is applied to M with a large number of uninformative features, which are then removed, it can create a clearer and more understandable structure, and improve the predictive performance (supervised).

Principal component analysis

Principal component analysis (PCA) was introduced in 1901, developed in 1933, and used widely when the electronic computer emerged [171]. It is important that all n included in a PCA are rational and similar in some measurements as non-proportional features are difficult to interpret after a PCA has been performed [171]. PCA has been said to essentially “flatten” matrix M of n into a new subspace and return Principal components (PCs), or D , ordered by variance (most important) [166, 171]. PCA for DR or feature elimination includes guidelines. For feature elimination, looking at PCs 1 & 2 feature variance, those which contribute the least to the new subspace can be removed.

For DR the first two PCs are most informative and are often used as a straightforward representation of M [166, 171], however this does not mean to exclude other D . Each PCs have an eigenvalue, representing the variance [166] and cut-off for cumulative variance depends on the type of PCs: $D=2$ is enough if PCs 1 & 2 contribute to $>90\%$ however if $D = 20$ then using PCs 1, 2, 3, & 4 (with a cumulative variance of approximately 60%) would be enough [171]. Eigenvalues can be plotted as a scree plot (similarly to

elbow plot in Subfigure 1.7a, page 47) [166, 171] and one should hope that PCs 1 & 2 are representing the most variance, rather than low variance across all PCs. If there is a straight line present, then one can cut-off at the most left PCs. If PCs 3, 4, & 5 are similar in value then this can indicate an error [171], known as Cattell’s formulation (1966) [178]. Additionally, PCs with variance less than 1 are not worth retaining [171], known as Kaiser’s rule (1960) [179].

1.6.4 Natural language processing

The simultaneous emergence of modern linguistics and computational advancements in the 1960s is when Natural language processing (NLP) unfolded [127, 180] for language interpretation and text analysis. In NLP, a computer not only needs to understand written language, but also the subject, context, and sentence structure [127, 180], as words can have multiple degrees of similarity [181, 182]. Natural language is highly expressive and it can be difficult for a computer to process [127, 134].

There are various preprocessing steps in NLP, including: lowercase, removal of special characters, tokenize (separating words), removal of stopwords (no semantic meaning, e.g. “the” or “and”) [127, 166]. Other preprocessing steps includes “stemming” (removing the stem of terms) or “lemmatisation” (converting term to base form) can both improve output data count as terms are matched better [67, 127, 166]. All of these steps can help the computer understand easier and analysis is more tractable.

Stemming and lemmatisation

- Stemming produces a word’s stem: “painful, pained, pains” all commonly have the stem “pain”.
- Lemmatisation produces a word’s lemma: “am, are, is” all commonly have the lemma “be”.

NLP methods in the clinical and patient domains are valuable [183] as written information is accurate and reliable as they are first-hand accounts from the individual directly [183, 184]. These immediate accounts reveal and reflect the complexity of clinical practice from the perspective of the clinician and patient.

In comparison, EHR are static due to the ease and convenience of entering records as medical codes [183, 184] instead of descriptive accounts, plus they are restricted and access is difficult [139]. There has been NLP analyses using EHRs looking into mental illnesses, for example in 2017 Jackson et al. [185] looked at symptoms from EHR discharge summaries through creating multiple models of a set of 50 symptoms from severe mental illnesses [185]. Jackson et al. [185] used a supervised ML algorithm to predicting binary symptomatic texts. From the corpus, the authors extracted mental illness symptoms in 87% of patients with a severe diagnosis and 60% of patients with a non-severe diagnosis [185]. To conclude their study, Jackson et al. [185] stated that symptoms were present across the various diagnosis rather than present in a few diagnoses. The authors summarised by discussing how their work demonstrates how manual curation and extracting of symptoms from clinical discharge summaries can be further investigated into automatic techniques and facilitates the secondary use of these models in research.

Clinical text analysis can advancing the medical domain, however NLP analysis benefit with manual observations and verified by a relevant expert [173, 183]. This is because interpretation is best from the involved domain experts, e.g. clinical work is rarely conducted without clinicians and other experts. The information can be better understood, or results are already known, and so evaluation is beneficial [183].

Annotation tools

There has been progress with clinicians using natural language in clinical care settings [33], and many tools have been developed to try and overcome the gaps in clinical and natural language. One such tool, NoteAid by Chen et al. [186], is a web-based interface for patients to use in order to essentially “translate” their EHR information into layperson terms - when they presented to physicians, they reported mixed feedback however no response was reported from the perspective of the patient. There also exists CogStack Jackson et al. [187], which initially was designed for EHR extraction, however they saw opportunities for CogStack and the lay user: improving an organisation’s data.

Integrating natural language and patient-preferred terms into research means there is also a need for tools to conduct NLP tasks, including text mining, named entity recognition (NER), and more. There are numerous tools for NLP annotation tasks:

- textpresso was developed in 2004 by Müller et al. [188], which consisted of numerous NLP tasks, specifically text mining. The tool is not publicly available⁹ (<http://www.textpresso.org/tpc>) however did allow use of ontologies.
- BRAT was developed in 2012 by Stenetorp et al. [189] and is a web-based (<http://brat.nlplab.org>) tool allowing NER tasks.
- NILE (2013) is an NLP program aimed for EHRs, developed by Yu et al. [190]. As of 2022, users need a .gov (or similar) email for use (<https://celehs.github.io/NILE.html>).
- tagtog is an interactive web interface tool developed in 2014 by Cejuela et al. [191]. As of 2022, users need to sign-up (<https://www.tagtog.net/>) to their agreement before use.
- Stanford CoreNLP was developed in 2014 by Manning et al. [192] and is a suite (<https://stanfordnlp.github.io/CoreNLP/>) that allows for NER and other NLP tasks using tags from ontologies, however manual work is needed first to extract the tags from ontologies.

Despite all these tools, none utilise biomedical ontologies or medical terminologies and so lots of work is needed to apply ontological frameworks. This lack of ability in tools restricts research [124]. As previously discussed by Leonelli [124], ontologies are designed in such a complex manner that researchers may avoid using them and since tools do not exist to use ontologies easily, this excludes researchers who are not computational.

Patient-generated data

There is a lack of openly available data, specifically patient texts, for secondary research use, and if research exists there lacks information and detail in order to reproduce data. Additionally, available data can lack in reasonable size (either too small or too large) introducing textual bias [183]. Bias can also be present due to the issue of multiple texts for a single patient [183], and whilst methods exist to overcome this (merging or removal of irrelevant texts), all texts can be valuable for developing a gold-standard dataset.

⁹as of March 16, 2022.

To extract valuable patient knowledge from unstructured text, NLP is required [165]. Patient voice can be in the form of PROMs (e.g. speech transcripts), EHRs (although mentioned that clinician-generated sources may be limited in patient experiences, a clinician may include notes about the patient) and social media [183]. Despite these sources, social media texts (and similar) do not perform well in classification models due to many reasons (noise, bias, unknown patient-preferred terms), however, there are possibilities in combining clinical and patient data which can advance research [183].

Accessing and extracting information from unstructured text is difficult as texts are unlabelled and noisy: adapted preprocessing and query methods are needed [67, 77]. For text mining efforts, having terms and their synonyms are valuable, e.g. using only clinical terms means we may miss sentences with a synonym (lay) and so less data is available for analysis [127]. NLP can be used for synonym curation methods, which includes statistical techniques, e.g. vectorisation (see section 1.6.5, page 56).

In 2014, Opitz et al. [67] conducted a study looking at terms that patients use on online forums for breast cancer. The authors extracted QoL-related text to better understand the important concerns in breast cancer patients, and importantly they note that retrieving lemmas improves text mining efficiency, and detect unknown words or synonyms. Opitz et al. [67] looked at co-occurrences in the text and noted some rare occurrences of terms, which suggested the effect on patients' lives [67]. The authors described the importance of patient synonyms for text mining objectives in order to better understand the patient and the effect of QoL on their daily lives. Another study looked at social media texts to assess discussions of bipolar disorder: Sekulić et al. [77] (2018) predicted users with bipolar disorder (training) and then carried out a feature analysis to determine which linguistic features are good predictors of the disorder (testing). Sekulić et al. [77] filtered out users with <1000 tokens (words) and extracted 5 million posts (from 4,619 unique users) between 2005 and 2018 from the source database which contained "I am diagnosed with bipolar", plus other paraphrased versions. Their control group were similar filters but ensured they were a group of mentally healthy people. Sekulić et al. [77] summarised that

terms which were predictors included “anger”, “sadness”, and “anxiety”, which increased the use of words related to sadness, while the control group uses more anger-related words.

Sentiment

Sentiment, or opinion-mining, is a method to gain understanding of a patient’s experience [25, 54] as we can see their perspective of their disease(s), treatments, and/or coping mechanisms. Sentiment analysis is one technique to understand QoL and done via computational techniques to process unstructured text [180]. Sentiment can be organised as categorical, nominal data: positive, neutral, and negative.

Sentiment analysis as a classification problem (supervised) learns on labelled data [25, 180] to predict unseen data. On the other hand, unsupervised ML includes rule-based techniques where particular terms have a particular score. There are numerous issues with both supervised and unsupervised sentiment analysis techniques. With supervised algorithms to predict sentiment score of text, the manual annotation of data can be a time-consuming task and annotation can be subjective. On the other hand, unsupervised techniques may not score new data appropriately due to the pretrained domain, e.g. common domains includes social media texts or movie reviews.

Sentiment analysis tasks perform best when terms are standardised, e.g. all terms or synonyms for “joy” are the same format [72]; and Denecke and Deng [25] discussed how a patient using the term “severe” rather than “slight” which when picked up by sentiment algorithms can make a difference in the QoL context, e.g. pain.

There have been numerous investigations into sentiment of individuals across numerous types of data sources. A 2015 study looked at patient sentiment of those with kidney issues [193] using ontologies to annotate text. Burckhardt and Padman [193] discussed the drawbacks of using biomedical ontologies due to them being medical-based and do not retain the lay-synonyms for text mining. In their conclusion, the authors stated how there was an increase of posts over time, however, not all information was accurate [193].

In 2016, Gkotsis et al. [194] looked into social media posts from Reddit to investigate 16 mental health topics from users. Reddit is a social media platform of networked communities (“subreddits”). Using various NLP methods, the authors discussed how many symptoms and specific subjects were shared across the 16 mental health topics, e.g. many conditions. In their sentiment analysis, using a rule-based technique based on other social media domain, Gkotsis et al. [194] concluded that negative sentiment was very high across the data. Then in 2017 the authors conducted a follow-up study with an analysis to classify posts which included one of 11 mental health disorders [195]. The study initially included some manual annotation to ensure posts were relevant for the disorder theme and concluded that classifying user-generated posts to characterise a group of illnesses can target interventions [195].

Lim et al. [196] also used a rule-based sentiment analysis in 2017 to look at infectious diseases, mined social media, and observe the relationship of the disease to symptoms. Authors discussed the most frequent terms and how the analysis potentially displays the disease within a short time-frame [196]. In 2018, Zhang et al. [197] looked into sentiment analysis of Twitter posts. Twitter is one of the most popular social media platforms with a rich source of real-world insight into the medical domain and patient behaviours. Zhang et al. [197] studied posts (“tweets”) comparing chemotherapy-related discussions of both patients and clinicians to assess their own perceptions. They performed sentiment analysis, using a pre-trained model and stated the accounts of clinical professionals were neutral regarding the side effects of chemotherapy. Whereas in patients, nausea was positive and fatigue was negative [197].

Davcheva [54] conducted sentiment analysis on a public forum, looking at mental health to shed light on various coping mechanisms. Authors noted that the average anxiety sentiment was more positive when discussing therapy; ADHD¹⁰ with exercise; OCD¹¹ with therapy; bipolar disorder with meditation; and more. However, autism had

¹⁰attention-deficit/hyperactivity disorder.

¹¹obsessive compulsive disorder.

a negative relationship with meditation; PTSD¹² with house pets; and more [54]. Finally in 2020 Gkotsis et al. [68] looked into the internet usage and social media content of care givers specifically topics relating to dementia. The authors first text mined 2101 posts using terms which included “suffer”, “diagnose”, and similar; then followed with a manual annotation. Gkotsis et al. [68] concluded that many users were females, aged 20-40, and posts related to their parent or grandparent, specifically a female relative. Additionally, topics that were discussed included advice relating to legal support [68].

1.6.5 Semantic similarity and vectorisation

Semantic similarity a measure of relatedness between two entities by the similarity in meaning [132]. A computer being able to extract semantics, ambiguity¹³, or understanding grammar, results in NLP algorithms being able to understand a lot of important information from sentences [127, 180, 181].

Semantics can be calculated via Resnik’s semantic similarity measure [198, 199], developed by Resnik [198] in 1995. Resnik relies on node relations and information content: in the context of this thesis Resnik semantics investigates relationships between two phenotypes to determine the similarity via the information they share [200]. Each phenotype is calculated based on occurrence frequency. Some issues with Resnik’s similarity include not being able to distinguish the information between two phenotypes due to similar information content [200].

We can also calculate semantic similarity through representing words in a numerical space; this is vectorisation. Vectorisation includes assigning a number to each term, each term in the vector space has a distance score from another term [166, 180, 201] - an embedding model consists of multiple vector space layers representing numerous iterations each term. These embedding models can be distributed and shared online for secondary research use. Vectorisation has been said to be both supervised and unsupervised ML due to not needing training or testing datasets, but the model still needs to learn from itself

¹²post-traumatic stress disorder.

¹³a term with more than one meaning.

[201, 202]. Figure 1.8 illustrates an example of plotting a 2D of an embedding model via DR consisting of medical terms: diseases, symptoms, and therapies grouped together.

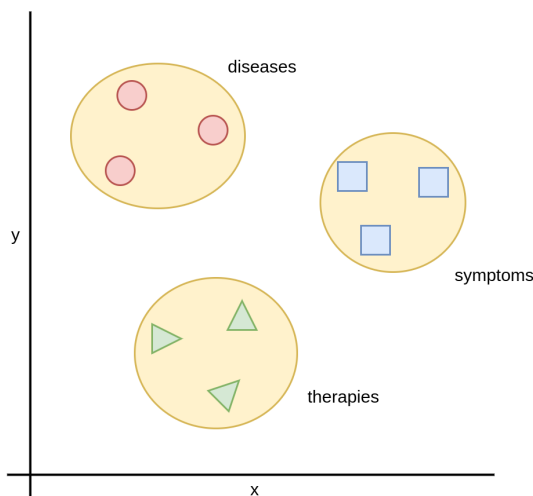


Figure 1.8: An example of plotting an embedding model in which distances between terms illustrate how similar terms are. We can expect patients to discuss symptoms in a similar way (e.g. synonyms or sentence structure), but a particular symptom being closer to a disease or therapy can indicate a strong association.

Word vectorisation was introduced around 1986 by Rumelhart et al. [203], who initially presented the idea of layered statistical networks with inputs at the bottom, intermediate layers, and outputs at the top [203], then the output state is determined. This initial idea was expanded over time into deep statistical models [180, 204]. Rumelhart et al. [203] introduced the idea of vectorisation stating that it can be applied to different tasks to show internal representations of data in a vectorisation space.

Vectorisation can be used as a form of synonym curation [205] or uncovering novel word associations. Building models that learn useful representations of clinical text is a challenge [206, 207], firstly one obstacle is set by the problem of unstructured text. In addition to bias in that building embedding models can produce demographic, domain, or intended task bias [201]. For example, some clinician-generated texts are long and their words are interdependent [206, 208], so these methods cannot capture the long-range dependencies needed to capture clinical meaning.

Apart from their direct use in ML algorithms [202], embedding models in the medical domain offer the potential for integrating data, such as combining word embeddings with

images or vectors derived from structured patient data. Multiple methods have observed that the size of the training window is a problem [204, 209] - in this, it can be an issue if source data is too large. Studies, including one from Gururangan et al. [210], discuss pretraining a model is the best methodology as the model consists of some prior context and multiple domains are more cost-effective: a model with a lot of data from various domains, in addition to fine-tuning, can struggle to encode the complexity of semantics, in comparison to a single textual domain.

Algorithms

There are multiple vectorisation algorithms that use vector spaces for developing embedding models [180, 201, 209] for conducting semantic similarity.

Word2Vec

The Skip-gram model was introduced in 2013 by Mikolov et al. [182] and is a method for learning vector representations of words from large amounts of unstructured text data and training does not involve dense matrix multiplications, making it extremely efficient [182, 209]. The Skip-gram model attempts to predict words based on others in the same sentence. Each word is an input with a projection layer. The model then predicts words within a certain range before and after the current word [182]. The continuous Bag-of-Words (CBOW) model includes all words projected into the vector space in the same position so their vectors are averaged [180, 182]. These vectors allow computers to react to the numeric structure to weave out the information based on text structure [180].

Word2Vec is an extension of the Skip-gram model with CBOW [209] by looking at frequently used words that cross a threshold to weave-out unimportant terms. Mikolov et al. [209] state that parameter selection is complicated and multiple tests are needed, such as the size of vectors, subsampling rate, or the size of the training window.

The use of embeddings offers the advantage of standardising the input to any machine learning algorithm and removing some of the complexity of the underlying data. Doc2Vec [211] for example, an approach closely related to Word2Vec [209], can capture the essence

of a paragraph of text in a numerical vector while approaches are able to compress and embed structured information [201], such as from EHR records. Onto2Vec was introduced in 2018 and essentially uses the biological knowledge that is represented in ontologies for vectorisation [135]. Onto2Vec learns high-quality vectors from biomedical ontology concepts [135]. Such an example is using it for similarity-based prediction of interactions between proteins that can be used to identify particular relations between proteins [135]. An extension of Word2Vec, called fastText, can handle out-of-vocabulary new additions with n-grams [201, 212].

BERT

Bidirectional Encoder Representations from Transformers (BERT) was introduced in 2018 by Devlin et al. [213] as a new form of word vectorisation. BERT was designed for deeper representations of unlabelled text through bidirectional training: left and right contexts in all layers of vectorisation models and so context-dependent. Authors argued that current other techniques of vectorisation restrict the ability to fine-tune models in later stages and so BERT has more parameter tuning and abilities [213], however, these abilities work best on smaller datasets. Moreover, BERT's abilities are not well understood in that there is redundancy in the information encoded and so the same patterns are repeated - this is due to over-parametrised models [214].

In 2020, Kinkead et al. [215] developed a ML model of random forests to compare with the neural network BERT to observe if they could automate the assessments of online health questionnaires. Their tool consisted of 15 questions for the lay-person to evaluate treatment options available to the patient in regards to online information via social networks or engines [216]. The authors work produced results indicating that BERT outperformed random forests, however not better than a manual assessment [215].

Clinical embedding models

Many embeddings have been produced from clinical text for secondary research use. Many of these embeddings are trained on EHR or clinical letters. A paper discussed

the importance of clinical note representations in 2018 by Boag et al. [217] in how clinical notes are unstructured and unlocking the full potential of them are challenging yet have untapped power. Clinical notes provide richer pictures of patients as they describe symptoms, diagnostic tests for diagnosing, current medical therapies, and patient history [217] and are valuable to all care staff as short summaries to capture patient events. Clinical notes use abbreviations, jargon, and have unusual grammatical structure.

In 2010, Chan et al. [218] used EHR plus imaging to create patient similarities using 14 similarity measures to determine similarities between patients with hepatocellular carcinoma (one of the most common types of liver cancer) regarding survival time. Authors used ML to propose algorithms, however, the stated more analysis was needed to better separate patient similarity pairs [218].

In 2016, the embedding model DeepPatient was introduced by Miotto et al. [219], who used structured information from EHR records and assessed the probability of patients developing a particular disease(s). Each disease included needed at least 10 patients with that disease and they manually removed diseases that could not be predicted due to social behaviours (e.g. HIV), external causes (e.g. injuries, poisoning), or general (e.g. “other form of cancers”) [219]. These steps were taken to decrease training sample size and uninformative patients. The authors trained their model on 200,000 different patients with at least five records and looked at 78 disease predictions. Their receiver operating characteristic (ROC) curve scores ($>60\%$) included “cancer of testis”, “sickle cell anemia”, and “cancer of prostate” - in contrast with some diseases which lacked in predictive power ($<60\%$): “hypertension”, and “diabetes mellitus without complications”, which authors stated could be due to diagnosis methods [219].

One embedding model source based on clinical letters is ClinicalBERT, introduced in 2019 by Huang et al. [206], built from 100,000 MIMIC-III [220] clinical notes of intensive care unit (ICU) patients. ClinicalBERT learns deep representations of clinical text, which can uncover clinical insights and provides a useful resource for examining clinical relationships, secondary research use, and semantic similarity [206]. ClinicalBERT being

pre-trained on clinical notes provides significant clinical value [206] as they are an insight into the clinicians working in fast-paced intense environments, who need to make concise decisions under time constraints, adding to a clinician’s workload [206].

Medical Information Mart for Intensive Care III (MIMIC-III)

A set of patient records from a large database, freely available, consisting of 38,597 ICU critical care patients between 2001 and 2012 [220], with a total of 58,976 unique hospital admissions.

Huang et al. [206] process 100,000 clinical notes to create an embedding model and assign risk scores to patients for an evaluation study of whether the patient will be readmitted within 30 days. They acknowledge the limitations of building models within a clinical domain due to word dependencies. Each clinical note is input into the model, which represent a collection of tokens, they performed using a cross-five-folds validation for independent runs [206]. ClinicalBERT presents that when compared to BERT, it performs better with clinical sentence prediction: highlighting the need for models tailored to clinical domains. When comparing to a set of 20 medical term pairs which clinicians rated similarity [200], ClinicalBERT scored 0.670 with BERT and 0.553 with Word2Vec [206]. ClinicalBERT, when using DR, they present a cluster with heart-related terms, showing the model captures semantics [206].

Other BERT models include, SciBERT [221], ClinicalBERT¹⁴ [222], and BioBERT [223]. SciBERT was introduced in 2019 by Beltagy et al. [221] and was trained on a random sample of 1.14M full-text papers from Semantic Scholar for a large model of labelled scientific data, with the purpose of NLP tasks. The other ClinicalBERT, developed in 2019 by Alsentzer et al. [222], was aimed for more generic clinical text, specifically discharge summaries, which uses 2M discharge notes in the MIMIC-III database. BioBERT, by Lee et al. [223] in 2020, was pre-trained on PubMed abstracts and PubMed Central (PMC) full-text articles, fine-tuned for biomedical text mining tasks, such as NER, question answering, and relation extraction.

¹⁴not to be confused with Huang et al. [206] ClinicalBERT.

Another Word2Vec model based on clinical, natural language was recently created, yet not available, by Lau et al. [60]. Lau et al. [60] measured how clinical language is semantically similar to “Ceiling of Treatment” and how it relates to prognosis and discovered the contextual meaning in regards to “end of life”. The authors used Word2Vec for n-gram observations and then given a set of root n-grams (e.g. “ceiling of care”, “withdrawal of care”, or “end of life care”), the top 10 most similar n-grams to each of the roots were collected based on the cosine distance between embeddings, these were presented to healthcare professionals [60]. Lau et al. [60] presented the ways clinician, natural language discuss patients in these scenarios from EHR. A final note from Lau et al. [60] was that they noted how clinicians reuse many texts (“copy and paste”) or are unnecessarily verbose.

Lay embedding models

Retraining clinical models, consisting of a standardised clinical language, with patient natural language domains, can bridge two gaps. The patient natural language domain is underutilised [3]. The patient voice, expressed over social media, forms an invaluable resource for understanding how patients discuss their conditions in a manner that is not typically captured in clinical related notes; patients are the experts of their experiences [3, 15, 31]. There are some patient biomedical embeddings available, however, they are trained in their own way and so cannot be systematically compared to clinical text.

CC BERT, was developed from “Chief Complaints” [224]. Chief Complaints are reasons why a patient went seeking medical care. This model was derived from a set of patient-generated texts but included nurses choosing the most structured complaint patient text and only choosing text which were appearing more than 10 times. In their study, they stated that there were only 795 unique patient complaints and noted their dataset for inclusion was relatively “small” [224]. Whilst CC BERT was developed using texts which appeared more than 10 times (so well-known associations were stronger) the lack of including under 10 complaints could possibly be excluding important associations.

Additionally, these complaint texts under 10 could be relevant with patient-preferred terms and so there could be a loss of power. Overall the work from Valmianski et al. [224] demonstrated that embedding models, from patient-generated text, performs well when using a clinical pre-trained model - however, the data is not accessible.

BioReddit was created in 2019 by Basaldella and Collier [225]. Basaldella and Collier [225] chose 68 medical subreddits (topics) from Reddit, gaining approximately 800,000 user posts between 2015 and 2018. These topics include general health (“r/AskDocs”) and disease-specific (“r/cancer” or “r/asthma”). A subset of these texts is available as the corpus, COMETA [226], with approximately 20,000 posts. They generated embeddings and chose models which were popular, better performance on relative low resource requirements [225]. To evaluate the BioReddit embeddings, the authors had professionals manually annotate 4,800 threads from a forum, HealthUnlocked using NER tags, such as Anatomy, Disease, and Procedure. Their results show that patient-derived texts improve NER NLP tasks and embeddings trained on biomedical scientific literature are not guaranteed to be effective, and so retraining is valuable. Finally, the authors note the importance of future work to compare to non-user generated scientific data [225].

1.6.6 Machine learning clinical advances

ML has already been employed to advance clinical practice in various ways. Many studies have conducted unsupervised clustering on patients with various diseases, with multiple algorithms, in order to seek new insight into biomarkers or underlying relationships [88, 91, 177, 227–231]. Studies have looked into cardiovascular diseases [177, 227, 228, 232], asthma [91], imaging [229], and blood markers [88, 230] to reveal novel correlations. These studies note various features which characterise clusters, such as WBCs [91, 230], by most healthy or at risk [228], genetic sex [229], or external factors [177]. Other studies reflect patient sentiment on clusters using K-means to observe correlations [231], such as inclusive features or other external, e.g. medication. Of these studies, K-means and hierarchical are both most commonly used, with some studies only focusing on one, whilst others

use both algorithms. Some studies have revealed novel biomarkers, such as depression [229, 230], or the relationship between inflammation and associations [177, 230], such as inflammation with kidney stress compared to the relationship with cancer. Other unsupervised techniques have been done, such as DR analysis with PCA to observe blood markers as diagnostic tools [52, 122].

Supervised studies through clinical trials or associations have also revealed biomarkers [22, 23, 95, 115], using regression analysis. Many go as far back as 1995 in that signals for IBD were revealed, including CRP [95] or a history of smoking [115]. Some include regressions which use sentiment as the endpoint to observe mental health indicators [23] or relationships with symptoms [22], such as fatigue. Using ML for classifying user-generated text can characterise cohorts of patients who have a particular disease [185, 195], which can improve understanding and target specific interventions. Specifically, in patient-generated text, ML has been used to look at the rare occurrences of terms for synonyms [67]. Additional synonyms in vocabulary improves NLP tasks, including text mining [193]. Many NLP annotation tasks [233] have built upon a foundation for computers to better understand semantics and clinical domains.

In 2001, Turney [234] investigated automatically recognising synonyms with unsupervised ML by measuring the similarity of pairs of words from student examinations: their algorithm achieved a score of 74% and the latent had a score of 64% [234].

Sentiment analysis has previously been conducted in clinical work [25] looking at sentiment relating to infectious diseases posted by clinicians via social media [196]. Other studies have looked into patient sentiment in regards to particular diseases, such as infectious [196], mental [54], and chronic BMI [27], to observe the relationship of patient sentiment to the medical term of interest and deeper associations to better understand a patient's opinion on their health. Whilst sentiment analysis can be conducted on patient text, clinician-generated professional text has also been demonstrated to understand the side effects of chemotherapy [197], such as which particular symptoms were written about more negatively and can be further investigated to improve understanding of QoL and

develop better therapy plans or better understand coping mechanisms [54].

With vectorisation, very few embedding models have been released for secondary research [219], some developed from EHR (structured data), imaging data [218], research articles [223], clinical letters [60, 206] and discharge summaries. Much fewer are openly available derived from patient-generated texts, one includes documents based on patients describing why they sought medical care [224]. All forms of embedding models can observe semantic similarities and co-occurrences in various medical terms of interest, such as cancer [218] or particular phrases [60], in order to apply to NLP tasks (e.g. NER [223, 224]) and prediction studies, such as predicting re-admissions to hospitals [206]. Vectorisation has been further exploited with clustering on the numerical vector space [235, 236], in addition to using clustering for synonym curation [205] - both to observe clusters of particular terms and deeper associations.

1.7 Challenges and research questions

This chapter has introduced the main components covered this thesis; inflammation and the patient voice. This background research sets the scene for the gaps in existing work for investigating the patient voice. Existing ontologies do incorporate synonyms, however there is a lack of synonyms related to the patient vocabulary and whilst there has been work to combat this, there is minimal work done to better utilise and integrate them in biomedical research. Moreover, work has investigated chronic inflammatory biomarkers and separately patient QoL of chronic diseases, there is insufficient understanding how these biomarkers interlink with the patient voice.

I will explore the patient voice pertaining to two inflammatory diseases, Uveitis and Inflammatory Bowel Disease (IBD). Both are an IMID and often undifferentiated. This thesis investigates patients' chronic inflammatory condition influences quantitative sentiment via NLP, semantic characterisations, and clustering.

1.7.1 Lack of lay synonyms mappings to clinical terms

In recent years, there have been articles calling for the use of natural language in clinical practice rather than strict, formalised clinical terms [12, 33]. This can help clinicians become more familiar with patient-preferred terms for improved communication standards, and overcome the issue of patients misunderstanding medical concepts [11, 30, 31, 34].

Despite classification and sentiment analysis tasks benefiting from standardised synonyms [72], we need to first understand the synonyms patients are using. Patient-generated online text is relatively inaccessible for extraction and analysis because they are unstructured. Ontologies have demonstrated success in addressing challenges presented by unstructured text data [125, 130]. In prior work, ontologies have demonstrated the improvement of understanding diseases through text mining ontological phenotypes/classes [237], and various disease networks have been explored to present meaningful clusters based on semantic similarity of diseases [46, 238]. However, ontologies are lacking patient-preferred terms, impeding the ability to successfully text mine and analyse patient sentiment for an insight into QoL. Formally encoding and integrating patient-preferred lay synonyms of formalised and standardised clinical terms within an ontology [12] may improve our ability to classify patients in meaningful disease groups [130].

Many ontologies are designed in such a complex manner that researchers may avoid using them and the lack of tools excludes researchers who are not computational [124]. Many ontologies are not easy-accessible via repositories [124] and the lack of ability to contribute to an ontology’s development may restrict researchers. These large scale biomedical ontologies are updated in their own time-intervals, in addition to the possible rejection of suggestions. Medical terminologies are more strict as they are developed and maintained by governing bodies, e.g. ICD maintained by World Health Organization (WHO) [153] and Read codes by the NHS. There has been work previously discussing ocular IMIDs in ontologies: Sergouniotis et al. [131] discussed how HP and ORDO are two ontologies that can make a foundation for ocular IMIDs, this was one of the first efforts to scale for providing terminology standards for the rare eye disease community.

However, after working closely with domain experts in ophthalmology, it was revealed that ocular IMIDs are not standardised and there are an overwhelming number of contradictions, both structural and conceptual, in current biomedical ontologies and medical terminologies. Whilst the biomedical ontologies and terminologies include ocular IMIDs concepts relevant to their domain, none comprehensively define all the important clinical concepts needed for meaningful application in relation to a particular disease and its presentation, investigation, diagnosis, management and impacts. The domain experts described many relationships defined in the ontologies were in some cases are not 100% true for ocular IMIDs, and that they needed to be corrected - some limitations are described in Table 1.1. These limitations need to be overcome and be a better standard for research.

Ontology	Limitations
UBERON	- pupil part of some iris - iris part of some uvea
UBERON	- chorioretinal region part of some eyeball camera-type eye
DOID	- anterior uveitis subclass of some panuveitis - posterior uveitis subclass of some panuveitis

Table 1.1: Structural issues in current biomedical ontologies in regards to ocular immune-mediated inflammatory diseases, as highlighted by domain experts.

A first example in HP that intermediate uveitis HP:0012124 has no subclasses [5] and most disease associations were missing. In DOID the only subclass of posterior uveitis DOID:12574 is birdshot chorioretinopathy DOID:0111079 [141]. ORDO was designed for rare diseases and so was more developed in disease associations [143], however key associations were missing. Domain ontologies can cover a too-broad topic, for example, UBERON includes other species [142], this is where application ontologies can be useful [33, 124] and retaining the most important entities [149].

Compared to medical terminologies, similar issues are evident: Read codes are specifically aimed for GPs [154, 155] and lack detail for rarer ocular IMIDs. Despite variations of uveitis existing in ICD-10, such as iridocyclitis H20 or tuberculosis of eye A18.5, the concept alone for uveitis is absent [153, 160], meaning EHR fail to be complete and the absence of descriptions and imprecision in coding means users can interpret uses of ICD-10 differently and so their analysis and reproducibility is affected [160].

Finally, despite some existing biomedical ontologies containing layperson terms for ocular IMIDs, they are lacking. Specifically, in HP synonyms are annotated as axioms, which is not computationally correct: they should be annotated as metadata synonyms to make ontologies easier to manipulate. Despite expansion of HP with lay synonyms [30], none exist for uveitis-related concepts. The concept posterior uveitis in both HP `HP:0012123` and ORDO `ORDO:280892` has “choroiditis” as a synonym [5, 143], which is a subtype of posterior uveitis. In DOID, posterior uveitis `DOID:12574` has the exact synonym “Uveitis, posterior”, but the structure is incorrect as DOID infers that posterior uveitis is a subclass of panuveitis `DOID:12030` [141]. In PATO [134], inflexible `PATO:0001545` contains definitions and synonyms, yet watery `PATO:0002408` is briefer and no synonyms. In SNOMED-CT, posterior uveitis `43363007` has no synonyms [156].

Whilst the ophthalmologists recognise a vocabulary set for uveitis [103], it is deficient in other areas of ophthalmology. Additionally, an application ontology exists for birdshot chorioretinopathy (a type of posterior uveitis) [239] but the domain experts stated it is not well represented. Furthermore, other ocular IMIDs have very little qualitative research undertaken [3] and no disease-specific PROMs has been developed for uveitis [26] plus other ocular IMIDs.

Therefore, the development of an ontology is needed to meet the standards of clinical practice for ocular IMIDs. To enhance the ontology of patient-preferred terms as synonyms, I will extract lay terms from online patient forum conversations using a novel methodology, the statistical term frequency–inverse document frequency (TF-IDF) technique for ranked terms, based on frequency in the corpus, to be curated and domain experts annotating those for inclusion. In clinical research it is common practice to annotate manually, which is extremely time-consuming and so manual synonym curation can be compared to the statistical TF-IDF computational technique. In order to demonstrate the ontology with lay synonyms, application involves NLP NER annotation will present term frequencies, and sentiment analysis will provide insight into patient opinions regarding their type of ocular IMID - all a perspective into the role patient fora play on the patient or carer.

This thesis will demonstrate why clinical terms and their patient-preferred (lay) synonyms should be used in patient voice text mining and NLP tasks. Ontologies are a useful tool when text mining a corpus, they aim to standardise vocabulary for a domain which can be useful to identify associations in text [125, 127, 135, 138].

1.7.2 Lack of open embedding models for interpreting of the patient lexicon

Semantic similarity has been proposed and shown to better understand the lexicon and vectorisation is a solution to bridging the gaps in semantic similarity for clinical practice [205–207]. Embedding models for secondary research use is lacking and although research exists that build these embedding models, there is an absence of open source models. Those which are available are mainly derived from clinical text and whilst valuable, the absence of patient-generated models hinders research. Publicly-available resources derived from online patient voices via embedding models has not been fully exploited yet.

Building embedding models from scratch involves many problems: firstly training a model is computationally expensive and time-consuming as one may need to fine-tune parameters. A corpus too large or small can hinder results: domains can be over- or under- represented introducing bias. Using unigrams and more n-grams (bigrams, etc.) is a good resource for classification tasks [72] and so looking into different n-gram semantic similarities can benefit semantic investigations.

To better understand semantic similarity, three objectives can converge upon a solution: using a preexisting model derived from clinical text as it is difficult to access clinical records, secondly training models solely consisting of patient-generated text, and finally retraining the existing clinical model with the patient text to build a model consisting of two domains. These embedding models will allow systematic comparisons in a unified fair experiment to investigate semantic similarity: to understand the value of the clinician and patient voice and how clinicians and patients use terms in similar or different ways.

The existing ClinicalBERT model is procured from clinical letters, easy to access, and provide thorough detail regarding input training size and model parameters [206]. One source of the patient voice will be derived from patient forum, and the other from social media: using two sources of patient-generated conversations can reveal similarities or differences of patient terminology from two sources.

All models will be applied to inflammatory diseases to understand the context of semantics. Models built from clinician-generated and patient-generated text, using the same unbiased building methodology will allow for semantic characterisations. The combined model could aid in the discovery of novel insights, in particular, the domain merges two valuable sources together and captures the focuses, concerns, and conceptions of both clinicians and patients.

1.7.3 Lack of understanding physiological markers adjacent to quality of life

Many biomarkers exist for inflammatory stress, including elevated CRP, WBC leukocyte subpopulations: neutrophil (acute) and lymphocyte count (chronic). Numerous ML techniques have been applied to clinical data to seek out novel biomarkers for various inflammatory diseases, including IBD. However many of these experiments lack detailed descriptions of analysis [168] and interpretation of results - there has been a need for better explanations as research can be under or over-estimated.

There is also a need for better diagnostic methods for IBD, shortening the diagnosis time, and recognising the patient sentiment and QoL. Quantitative sentiment and QoL patient feedback are underutilised and there is a lack of incorporation in biological investigations. Patient QoL and IBD needs to be studied to better understand the relationship of biomarkers and the patient voice. Blood tests are preferred by the patient, as they are less invasive, and investigations for new blood biomarkers via blood tests provides a solution for improving patient medical investigations as IBD needs novel biomarkers [94].

To meet this challenge, I can use the UK Biobank (UKB) [240]. The UKB is a large repository in the UK of clinical data. Between 2006-2010 recruitment took place and approximately 500,000 people aged 40-69 are stored in the UKB. Since its inception, the UKB has been consistently updated with new data [241]. When a patient is recruited to the UKB, they attend a recruitment centre where baseline measurements are taken, such as weight, or blood tests. Demographic, sociodemographic, and psychosocial details are available, such as genetic sex, if breastfed, mental health, and current lifestyle: diet/smoking. The UKB also includes synchronised hospital records of diagnosis [240]. All these types of data can reveal inner, hidden subgroups of IBD patients not otherwise known and QoL-related features can be reflected.

I will meet the challenge of linking blood biomarkers to the patient voice by using unsupervised ML clustering to investigate biomarkers in an IBD cohort: minor signals of abnormal blood measurements and the relationship with sentiment can improve clinician knowledge of IBD - results reflecting the literature is a foundation to rely upon potential novel results. Only a small number have investigated IBD in isolation, specifically UC, with small cohorts [52, 122, 123]. To identify other potential blood biomarkers and comprehend the relationship with the patient sentiment, I will develop an unsupervised ML clustering workflow, adjacently using various clustering algorithms and methods reflecting literature, to observe how clustering is computationally expressing patients with IBD and projecting QoL features to investigate underlying patterns.

The aim is for a clustering algorithm to separate patients with IBD, or not-yet diagnosed, to systematically compare clusters of patients based on inner groups and then uncover what characterises these groups, which could indicate disease activity and/or patient QoL. All a preliminary step into uncovering underlying relationships of patients with IBD however with future developments with clinical interpretation can it be fully exploited: we can understand what these sentiment signals mean for those with IBD, or those who will be diagnosed.

1.8 Summary of contributions

1. To address the first challenge and building a resource of patient-preferred (lay) terminology for clinical use, I created a computational, application ontology for ocular IMIDs, the first of its kind in ophthalmology. The Ocular Immune-Mediated Inflammatory Diseases Ontology (OcIMIDo) was developed using a clinical consensus document as a foundation and close collaboration with clinical domain experts to ensure a robust and standardised ontology for ocular IMIDs. OcIMIDo was enhanced with patient-preferred terms via domain experts curating important terms derived from the novel methodology and computationally faster (compared to manual annotation) statistical TF-IDF, which was applied to online patient forum conversations. Some patient-preferred terms overlapped with biomedical ontologies showing patients are using clinical terms, however, many lay terms were not present in these biomedical ontologies. Application of the ontology proved incorporating patient-preferred terms as clinical synonyms yields more data extraction from text mining tasks and so a more fruitful analysis. NLP annotation of term occurrences saw more discussions of particular disorders and treatments; with sentiment analysis revealing insight into the role a forum plays.
2. Multiple sources of the patient voice, including clinical letters, was explored with semantic similarity vectorisation. I used a preexisting embedding model derived from clinical letters, ClinicalBERT, to build models consisting of the patient voice using the same methodology, such as model-building parameters. This overcomes the main challenge of training models from scratch: avoiding computationally expensive tasks and potentially over- or under- representing a domain. I developed single-domain models, derived from patient-generated texts, one acquired from patient forum conversations and the other from social media content. Additionally, due to the lack of embedding models for secondary research use, an embedding model for secondary research use has been released for open source communities: combining two domains of clinical

letters and patient forum conversations. All models were systematically compared via multiple techniques to reveal semantic similarity differences of the clinician and patients, revealing characterisations of these generated texts, and priorities in the individual.

3. Finally, I explored the relationship between blood biomarkers and the patient voice via quantitative sentiment (QoL) feedback. Using unsupervised ML, I clustered blood markers of patients with IBD, from the UK Biobank. The cohort of patients with IBD was split into three time-stamped groups for systematic comparisons to observe the sentiment of those currently with IBD, those who were not-yet diagnosed, and those who were diagnosed at baseline but were also to be diagnosed in the future. Methodology included extensive clustering and cluster validation techniques: feature elimination, DR, clustering, and evaluation of clusters. Clusters unveiled insight into the presence of inflammatory stress and the relationship with QoL and how WBC, smoking, and sentiment could play a larger role in the IBD journey.

2 THE OCULAR IMMUNE-MEDIATED INFLAMMATORY DISEASES ONTOLOGY

The work in this chapter has been published as of 2021 in Computers in Biology and Medicine *Development and application of the ocular immune-mediated inflammatory diseases ontology enhanced with synonyms from online patient support forum conversation* by Pendleton et al. [242].

2.1 Introduction

Uveitis and other ocular immune-mediated inflammatory diseases (IMIDs) need to be standardised and there is an unmet need for capturing the patient voice. The patient voice allows a deeper understanding of patient sentiment in regards to their phenotype, complications, and medical therapies.

Sentiment analysis tasks benefit from synonyms [72] but firstly there needs to be a vocabulary of the synonyms patients are using. A resource of patient-preferred synonyms is online patient forum conversations, however these are inaccessible due to being unstructured. Ontologies have previously demonstrated great success in addressing the challenge presented by unstructured text [125, 130] yet previous investigations of current biomedical ontologies and medical terminologies have revealed gaps, both structural and conceptual.

This chapter covers the solution of building an application ontology for ocular IMIDs to standardise this clinical vocabulary as a foundation and act as a repository for enhancement of corresponding patient-preferred terms. To capture these lay terms, online patient forum

conversation is text mined to apply a statistical technique for ranked terms. Formally encoding and integrating patient-preferred lay synonyms of formalised and standardised clinical terms within an ontology improves text mining techniques and better insight into sentiment analysis of meaningful groups [12, 130].

Capturing patient-preferred terms aids in understanding the patient voice. To demonstrate the value of the new application ontology, I use NLP techniques to explore the patient voice and quality of life (QoL) of specific ocular IMID, complication, associated systemic disease, and treatment. A deeper knowledge of these and QoL can potentially lead to improved condition management and positive QoL interventions.

2.2 Data and methods

2.2.1 Building the ontology

A foundation for the ontology was derived from the Royal College of Ophthalmology Uveitis Clinical Dataset (RCOphth), a set of clinical guidelines and consensus document, developed by ophthalmologist specialist team Denniston et al. [243]. With the RCOphth, I curated a structure for the initial construction of the ontology from the numerous tables provided, Figure 2.1 (page 76) shows an example screenshot. From this figure I was able to determine that, for example, posterior uveitis occurs in the choroid or retina and has subtypes: choroiditis, retinitis, and more. The consensus document included tables for numerous categories, including types of sight-threatening complications, disease associations, signs and symptoms, medical therapies, and more.

The ontology was developed using Protégé [244], a free, open-source framework. Whilst I created the foundation, domain experts (ophthalmologists) searched through the Uberon Multi-species Anatomy Ontology (UBERON) [142] for curation of anatomical concepts for expansion and matching RCOphth. The experts additionally curated a list of relevant ocular IMIDs concepts to be included by scanning through biomedical ontologies, such as Disease Ontology (DOID) [141], Human Phenotype Ontology (HP) [5], Orphanet

Value	Definition: <i>Uveitis in which the primary site of inflammation is</i>	Subvalue
Anterior	Anterior chamber	
Intermediate	Vitreous	Pars planitis
		Non-pars planitis
Posterior	Choroid or retina	Choroiditis
		Chorioretinitis
		Retinochoroiditis
		Retinitis
		Neuroretinitis
Panuveitis	Anterior chamber, vitreous and retina or choroid	

Figure 2.1: Screenshot of a table from the Royal College of Ophthalmology Uveitis Clinical Dataset by Denniston et al. [243], showing types of uveitis, where the site of inflammation occurs, and subtypes.

Rare Disease Ontology (ORDO) [143], Phenotype and Trait Ontology (PATO) [134], and medical terminologies ICD-9 & ICD-10 [153], SNOMED-CT [156], and Read codes [155]. Many terms were deemed not relevant to ocular IMIDs by the experts, such as iris nevus HP:0000525 (eye freckle) or insect bolwig organ UBERON:0000970 (subclass of eye UBERON:6005805). I implemented this list of concepts and x-ref in the ontology, which the ophthalmologists then reviewed the ontology to ensure the content and information is a high standard representation of their domain.

Some limitations previously mentioned (Table 1.1, page 67) were fixed for inclusion in the ontology. Additionally, the domain experts brought to attention that many relationships that could be defined were actually not 100% true and so it was decided to only include relationships in the ontology that were a 100% true case - these relationships definitions were extracted from Relation Ontology (RO) [245], Basic Formal Ontology (BFO) [246], and more. The axioms were automatically populated based on the structure, or from expertise, such as equivalent or disjoint classes, (e.g. high intraocular pressure “EquivalentTo” ocular hypertension).

With an ontology completed, the expansion of patient-preferred terms was to be included.

2.2.2 Patient-preferred terms

The domain experts had clinical-research funding from Olivia’s Vision (OV) (<https://www.oliviasvision.org/>), a UK-based charity, established in 2010 as a support platform for patients (or care providers) who have uveitis. With this funding relationship, the domain experts had access and permission to download the charity’s online forum.

To identify patient-preferred terms from the OV forum for inclusion in the ontology, I used a novel, semi-manual curation technique: splitting the text into training and test sets for validation of term frequency. Although the data was split into training and test sets, it should be highlighted that this methodology was for validation purposes and no ML algorithm was implemented: instead NLP and statistical techniques were used.

Whilst treating each thread as a unique observation, the equal frequency binning technique was used: threads were divided into three size classes, based on the number of posts in each and the stratified training set was generated using a random 80% of threads from each bin, with the remaining 20% in each bin as the test set. Using the Natural Language Toolkit (NLTK) [247] in Python [248], each thread in the training set was formatted with the following steps: lowercase, removal of stop words, and stemming for standardisation. See the pseudo-code example in Listing 2.1 (below).

```
1 import json
2 data = dict(text.json)
3 for thread, listposts in data.items():
4     for post in listposts:
5         regex('[^a-zA-Z0-9]+', ' ', post)
6         wordtokenised = post.lower().split()
7     for word in wordtokenised:
8         remove stop words
9         stem words
10 if tf-idf first run:
11     continue
12 else:
13     remove ontology terms (including synonyms)
```

Listing 2.1: Coding example of the text preprocessing steps for cleaning the forum.

To extract all informative terms from the training set, a statistical ranking, text feature extraction function was used: term frequency-inverse document frequency (TF-IDF) using the Scikit-learn [249] package. Every word in the training set is ranked based on frequency across the corpus and then overall ranking tells us how “important” it is - words with a higher score represent a stronger importance (Appendix Equation A.2, page 191). With this, users (or in the context of this chapter the domain experts) can learn and interpret important terms [166, 180]. With the output of ranked terms, domain experts went through the list looking at only those scored above 1.0 and deemed some terms useful in terms of ocular IMIDs. These useful terms were either annotated as a synonym with a class mapping, or annotated as a new class. All these annotated terms were included in the ontology and added to the vocabulary.

Myself and the team of experts proceeded in an iterative way,

1. After the TF-IDF, all ontological terms in addition to the newly added synonyms from the previous TF-IDF, were removed from the training set.
2. I reran TF-IDF in order for the words to be reweighed. These newly reweighed terms were again sent for review to the domain experts for another observation for relevant synonyms for inclusion in the ontology.

These two steps were repeated until the domain experts identified no additional informative terms. An important thing to note is that the domain experts manually consolidated misspellings into correctly spelt terms during the evaluation of TF-IDF metrics at each stage, with both corrected and misspellings being included in the ontology.

2.2.2.1 Validation

To validate this process of using TF-IDF for synonym curation, three main tasks were conducted for comparison,

1. First the TF-IDF was performed on the OV test set, which domain experts compared to the training set to identify if any “important” terms were missed - whilst timing this activity.
2. Secondly, two other ophthalmologists, who were not involved in the development of the ontology, independently spent a timed 60-minute period reviewing the forum conversations. The two ophthalmologists manually highlighted as many words of

relevance to ocular IMIDs as possible. All of the terms curated from the two ophthalmologists were compared to the terms curated as synonyms manually identified from the TF-IDF. The time taken to curate terms by-hand manually against the time taken to curate terms from a TF-IDF were also compared. Additionally the times were extrapolated to obtain an estimate of how long it would take to curate synonyms from the entire dataset.

3. Third, a TF-IDF was performed on another online uveitis-based support forum, USA-based Uveitis Foundation (<https://uveitis.org/>) which the domain experts gained permission. Uveitis Foundation (UVE) is another support organisation for the treatment and education of ocular IMIDs.

2.2.2.2 Annotation

The frequency of ontological terms across both OV and UVE fora sets were compared to observe geographical differences. Annotation was conducted using the Stanford CoreNLP [192] suite with the **RegexNER** annotator to tag named entities (NER) with the appropriate concept identifier, for example “blurred vision” and synonym “blurry” have the same IRI.

2.2.3 Sentiment analysis

Using TextBlob [250] and VADER [251], natural language sentiment scores were obtained for each post in OV and UVE. Both TextBlob and VADER are open source tools that perform sentiment analysis and other NLP-related tasks. Both tools use a lexicon-based method to estimate sentiment, with the advantage of not needing any training data as they are pre-trained models developed from NLTK [247], thus avoiding the time-consuming task of manually labelling data. Both are rule-based sentiment analysis techniques [180] meaning particular terms already have a score (as described in section 1.6.4, page 54).

Both TextBlob and VADER have been widely used in sentiment analysis of online public free text data, such as research looking into Twitter posts for discussions about chemotherapy [197, 252]. There is a lack of consensus regarding which tool is better in which context, with some studies favouring VADER (e.g. for social media data) and others favouring TextBlob [253].

With the results, an exploratory statistical analysis was performed using standard statistical software, Stata (release 13.1) [254]. The data was sorted by retaining the most positive and negative (25%) sentiment and dropped neutral scores, for each forum and sentiment tool independently. The data was explored via single and multiple logistic regression (binomial) models for the odds ratio (OR) of a score in the most negative 25th percentile, selecting a range of different clinically relevant predictor variables, such as ocular IMIDs subtypes, systemic disease associations, ocular complications, and systemic treatments, for illustration. The most parsimonious model was identified, which had the smaller value of the Akaike’s Information Criterion (AIC) and the Bayesian Information Criterion (BIC), using a stepwise backward elimination approach. A receiver operating characteristic (ROC) curve was used to measure the discriminative performance of the model, obtaining global p-values using the likelihood ratio test.

2.3 Results

2.3.1 Description of the ontology

The Ocular Immune-Mediated Inflammatory Diseases Ontology (OcIMIDo) is open-source (<https://github.com/sap218/ocimido>) and currently v1.2.0¹, it has been fully documented following the MIRO² guidelines [255]. Protégé’s Pellet reasoner [244] was used to ensure a coherent and consistent ontology.

Each concept in OcIMIDo was assigned a unique identifier (the IRI), such as uveitis `OCIMIDO:00213`. OcIMIDo contains 661 classes, with 210 classes extracted from the foundation document, RCOphth [243]. OcIMIDo has eight top-level classes, three of these included classification `OCIMIDO:00001`, complications `OCIMIDO:00003`, and therapeutic interventions `OCIMIDO:00004`. Top-level class, therapeutic interventions, contained types of therapy `OCIMIDO:00391`, with subclasses medical therapy `OCIMIDO:00020` and surgical

¹it should be noted that the ontology was expanded after the emergence of COVID-19 due to new findings.

²read the full documentation here: <https://sap218.github.io/ocimido/MIRO>.

therapy OCIMIDO:00021. In Table 2.1 the OcIMIDo class, x-ref, and synonym count from the different sources are summarised.

Source	Class count	x-ref count	Synonym count
RCOphth [243]	210	0	9
DOID [141]	3	105	0
HP [5]	67	158	0
ORDO [143]	3	63	0
PATO [134]	14	28	0
UBERON [142]	61	63	0
RO [245]	0	2	0
BFO [246]	0	2	0
SYMP*	0	9	0
SCDO [256]	0	1	0
ICD-10 [153]	0	107	0
ICD-9 [153]	0	25	0
Read codes [155]	0	170	0
SNOMED-CT [156]	2	398	0
OV	21	0	187
Expert Opinion	280	0	5
TOTAL	661	1131	201

Table 2.1: Summary of Ocular Immune-Mediated Inflammatory Diseases Ontology sources: class count, database cross reference (x-ref), and synonyms.

*Symptom Ontology (SYMP) <https://bioportal.bioontology.org/ontologies/SYMP/>.

Along with the 661 classes, there are a total of 1661 relationships (including axioms), which can be observed in Figure 2.2 (page 82), created with WebVOWL³ [257] with relationship count further detailed in Table 2.2.

Relationship	Count	x-ref	Example	Notes
adjacent to	62	RO:0002220	choroid is adjacent to the sclera	Anatomical descriptor
part of	104	BFO:0000050	choroid is part of the uvea	Anatomical descriptor (inverse always true)
has part				
characterised by	20	SCDO:0000662	sarcoidosis is characterised by granulomatous histopathology	Disease definition (unidirectional)
investigated by	84	None applicable	reiter's investigated by HLA-B27	Investigations included in RCOphth [243] (inverse always true)
investigation for				
is a	657	None applicable	oral is a route of administration	Subclass relation (unidirectional)
occurs in	57	BFO:0000066	retinitis occurs in the retina	Pathological descriptor (unidirectional)
treatment of	16	RO:0002606	anti-viral is a treatment of viral infection	(unidirectional)

Table 2.2: Examples of constructed relationship types contained in the OcIMIDo, with descriptions and x-ref.

There are a total of 2851 annotations in OcIMIDo. With 1131 as x-ref to biomedical ontologies and medical terminologies (n=700, of which 398 were to SNOMED-CT), and 187 being synonyms from OV (e.g. generic, proprietary drug names, and commonly used abbreviations). In Table 2.3 (page 83) OcIMIDo different annotation frequency (labels, comments, x-ref, sources from which data were extracted, synonyms, and sources

³it should be noted that as of date-accessed 30-11-2021 WebVOWL is no longer available, however as of date-accessed 01-03-2022 was available.

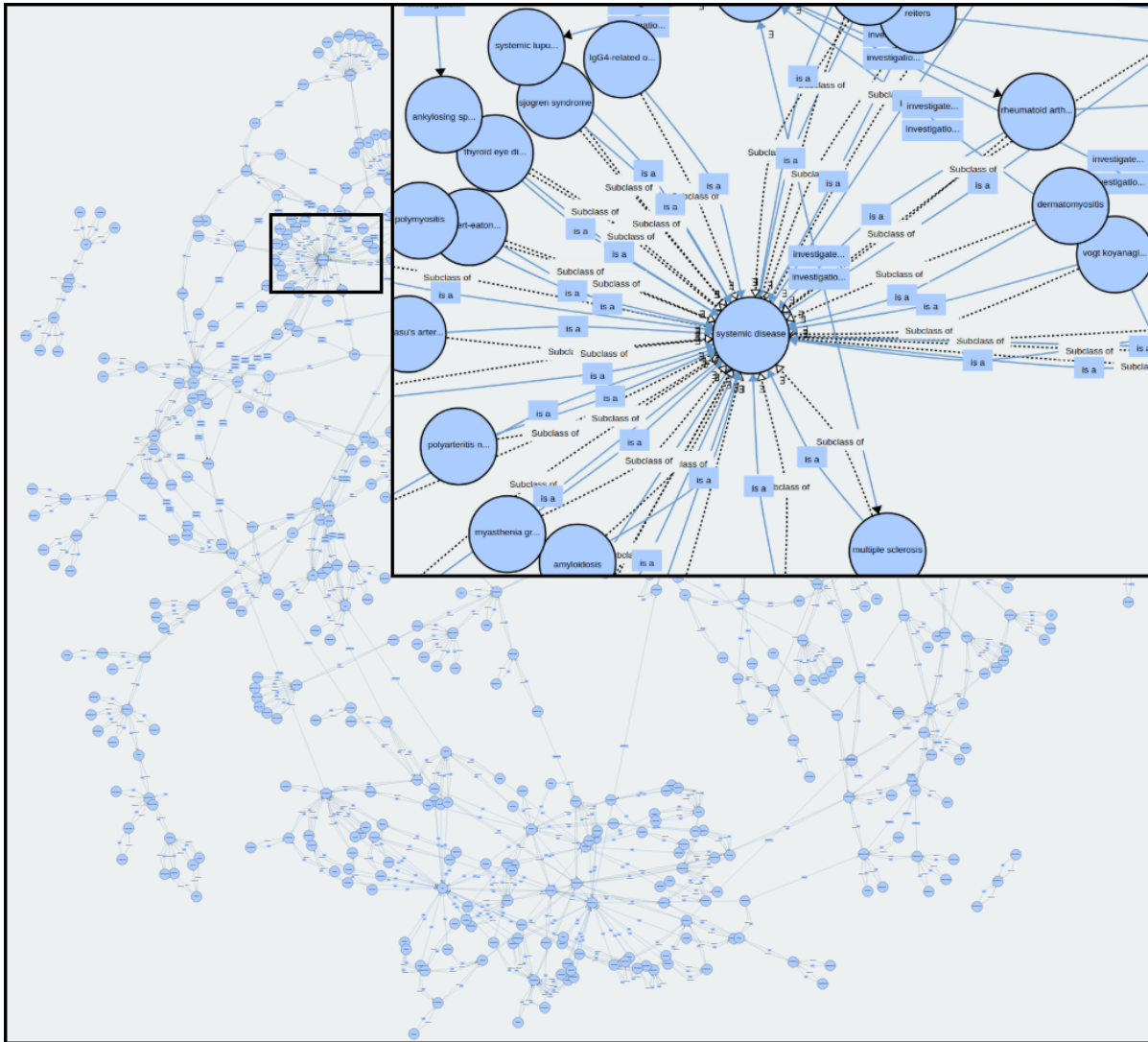


Figure 2.2: Graphical representation of the Ocular Immune-Mediated Inflammatory Diseases Ontology, with a close-up of systemic, non-infectious disease associations. Each node is a class (e.g. “multiple sclerosis”) and each edge is a relationship, with dotted lines defining “subclass of” and solid lines defining constructed relationships, labelled with boxes (e.g. “investigated by”).

of synonym extraction) are summarised, and illustrates these annotations with respect to one disease, Vogt-Koyanagi-Harada disease OCIMIDO:00108.

Some limitations mentioned in the literature review (Table 1.1, page 67) in which domain experts mentioned issues in current biomedical ontologies have been corrected in OcIMIDO, which can be seen in Table 2.4.

Annotation	Ontology Syntax	Count	Illustration of annotation
label	<code>rdfs:label</code>	623	Vogt Koyanagi Harada disease
definition	<code>rdfs:comment</code>	315	Rare disease that affects several parts of the body, including the eyes, ears, nervous system, and skin
x-ref	<code>rdfs:databaseCrossReference</code>	948	D0ID:12297, ICD10:H208
source of entity	<code>rdfs:extractedFrom</code>	381	RCOphth
synonym	<code>rdfs:hasBroadSynonym</code>	191	VKH
source of synonym	<code>rdfs:synonymExtraction</code>	140	Olivia's Vision

Table 2.3: Summary of the different annotations within OcIMIDo and illustration using Vogt-Koyanagi-Harada disease OCIMID0:00108.

Ontology	Limitations	Corrections
UBERON	- pupil part of some iris - iris part of some uvea	- pupil part of some anterior segment of eyeball - iris part of some anterior uvea - anterior uvea part of some uvea - uvea part of some anterior segment of eyeball
UBERON	- chorioretinal region part of some eyeball camera-type eye	- chorioretinal region part of some posterior segment of eyeball
DOID	- anterior uveitis subclass of some panuveitis - posterior uveitis subclass of some panuveitis	- anterior uveitis is a some uveitis - posterior uveitis is a some uveitis - panuveitis is a some uveitis

Table 2.4: Structural issues in current biomedical ontologies in regards to ocular IMIDs as highlighted by domain experts, with their corrections in OcIMIDo.

2.3.2 Enhancing ontology with lay terms

Forum extracted consisted of:

- The Olivia's Vision (OV) forum contained 2176 posts⁴, including 416 primary topic threads, with an average of 5.2 posts (standard deviation 3.9) per thread, ranging from 1 to 16.
- The Uveitis Foundation (UVE) forum contained 6855 posts⁵, split into 1488 primary topic threads, with an average of 4.6 posts (standard deviation 3.2) per thread, ranging from 0 to 10.

Together there were 1904 threads and 9031 individual forum posts.

The TF-IDF on OV training (334 threads/1731 posts/6032 words) resulted in an additional 208 new terms for inclusion in the ontology: 187 synonyms and 21 classes (see Table 2.1, page 81). See Appendix Table B.1 (page 194) for the synonym/class mappings in OcIMIDo, with an example of one of the synonyms in Table 2.3. It took less than 5

⁴March 14, 2019.

⁵October 6, 2019.

minutes to manually check the list of 653 ranked terms with scores ≥ 1.0 , and 30 minutes for all words. After five complete TF-IDF rounds, with the sixth identifying no additional terms. Many words were considered irrelevant by the domain expert, for example “time”, “day”, and “hospital”. Examples of words added as synonyms includes, “drop”, “steroid”, “humira”, “flare”, “treatment”, and “pressure”.

2.3.2.1 Validation

First, the TF-IDF was performed on the OV test set (82 threads/428 posts/2581 words), taking 0.6 seconds, which domain experts spent 13 minutes reviewing and comparing to the training set: to identify any missed. The domain expert found 46 new terms, of which 29 were not “new” synonyms and the other 17 are non-relevant. The distribution of a selection of tagged OcIMIDo classes appears very similar in the training and test sets, as seen in Appendix Figure B.1 (page 195). For example, 74% of posts in both sets mentioned “uveitis”, 27% “methotrexate”, and 22-23% mentioned “corticosteroids”. In total, using the TF-IDF approach to identify terms for inclusion in the ontology from all 2159 posts in OV, took approximately 3 seconds of computer time, plus approximately 43-minutes of domain expert time to scan all 8613 terms.

Secondly, two other domain experts, not involved in the development of the ontology, independently spent a timed 60-minute period manually highlighting as many words of relevance from OV. Clinician one identified 103 unique terms in 376 posts, 54 of these were already included as classes and 30 were synonyms found in the TF-IDF on the training set (19 remaining). Whereas clinician two identified 100 unique terms in 69 posts, 37 of these were already classes and 20 were synonyms found in the TF-IDF on the training set (43 remaining). Of the combined remaining 62 terms, only 10 overlapped, with only 6 additional terms were useful: “attack”, “bouts”, and “bloodshot”.

Third and final, a TF-IDF was performed on the full OV and UVE fora (approximately 12 seconds) in which the domain expert identified the only non-overlapping items were region-specific drug brand names, for example “xibrom” (drug used in North America).

Extrapolating, to explore and extract informative terms from all OV and UVE 9031 posts manually, it is estimated that it would take 24 hours for clinician one, 131 hours for clinician two, in comparison to a computer can take approximately 12 seconds for the TF-IDF to run and 2 hours for the domain expert to scan through 24,312 ranked terms. This illustrates that TF-IDF analysis, with expert curation of a ranked list of items, can be a much faster approach for synonym retrieval when compared to a purely manual curation approach, which is common in medical research.

Impact

Adding patient-preferred synonyms increased the number of posts identified to be relevant to the sentiment analysis on account of containing tagged classes. For example, using “methotrexate” with synonyms “mtx”, “mxt” (a common misspelling based on ranking from domain expert opinion), and “amethopterin” increased the post count by 80%, from 185 to 333; using “adalimumab” with synonym “humira” increased the post count by 3014%, from 7 to 218. See Figure 2.3 (page 86) for an illustration of some terms data increase from addition synonyms.

2.3.2.2 Annotation

Exploration of the 9014 posts from the two support fora discussions provided insight into aspects of inflammatory eye disease and its management of interest or concern to the patients using these two patient support fora, see Table 2.5 (page 87). Discussions of the subtypes of inflammatory eye disease differed between OV and UVE. For example, anterior uveitis was mentioned by 5.7% (n=124) OV posts and 8.4% (n=577) UVE posts ($p<0.001$), whereas optic neuritis was mentioned by no posts in OV but 1.0% (n=70) in UVE ($p<0.001$). Discussions of complications was frequent in both fora. For example, cataract was mentioned in 10.0% (n=218) OV posts and 7.1% (n=483) UVE posts ($p<0.001$), and glaucoma was mentioned in 3.7% (n=81) OV posts and 5.3% (n=363) UVE posts ($p=0.003$). Systemic diseases were also frequently mentioned. For example, juvenile idiopathic arthritis was mentioned in 4.5% (n=97) OV posts, and 1.5% (n=105)

Term	WITHOUT SYNONYMS		WITH SYNONYMS		
	Posts	Sentiment	Synonyms	Posts	Sentiment
Biologic therapies	Adalimumab	7 	1	218	
	Infliximab	64 	1	83	
Immunomodulatory therapies	Methotrexate	185 	3	333	
	Mycophenolate	11 	2	41	
Clinical phenotypes	Anterior uveitis	70 	1	122	
	Panuveitis	18 	1	32	

Figure 2.3: Illustration of increased text mining efforts with the addition of synonyms for a given class in OcIMIDo. For example, adding a synonym for panuveitis increased the number of posts mentioning this inflammatory disorder by 14.

UVE posts ($p < 0.001$); whilst inflammatory bowel disease was mentioned in 0.5% ($n=11$) OV posts and 2.2% ($n=152$) UVE posts ($p < 0.001$). Posts often included treatments being used or considered, and some differences may result from differences in treatment preferences and availability in the UK and USA. For example, adalimumab was mentioned in 10.0% ($n=218$) OV posts and 7.2% ($n=495$) UVE posts ($p < 0.001$), and mycophenolate mofetil was mentioned in 1.9% ($n=41$) OV posts and 5.2% ($n=356$) UVE posts ($p < 0.001$). Newer biologic therapies, including ustekinumab, tocilizumab, certolizumab, had few mentions in both fora.

2.3.3 Application into sentiment

The TextBlob and VADER methods yielded very different median sentiment scores in the full combined dataset, of 0.12 (IQR 0.01-0.24) and 0.77 (IQR 0.08-0.93), respectively.

OcIMiDo class	OV % (n)	UVE % (n)	χ^2 p-value	Most negative + positive 25th posts, Odds of post expressing negative sentiment OR (95% CI), p-value	
				TextBlob (n=4480)	VADER (n=4557)
clinical phenotypes					
anterior uveitis	5.7 (124)	8.4 (577)	p<0.001	2.3 (1.7–3.1), p<0.001*	1.3 (1.0–1.5), p=0.027
intermediate uveitis	1.8 (38)	5.8 (397)	p<0.001	1.5 (1.0–2.1), p=0.041	0.8 (0.6–1.1), p=0.145
macular oedema	1.4 (31)	2.1 (144)	p=0.046	1.3 (0.7–2.4), p=0.389	0.5 (0.3–0.7), p=0.001
optic neuritis	0 (0)	1.0 (70)	p<0.001	1.1 (0.5–2.2), p=0.895	0.9 (0.5–1.7), p=0.829
posterior uveitis	0.9 (20)	2.6 (179)	p<0.001	1.4 (0.8–2.3), p=0.263	0.5 (0.3–0.7), p<0.001
panuveitis	1.5 (32)	1.7 (113)	p=0.565	1.8 (1.0–3.5), p=0.070	0.6 (0.4–1.0), p=0.036
scleritis	0.1 (2)	2.7 (184)	p<0.001	1.9 (1.2–3.2), p=0.007	1.3 (0.9–1.9), p=0.207
associated systemic diseases					
ankylosing spondylitis	0.7 (16)	1.6 (109)	p=0.003	1.3 (0.6–2.5), p=0.515	0.8 (0.5–1.3), p=0.370
behcet's disease	0.3 (7)	1.3 (92)	p<0.001	0.7 (0.4–1.4), p=0.290	0.7 (0.4–1.2), p=0.214
birdshot	0.60 (13)	1.01 (69)	p=0.080	0.8 (0.3–1.7), p=0.516	0.4 (0.2–0.7), p=0.001
cat scratch disease	1.0 (1)	0.2 (8)	p=0.356	8.2 (1.0–67.5), p=0.050*	1.2 (0.5–2.7), p=0.702
inflammatory bowel disease	2.2 (152)	0.5 (11)	p<0.001	1.3 (0.7–2.3), p=0.351	1.0 (0.6–1.4), p=0.844
juvenile idiopathic arthritis	4.5 (97)	1.5 (105)	p<0.001	1.7 (1.0–3.0), p=0.052*	0.6 (0.4–0.9), p=0.017
multiple sclerosis	0.4 (8)	1.7 (119)	p<0.001	2.1 (1.0–4.2), p=0.049	1.0 (0.6–1.5), p=0.842
psoriasis	0.3 (7)	1.7 (113)	p<0.001	0.9 (0.5–1.9), p=0.813	1.1 (0.7–1.8), p=0.688
rheumatoid arthritis	0.7 (16)	1.1 (78)	p=0.107	2.8 (1.1–7.2), p=0.029*	1.2 (0.7–2.1), p=0.453
sarcoidosis	0.1 (1)	0.3 (19)	p=0.046	2.0 (0.4–10.7), p=0.437	0.5 (0.1–1.6), p=0.244
complications					
cataract	10.0 (218)	7.1 (483)	p<0.001	1.6 (1.2–2.2), p=0.003*	1.5 (1.1–2.0), p=0.018
epiretinal membrane	0.28 (6)	0.93 (65)	p=0.002	2.7 (0.9–8.5), p=0.089	2.4 (0.7–8.0), p=0.145
glaucoma	3.72 (81)	5.3 (363)	p=0.003	1.8 (1.3–2.5), p=0.001	1.3 (0.9–1.8), p=0.231
ocular hypertension	1.19 (26)	0.53 (36)	p=0.001	2.5 (1.0–6.4), p=0.063	0.2 (0.1–0.5), p=0.238
retinal detachment	0.6 (13)	0.47 (32)	p=0.451	1.0 (0.3–3.4), p=0.975	0.2 (0.1–0.5), p=0.001
treatments					
adalimumab	7.22 (495)	10.02 (218)	p<0.001	1.5 (1.2–1.9), p=0.001	0.9 (0.7–1.0), p=0.131
bevacizumab	0.23 (5)	1.27 (87)	p<0.001	1.1 (0.5–2.3), p=0.888	0.2 (0.1–0.4), p<0.001
infliximab	3.81 (83)	5.25 (360)	p=0.007	1.3 (0.9–1.7), p=0.118	0.7 (0.6–0.9), p=0.019
intravitreal triamcinolone	0.09 (2)	0.64 (44)	p=0.002	7.9 (1.0–64.4), p=0.055*	1.1 (0.5–2.4), p=0.742
methotrexate	15.3 (333)	13.6 (937)	p=0.056	1.5 (1.2–1.9), p<0.001*	0.8 (0.7–1.0), p=0.003
mycophenolate mofetil	1.88 (41)	5.16 (356)	p<0.001	1.7 (1.2–2.5), p=0.004	0.6 (0.4–0.7), p<0.001
oral prednisolone	6.2 (135)	4.8 (329)	p=0.010	2.2 (1.5–3.4), p<0.001*	1.3 (1.0–1.7), p=0.040
rituximab	0.09 (2)	1.15 (79)	p<0.001	0.7 (0.4–1.3), p=0.247	0.2 (0.1–0.5), p<0.001
topical steroid	6.0 (130)	3.3 (224)	p<0.001	2.0 (1.3–3.1), p=0.002*	1.5 (1.1–2.0), p=0.014
TOTAL	100 (2176)	100 (6855)			

Table 2.5: Frequency count of a selection of OcIMiDo class items in all 9031 posts from the two patient fora, with sentiment score, and chi-squared (χ^2) p-values.

*odds ratio (OR) from Multivariable logistic regression model, also adjusted for the first post (versus replies to a thread). Otherwise, OR are for single variable analysis.

Key: Olivia’s Vision (OV), Uveitis Foundation (UVE).

The median TextBlob sentiment score was 0.10 (IQR 0 to 0.20) for the 2176 OV posts, and 0.13 (IQR 0.02 to 0.25) for the 6855 UVE posts, with sentiment scores ranging from -1.00 (most negative) to +1.00 (most positive). The median VADER sentiment score was 0.69 (IQR 0 to 0.91) for OV posts, and 0.79 (IQR 0.15 to 0.94) for UVE posts. In both multivariable models the odds of a post expressing negative (25th percentile) language sentiment were significantly associated with the first post in a thread, compared to replies (OR 3.3, 95% CI 2.8 to 3.9, p<0.001 for TextBlob and OR 3.3, 95% CI 2.9 to 3.9, p<0.001 for VADER), and posts mentioning treatment with oral prednisolone (OR 2.2, 95% CI 1.5–3.4, p<0.001 for TextBlob; OR 1.3, 95% CI 1.0 to 1.7, p<0.001 for VADER).

Using 50% of the data from each forum, after splitting into quartiles of TextBlob and VADER scores separately, the ROC curve were compared for the multivariable models of the odds of a post expressing a negative sentiment using either TextBlob (AUC=0.64, n=4480 posts, Table 2.5) or VADER (AUC=0.65, n=4557 posts). Furthermore, using

data from these sentiment analysis tools, which yielded very different median scores, resulted in models with markedly differing significant associations with OcIMiDo class items. TextBlob median was closest to 0, whilst VADER scores near 1 indicating error towards a ceiling effect, this may be due to the data each were trained on or that VADER's method focuses on individual words and ignoring that word context in which it is used. Figure 2.4 shows the TextBlob sentiment scores associated with posts mentioning an illustrative selection of different classes (and their synonyms) within the ontology - VADER sentiment plots can be observed in Appendix Figure B.2 (page 196).

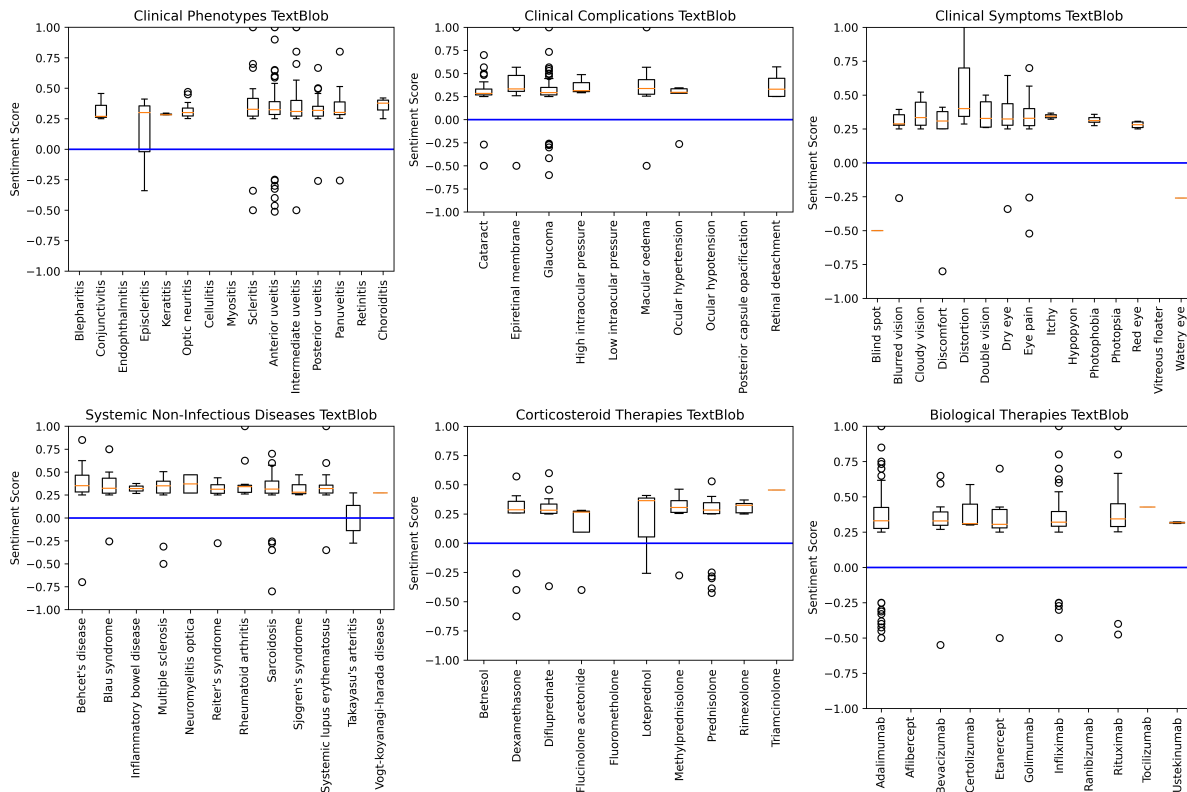


Figure 2.4: Box plots of TextBlob natural language sentiment scores for a range of clinical phenotypes, complications, symptoms, systemic non-infectious diseases, and treatments - combining data from Olivia's Vision and Uveitis Foundation; scores (not normalised) for each class (including synonyms) ranged from -1.00 (most negative language) to +1.00 (most positive language).

2.4 Discussion

In this chapter, I presented and validated a new application ontology, for ocular IMIDs: the Ocular Immune-Mediated Inflammatory Diseases Ontology (OcIMIDo) (version 1.2). OcIMIDo is freely available to the community as an open-source tool, via GitHub⁶ and BioPortal⁷ [147]. OcIMIDo organises 661 classes into high-level concepts, which includes disease activity, anatomical classification, clinical features, core investigations, therapeutic interventions and complications plus functional impacts. All of these superclasses were described from domain expert knowledge for the best structural representation.

As described in the methods, OcIMIDo contains only a small number of classes (other disease area-specific) from other existing well-known biomedical ontologies in general medicine, uncovered from the same domain experts. These classes have appropriate x-ref (Table 2.1, page 81). Furthermore, OcIMIDo provides an implicit axiomatic structure for data annotated with SNOMED-CT [156], ICD-10 [153], or Read codes [155], as these are included as x-ref in the ontology.

These biomedical and structured vocabulary classes alone enhance the value of OcIMIDo for searching and curating unstructured clinical data or publications. Clinician researchers in other medical specialities, such as rheumatology, gastroenterology, infectious disease, and neurology, may find OcIMIDo a useful resource for advancing the understanding of the inflammatory ocular manifestations of the systemic diseases they manage as ocular IMIDs have strong associations with infectious and autoimmune aetiology [26, 99, 100].

I have reported a novel NLP-guided synonym curation method to identify patient-preferred terms, with help from domain experts. The identified synonyms were included in the ontology as synonym annotations to enhance it. This addresses the unmet need to better align computational tools with clinical and patient-preferred language [3, 30]. These described uveitis patients are using specific terms to describe their conditions, abbreviations for medications, or misspellings which could be missed by computational

⁶<https://github.com/sap218/ocimido>.

⁷<https://bioportal.bioontology.org/ontologies/OCIMID0>.

techniques. Additionally, it is important that we discovered some terms that clinicians did not previously consider: future knowledge of these terms and how they are being used can improve clinical care.

The OcIMIDo-based annotations confirmed that both online fora provide a space for patients to discuss a broad range of issues relating to inflammatory eye disease. The differences that were present between the fora may relate to geographical groups or patients targeted: UVE is hosted by a USA-based charity and serves patients with a wide range of ocular inflammatory disorders specifically uveitis, however not limited to other disorders optic neuritis and scleritis; whereas OV is hosted by a UK-based charity catering especially to paediatric uveitis patients and so many users could be children or parents/guardians/carers aiming to find support about their child’s condition.

By illustrating that OcIMIDo, used alongside NLP approaches, facilitates standardised extraction for analysis of real-world unstructured online patient forum posts, it highlights how quantitative data can be obtained from this valuable, untapped patient-reported resource. There is increasingly urgent recognition of the need to better understand and integrate the patient voice in clinical care pathways, research priority setting, and the development of meaningful outcome measures [3]. It is proposed that patient support fora could play a particularly helpful role in identifying issues of concern and advancing patient-centred care for rarer diseases.

2.4.1 Implications

Sentiment analysis using both TextBlob and VADER revealed that first posts were significantly more negative than replies; providing insight into the supportive role that online fora play for the posting user or other user replies⁸. A brief manual review of posts and sentiment scores indicated that sentiment analysis at the level of an entire post provided only limited insight; individual posts frequently contained discussion of a number of disease presentations and treatments (multiple ontology classes), with both positive and negative

⁸e.g. a user may start a thread confiding in their worst fears about their ocular complications, and users who reply may offer positive support.

language sentiments expressed in relation to these.

In Figure 2.4 (page 88), the sentiment analysis box plots visualise positive and negative associations with OcIMIDo classes. In such that some ocular IMIDs had a slightly higher positive score than others (e.g. anterior uveitis or choroiditis); in complications cataracts were more negative; in symptoms distortion was positive; in systemic non-infectious associations Behçet’s disease is slightly more positive whilst Takayasu’s arteritis is more negative. Whilst these are all implications, we cannot fully understand the sentiment behind it without knowing the full context of the patient, e.g. age, other health-risk factors, or if discussing medication outcomes.

The exploratory regression models had a reasonable fit (TextBlob AUC=0.64, VADER AUC=0.65), given the lack of individual patient data. However, the multiple significant associations between the most negative 25% of post sentiment scores and various ontology classes differed comparing VADER and TextBlob, suggesting that these tools are not measuring sentiment in a concordant way. Results revealed significant associations between the most negative 25% of post sentiment scores and various ontology classes, including anterior uveitis ($p<0.001$), steroid treatment (via multiple routes of administration, p -values ranging from $p<0.001$ to $p=0.055$), methotrexate ($p<0.001$), development of cataract ($p=0.003$), and various associated systemic diseases (e.g. rheumatoid arthritis $p=0.029$). These associations align with a priori clinical expectations.

2.4.2 Limitations and future endeavours

Only a small number of clinical specialists with domain expertise were involved in the development of OcIMIDo. However having OcIMIDo publicly online, myself and the team of ophthalmologists anticipate that the ontology will be further revised and expanded: being openly discussed if any suggestions or structural issues are present. In relation to the ontology development, the Royal College of Ophthalmology Uveitis Clinical Dataset (RCOphth) consensus document [243], used as the ontology foundation, lacked detail on relationships. Whilst the data was available in tables (Figure 2.1, page 76) and we can

extrapolate a structure of subclasses, further relationship information in the RCOphth would have been beneficial.

Secondly, as a data resource, unstructured text has multiple limitations. It was discussed using spell-checking modules in Python when preprocessing the data, however, these dictionaries did not recognise many terms (e.g. medications) which were curated as synonyms. In future endeavours, low-scored TF-IDF terms could potentially be addressed by distance measure techniques to reweigh terms. Also, despite a novel NLP-guided synonym curation technique established, in future further developments this technique can be enhanced so that synonyms can be included in ontologies without domain experts assistance and with some matching algorithm.

Thirdly, there were limitations of using natural language sentiment analysis to illustrate an application of the ontology. Despite extracting each post individually without reference to usernames, some posts may not be independently and may depend on other context⁹. Although some discussions are mentioned in implications, the emergent clinical phenotypes suggested for each forum and modelled associations should be interpreted cautiously as they were purely exploratory, as suggested. Additionally, as mentioned the study revealed that the two commonly used sentiment analysis tools yielded very different sentiment scores. Further work is needed to explore the possible reasons for this - or possible future work into creating a supervised ML model of labelled data.

It is important to highlight the need for manual efforts: building of ontologies, annotation of synonyms, and checking sentiment scores - as they still play an important role in research for in-depth validation analysis. It has been illustrated the simple instantiating of ontology classes in an unstructured dataset, but additional tools will be needed for more nuanced analysis [258]. For example, a given class term, such as ‘methotrexate’, might be mentioned in different contexts, such as, ‘history of methotrexate use’, ‘current methotrexate use’, ‘allergy to methotrexate’, or ‘no prior methotrexate use’ - linking back to how some concepts in SNOMED-CT are mismatched in context [33]. One possible

⁹i.e. one patient may have posted multiple times, or replied a similar response on multiple occasions.

method to overcome this is using context disambiguation tools in Stanford CoreNLP such that we could determine information such as negation. However, since research into the online forum is still quite a new area of exploration, analysis is needed to observe how pre-trained models work on new unstructured, text data.

In the future, revisions and expansions of OcIMIDo will be conducted. Further defined robust relationships and axioms will be included, in addition to an expanded set of classes relating to advances in the clinical domain; including evidence-based investigations or additional disease associations. These new classes and inter-relationships will include extensive x-ref to appropriate biomedical ontologies and medical terminologies, such as SNOMED-CT [156] and the soon-to-be adopted ICD-11 [153]. The value of more descriptive axiomatisation is that it will further enable ontology-based analysis techniques, such as semantic similarity, which can be used to risk-stratify patients. Synonyms will also be further considered into classifying as exact (precise alternatives), narrow (more specific), broad (general), or related (associated terms), additionally appropriate layperson. These synonyms can be identified from a wide range of sources: crowd-sourcing applications or additional online patient support groups - all for different ocular and systemic IMIDs.

Further potential applications include exploring clinical phenotypes and genetic phenotypes in biobank databases [240], text-mining both open source and EHR data, performing cluster analysis to study underlying relationships in patients with a given disease, and developing ontology-guided predictive algorithms [259].

Developing OcIMIDo was part of a wider program of work by the research group of the domain experts, which includes the development of a semantic framework to fully capture health-related QoL. In future research, our team aims to conduct more advanced applications of OcIMIDo and QoL together, to improve clinical understanding of the impacts of inflammatory eye disease plus treatment on patients, and to identify novel content for inclusion in patient-reported outcome measures, thereby helping to overcome one of the key barriers to their development.

2.4.3 Conclusions

I have demonstrated that the development of an ontology for a specific disease area, enriched with patient-preferred synonyms, provides a potential solution to the challenges arising from the expanding volume of valuable but unstructured patient-reported text data from patient forum conversations. I presented the importance of these lay terms for text-mining and sentiment analysis. My approach to developing and validating an ontology, using online patient support fora as a data source to permit incorporation of the patient voice, is readily applicable to other areas of medicine.

One of the major fundamental components of this chapter is remembering, by Noy and McGuinness [129], that *there is no single correct ontology for any domain* and whilst OcIMIDo advances the nomenclature for ocular IMIDs, others may disagree: which is why having it publicly online benefits the medical community in that we can work together to improve these systems.

Summary

1. Addressed the challenge of the lack of standardisation for ocular IMIDs through the development of an application ontology, Ocular Immune-Mediated Inflammatory Diseases Ontology (OcIMIDo).
2. OcIMIDo was enriched with patient-preferred terms, derived from UK-based Olivia's Vision online forum, as synonyms mapped to their clinical term.
3. These mappings of the patient vocabulary to clinical counterparts provides a solution to the gaps in terminology.
4. Synonym curation overcame the problem introduced by unstructured text by a semi-manual novel workflow utilising statistical computational resources.
5. Computational methodology for synonym curation improves alternative clinical practices of manual annotation.

6. Some patient-preferred terms were not known to the domain experts including terms not in existing biomedical ontologies.
7. Text from geographically different locations included regionally-dependent lay terms, such as UK and USA medication brands, which must be considered when mining text from the internet.
8. This work shows the utility of including patient-preferred terms in text mining: we can capture more data for downstream analysis.
9. Application of annotation and sentiment analysis with a selection of OcIMIDo classes to forum conversation provided insight into the role these communities play for patients or carers.

The work in this chapter influenced the development of a new tool for NLP tasks with the ability to manipulate ontologies. It has been published as of 2020 in the Journal of Open Source Software *Jabberwocky: a toolkit for ontologies* by Pendleton and Gkoutos [260]. As of 2021, Jabberwocky uses spaCy [261], a state-of-the-art, advanced NLP tool.

3 THE PATIENT VOICE AS A CLINICAL NARRATIVE VIA EMBEDDINGS

The work in this chapter is open source for secondary research use: *Combined Ontology for Inflammatory Diseases*, Zenodo 2021 by Pendleton [262]; and *PatientINF*, HuggingFace 2022 by Pendleton [263].

3.1 Introduction

Semantic similarity builds upon the patient voice in understanding the sort of terms patients are using and why they are similar to another. During the previous experiment, a large number of terms were added to OcIMIDo from the patient forum conversations: vocabulary building on classes and synonyms. Many terms were not known to the domain experts. This has revealed gaps in the biomedical domain of the sort of terms patients are using and can be explored with semantic similarity. Previous work has found that patients mention things amongst themselves that they would not to clinicians, or which clinicians are unfamiliar with [31, 34].

Semantic similarity can be encapsulated via vectorisation. Vectorisation numerically represents text in an embedding model and these models provide an explanation of semantics. However, there is an absence of embedding models derived from patient-generated texts hindering secondary research use. Some clinically relevant models are available and derived from clinical text. Whilst valuable for the clinical domain, they have not been fully exploited and systematically compared to patient text.

To better understand semantic similarity in patient terms, three objectives can build upon a solution: using a preexisting model derived from clinical text as it is difficult to access clinical records, secondly training models solely consisting patient-generated text, and finally retraining the existing clinical model with the patient text.

The existing ClinicalBERT [206] is easy to access and information is available regarding input training size and model parameters. Two patient-generated models will be compared to observe the differences in patient-generated conversations: derived from forum and the other from social media. I anticipate these models will provide insight into patient priorities and how they may differ across different platforms.

Using clinical and patient-derived texts that cover a wide range of medical-related entities and relatedness of medical entities [132]. Inflammation is the body’s first response to threat, generally present in many cases, and has a complex relationship with a wide range of diseases [2]. Inflammatory diseases being explored as a case study allows for investigating associations found amongst the different contextual embeddings through a combination of manual evaluation, clustering, and semantic analysis. With application to inflammatory diseases, for which a new ontology is introduced to cover a wide range of inflammatory diseases, enhanced with important findings from the experiment. These aims allow a better understanding of patient semantic similarity, but also a systematic comparison to clinician-generated texts. All to improve communication standards.

3.2 Data and methods

3.2.1 Development of an ontology for inflammation

There is a previous source of work conducted on the potential development of an ontology for inflammation in 2004 by Gangemi et al. [264]. This work focuses on biological processes from the Gene Ontology [140]. However, when looking at citations via Google Scholar¹, none discuss the development of such ontology.

¹March 16, 2022.

Using two sources of comprehensive lists of inflammatory diseases, I curated an overlapping list of inflammatory diseases - screenshots of the lists can be seen in Figure 3.1, which I used ICD10data.com² (Subfigure 3.1a) [265] and ICD List³ (Subfigure 3.1b) [266].

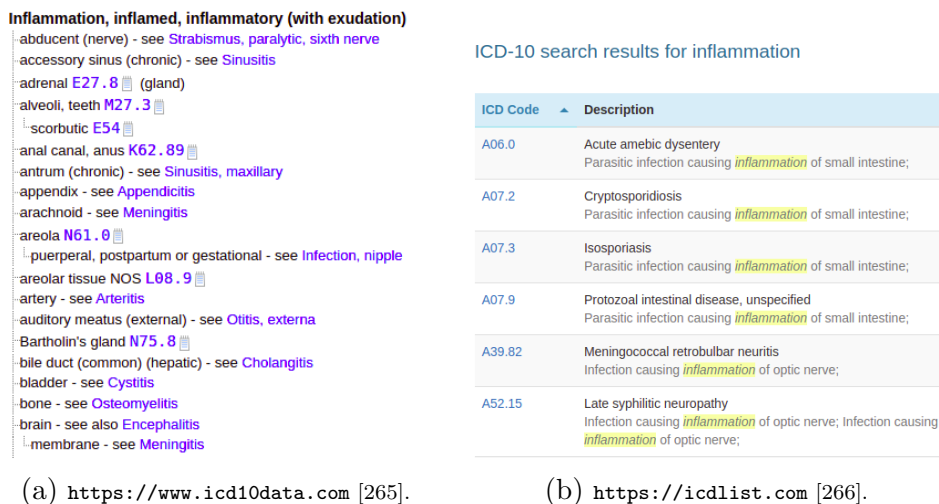


Figure 3.1: Screenshots of the ICD-10 comprehensive lists of inflammatory diseases that were included in ontology development.

Using Protégé [244], I created an ontology for inflammatory diseases with the initial structure derived from the two ICD-10 lists of inflammatory diseases, which were mapped to ICD-10 terms and organised in a similar taxonomic structure for a foundation. For example, digestive inflammatory conditions were in the same digestive superclass.

Although this ontology is a domain ontology for inflammatory diseases, it is an enhanced terminology system: a knowledge structure of terms with a logical interface, and extensive synonyms. Whilst DOID and ICD-10 exist, the domain experts revealing structural issues and gaps of knowledge in ocular IMIDs arose issues using biomedical ontologies and terminologies for such specific analysis. However, the appropriate database cross reference (x-ref) were included, in addition to a search of the NHS (<https://www.nhs.uk/>) site for the inclusion of symptoms that they report, such as reporting itchy and flaky eyelids for blepharitis.

²https://www.icd10data.com/ICD10CM/Index/I/Inflammation%2C_inflamed%2C_inflammatory

³<https://icdlist.com/?t=icd10&s=inflammation>

3.2.2 Forum and creation of embeddings

To systematically explore the patient voice, in this chapter I look at using a preexisting clinical embedding model to develop models that consist of patient-generated text. Using a preexisting model avoid the issues mentioned in the literature regarding model building, such as the issue with the training window size [204, 209]. Additionally, retraining a model that already consists of a vocabulary or defined parameters can control issues of computational resources: avoiding training tasks of multiple runs for fine-tuning metrics or overcoming the semantic complexities [210].

This chapter consists of two main investigations - in an opening study, the preexisting clinical model was compared to a model derived of patient forum texts. The second study consists of comparing two models of patient-generated texts: one being the same forum data and other social media conversations.

3.2.2.1 *ClinicalBERT*

Whilst this thesis is observing the patient’s voice, the clinician’s voice is also important. A source of the clinician’s voice can be derived from numerous sources, such as literature, EHRs, and clinical letters. Huang et al. [206] ClinicalBERT (CB) embedding model is derived from clinical letters, which are a representation of the Medical Information Mart for Intensive Care III (MIMIC-III) [220]. MIMIC-III is a set of patient records from a large database consisting of 38,597 ICU critical care patients (2001-2012) [220], with a total of 58,976 unique hospital admissions. Whilst ICU is based on critical conditions, inflammation is highly multi-morbid and the body’s first response to injury or continuous recovery [2, 81]; and so this model has the opportunity to represent inflammatory diseases.

3.2.2.2 *PatientINF*

The main focus of this chapter is exploring the patient voice via online, public forum conversations for patients with a wide range of issues. This model, PatientINF (PI), is derived from Patient UK (<https://patient.info/>). Patient UK was developed in 1996

by Dr Tim Kenny and Dr Beverley Kenny, as an online health information directory. The success of this website led to creating open forums [267]. Patient UK consisted of medical professionals, who originally developed resources for GPs delivering care towards their patients. Now, it is aimed for the patients, with clinically correct information, produced and shared by clinical professionals derived from literature [267]. Patient UK have an open forum board, not moderated or reviewed by clinicians, free for patients to share experiences, support each other, and create communities.

Patient UK’s forum board consists of a wide variety of topics, such as general health, medication, or a specific inflammatory disease, e.g. arthritis. These diseases can vary from being little in severity or extremely severe: in all subjective yet important to the patient. This data is very different to CB. The CB model is derived from MIMIC-III clinical letters from an ICU environment, however comorbidities are common [19, 20] and so clinical letters may consist of multiple disease descriptions. Additionally, CB has advantages as clinical letters are difficult to access.

3.2.2.3 *BioReddit*

In the second study (mentioned above), I compare PI to another set of patient-generated texts derived from social media. Using the COMETA corpus, released by Basaldella et al. [226], an embedding model will be developed. Initially the embedding model produced by Basaldella and Collier [225], BioReddit, was intended to be used however there lacked data development information, for example: post count, average word count, or embedding parameters; and so a fair comparison of models may not be possible. This version of BioReddit I aim to develop, will be referred to as, *modified*BioReddit ($_m$ BR).

3.2.2.4 *Design of two experiments*

For each of the two comparison studies, the preexisting ClinicalBERT (CB) methodology was used for text preprocessing and model building (such as parameters) in order to compare the different models with the least bias.

In the first study, CB is compared to PI whilst the second study consists of a comparison between m BR and PI. Both studies will also include the development of a model derived from both datasets, for example CB will be enhanced with PI to create ClinicalBERT + PatientINF (CB + PI). Moreover m BR and PI will be processed to developed the combined *modified*BioReddit + PatientINF (m BR + PI).

Using the same methodology as CB means all models were developed in the same manner. Huang et al. [206] describes the process of creating CB in vast detail, including text preprocessing methods, average word count, size of input, and model parameters/metrics (e.g. number of **epochs** or matrix **size**). By following these steps, the experiments avoid bias of possibly over-representing a domain and promotes fairness in the experimental design and comparisons.

Text preprocessing, as described by Huang et al. [206], in that CB was created:

1. each clinical letter was treated separately,
2. lowercase,
3. line breaks and special characters (such as brackets or full-stops) were removed,
4. finally, each clinical letter was tokenised.

The embedding models that consist of patient-generated text will present different dimensions of the patient voice, furthermore insight into the clinician voice will provide knowledge of domain priorities. These comparisons of embedding models will observe differences in semantic similarity, how well these models capture synonyms, and priorities in text within these different environments. Throughout this chapter, I will refer to the following abbreviations in Table 3.1:

Abbr.	Full name	Source	Description
CB	ClinicalBERT	Huang et al. [206]	embedding model derived from clinical letters
m BR	<i>modified</i> BioReddit	COMETA [226]	social media medical embedding model
PI	PatientINF	Patient.info [267]	patient-generated forum discussions
CB + PI	ClinicalBERT + PatientINF	[206] + [267]	the CB embedding model retrained on PI
m BR + PI	<i>modified</i> BioReddit + PatientINF	[226] + [267]	model derived from both m BR and PI

Table 3.1: Notation of the data sources and embedding models used in this chapter.

Statistical analysis

When the embedding models are developed in both experiments, I perform a series of tasks and statistical analysis for comparisons.

1. Firstly an observation is conducted into the models: comparing the sort of semantic similarities and differences. This includes a t-distributed stochastic neighbor embedding (t-SNE), a type of dimension reduction (DR) that performs well on word embeddings, as it is a non-linear technique [268] and can represent the complex relationship between terms. The t-SNE learns the mappings of terms from the original vectorisation matrix via the distance probability of terms [268, 269].
2. Secondly a Wilcoxon signed-rank test is conducted to observe the semantic similarity scores of terms on the multi-domain models against the single-domain model. Also a receiver operating characteristic (ROC) investigates whether the models could capture synonyms and if so, how well.
3. Third, following the Pearson correlation coefficient experimentation from Huang et al. [206], in which I look at how each embedding model similarities compare to physician annotations by Pedersen et al. [200].
4. Finally, as a form of validation I ran the analysis on multiple seeds in expectation of similar results.

3.3 Results

3.3.1 Combined Ontology for Inflammatory Diseases summary

The Combined Ontology for Inflammatory Diseases (COID) is open-source (<https://github.com/sap218/coid>) and currently v1.0.0, it has been fully documented following the MIRO⁴ guidelines [255]. Protégé’s Pellet reasoner [244] was used to ensure a coherent and consistent ontology.

COID was built using ICD-10 as a foundation, based on anatomical structure, with each concept in COID assigned a unique identifier (the IRI), such as hepatitis `COID:00256`. COID contains 1185 classes and organises 365 inflammatory-related concepts into 10 anatomical superclasses, e.g. “inflammation of visual system” `COID:00006`, as seen in Figure 3.2. This ontology allows for semantic characterisations of disease subgroups for the embedding models which are explored later.

There are a total of 8439 different relationships and axioms: 2067 subclasses, 797 symptoms reported by the NHS (<https://www.nhs.uk/>), and 78 patient-reported (new

⁴read the full documentation here: <https://sap218.github.io/coid/MIRO>.

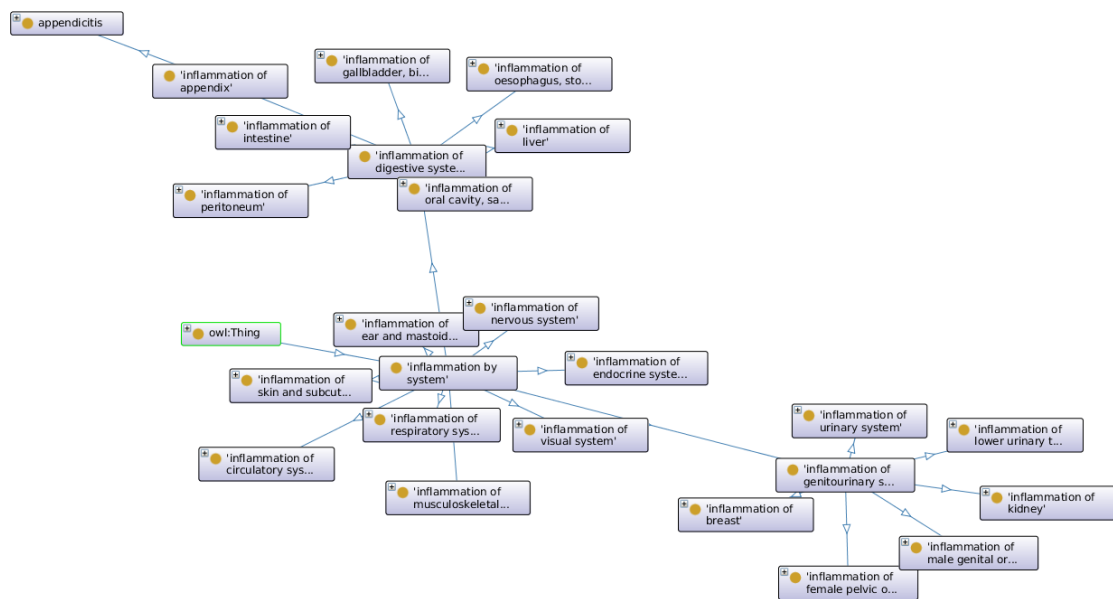


Figure 3.2: Graphical representation of Combined Ontology for Inflammatory Diseases superclasses, showing the major anatomical systems, as seen in ICD-10. With an expansion of inflammation of digestive and genitourinary system.

defined relationship C0ID:00942) symptoms, derived via TF-IDF. A full perspective of C0ID can be observed in Appendix Figure C.1 (page 197), created with WebVOWL [257].

There are 5164 annotations: 1763 annotations alone were x-ref of ICD-10 (n=825), SNOMED-CT (n=304), and HP (n=43) - as seen in Table 3.2 (page 104). From x-ref to other biomedical ontologies, terms were extracted as types of synonyms, such as 48 terms are “exact” synonyms, 11 are “broad”, and 397 lay terms from the TF-IDF of PI.

3.3.2 Existing embedding models and description of data

Downloaded data sources are described in Table 3.3 (page 104), as a note, model building was conducted using Gensim [270] a Python library [248].

ClinicalBERT (CB) Word2Vec and similar n-gram model, fastText, are both freely available (<https://github.com/kexinhuang12345/clinicalBERT>) with detailed descriptions of model building [206]. The *modified*BioReddit (*m*BR) COMETA corpus dataset was available online (<https://github.com/cambridgeltl/cometa>) and downloaded [226].

Source	x-ref count
ICD-10 [153]	825
SNOMED-CT [156]	304
ICD-9 [153]	267
Read codes [155]	196
NCIT*	61
SYMP**	43
HP [5]	43
DOID [141]	18
PATO [134]	3
ORDO [143]	1
UBERON [142]	1
BFO [246]	1
TOTAL	1763

Table 3.2: Summary of Combined Ontology for Inflammatory Diseases (COID) sources: class count, database cross reference (x-ref), and synonyms.

*National Cancer Institute Thesaurus (NCIT) <https://ncit.nci.nih.gov/ncitbrowser/>.

**Symptom Ontology (SYMP) <https://bioportal.bioontology.org/ontologies/SYMP/>.

Data source	Text count	Word average
ClinicalBERT (CB)	100,000 clinical letters	237.41
<i>modified</i> BioReddit (<i>m</i> BR)	20,014 social media posts	23.64
PatientINF (PI)	2,954,802 forum posts*	88.48

Table 3.3: Description of the embedding models and data available to use for analysis.

*October 17, 2021.

Finally, I extracted all of Patient UK’s forum board with the BeautifulSoup [271] library in Python. PatientINF (PI) is derived from 1369 topics⁵. Each topic forum has multiple threads, with each thread consisting of having posts (comments) and replies.

As an application for this chapter, both study experiments looked into 14 inflammatory topics of interest, as seen in Table 3.4. As a final note, lay synonyms were added to COID from a TF-IDF of the 14 inflammatory topics of interest following the methods from chapter 2, OcIMIDo [242].

3.3.3 Experiment one: clinical and patient differences

For a fair comparison with CB (clinical note count and average word count), I first limited word count per post of PI to 100. Secondly, I limited the number of posts per topic to 73 (100,000 letters / 1369 forum topics), so for those topics which had less than 73, all were used. In all gives an end total of 91,751 PI posts to be included, with an average

⁵October 17, 2021.

Condition	System	Description	Onset
appendicitis	digestive	inflammation of the appendix	acute
arthritis	musculoskeletal	inflammation of the joints	chronic
atrophic vaginitis	genitourinary	inflammation of the vaginal walls	chronic
blepharitis	visual	inflammation of eyelids	chronic
cellulitis	skin	inflammation of deeper skin layers	acute
cystitis	genitourinary	inflammation of the urinary tract	acute
folliculitis	skin	inflammation of hair follicles	acute*
hepatitis	digestive	inflammation of the liver	acute or chronic
plantar fasciitis	musculoskeletal	inflammation of the plantar fascia	acute*
rhinitis	respiratory	inflammation caused by allergen	acute or chronic
scleritis and episcleritis	visual	inflammation of the white of the eye	chronic
sinusitis	respiratory	inflammation of the sinuses	acute*
ulcerative colitis	digestive	inflammation of the large intestine	chronic
vasculitis	skin	inflammation of the blood vessels	acute or chronic

Table 3.4: The 14 inflammatory terms of interest for analysis of the embedding models.
 *these can be recurring acute issues.

word count of 223.44. When building the models, the same parameter configurations were used, such as `epochs=20`, `matrix size=100`. CB consisted of 44,082 terms in the model, PI consisted of 30,157, and CB + PI had 58,865 terms.

t-distributed stochastic neighbor embedding (t-SNE)

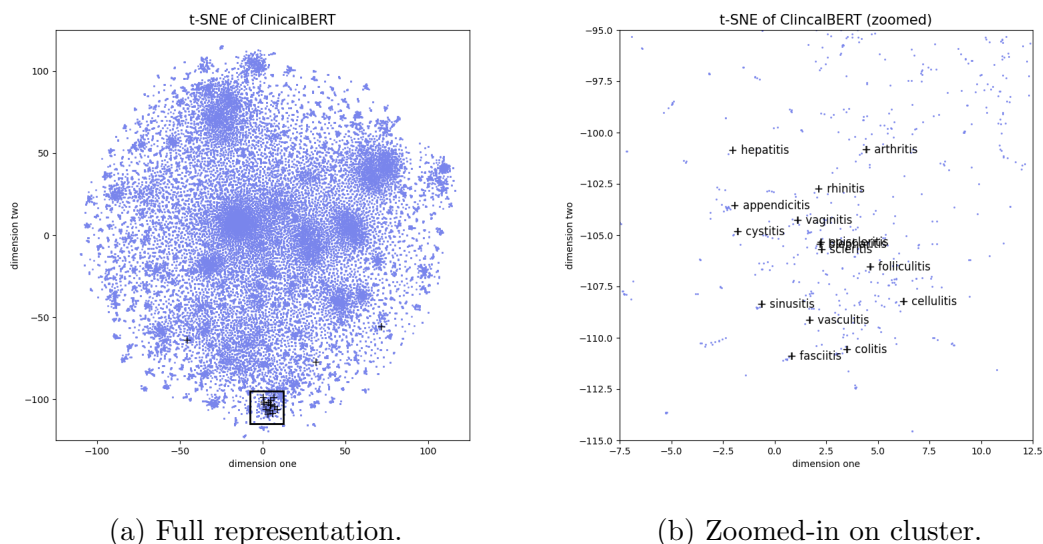


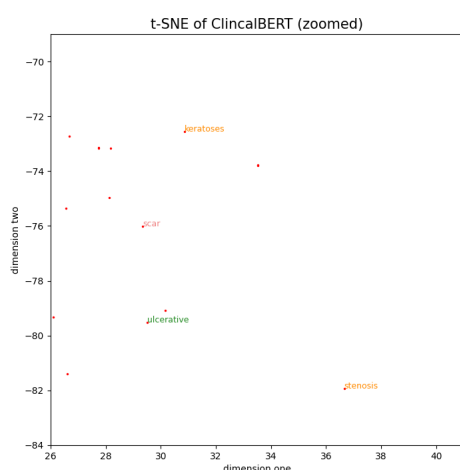
Figure 3.3: A t-SNE of CB. Each node is a term and distances represent the measure of association.

ClinicalBERT (CB)

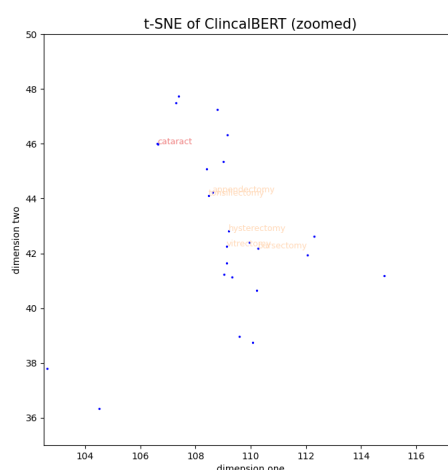
In Figure 3.3 (page 105), all inflammatory terms of interest are together with the exception of “atrophic”, “plantar”, and “ulcerative”. We see blepharitis, scleritis, and episcleritis are overlapping: all visual inflammatory disorders.

When filtering by COID terms and highlighting some components of superclasses, Figure 3.5 (page 107) shows that **inflammatory**, **non-inflammatory**, and **surgical therapies** terms are clustered in their separate categories. Whereas **symptoms** and **medications** are scattered clusters. Finally **impacts** (cause of inflammation), **common conditions**, **outcomes**, and **investigations** are scattered.

When zoomed in on Figure 3.5 dotted boxes, Figure 3.4 visualises a set of terms. The **red box**, Subfigure 3.4a, shows **non-inflammatory** conditions: keratoses (skin growth) and stenosis (narrowing of spine causing nerve pain) near scar (**outcome**) and “ulcerative” (colitis, **inflammation**). The **blue box**, Subfigure 3.4b, shows cataract (**outcome**) near types of **surgical therapies** - with appendectomy and tonsillectomy.



(a) Red highlighted.



(b) Blue highlighted.

Figure 3.4: Two close-up t-SNE plots of CB filtered with COID terms. Each node is a term and distances represent the measure of association.

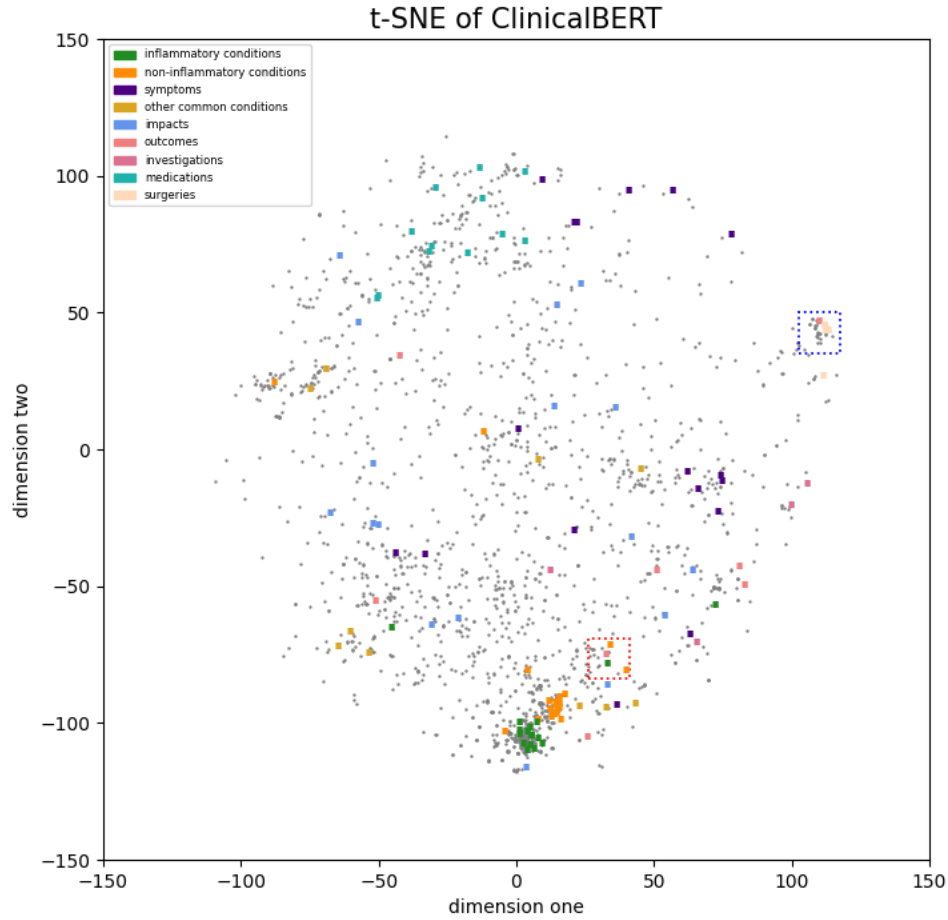
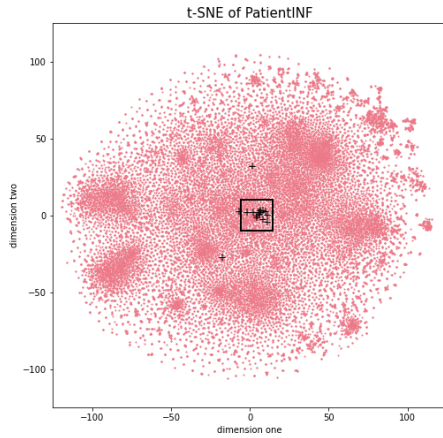


Figure 3.5: Full representation of CB t-SNE filtered using terms only in COID. Each node is a term and distances represent the measure of association.

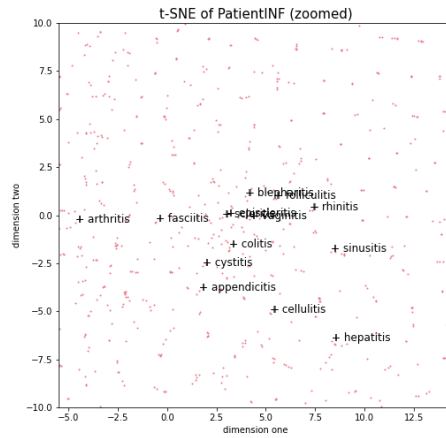
PatientINF (PI)

In Figure 3.6 (108), all inflammatory terms of interest are together with the exception of “plantar” and “ulcerative”, similarly to CB. Interestingly, vaginitis and scleritis are overlapping whilst blepharitis is separated from scleritis, unlike CB.

Filtering by COID terms and highlighting some components of superclasses, Figure 3.8 (page 109) shows that, similarly to CB, **inflammatory**, **non-inflammatory**, and **surgical therapies** are clustered. Interestingly, in contrast to CB, **investigations**, **symptoms**, and **medications** are more clustered together.



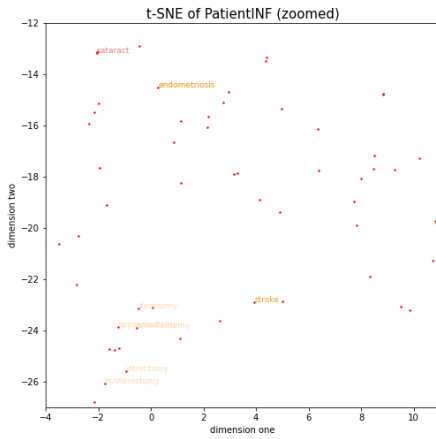
(a) Full representation.



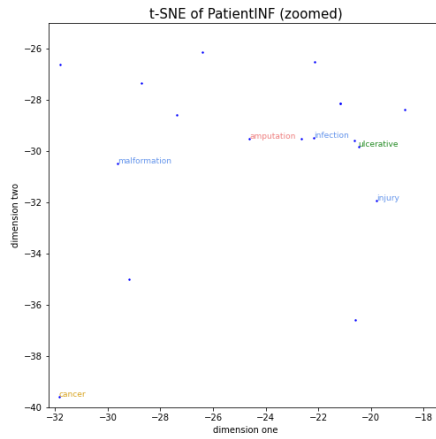
(b) Zoomed-in on cluster.

Figure 3.6: A t-SNE of PI. Each node is a term and distances represent the measure of association.

Zoomed in on Figure 3.8 dotted boxes, Figure 3.7 shows the **red box**, Subfigure 3.7a, endometriosis (**non-inflammatory**) near cataract (**outcome**), and cataract near **surgical therapies** (similarly to CB). The **blue box**, Subfigure 3.7b, shows malformation, infection, and injury (**impacts**) near amputation (**outcome**) and “ulcerative” (colitis, **inflammation**).



(a) Red highlighted.



(b) Blue highlighted.

Figure 3.7: Two close-up t-SNE plots of PI filtered with COID terms. Each node is a term and distances represent the measure of association.

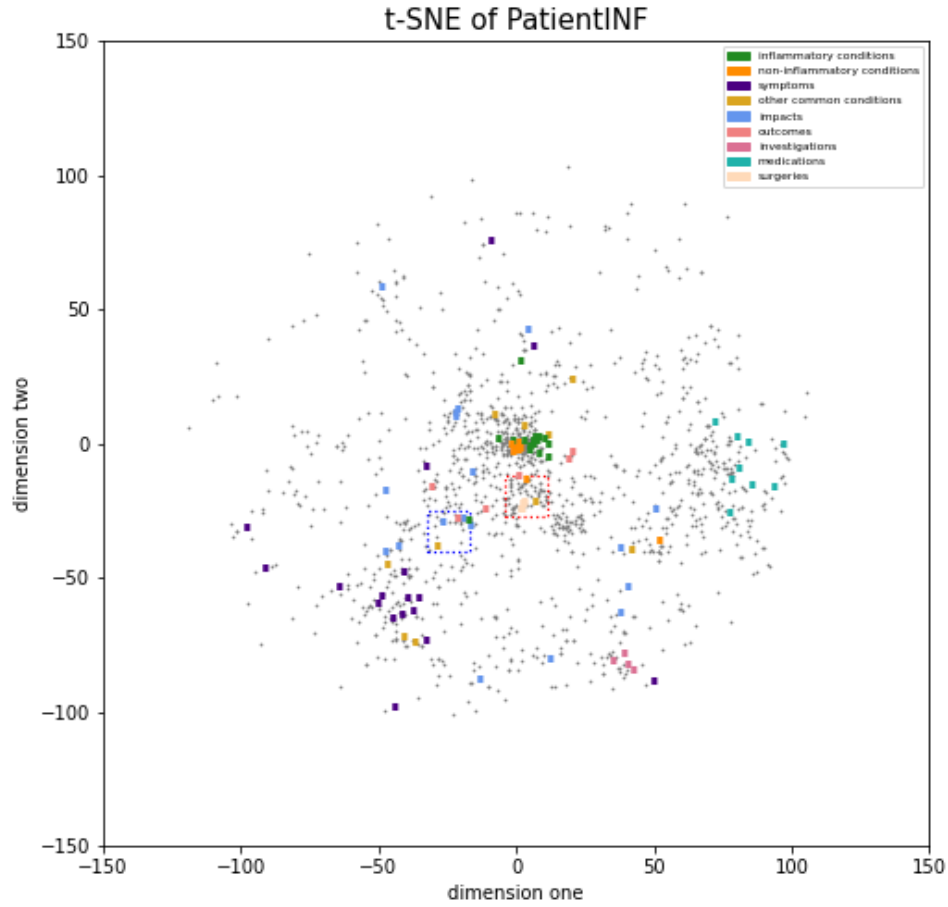
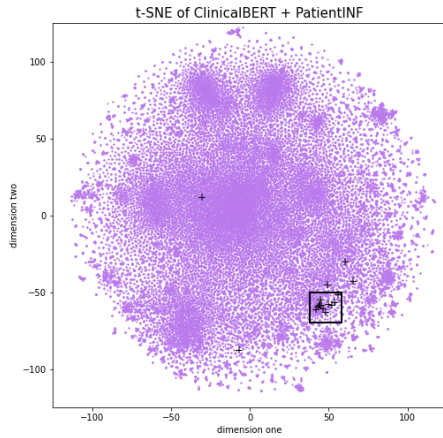


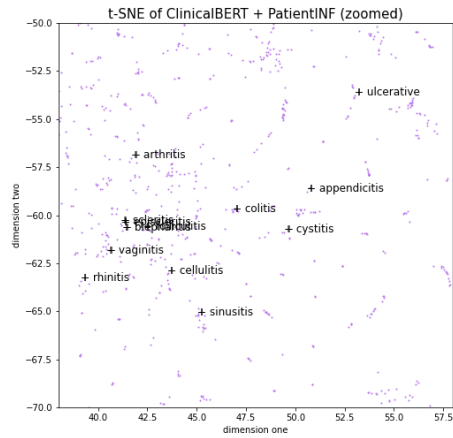
Figure 3.8: Full representation of PI t-SNE filtered using terms only in COID. Each node is a term and distances represent the measure of association.

ClinicalBERT + PatientINF (CB + PI)

Finally, in Figure 3.9 (page 110), the inflammatory cluster is much more varied: “atrophic” and “plantar” are again separated, however now fasciitis seems to be more separated from the other inflammatory conditions. Both colitis and “ulcerative” are closer together, and appendicitis and colitis are also closer (both digestive conditions). When filtering by COID terms and highlighting some components of superclasses, Figure 3.11 (page 111) shows that the patient voice has contributed to the clinical text as **symptoms** and **medications** are reflecting PI in that they are closer together.



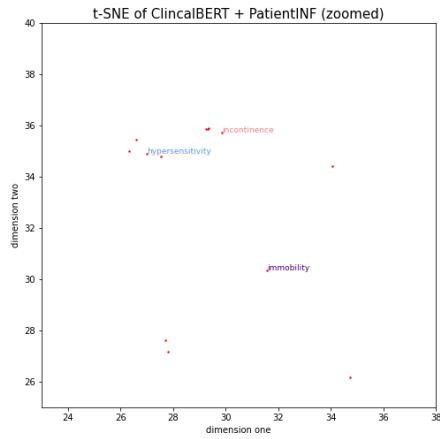
(a) Full representation.



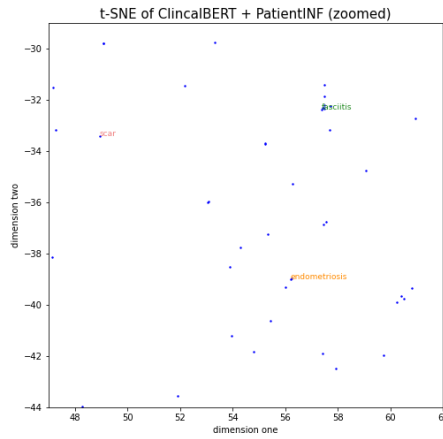
(b) Zoomed-in on cluster.

Figure 3.9: A t-SNE of CB + PI. Each node is a term and distances represent the measure of association.

When zoomed in on Figure 3.11 dotted boxes, Figure 3.10 shows the **red box**, Subfigure 3.10a, that hypersensitivity (**impact**), incontinence (**outcome**), and immobility (**symptom**) together. The **blue box**, Subfigure 3.10b, shows endometriosis (**non-inflammatory**), scar (**outcome**), and fasciitis (**inflammation**) together.



(a) Red highlighted.



(b) Blue highlighted.

Figure 3.10: Two close-up t-SNE plots of CB + PI filtered with COID terms. Each node is a term and distances represent the measure of association.

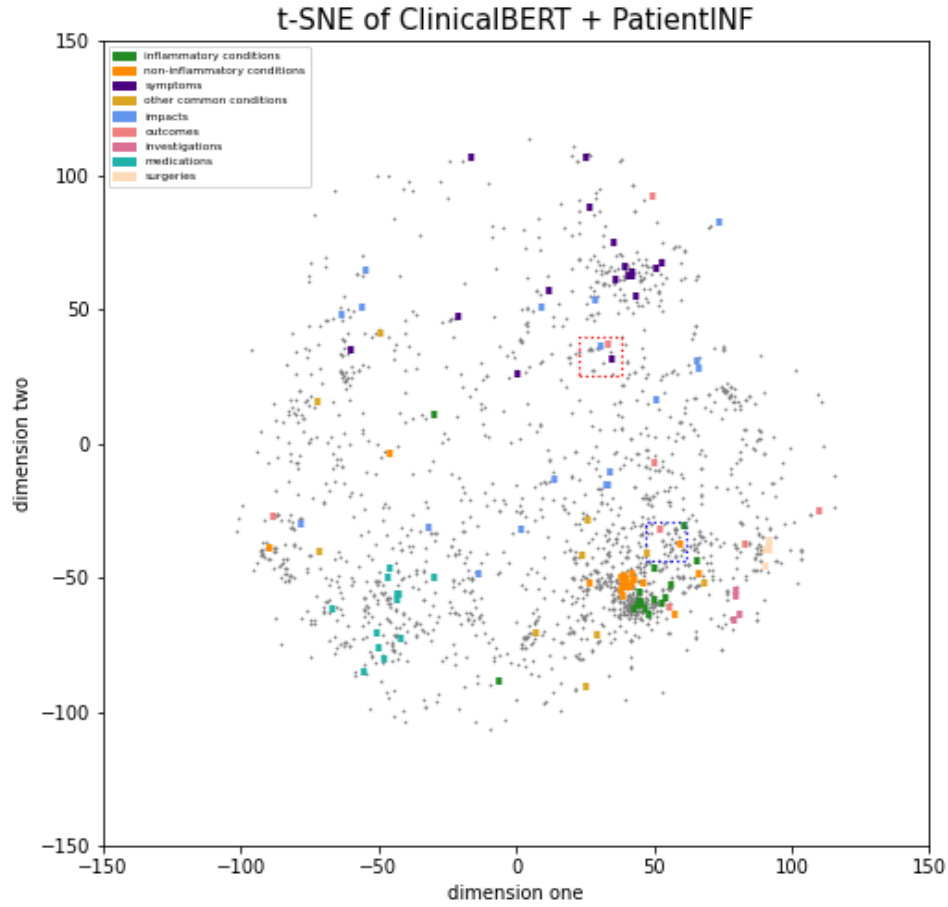


Figure 3.11: Full representation of CB + PI t-SNE filtered using terms only in COID. Each node is a term and distances represent the measure of association.

Semantic similarities

In this section, I look into the top similar terms of our 14 terms of interest and attempt to characterise them, in addition to using Vec2graph [272] to visualise and represent the most similar terms. See Appendix Section C (page 198) for additional characterisations.

Appendicitis

In CB the most similar terms are cholecystitis (inflammation of the gallbladder) and misspellings: “cholecystitis”, “cholcystitis”, “cholecysitis”, “choleycystitis”, and more. Then followed by cystitis and its misspelling “cystitis”. In PI the most similar terms

are “appendicitus” (misspelling), “appendicolith” (calcified deposit within the appendix), diverticulitis (inflammation of pouches in intestines), endometriosis (non-inflammatory tissue outside the uterus causing pain), and pancreatitis (inflammation of the pancreas). In CB + PI, the most similar terms are “appendicitus”, cholecystitis’ misspellings: “cholecytitis”, “cholecycitis”, and “cholecysitis”; and “choleangitis” (misspelling of cholangitis, inflammation of the bile duct).

Arthritis

In CB the most similar terms are “osteroarthritis” and its correction osteoarthritis, polyarthritis, and misspellings: “arthriitis” and “athritis” - all types of arthritis. In PI, the most similar terms are osteoarthritis, spondylitis (inflammation of spinal bones), and misspellings: “arthritus”, “arthritis”, “athritis”, “arthiritis” and more. In CB + PI, the most similar terms are misspellings: “arthriitis”, “arthiritis”, “athritis”, “arthritus”, and more. Then followed by “osteroarthritis” and its correction osteoarthritis, and finally polyarthritis. Reflecting the best of both CB and PI.

Blepharitis

In CB the most similar terms are iritis (type of anterior uveitis), “puritis” (misspelling of pruritus, itchy skin), scleritis, “arthritis” (misspelling of arthritis), and sialadenitis (salivary gland infection). In PI the most similar terms are episcleritis, gingivitis (inflammation of the gums), itis, iritis, and “belanitis” (misspelling of balanitis, inflammation of the penis). In CB + PI, the most similar terms are scleritis, episcleritis, iritis, “arthritis”, and “puritis”. Figure 3.12 illustrates the semantic networks.

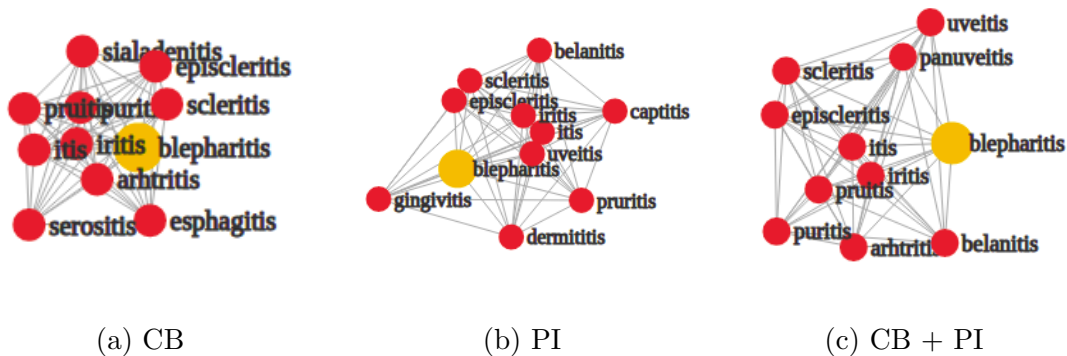


Figure 3.12: Vectorisation representation of blepharitis.

Rhinitis

In CB the most similar terms are “tinnitis” (misspelling of tinnitus, known for ringing in the ears), retinitis, linitis (type of stomach cancer), “puritis”, vulvovaginitis (inflammation of both the vagina and vulva), mononeuritis (disorder of the nervous system), and vaginitis. In PI the most similar terms are “rhinitus”, vaginitis, “belanitis”, “adnitis” (misspelling of adeniti, inflammation of a gland), and “rhinnetis”. The similarity of vaginitis and rhinitis is present in both CB and PI, but interestingly balanitis is also similar in PI, another inflammatory condition of the genitourinary system. In CB + PI, the most similar terms are “rhinitus”, linitis, retinitis, “rhinitis” (misspelling not present in CB or PI), “belanitis”, and chorioretinitis.

Scleritis and episcleritis

In CB the most similar terms are blepharitis, “puritis”, iritis, sialadenitis, “arthritis”, and mastitis (inflammation of breast tissue). In PI the most similar terms are iritis, neuritis (inflammation of the optic nerve), itis, uveitis, and blepharitis - all ocular IMIDs. Followed by myelitis (inflammation of the spinal cord) and gastroenteritis (type of IBD). In CB + PI, the most similar terms are itis, iritis, blepharitis, polyarteritis (inflammation of arteries), retinitis, angiitis (inflammation of the walls of small blood vessels).

Sinusitis

In CB the most similar terms are misspellings: “sinusisitis”, “sinusiitis”, and “sinuisitis”. Followed by pansinusitis, mastoiditis (infection of the bone behind the ear), mastitis, and otomastoiditis. In PI the most similar terms are “sinusitus”, uveitis, “adnitis”, “sinusus”, itis, “tinnitis”, and iritis. In CB + PI, the most similar terms are “sinusiitis”, “sinusisitis”, “sinusitus”, “sinuisitis”, “sinusus”, and more. Followed by pansinusitis and “mucusitis” (misspelling of mucositis, inflamed mouth - not present in either CB or PI).

Figure 3.14 illustrates the semantic networks.

Ulcerative colitis

In CB the most similar terms are pancolitis, ileitis (inflammation of one or more sections of the intestine), “colitis”, diverticulitis, “coliitis”, and ileocolitis (type of Crohn’s

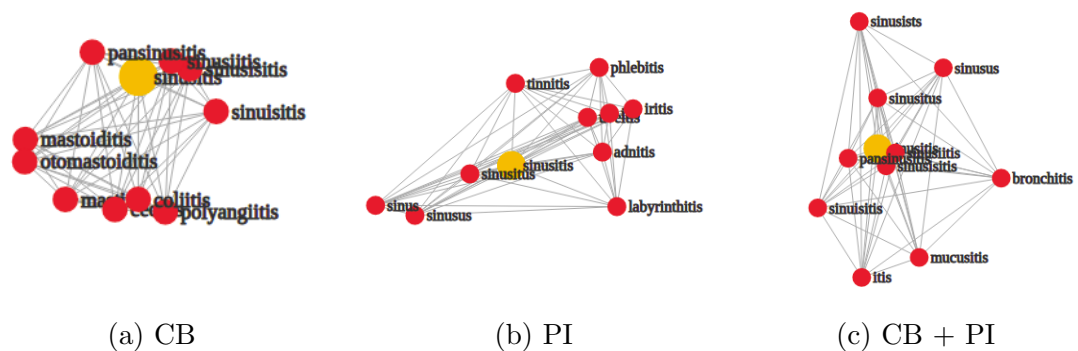


Figure 3.14: Vectorisation representation of sinusitis.

disease (CD): IBD) - all types of IBD. In PI the most similar terms are ulcerative, colitis, ulceration, iritis, and bronchiolitis (lower respiratory tract infection). In CB + PI, the most similar terms are ulcerative, colitis, pancolitis, proctocolitis, “cholitis”, and ileocolitis.

Systematic comparisons

Paired comparison test

A Wilcoxon signed-rank test (W) was conducted to observe scores of similar terms from CB to CB + PI, and PI to CB + PI, to test the significance of scores changing from the expansion of the models. In Table 3.5 (page 116), p-value for cellulitis (CB to CB + PI) was large.

A receiver operating characteristic (ROC)

In Figure 3.15 (page 116), I performed a ROC to observe how well the embedding models captured the 14 inflammatory terms of interest and their synonyms. In Subfigure 3.15a the terms and synonyms from COID are from ICD-10, biomedical ontologies, and from the TF-IDF from chapter 2 [242]. In Subfigure 3.15b the terms and synonyms from COID are expanding using terms of interest and their 100-most similar terms from the embedding models.

Condition	CB to CB + PI W p-value	PI to CB + PI W p-value
appendicitis	p=0*	p=0*
arthritis	p=0*	p=0*
atrophic vaginitis	p=2.018e-109	p=0*
blepharitis	p=1.790e-214	p=0*
cellulitis	p=0.832	p=0*
cystitis	p=0*	p=0*
folliculitis	p=1.973e-101	p=0*
hepatitis	p=0*	p=0*
plantar fasciitis	p=0*	p=0*
rhinitis	p=0*	p=0*
scleritis and episcleritis	p=0*	p=0*
sinusitis	p=2.217e-19	p=0*
ulcerative colitis	p=1.833e-267	p=0*
vasculitis	p=0*	p=0*
ALL	p=0*	p=0*

Table 3.5: The inflammatory terms of interest and their Wilcoxon signed-rank test (W) p-values from the single-domain embedding models to ClinicalBERT + PatientINF (CB + PI).

*computed value so small it output as 0 (zero).

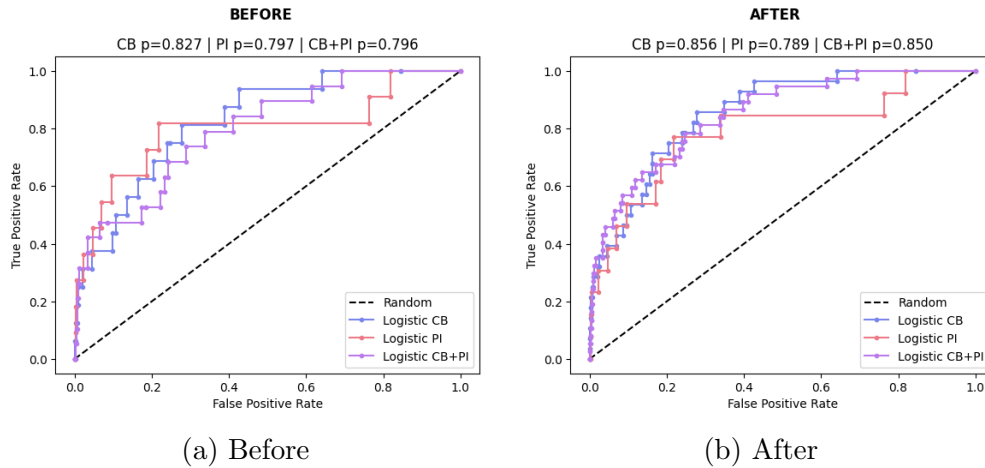


Figure 3.15: Plotting receiver operating characteristic (ROC) curve of synonyms captured from the embedding models using the 14 inflammatory terms of interest.

We can see that using the previous methods from OcIMIDo, Subfigure 3.15a, CB p=0.827, PI p=0.797, and CB + PI p=0.796. The CB models perform best, which can be expected due to COID majority consisting of synonyms from other biomedical ontologies and little layperson. However, Subfigure 3.15b presents the value of gaining synonyms from other methods too: embedding models, as CB p=0.856, PI p=0.789, and CB + PI p=0.850. The CB model improves, showing more synonyms were extracted of clinical domain, however PI decreases and CB + PI significantly increases: showing this combined model captures both lay and clinical synonyms.

Pearson correlation coefficient

Following the Pearson correlation coefficient (r) experimentation from ClinicalBERT’s Huang et al. [206] paper, I look at how each embedding model similarities compare to clinical annotations by Pedersen et al. [200] in Table 3.6.

Term 1	Term 2	Physician	CB	PI	CB + PI
renal failure	kidney failure	1.0	1.0	1.0	1.0
heart	myocardium	0.767	0.120	0.275	0.173
stroke	infarct	0.667	0.736	0.247	0.445
abortion	miscarriage	0.667	0.452	0.347	0.587
delusion	schizophrenia	0.667	0.488	0.187	0.442
congestive heart failure	pulmonary edema	0.667	0.551	0.420	0.539
metastasis	adenocarcinoma	0.567	0.869	0.450	0.743
calcification	stenosis	0.567	0.552	0.535	0.583
diarrhea	stomach cramps	0.433	0.465	0.780	0.805
mitral stenosis	atrial fibrillation	0.433	0.277	0.475	0.234
chronic obstructive pulmonary disease	lung infiltrates	0.433	0.398	0.489	0.474
rheumatoid arthritis	lupus	0.333	0.692	0.719	0.876
brain tumor	intracranial hemorrhage	0.333	0.627	0.395	0.540
carpel tunnel syndrome	osteoarthritis	0.333	0.712	0.451	0.475
diabetes mellitus	hypertension	0.333	0.763	0.543	0.827
acne	syringe	0.333	0.333	0.316	0.100
antibiotic	allergy	0.233	0.224	0.466	0.443
cortisone	total knee replacement	0.233	0.0	0.461	0.371
pulmonary embolus	myocardial infarction	0.233	0.212	0.723	0.426
pulmonary fibrosis	lung cancer	0.233	0.358	0.652	0.613
cholangiocarcinoma	colonoscopy	0.100	0.381	0.245	0.424
lymphoid hyperplasia	laryngeal cancer	0.100	0.504	0.582	0.667
multiple sclerosis	psychosis	0.0	0.683	0.214	0.372
appendicitis	osteoporosis	0.0	0.445	0.584	0.436
rectal polyp	aorta	0.0	0.317	0.168	0.299
xerostomia	alcoholic cirrhosis	0.0	0.288	0.469	0.175
peptic ulcer disease	myopia	0.0	0.356	0.117	0.141
depression	cellulites	0.0	0.059	0.0	0.0
varicose vein	entire knee meniscus	0.0	0.333	0.323	0.275
hyperlipidemia	metastasis	0.0	0.393	0.388	0.204
Pearson correlation coefficient			0.430	0.322	0.485

Table 3.6: Pearson correlation coefficient table comparing normalised physician scores by Pedersen et al. [200] to ClinicalBERT + PatientINF (CB + PI).

Looking at the results, I can conclude some interesting observations, firstly all models agree kidney and renal failure are most similar and most low-scored terms from the physician are actually ranked higher in similarity, with the exception of depression (term 1) and cellulites (terms 2).

In CB, diabetes mellitus (term 1) and hypertension (term 2) was scored much higher compared to the physician, and heart (term 1) and myocardium (term 2) was much lower compared to the physician.

In PI, diarrhea (term 1) and stomach cramps (term 2) was scored much higher compared to the physician and CB, and delusion (term 1) and schizophrenia (term 2)

was much lower compared to the physician and CB.

Finally, in CB + PI, rheumatoid arthritis (term 1) and lupus (term 2) were higher compared to the physician, CB, and PI; and acne (term 1) and syringe (term 2) were lower compared to the physician, CB, and PI. Moreover, diarrhea (term 1) and stomach cramps (term 2) were much higher compared to the physician and CB, and slightly more than PI. When looking more into this, CB similarities are misspellings or nausea, PI includes misspellings too but also symptoms: terms such as constipation, bloated, heartburn, and vomiting; and CB + PI includes CB misspellings but also PI symptoms.

Repeating with multiple seeds

When repeating the experiment using different seeds, we can see in Table 3.7 that the r varies by each seed, however, we do see that in PI the r seems to vary between 0.206 and 0.364 whilst in CB + PI the r varies between 0.381 and 0.502.

Seed	PI	CB + PI
1	0.206	0.381
2	0.322	0.485
3	0.264	0.432
4	0.364	0.502
5	0.245	0.403
Average r	0.280	0.441

Table 3.7: Pearson correlation coefficient (r) table comparing different seeds to present consistency of results in ClinicalBERT + PatientINF (CB + PI).

Seed mapping: 1=1, 2=123, 3=808, 4=6298948, 5=2014428590885440.

3.3.4 Experiment two: patient and patient differences

In this study, I aimed to compare two different patient-generated embedding models, m BR and PI, in addition to the combined m BR + PI. As previously mentioned, I used Huang et al. [206] (CB) text and model preprocessing methods, e.g. lowercase and `epochs=20`. The m BR model (COMETA [226]) consisted of 20,014 posts (Table 3.3, page 104), with an average word count of 23.64 and so for PI since the average word count was higher, there was no limit, but a cut-off of posts per topic of 14 (20,014 / 1369) - giving an end total of 19,166 PI posts with an average of a 100.3 word count. The models' vocabulary consisted of m BR 4746, PI 9486, and m BR + PI 10,232 terms.

t-distributed stochastic neighbor embedding (*t*-SNE)

*modified*BioReddit (*m*BR)

When filtering by COID terms and highlighting some components of superclasses, Figure 3.16 shows that majority of categories are clustered together, with the exception of **impacts**, which are quite scattered.

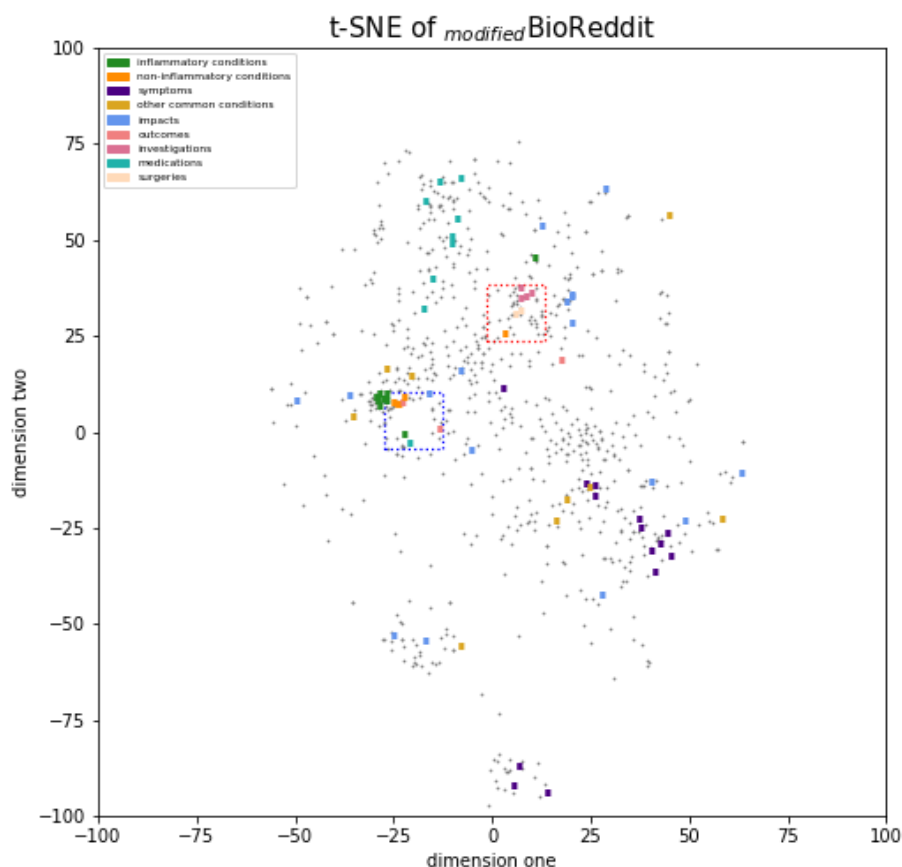
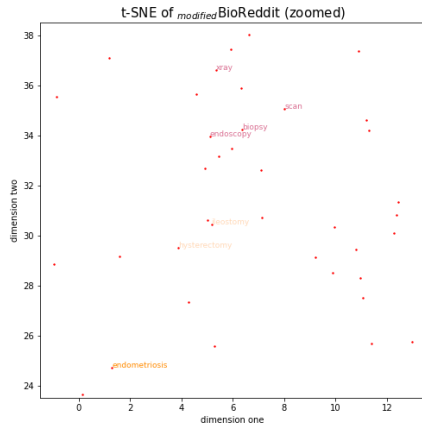


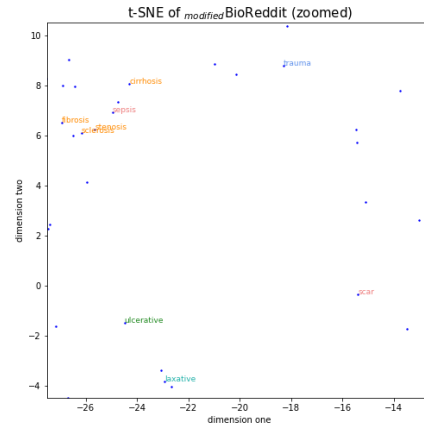
Figure 3.16: Full representation of *m*BR t-SNE filtered using terms only in COID. Each node is a term and distances represent the measure of association.

When zoomed in on Figure 3.16 dotted boxes, Figure 3.17 (page 120) shows the sort of terms clustered together. The **red box**, Subfigure 3.17a, shows **investigations** near **surgical therapies**, and endometriosis (**non-inflammatory**). The **blue box**, Subfigure 3.17b, shows laxative (**medications**) near ulcerative (**inflammatory**) and scar (**outcomes**). Additionally,

trauma (impacts) is near non-inflammatory and sepsis (outcomes).



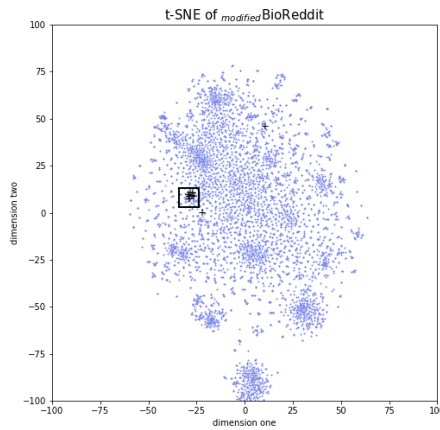
(a) Red highlighted.



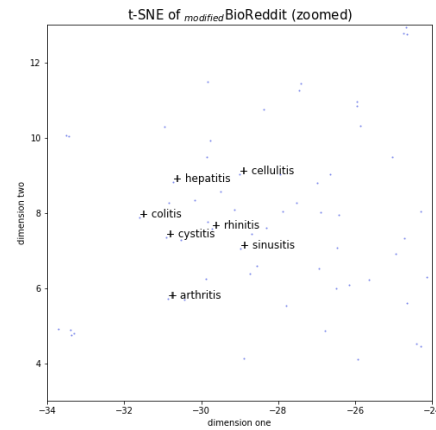
(b) Blue highlighted.

Figure 3.17: Two close-up t-SNE plots of m BR filtered with COID terms. Each node is a term and distances represent the measure of association.

In Figure 3.18, inflammatory terms of interest are clustered together, such as respiratory inflammatory disorders (sinusitis and rhinitis). However the exception of “ulcerative” (colitis) and “plantar” (fasciitis, which is not a term present in COMETA, including other terms of interest, such as episcleritis).



(a) Full representation.



(b) Zoomed-in on cluster.

Figure 3.18: A t-SNE of m BR. Each node is a term and distances represent the measure of association.

PatientINF (PI)

Filtering by COID and highlighting some components of superclasses, Figure 3.19 shows similarities with $_m$ BR, however **investigations** and **outcomes** are more scattered.

When zoomed in on Figure 3.19 dotted boxes, Figure 3.20 (page 122) shows the **red box**, Subfigure 3.20a in that IBS (**common condition**), hysterectomy (**surgical therapy**), and infliximab (**medication**). The **blue box**, Subfigure 3.20b, shows hypertension (**common condition**), amputation (**outcome**), and malformation (**impact**) together.

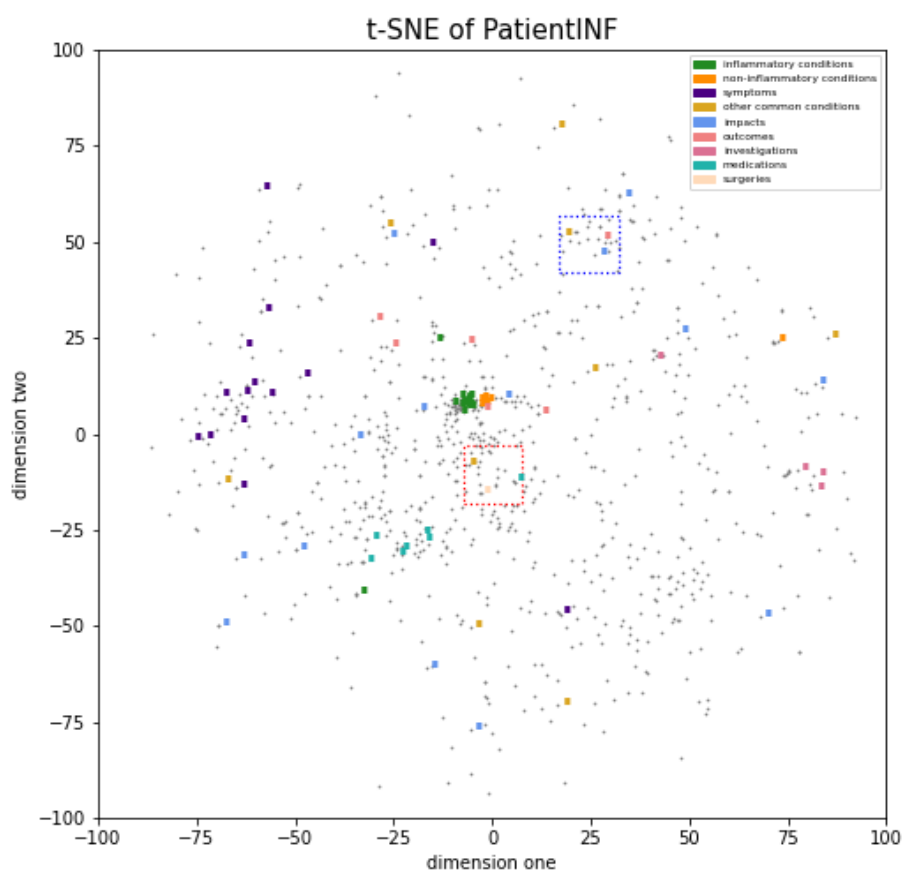
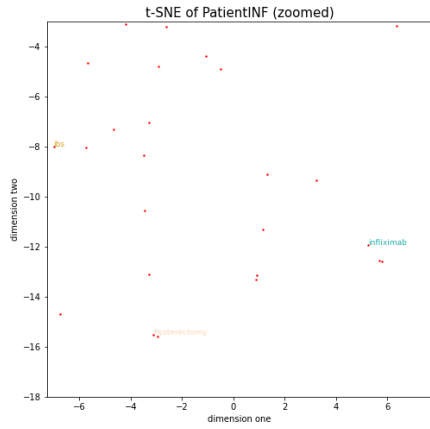
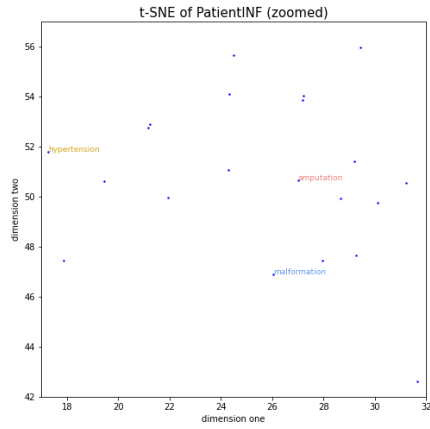


Figure 3.19: Full representation of PI t-SNE filtered using terms only in COID. Each node is a term and distances represent the measure of association.



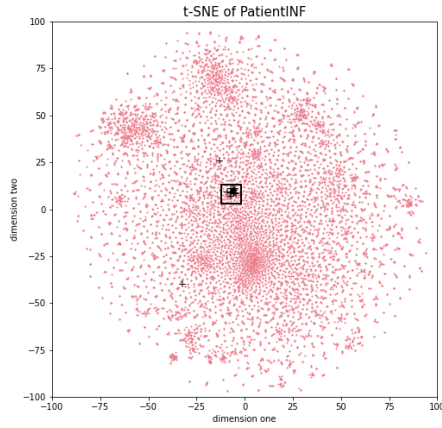
(a) Red highlighted.



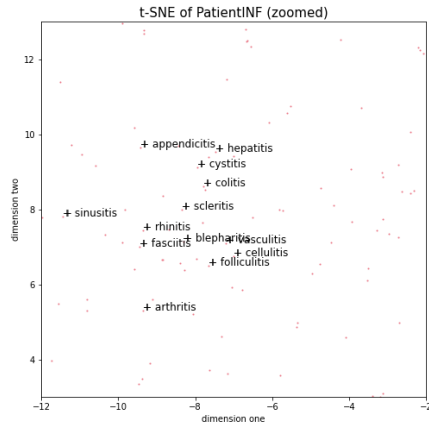
(b) Blue highlighted.

Figure 3.20: Two close-up t-SNE plots of PI filtered with COID terms. Each node is a term and distances represent the measure of association.

In Figure 3.21, all terms are clustered, with the exception of “ulcerative” (colitis) and “plantar” (fasciitis) - reflecting m BR. Terms which were not previously present in m BR are present in PI. Digestive conditions (appendicitis, hepatitis, and colitis) are near, however close to cystitis (genitourinary).



(a) Full representation.



(b) Zoomed-in on cluster.

Figure 3.21: A t-SNE of PI. Each node is a term and distances represent the measure of association.

$modifiedBioReddit + PatientINF ({}_mBR + PI)$

When filtering by COID terms and highlighting some components of superclasses, Figure 3.22 shows that **inflammatory**, **non-inflammatory**, and others are clustered together. However it seems that **medications** are more separated.

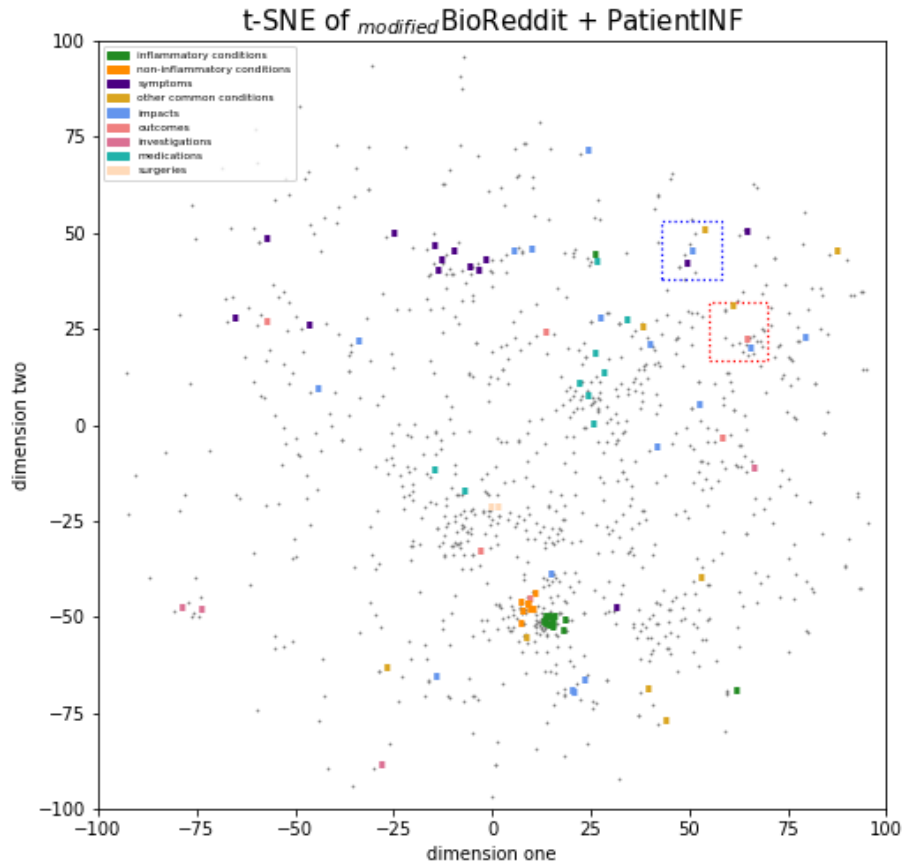
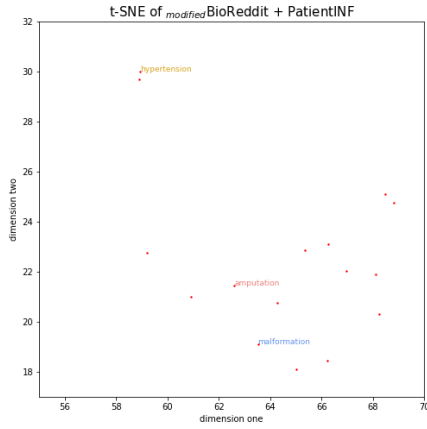
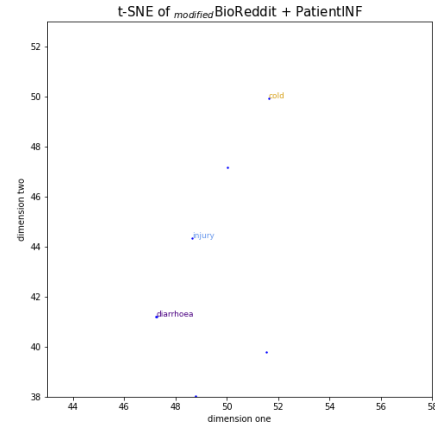


Figure 3.22: Full representation of ${}_mBR + PI$ t-SNE filtered using terms only in COID. Each node is a term and distances represent the measure of association.

When zoomed in on Figure 3.22 dotted boxes, Figure 3.23 (page 124) shows some terms clustered. The **red box**, Subfigure 3.23a, reflects Subfigure 3.20b (page 122) of PI: showing these terms (**hypertension**, **amputation**, and **malformation**) clustered together. The **blue box**, Subfigure 3.23b shows diarrhoea (**symptoms**), injury (**impacts**), and cold (**common conditions**) clustered together.



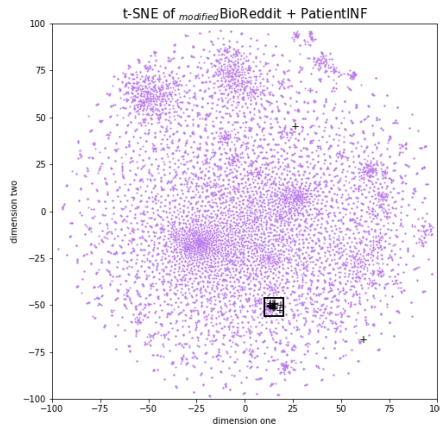
(a) Red highlighted.



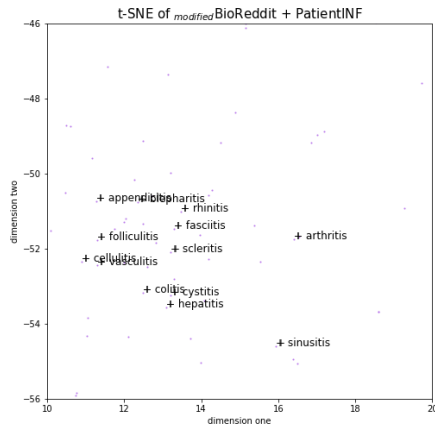
(b) Blue highlighted.

Figure 3.23: Two close-up t-SNE plots of $mBR + PI$ filtered with COID terms. Each node is a term and distances represent the measure of association.

In Figure 3.24, all inflammatory terms of interest are clustered together with the exception of “plantar” and “ulcerative” - reflecting both mBR and PI . Interestingly, appendicitis and blepharitis are very near; rhinitis and sinusitis have separated; and appendicitis has separated from colitis and hepatitis.



(a) Full representation.



(b) Zoomed-in on cluster.

Figure 3.24: A t-SNE of $mBR + PI$. Each node is a term and distances represent the measure of association.

Semantic similarities

Looking into the top similar terms of the 14 terms of interest, below I characterise and represent the semantic similarity using Vec2graph [272]. See Appendix Section C (page 200) for additional characterisations.

Appendicitis

In $_m$ BR the most similar terms are UTIs (abbreviation for urinary tract infection), “tendinitis” (misspelling of tendonitis: inflammation or irritation of a tendon), dermatitis (general term that describes a common skin irritation), tonsillitis (inflammation of the tonsils), diverticulitis, pancolitis, and colitis. In PI the most similar terms are appendix, cystitis, blepharitis, osteomyelitis (inflammation occurring in the bone), spondylitis, scleritis, and STIs (possibly an abbreviation for sexually transmitted infection). In $_m$ BR + PI, the most similar terms STIs, cystitis, blepharitis, spondylitis, bronchitis (infection of the main airways of the lungs), colitis, and appendix.

Atrophic vaginitis

In $_m$ BR the most similar terms are rhinitis, costochondritis (inflammation of the cartilage that joins ribs to breastbone), uveitis, cystitis, and osteomyelitis. In PI the most similar terms are rhinitis, arachnoiditis (inflammation of the arachnoid, membrane that surrounds nerves of the spinal cord), scleritis, urethritis (inflammation of the urethra), and osteomyelitis. In $_m$ BR + PI the most similar terms are urethritis, rhinitis, scleritis, meningitis (inflammation of the lining around brain and spinal cord), uveitis, and arachnoiditis.

Plantar fasciitis

In $_m$ BR the most similar terms are prostatitis (inflammation of the prostate gland), osteomyelitis, spondylitis, osteoarthritis, cellulitis, rhinitis, UTIs, uveitis, costochondritis, and cystitis. In PI the most similar terms are fasciitis, plantar, rhinitis, cystitis, blepharitis, proctitis (inflammation of the rectum), UTIs, hidradenitis (skin condition that causes abscesses and scarring), hepatitis, and planus (Lichen planus: a rash). In $_m$ BR + PI the most similar terms are fasciitis, plantar, rhinitis, uveitis, blepharitis, proctitis, cystitis, meningitis, UTIs, and scleritis. Figure 3.25 (page 126) illustrates the semantic networks.

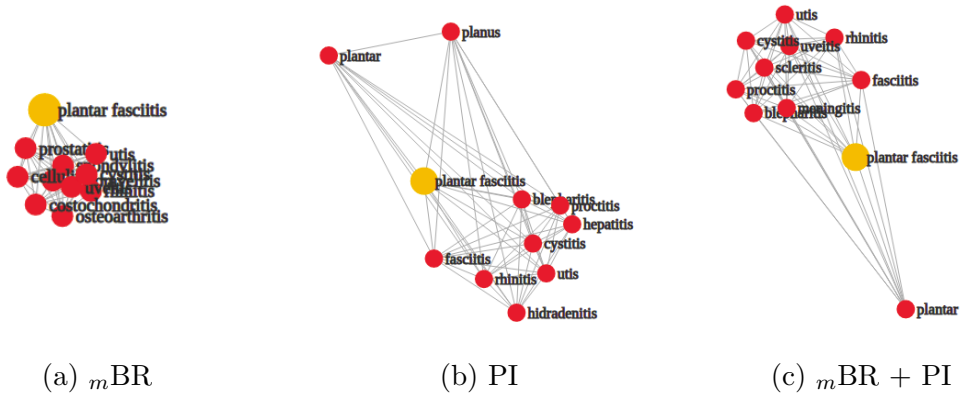


Figure 3.25: Vectorisation representation of plantar fasciitis.

Rhinitis

In mBR the most similar terms are uveitis, dermatitis, osteomyelitis, cystitis, colitis, and UTIs. In PI the most similar terms are hidradenitis, osteomyelitis, tendonitis, cystitis, arachnoiditis, and UTIs. In $mBR + PI$ the most similar terms are uveitis, hidradenitis, “tendinitis” (misspelling of tendonitis), osteomyelitis, pancolitis, and UTIs.

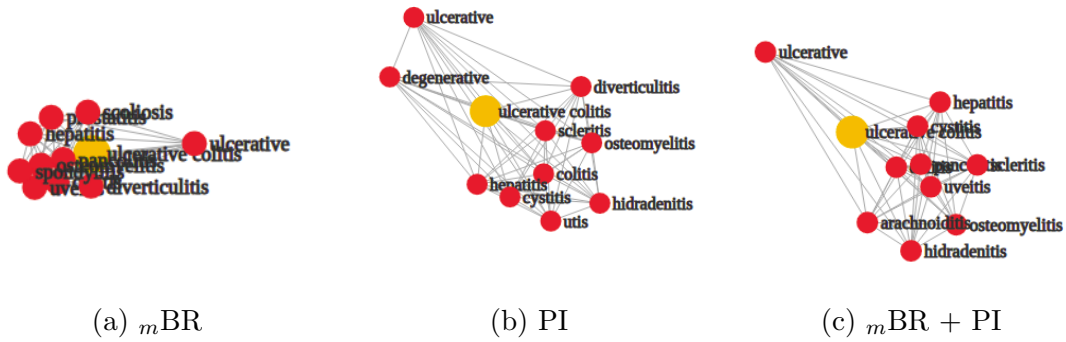


Figure 3.26: Vectorisation representation of ulcerative colitis.

Ulcerative colitis

In mBR the most similar terms are colitis, pancolitis, osteomyelitis, hepatitis, uveitis, diverticulitis, prostatitis, spondylitis, ulcerative, and scoliosis (curvature of the spine). In PI the most similar terms are ulcerative, colitis, hepatitis, cystitis, scleritis, degenerative, osteomyelitis, diverticulitis, hidradenitis, and UTIs. In $mBR + PI$ the most similar terms

are ulcerative, colitis, pancolitis, cystitis, uveitis, hepatitis, hidradenitis, osteomyelitis, arachnoiditis, and scleritis. Figure 3.26 illustrates the semantic networks.

Systematic comparisons

Paired comparison test

A Wilcoxon signed-rank test (W) was conducted of $_m\text{BR}$ to $_m\text{BR} + \text{PI}$, and PI to $_m\text{BR} + \text{PI}$, to observe similarly score changes with the expansion of the models. In Table 3.8, p-values of appendicitis ($_m\text{BR}$ to $_m\text{BR} + \text{PI}$) and vasculitis (PI to $_m\text{BR} + \text{PI}$) were large.

Condition	$_m\text{BR}$ to $_m\text{BR} + \text{PI}$ W p-value	PI to $_m\text{BR} + \text{PI}$ W p-value
appendicitis	p=0.166	p=7.030e-41
arthritis	p=7.209e-14	p=1.891e-209
atrophic vaginitis	p=7.654e-156	p=5.374e-34
blepharitis	p=4.889e-05	p=1.977e-21
cellulitis	p=0.003	p=6.807e-272
cystitis	p=2.872e-74	p=4.933e-20
folliculitis	p=1.834e-115	p=1.307e-12
hepatitis	p=0.0003	p=0.004
plantar fasciitis	p=0*	p=0.0003
rhinitis	p=0.001	p=1.809e-05
scleritis and episcleritis	p=1.021e-56	p=6.486e-05
sinusitis	p=8.325e-22	p=9.288e-12
ulcerative colitis	p=6.428e-17	p=0*
vasculitis	p=8.253e-64	p=0.167
ALL	p=4.203e-24	p=8.460e-21

Table 3.8: The inflammatory terms of interest and their Wilcoxon signed-rank test (W) p-values from the single-domain embedding models to *modified* BioReddit + PatientINF ($_m\text{BR} + \text{PI}$).

*computed value so small it output as 0 (zero).

Pearson correlation coefficient

Following the Pearson correlation coefficient (r) experimentation from ClinicalBERT’s Huang et al. [206] paper, I look at how each embedding model similarities compare to clinical annotations by Pedersen et al. [200] in Table 3.9 (page 128).

Looking at the results, I can conclude some interesting observations, firstly all models agree kidney and renal failure are most similar, however, interestingly $_m\text{BR} + \text{PI}$ is 0.9.

In $_m\text{BR} + \text{PI}$, metastasis (term 1) and adenocarcinoma (term 2) was scored higher than the physician, with $_m\text{BR}$ scoring is almost a 100% match. On the other hand, in $_m\text{BR} + \text{PI}$ stroke (term 1) and infarct (term 2) was almost a 0% match.

Term 1	Term 2	Physician	mBR	PI	$mBR + PI$
renal failure	kidney failure	1.0	1.0	1.0	0.9
heart	myocardium	0.767	0.149	0.436	0.414
stroke	infarct	0.667	0.428	0.028	0.0
abortion	miscarriage	0.667	0.335	0.561	0.504
delusion	schizophrenia	0.667	0.605	0.300	0.202
congestive heart failure	pulmonary edema	0.667	0.558	0.294	0.261
metastasis	adenocarcinoma	0.567	0.924	0.455	0.657
calcification	stenosis	0.567	0.719	0.598	0.476
diarrhea	stomach cramps	0.433	0.760	0.630	0.593
mitral stenosis	atrial fibrillation	0.433	0.811	0.751	0.675
chronic obstructive pulmonary disease	lung infiltrates	0.433	0.584	0.545	0.531
rheumatoid arthritis	lupus	0.333	0.888	0.527	0.658
brain tumor	intracranial hemorrhage	0.333	0.638	0.612	0.542
carpel tunnel syndrome	osteoarthritis	0.333	0.811	0.514	0.648
diabetes mellitus	hypertension	0.333	0.462	0.575	0.587
acne	syringe	0.333	0.285	0.649	0.405
antibiotic	allergy	0.233	0.308	0.462	0.437
cortisone	total knee replacement	0.233	0.536	0.339	0.3460
pulmonary embolus	myocardial infarction	0.233	0.430	0.303	0.239
pulmonary fibrosis	lung cancer	0.233	0.722	0.675	0.708
cholangiocarcinoma	colonoscopy	0.100	0.237	0.226	0.167
lymphoid hyperplasia	laryngeal cancer	0.100	0.763	0.584	0.523
multiple sclerosis	psychosis	0.0	0.628	0.344	0.375
appendicitis	osteoporosis	0.0	0.871	0.425	0.285
rectal polyp	aorta	0.0	0.0	0.324	0.232
xerostomia	alcoholic cirrhosis	0.0	0.724	0.596	0.556
peptic ulcer disease	myopia	0.0	0.403	0.422	0.428
depression	cellulites	0.0	0.143	0.0	0.146
varicose vein	entire knee meniscus	0.0	0.430	0.378	0.376
hyperlipidemia	metastasis	0.0	0.622	0.585	0.602
Pearson correlation coefficient			0.220	0.267	0.255

Table 3.9: Pearson correlation coefficient table comparing normalised physician scores by Pedersen et al. [200] to *modified*BioReddit + PatientINF ($mBR + PI$).

To compare these results with CB (Table 3.6, page 117), diarrhea (term 1) and stomach cramps (term 2) was scored 0.433 by the physician, 0.465 by CB, 0.780 PI, and 0.805 CB + PI. However in this experiment, PI scores these two terms as 0.630, mBR 0.760, and $mBR + PI$ 0.593. These are interesting results as whilst PI is higher than the physician in both experiments, CB is lower yet mBR is higher. Moreover, CB + PI had a higher similarity score compared to $mBR + PI$.

As previously investigated, CB similarities for diarrhea and stomach cramps included misspellings or nausea; PI also had misspellings but more symptoms: such as bloated, and vomiting; and CB + PI included CB’s misspellings but also PI symptoms (mixture of both). In this experiment, there are fewer misspellings: in mBR , there are symptoms such as dizzy, fatigue, and bleed; PI had nausea, bloating, and pain; and $mBR + PI$ included vomiting, constipation and again pain.

Repeating with multiple seeds

When repeating the experiment using different seeds, we can see in Table 3.10 that the r varies by each seed, however slightly consistent, excluding seed 3 & 4.

Seed	PI	$_m\mathbf{BR} + \mathbf{PI}$
1	0.212	0.233
2	0.267	0.255
3	0.136	0.143
4	0.067	0.194
5	0.234	0.290
Average r	0.183	0.223

Table 3.10: Pearson correlation coefficient (r) table comparing different seeds to present consistency of results in *modified*BioReddit + PatientINF ($_m\mathbf{BR} + \mathbf{PI}$).
Seed mapping: 1=1, 2=123, 3=808, 4=6298948, 5=2014428590885440.

3.4 Discussion

In this chapter, I have investigated the patient voice from online forum conversations, and compared to clinical letters and other patient-generated texts, from social media. Firstly, I developed the Combined Ontology for Inflammatory Diseases (COID), based on ICD-10 and similarly to DOID (Disease Ontology) for disease however specific to inflammation and extends to symptoms, medical therapies, and more. COID (version 1.0) is available as an open-source tool, via GitHub⁶ and BioPortal⁷ [147].

Using the preexisting ClinicalBERT [206] embedding model representing MIMIC-III [220] ICU clinical letters, I avoided the time-consuming and computational-heavy job of training a model with the intensive parameter-tuning task. Additionally, this avoids the difficult task of gaining access to difficult-to-reach clinician-generated texts. Huang et al. [206] had detailed descriptions regarding the number of clinical letters included, average word count, and embedding parameters. Using the same text preprocessing methods, building the PatientINF and ClinicalBERT + PatientINF (CB + PI) models meant three separate models were initialised and trained in the same way: with similar size input, word count, and matrix size.

⁶<https://github.com/sap218/coid>.

⁷<https://bioportal.bioontology.org/ontologies/COID>.

Despite the initial aim to use BioReddit [225], there lacked information and original data and so in good practice, it could not be used: BioReddit could be over or under-representing the domain. Instead, using the COMETA corpus [226], a subset of BioReddit, was beneficial as I used the same methods as ClinicalBERT to create *modified*BioReddit, PatientINF, and the *modified*BioReddit + PatientINF (*m*BR + PI) models.

These embedding models, specifically the retrained combined models, provide insight into the patient voice and semantic similarity from two domains in that CB + PI merges clinician and patient-generated texts, and *m*BR + PI merges two different patient-generated texts: forum and social media. Using fastText allows n-gram information to be retained in the models, despite computational resources.

The statistical analysis of these models revealed deep insight into clinical and patient-generated texts. The t-SNE shows the differences and similarities in that consistently inflammatory and non-inflammatory terms cluster together, with medications and symptoms. An important aspect of this study is the characterisation comparisons: despite manual comparisons of similar terms to an inflammatory term of interest, it shows the type of semantic scores in the models. Additionally, the Pearson correlation coefficient (r) experimentation, followed from Huang et al. [206], in which each embedding model similarities are compared to physician annotations by Pedersen et al. [200], we can see similarities between CB and the physician (two domain expert cohorts) and *m*BR and PI (two patient sources), but also differences. Despite r results differing between seed experiments, this form of validation presented somewhat consistent results.

The model ClinicalBERT + PatientINF (CB + PI) can be used as a source of secondary resource as it is openly available on Hugging Face (<https://huggingface.co/sap218/PatientINF>)⁸, a modern repository for ML datasets. With the publicly available data methods details, this model can be reused and retrained to compare or combine other domains.

⁸with extension documentation and scripts available on GitHub (<https://github.com/sap218/PatientINF>).

3.4.1 Implications

In all models, inflammatory terms, in general, were somewhat clustered together, with the exception of plantar fasciitis and ulcerative colitis. A possible reason for differences with plantar fasciitis specifically in PI (experiment one) is that semantic similarity revealed terms of body-related terms, such as “fascia”.

Experiment 1

In the first experiment, looking into comparing CB and PI, there were differences in that investigations category terms and the diseases were much closer in PI. Additionally in CB + PI symptoms came closer together. From the t-SNE, surgeries appearing close together reflects literature in that appendectomy and tonsillectomy have been reported to be risk factors for other diseases, such as both being common in patients with CD [273], and those who had a tonsillectomy subsequently developed appendicitis (resulting in an appendectomy) [274]. PI medication, symptoms, and investigations were much semantically closer compared to CB, but blepharitis was not as near to scleritis. The term stroke was near surgeries, and there is a relationship between strokes and non-major surgeries [275]. CB + PI reflected PI in that categories clustered closer, however, fasciitis was more separated from the other inflammatory conditions, yet “ulcerative” and “colitis” were much closer together in this model.

There were a surprising amount of misspellings in the most similar terms and PI used more abbreviations. The PI model consisted more of ocular IMIDs possibly due to visual issues and their major effect on quality of life. The combined CB + PI reflected mostly CB but did have some weighing from PI, however, did also include terms not most similar in the other two models. Many similarities reflect literature, in that CB consisted of other conditions, which can reflect misdiagnosis, diseases with similar conditions (for investigations), or manifestations.

There was little overlap between CB and PI, yet CB + PI contained majority CB similar terms and valuable terms derived from PI. From the manual characterisations of semantic similarities, in CB many clinical insights reflected literature in that [folliculitis]

two digestive conditions, diverticulitis and colitis are similar, which when looking into literature, folliculitis is a manifestation of IBD [276]; [rhinitis] both vaginitis and vulvovaginitis are manifestations of rhinitis [277, 278]; and finally [sinusitis] mastoiditis can be hard to diagnose as they present similar symptoms [279]. In PI, [rhinitis] both vaginitis and vulvovaginitis are manifestations of rhinitis [277, 278]; and [scleritis and episcleritis] there is known ocular complications and manifestations of IBD [280]. Some interesting associations which contribute to literature includes in CB [appendicitis] and cholecystitis are very similar yet rarely occur in union [281]; and in PI [appendicitis] has previously shown to be a complication of pancreatitis [282].

The Wilcoxon signed-rank test (W) shows the retraining changes semantic scoring significantly, with the exception of CB cellulitis. The r was higher in CB + PI, especially when validating with multiple seeds. But interestingly, PI r of diarrhea and stomach cramps was high showing patient voice differences - the physician scored diarrhea and stomach cramps lower than 50% similarity, which CB was almost in agreement, moreover, some clinical and physician scores did not match.

Experiment 2

Looking into comparing m BR and PI, two different sources of the patient voice, there were differences. In m BR, the t-SNE shows that categories were not as clustered together, however, PI and m BR + PI investigations were much more scattered. PI infliximab (medication) is used to treat autoimmune conditions, and being near IBS (irritable bowel syndrome), which is misdiagnosed with IBD, is interesting. In the m BR model, all categories were somewhat clustered together with exception of impacts (cause of inflammation). The term ileostomy being near endoscopy reflects literature in that both are used to diagnose/treat CD [107, 113]. m BR + PI embedding model revealed appendicitis and blepharitis are very near; rhinitis and sinusitis have separated; and appendicitis has separated from colitis and hepatitis. Additionally, there were more similarities and reflections to PI that the same patterns were found: hypertension, amputation, and malformation clustered together.

Reflecting PI from the first experiment, there was more use of abbreviations in m BR and PI patient models. Majority of the similar terms in m BR terms were repeated, such as uveitis, osteomyelitis, spondylitis, and pancolitis. m BR + PI consisted majority of terms from PI, such as appendicitis, plantar fasciitis, and rhinitis. Each model [arthritis] had ocular IMIDs which is known to have an association with arthritis; [atrophic vaginitis] in m BR and PI osteomyelitis is most similar, however in m BR + PI that power is lost and meningitis has a higher score; and finally [ulcerative colitis] pancolitis was most similar.

Interestingly, ocular IMIDs and digestive (specifically colitis) occurred in the models, reflecting literature that ocular diseases and digestive have a high association. There was also frequent mention of other conditions, such as osteomyelitis and cystitis. One major difference between patient-patient and clinical-patient model comparisons is the lack of misspellings, this may be due to more text data included in the first experiment. The W from the single-domain to combined models resulted in significant p-values, with the exception of m BR appendicitis and PI vasculitis. The r , when compared to the first experiment of CB, m BR had a higher similarity score of diarrhea and stomach cramps, in comparison to PI, which was still higher than CB. Despite PI having the higher r result, on average m BR + PI performed better during the seeds validation.

Overall, CB + PI is the best as it enhances the clinical domain with the patient voice from two large sources of text. Further retraining with similar sub-sampled datasets will only enhance the model.

3.4.2 Limitations and future endeavours

The work from Gangemi et al. [264], which described an inflammatory ontology and the biological processes, can be used in future to better improve COID to include biological elements.

Embedding models

Firstly, a comparison between CB, m BR, and PI would be better suited in a 7-way comparison⁹. However, from the experiments, COMETA was under-representing the domain and so a fair multi-model comparison would not be good practice: major bias would be introduced.

CB itself is limited in scope, both with respect to the role of the texts: ICU clinical letters. Clinical letters can represent discharge summaries or contextual descriptions for GPs. Additionally, being ICU restricts the domain to critical care patients. Despite training patient posts covering 1369 topics, this could over or under-represent medical, patient-generated texts. Whilst ICU can cover a wide range of critical care situations, many topics discussed by patients could be more general or less severe. The m BR has limitations due to lack of input data, there was repetition of terms from the semantic characterisations. However, the overwhelming misspellings in the CB study may represent too much text data.

In both experiments, the varied r scores from the seeding validation methods were lacking in consistent results. However, without proper clinical interpretation, these variations in scores are not necessarily positive or negative and results are very subjective. In future, despite both CB + PI and m BR + PI having consistently better averaged r than the single-domain models, a large-scale with computational resources could conduct a 100+ seed validation experiment, however, if r scores fluctuate with only 5 seeds then statistically they may continue with more tests.

3.4.3 Conclusions

I have demonstrated a systematic comparison of embedding models derived from various domains, both clinical and patient-generated. These models show that the clinical domain is confounded with misspellings but also semantics complete of terms being similarly used for purpose of patient summaries. When the clinical domain is expanded with lay

⁹CB, PI, BR, CB+PI, BR+PI, CB+BR, and CB+PI+BR.

terms, patient-generated conversation, the embedding model is more informative and expressive of the patient voice. Moreover, from a comparison of two patient-derived embedding models, numerous terms were repeated as well as ocular IMIDs plus digestive inflammatory conditions: possibly indicating these conditions have a higher priority in a patient's expression, and more specifically quality of life.

Summary

1. Addressed the challenge for open embedding models for secondary research, that integrate the patient voice:
 - Extracted the full Patient UK forum board of patient-generated text and released code online for others to utilise.
 - Built an embedding model derived from clinical letter and patient forum conversations, ClinicalBERT + PatientINF - released for secondary use.
 - Developed an application ontology, Combined Ontology for Inflammatory Diseases (COID), using ICD-10 as a foundation to meet the need for a domain terminology system of inflammation, with application potential.
 - COID enhanced with patient-preferred terms, some derived from vectorisation.
2. Overcame the obstacle posed training a model by scratch via using a preexisting embedding model.
3. Systematically compared the clinician voice and two sources of the patient voice for an insight into semantics and individual priorities: revealing novel semantics of terms and disease associations, with some reflecting literature.
4. Semantic characterisations of the models revealed clinician-generated models consisted of misspellings and patient-models consisted of abbreviations.
5. Analysis has bridged gaps in clinician-patient understanding of inflammatory conditions, allowing us to understand semantics, priorities, and word associations deeper.

4 A BIOLOGICAL PERSPECTIVE INTO INFLAMMATORY BOWEL DISEASE AND SENTIMENT

4.1 Introduction

Patients who have Inflammatory Bowel Disease (IBD) learn to live with their disease, specifically to control their inflammation. There are numerous methods to manage their condition - including medical treatments, maintaining a strict diet plan, or stopping smoking. For these patients, their biomarkers may be stable due to the regulation of their disease. For those who do not-yet have a diagnosis, their intestinal inflammation may be active and so causing many symptoms. To diagnose, there are rigorous investigations, however, very invasive and can take years. Whether or not a patient is having a flare-up, IBD can affect quality of life (QoL).

One solution for better diagnostic methods is looking into blood tests, since they are preferred by the patient and less invasive. Specifically for IBD, there is a need for identifying novel blood biomarkers [94] as many GPs only refer patients for further tests if there is an elevated C-reactive protein (CRP) in blood tests or issues in a stool sample, which is uncomfortable for the patient. Although studies have investigated biomarkers for IBD, or patient quantitative questionnaire responses, this thesis explores blood biomarkers and QoL simultaneously to investigate their relationship with IBD.

As previously discussed in section 1.7.3 (page 70), the UK Biobank (UKB) [240] presents a rich dataset of half a million UK patients. Patients are recruited and attend the

recruitment centre (dated), where they provide numerous measurements, such as basic information, blood measurements, and questionnaire responses. These patients in the UKB have synced hospital records, which can be filtered using ICD-10 codes [153]. These hospital records have a corresponding date that can be used to stratify patients into diagnosed, and those who were not-yet diagnosed at enrolment.

The UKB is a niche resource as it consists of blood marker data, including complete blood count (CBC), such as white blood cells (WBC) leukocyte subpopulations (e.g. lymphocyte and neutrophil) plus platelet count; and biochemistry (such as CRP and types of cholesterol). Additionally, there are QoL and other quantitative data, including a patient's current mental health status or history: happiness, diet, and smoking. Even with access to this data, there still lies the challenge of how to reveal any of these potential relationships. Numerous ML techniques have been applied to clinical data to seek out biomarkers for various inflammatory diseases, including IBD [52, 122, 123]. However many of these experiments lack descriptions of analysis [168] - there has been a need for better reproducibility and explanations/documentation as research can be under or over-estimated.

In this chapter, I present an unsupervised ML approach to cluster time-split cohorts of patients with IBD. All baseline patients, with chronic IBD from the UKB, will be preprocessed in the same manner (such as feature elimination) however, split into three separate groups: IBD patients diagnosed (IBD-D), IBD patients not-yet diagnosed (IBD-NYD), and IBD patients diagnosed with future diagnosis (IBD-D-FD). With each subset, dimension reduction (DR) is performed, various techniques for the optimal number of clusters K for individual clustering algorithms, and many evaluation methods. I conclude with statistical analysis to observe cluster differences of each time-stamped cohort in relation to possible novel blood biomarkers QoL markers, and their relationship.

The goal is being able to characterise and extract valuable information from the clustering and discover the underlying patterns, to better understand the relationship of biomarkers and QoL has on a patient with IBD, not-yet diagnosed, or those with

multiple IBD diagnoses. Blood biomarkers and adjacent QoL-related markers can direct potential novel insight into the patient voice, disease progression, and the role of both biological and external factors have on chronic disease development in evolving diagnostic methods for shortening diagnosis time. This study is a preliminary step into uncovering relationships of patients with IBD.

4.2 Data and methods

Due to the dates of attending the assessment centre, diagnosed date, and acute inflammation persisting for less than 6 weeks [2], it is important to filter out any patient with an acute diagnosis of IBD and investigate solely on those with chronic inflammation - if any acute inflammation is present then this may indicate an acute flare-up. As previously mentioned, inflammatory conditions can be acute or chronic, and ICD-10 codes represent both these acute and chronic conditions. Before analysis can be conducted, knowledge of which IBD ICD-10 codes is chronic is necessary. To overcome this, there are two sources of data that were annotated by teams of experts of each ICD-10 code whether acute or chronic. The two resources were derived from one group with the Healthcare Cost and Utilization Project [283] and another by Calderón-Larrañaga et al. [284].

For IBD, although Crohn's disease (CD) and ulcerative colitis (UC) are chronic conditions, their subtypes may not, and gastroenteritis (GE) is often recurrent. The diagnostic ICD-10 codes for IBD are K50-K52, specifically CD K50, UC K51 (and their subclasses, e.g. K51.4 is UC with inflammatory polyps), and GE K52. With each code annotated by experts being acute or chronic means I can determine which type of IBD code can be used in the study.

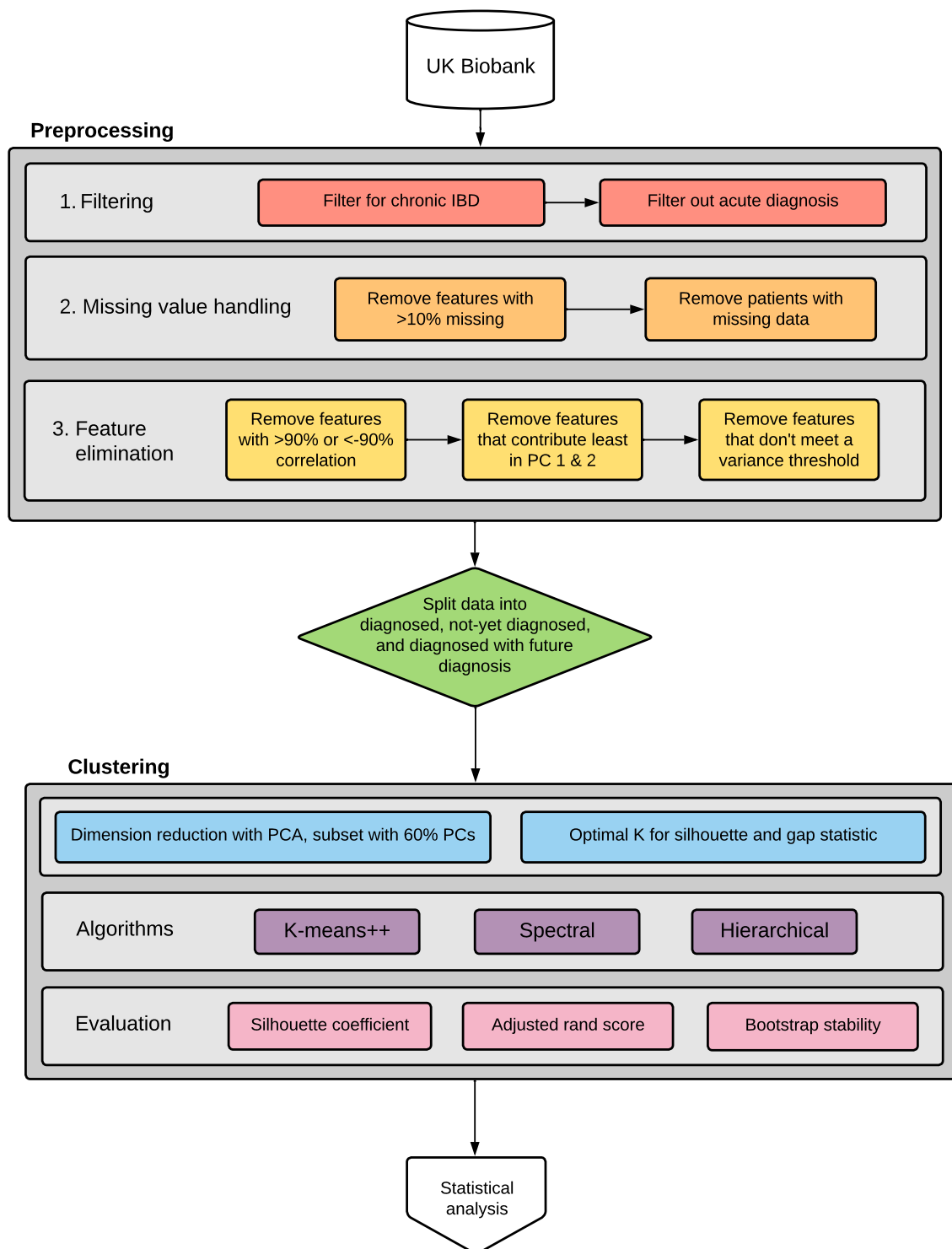


Figure 4.1: Workflow of data preprocessing and Machine learning techniques.

4.2.1 Workflow

From the literature of best practices, a ML workflow approach was developed. These steps follow data preprocessing techniques derived from studies in the literature that have conducted similar experiments. An illustration in Figure 4.1 (page 139) represents the workflow following a detailed account described below:

Workflow:

- Imported all data from UKB, removing any patient sample that did not have an ICD-10 code.
- Preprocessing steps for data wrangling:
 1. Filtering
 - From the two sources of chronic IBD ICD-10 codes [283, 284], patients were filtered to retain those who have been diagnosed with chronic IBD [177].
 - Filtered out any patient who was diagnosed within 6 weeks (± 0.115068 years) of their baseline to avoid acute inflammation.
 2. Missing value handling [166, 176] (section 1.6.2, page 47)
 - Removed features which had $>10\%$ missing data.
 - Removed patients with any missing continuous data values, e.g. BMI or blood markers.
 3. Feature elimination [166] (section 1.6.2, page 47)
 - Standardisation scaling performed to scale the data [166] (section 1.6.2, page 47).
 - Removed features with $>90\%$ and $<-90\%$ from Pearson correlation coefficient (r).
 - Initial Principal component analysis (PCA) (section 1.6.3, page 49) conducted to remove features contributing least in PCs 1 and 2 as these are the two largest explained dimensions [122, 171, 172, 177].
 - Removed features lower than the variance at threshold from Equation A.1 (page 191).
- Remaining patients split into three distinct groups - similarly to time split experiments from the literature of Porcelli et al. [21], Bengtson et al. [115], Welsh et al. [232], and Kikut et al. [114]:
 - **IBD patients diagnosed (IBD-D).**
 - **IBD patients not-yet diagnosed (IBD-NYD).**
 - **IBD patients diagnosed with future diagnosis (IBD-D-FD).**
- Clustering blood markers using multiple algorithms [172]:
 1. A PCA was performed for DR (section 1.6.3, page 48), retaining the most 60% contributing PCs as seen often in the literature [87, 91, 122, 166, 171, 172, 177, 227–229] whilst also considering Cattell’s formulation (1966) [178] and Kaiser’s rule (1960) [179] (section 1.6.3, page 48).
 2. With a maximal cluster of $K = 10$ both a silhouette and gap statistic were conducted [166, 229] for three clustering methods (section 1.6.1, page 44): K-means++, Spectral, and Hierarchical (agglomerative) as these three methods are used frequently used in literature [91, 95, 151, 166, 167, 170–172, 227, 229–231], the same methodology (DR, optimal K , etc.) can be applied, and the results can be visualised via scatter plots of PCs 1 & 2. Other clustering methods were not used due to the assumptions of these methods; for example, DBSCAN is used for data shapes that are arbitrary, has issue with parameter fine-tuning with dense data, and has performance issues with datasets of large dimensions [166].

3. Once the optimal K was retrieved for each clustering algorithm within each time-split dataset, clustering was performed.
 4. To evaluate the clustering (section 1.6.1, page 44), various investigations were done: a second silhouette coefficient [166, 174] for visualisation of cluster size fluctuations, an adjusted Rand index (aRi) score comparing each clustering method to observe statistical agreement of cluster assignments [175], and bootstrap to assess cluster stability [227].
- Finally, the investigation proceeded with statistical analysis in the R statistical language [285].
 - With continuous features (e.g. CBCs) an Analysis of variance (ANOVA) was conducted if $K > 2$, otherwise, a Wilcoxon signed-rank test (W) [114, 115, 228–230].
 - For categorical features (e.g. type of IBD) a chi-squared (χ^2) was conducted [115, 227, 229].
 - A permutation variable importance measure (PIMP) analysis [286, 287] was followed to better understand which features are most informative in a random forest (section 1.6, page 43) for cluster assignment [151, 166, 229].
 - Additionally, a regression (section 1.6, page 43) analysis was conducted to predict particular output variables within clusters [23, 118, 151, 166, 167, 169, 177, 227, 228, 230].
 - Finally, characterisations have been made for each time-split cohort results and possible interpretation in the discussion.

4.2.2 Features included

One of the initial tasks was selecting which features to include in the study. Firstly, important basic information was collected, including the date patients attended the assessment centre, corresponding age, and physical measures (BMI). Secondly, the various biological samples were investigated for inclusion, specifically blood assays results (for a list of blood features and their normal ranges in genetic sex, see Appendix Table D.1, page 202). Finally, various questionnaire responses, from the assessment visit, were retrieved. These ranged from psychosocial, medical history, lifestyle, and early life factors. See below for a list of all features included in this study.

Using the UKB Showcase: <https://biobank.ndph.ox.ac.uk/showcase/search.cgi>

■ Basic information

- Patient ID
- Genetic sex
- Genetic ethnic background
- Date when attended recruitment centre
- Age when attended recruitment centre
- ICD-10 code diagnosed at hospital
- Hospital date corresponding to diagnosed ICD-10 code
- Body mass index (BMI)

■ **Blood assays** <https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=100080>

● Complete blood count (CBC) <https://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=100081>

- Types of white blood cells (WBC)
 - White blood cell leukocyte count
 - Basophil count
 - Eosinophil count
 - Lymphocyte count
 - Monocyte count
 - Neutrophil count
- Types of red blood cells
 - Red blood cell erythrocyte count
 - Haemoglobin concentration
 - Mean corpuscular haemoglobin
 - Reticulocyte count ‡‡
- Platelets
 - Platelet count
 - Mean platelet thrombocyte volume

● Blood biochemistry <https://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=17518>

- Alanine aminotransferase
- Albumin †
- Alkaline phosphatase
- Apolipoprotein A †
- Apolipoprotein B ‡
- Aspartate aminotransferase
- C-reactive protein (CRP)
- Calcium †
- Cholesterol ‡
- Creatinine
- Cystatin C
- Direct bilirubin †
- Gamma-glutamyl transferase
- Glucose †
- Glycated haemoglobin HbA1c
- HDL cholesterol †
- IGF-1
- LDL direct †‡
- Lipoprotein †
- Oestradiol †
- Phosphate †
- Rheumatoid factor †
- SHBG †
- Testosterone †
- Total bilirubin
- Total protein †
- Triglycerides
- Urate
- Urea
- Vitamin D †

■ **Quality of life (QoL)** and other health-related factors <https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=100025>

- Bowel cancer screening
- Diet-related features
 - Bread intake
 - Fresh fruit
 - Processed meat
 - Salt
- Medical therapy-related features
 - Medication
 - Medication count
- Pain location
- Sentiment-related features
 - Ability to confide
 - Anxiety
 - Happiness
 - Health rating
 - Health satisfaction
 - Miserable
- Smoking-related features
 - Ever smoked
 - Maternal smoking
 - Smoking status

Key

- † Features removed due to >10% missing.
- †† Features removed from Pearson correlation coefficient (r).
- ‡ Features removed from Principal component analysis (PCA).
- ‡‡ Features removed from variance threshold.

4.3 Results

When running the methodology via Python [248] with the data structure package, pandas [288]; and the ML, statistical models package, Scikit-learn [249]; the following describes changes to the dataset during each process.

Results:

- Initially started with **502,461** patients and **77** features - 12 being CBC.
 1. Filtering
 - With **18** chronic IBD ICD-10 codes [283, 284] to **8759** patients.
 - Removing patients with acute diagnosis, filtered to **8110** patients.
 2. Missing value handling
 - Removed 14 features with >10% missing, e.g. calcium, testosterone, and more - see the list in section 4.2.2 (page 141) marked appropriately.
 - Removed remaining patients who had any missing data, leaving a total of **6604** patients.

3. Feature elimination

- Removed features with $>90\%$ and $<-90\%$ with Pearson correlation coefficient (r) - see section 4.2.2 (page 141) marked appropriately. Removed LDL direct due to high r with apolipoprotein B and cholesterol - see Figure 4.3.
- PCA performed to remove features contributing least in PCs 1 and 2. Two features removed: cholesterol apolipoprotein B - see the list in section 4.2.2 (page 141) marked appropriately; and Appendix Figure D.1 (page 203).
- Removed feature with least variance: reticulocyte count - see the list in section 4.2.2 (page 141) marked appropriately; and Appendix Subfigure D.2 (page 204).
- Remaining **6604** patients and **24** blood features - see Figure 4.2 for histograms of data split.
 - IBD patients diagnosed (IBD-D) **1813** - Subfigure 4.2a.
 - IBD patients not-yet diagnosed (IBD-NYD) **3860** - Subfigure 4.2b.
 - IBD patients diagnosed with future diagnosis (IBD-D-FD) **931** - Subfigure 4.2b.

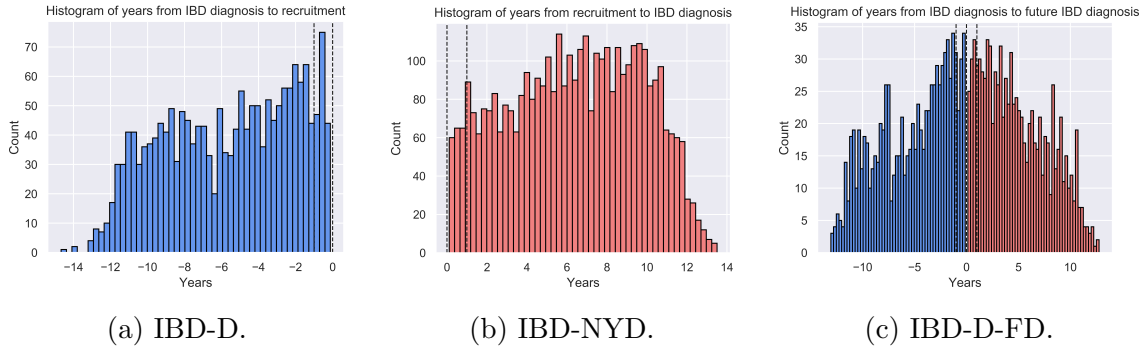


Figure 4.2: Histograms of a patient's diagnosis to recruit date.

4.3.1 Dimensionality reduction

Using PCA for DR, 24 blood features were reduced to: IBD-D $D=7$, IBD-NYD $D=7$, IBD-D-FD $D=6$. See Appendix Figure D.3 (page 205) for scree plots.

Validation was conducted using Cattell's formulation (1966) [171, 178] (PCs 3, 4, & 5 being a similar value indicates error) and using Kaiser's rule (1960) [171, 179] (PCs with a variance less than 1 are not worth retaining).

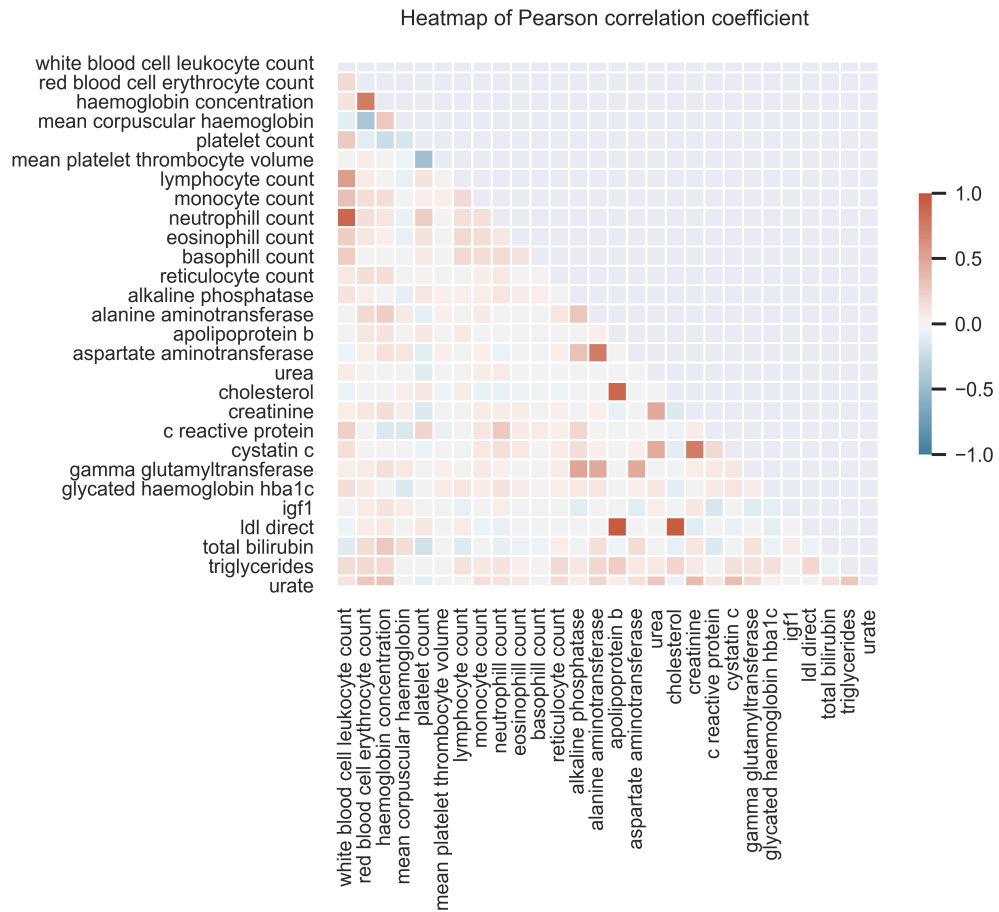


Figure 4.3: Pearson correlation coefficient feature elimination visualisation of feature relationships for removal.

4.3.2 Clustering

4.3.2.1 IBD patients diagnosed

For IBD-D, hierarchical clustering (Figure 4.4, page 146) presented best results in that a silhouette score (Subfigure 4.4a) recommended $K=7$ and the silhouette coefficient (Subfigure 4.4b) cluster sizes are each above the average silhouette scores (red dotted line) and despite minor size fluctuations, the clusters are somewhat evenly distributed. For K-means++ optimal $K=3$ and spectral $K=2$ - an aRi of K-means++ & hierarchical: 0.361. A bootstrapping cluster validation 0.627 (63%) and increases to 0.770 (77%) when removing two clusters of patient outliers ($K=4$ $n=5$, and $K=5$ $n=33$).

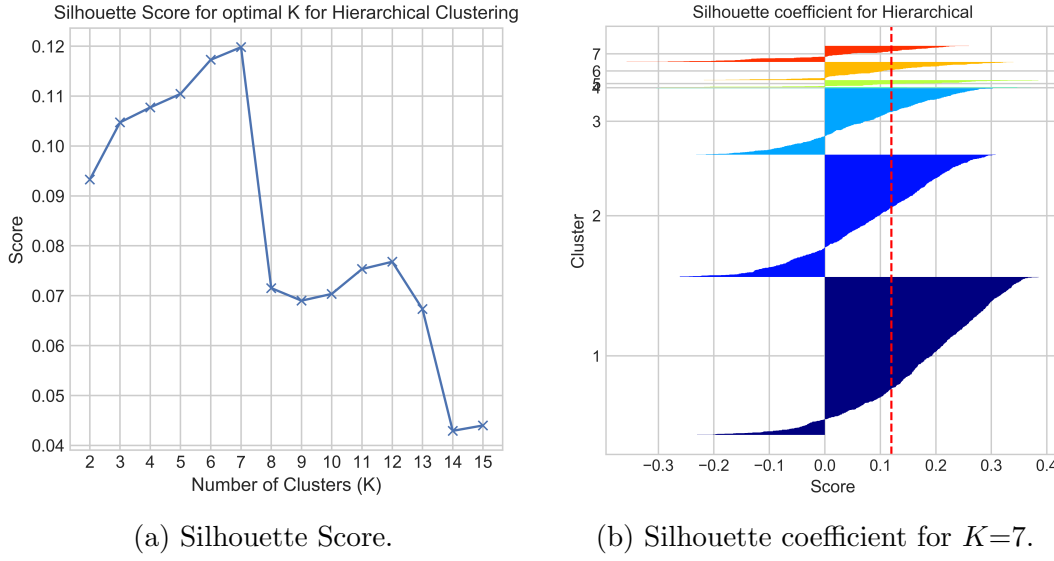


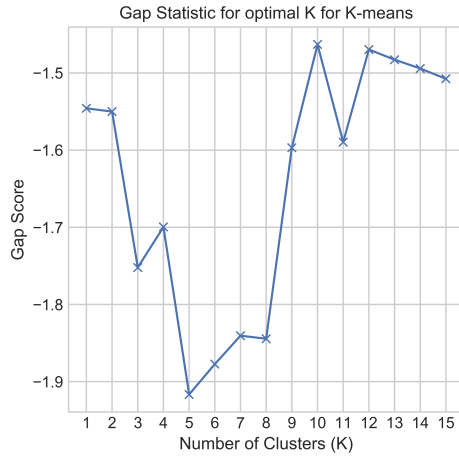
Figure 4.4: Hierarchical for IBD-D.

4.3.2.2 IBD patients not-yet diagnosed

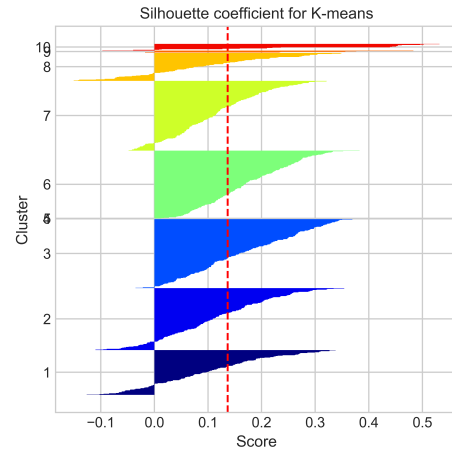
For IBD-NYD, spectral clustering presented best results in that a gap statistic was $K=3$ - whilst K-means++ also suggested $K=3$ from the gap statistic, the highest gap score was with Spectral. See Appendix Figure D.4 (page 206) for the gap statistic figures. An aRi of K-means++ & spectral: 0.621. A bootstrapping stability of 87% (0.873).

4.3.2.3 IBD patients with future diagnosis

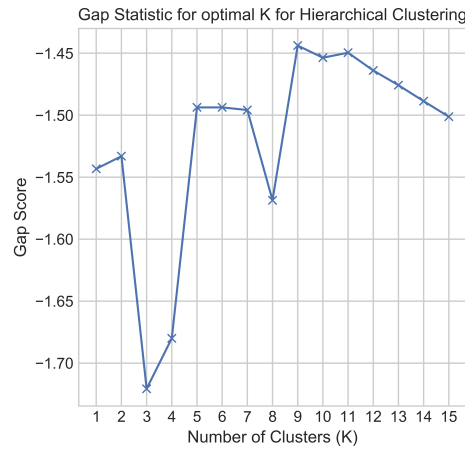
For IBD-D-FD, both hierarchical and K-means++ present promising clusters, see Figure 4.5: hierarchical $K=9$ and K-means++ $K=10$. Whilst the gap score for hierarchical is highest for $K=9$, the silhouette coefficient plot for K-means++ is more positive in that cluster sizes do not have wide fluctuations and average scores are positive. An aRi of K-means++ & hierarchical: 0.319. Bootstrapping for hierarchical 0.554 and K-means++ is 0.525. Overall, hierarchical will be used due to higher gap score and higher clustering stability (55%). When removing outliers, bootstrapping is 0.647 (65%) for $K=2$ ($n=23$), $K=5$ ($n=1$), $K=8$ ($n=7$), and $K=9$ ($n=1$).



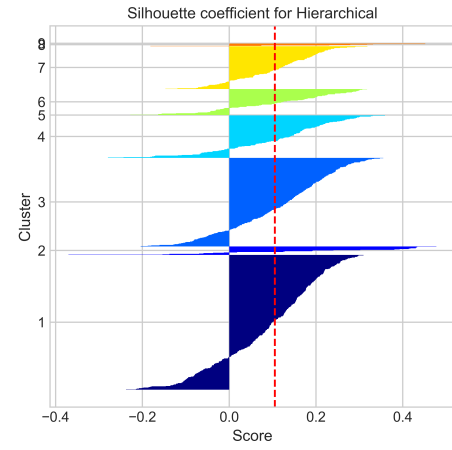
(a) K-means++ gap.



(b) K-means++ coefficient.



(c) Hierarchical gap.



(d) Hierarchical coefficient.

Figure 4.5: Gap statistic and silhouette coefficient for IBD-D-FD.

4.3.3 Statistical analysis

4.3.3.1 IBD patients diagnosed

See Table 4.1 (page 148) for averages of each feature in the assigned cluster, with ANOVA and x^2 p-values. See Figure 4.6 (page 149) for the corresponding box plots of means.

Feature	1	2	3	6	7	adj p-value ANOVA x^2	
size	736	570	309	85	75		
BMI	25.924	28.253	28.801	27.806	28.039	p=2.07e-08	
age	56.412	58.547	58.628	55.035	61.920	p=0.006	
diagnosed time	-5.589	-5.596	-5.482	-6.268	-5.093	p=1	
WBC leukocyte count	6.491	6.462	9.496	8.141	6.970	p=2.58e-32	p=9.13e-80
red blood cell erythrocyte count	4.233	4.713	4.605	4.404	4.213	p=0.011	p=1.59e-26
haemoglobin concentration	13.422	14.852	14.538	12.251	13.112	p=0.002	p=3.47e-69
mean corpuscular haemoglobin	31.820	31.606	31.628	27.919	31.208	p=8.51e-25	p=1.17e-50
platelet count	261.861	237.599	278.644	364.605	245.981	p=2.67e-14	p=1.38e-43
mean platelet thrombocyte volume	9.280	9.155	9.318	8.534	9.117	p=0.002	p=0.199*
lymphocyte count	1.810	1.606	2.336	1.630	1.496	p=1	p=1.38e-10
monocyte count	0.435	0.491	0.685	0.487	0.511	p=3.05e-05	p=0.559*
neutrophil count	4.043	4.164	6.125	5.814	4.735	p=3.15e-39	p=1.01e-45
eosinophil count	0.165	0.168	0.285	0.179	0.186	p=0.0002	p=0.001
basophil count	0.033	0.025	0.062	0.027	0.034	p=0.698*	p=5.16e-11
alkaline phosphatase	81.818	89.908	93.093	100.065	105.551	p=1.36e-19	p=1.88e-07
alanine aminotransferase	17.846	28.723	24.799	16.646	21.005	p=1	p=4.38e-11
aspartate aminotransferase	23.674	30.093	26.902	21.531	26.793	p=1	p=1.34e-22
urea	4.944	5.540	5.537	4.800	8.171	p=1.59e-38	p=3.45e-79
creatinine	65.598	79.521	75.266	68.820	105.780	p=2.88e-58	p=1.37e-90
CRP	2.313	2.732	4.861	16.890	7.467	p=2.55e-80	p=6.75e-49
cystatin C	0.865	0.975	0.980	0.964	1.376	p=1.57e-100	p=2.35e-75
gamma-glutamyl transferase	26.088	50.600	49.420	32.773	53.167	p=6.04e-09	p=4.51e-08
glycated haemoglobin HbA1c	35.369	36.269	38.939	38.602	41.165	p=7.98e-14	p=2.61e-09
IGF-1	21.405	22.031	21.852	19.187	21.780	p=1	
total bilirubin	8.503	11.010	8.569	6.435	8.531	p=0.129*	p=0.007
triglycerides	1.455	2.106	2.268	1.680	1.931	p=2.28e-07	p=3.93e-31
urate	259.796	352.823	335.402	294.424	415.007	p=3.15e-49	p=3.93e-12
NLR	2.564	2.949	3.035	4.818	4.118	p=9.13e-25	
ethnic background (W)	96%	97%	98%	88%	96%		p=0.391*
genetic sex	78% F	76% M	64% M	76% F	69% M		p=1.17e-98
comorbidity	12% S-T	28% K	30% K	11% D	16% I		p=0.003
IBS type †	92% w/o	95% w/o	7% (1% D)	9% (1% D)	5% (4% D)		p=0.095*
medication ††	11% I	55% O	2% L	37% P	38% A		p=3.46e-11
medication count	17% (1)	12% (0)	27% (4-6)	19% (10+)	27% (6-8)		p=2.05e-08
processed meat	31% (1)	1% (5)	30% (2)	28% (3)	8% (4)		p=4.83e-09
miserable (y)	50%	40%	46%	55%	50%		p=0.127*
anxiety (y)	68%	54%	60%	75%	50%		p=4.48e-06
ability to confide ‡	53% D	22% N	18% N	43% D	32% N		p=2.52e-05
health rating (poor)	10%	9%	17%	29%	29%		p=7.50e-13
happy (y)	97%	96%	92%	81%	90%		p=0.060*
health satisfaction (y)	77%	78%	65%	52%	55%		p=0.015
pain location (abdominal)	8%	5%	5%	6%	7%		p=0.001
smoking status (currently)	8%	7%	20%	6%	4%		p=1.24e-09
ever smoked (y)	63%	67%	74%	58%	59%		p=0.049

Table 4.1: Table of IBD patients diagnosed p-values for Analysis of variance (ANOVA) and chi-squared (x^2).

*p-value significant before adjusted for multiple testing.

Key

† w/o=without, D=with diarrhoea.

†† A=Aspirin¹, I=Ibuprofen², L=Laxatives³, P=Paracetamol⁴, O=Other.

‡ D=daily, N=never.

¹medication used to reduce pain, fever, or inflammation.

²medication used for treating pain, fever, and inflammation.

³medication used for loosening stools and increases bowel movements.

⁴medication used to treat fever and mild to moderate pain.

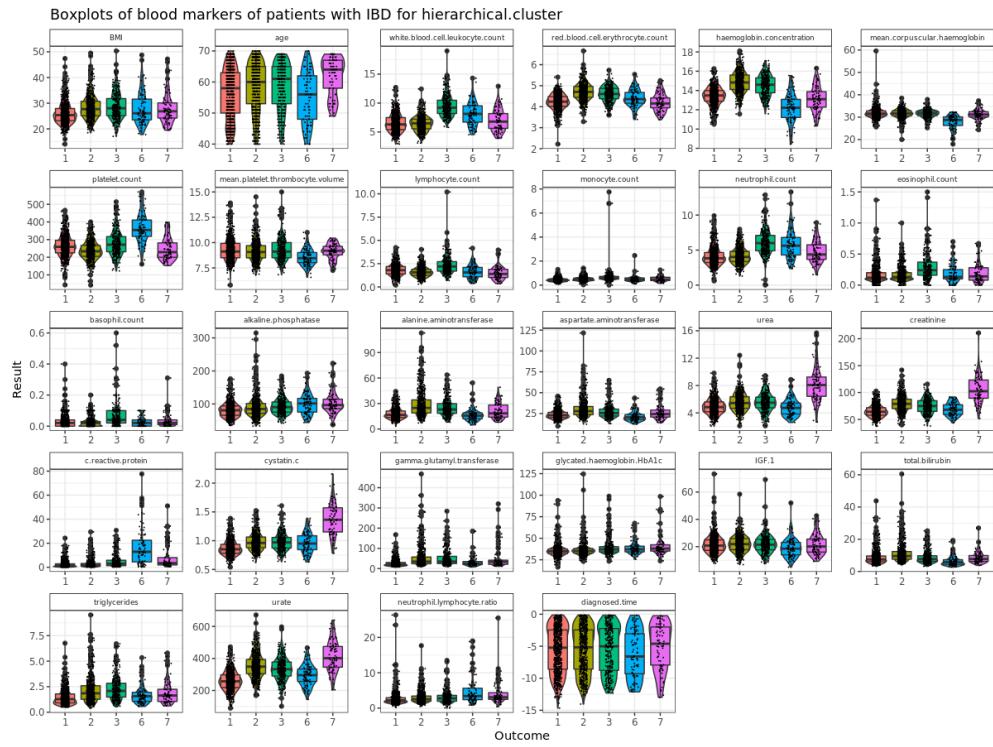


Figure 4.6: Figure of IBD-D, box plots of averaged blood markers.

Below in Figure 4.7 two QoL features, which p-value was less than 0.05, are visualised. Cluster 6 have 75% “yes” to having anxiety (Subfigure 4.7a) and 52% are unsatisfied (unhappy) with their health (Subfigure 4.7b).

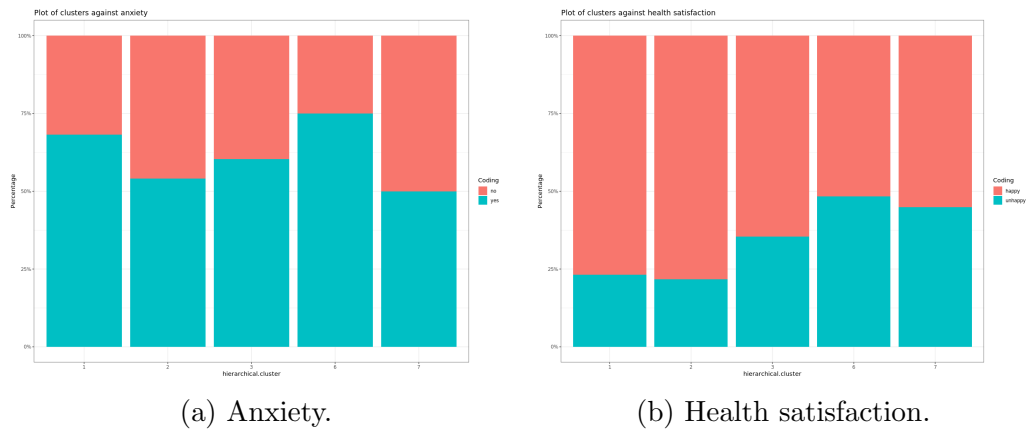


Figure 4.7: Bar plots of IBD-D.

4.3.3.2 IBD patients not-yet diagnosed

See Table 4.2 for averages of each feature in the assigned cluster, with ANOVA and x^2 p-values. See Appendix Figure D.5 (page 207) for the corresponding box plots of means.

In Figure 4.8, three smoking-related QoL features, which p-value was less than 0.05, are visualised alongside future IBD diagnosis. IBD diagnosis (Subfigure 4.8d) cluster 3 have 29% CD. Cluster 3 are also the strongest cohort of smokers: 22% currently smoke (Subfigure 4.8a); 69% have ever smoked (Subfigure 4.8b); and 34% have a history of maternal smoking (Subfigure 4.8c).

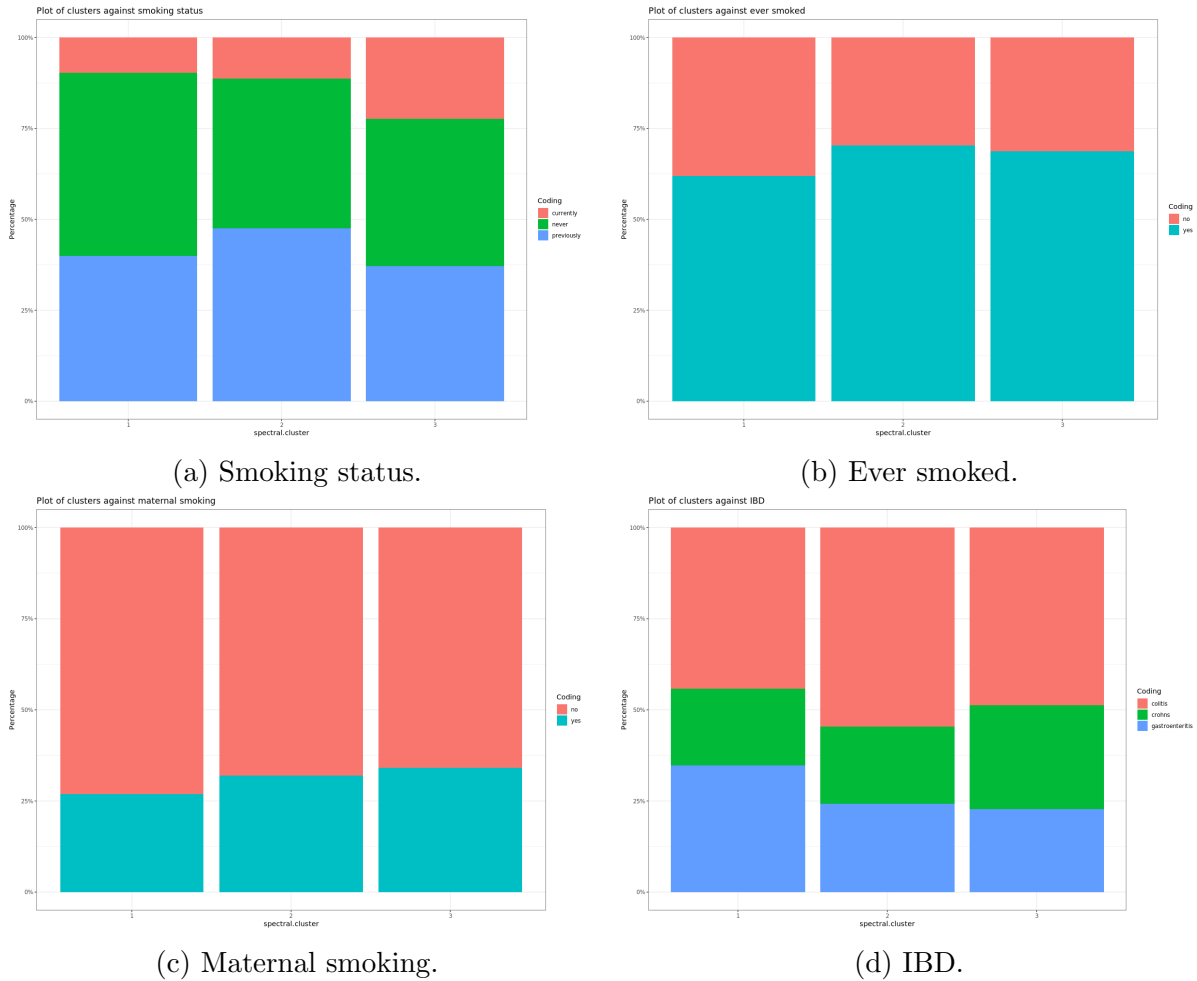


Figure 4.8: Bar plots of IBD-NYD.

Feature	1	2	3	adj p-value ANOVA x^2	
size	1265	1424	1171		
BMI	25.628	28.360	28.361	p=4.66e-45	
age	57.338	58.136	57.940	p=1	
not-yet diagnosed time	6.352	6.451	6.300	p=1	
WBC leukocyte count	5.906	6.868	8.825	p=0	p=9.55e-155
red blood cell erythrocyte count	4.221	4.764	4.445	p=1.28e-43	p=2.79e-40
haemoglobin concentration	13.374	15.033	13.655	p=5.94e-09	p=7.93e-56
mean corpuscular haemoglobin	31.751	31.637	30.792	p=6.73e-31	p=9.93e-18
platelet count	248.036	234.017	306.968	1.49e-115	p=2.60e-48
mean platelet thrombocyte volume	9.385	9.415	9.014	p=3.93e-15	p=1.20e-06
lymphocyte count	1.703	1.803	2.336	p=4.39e-78	p=5.00e-08
monocyte count	0.403	0.518	0.575	p=6.44e-63	p=0.001
neutrophil count	3.623	4.323	5.624	p=6.65e-279	p=6.93e-76
eosinophil count	0.141	0.188	0.233	p=9.04e-56	p=7.72e-07
basophil count	0.025	0.029	0.052	p=2.53e-42	p=4.64e-23
alkaline phosphatase	79.072	89.279	97.033	p=8.38e-49	p=3.44e-14
alanine aminotransferase	18.533	30.454	20.979	p=6.31e-05	p=4.41e-36
aspartate aminotransferase	24.143	30.313	24.127	p=1	p=1.23e-50
urea	5.167	5.745	5.394	p=0.001	p=0.0001
creatinine	65.248	82.233	69.340	p=4.52e-06	p=1.35e-16
CRP	1.857	2.573	6.561	p=4.11e-100	p=3.08e-91
cystatin C	0.859	0.993	0.955	p=8.19e-29	p=1.07e-19
gamma-glutamyl transferase	26.285	55.584	38.448	p=2.43e-08	p=5.17e-12
glycated haemoglobin HbA1c	34.869	37.011	38.273	p=7.01e-29	p=1.58e-19
IGF-1	21.000	21.829	19.841	p=6.13e-05	
total bilirubin	8.231	10.778	7.103	p=2.62e-08	p=5.35e-18
triglycerides	1.361	2.080	1.932	p=2.58e-46	p=1.16e-65
urate	262.051	359.074	303.382	p=1.72e-39	p=1.65e-42
NLR	2.351	2.685	2.777	p=5.77e-12	
genetic sex	83% F	82% M	69% F		p=6.56e-274
comorbidity	24% K	13% I	9% E		p=0.0001
IBS type †	90% w/o	7%	11% (2% D)		p=0.028
medication ††	14% I	19% A	21% P		p=2.40e-17
medication count	21% (1)	19% (0)	73% (2+)		p=5.90e-08
processed meat	35% (1)	35% (3)	30% (2)		p=8.90e-23
salt ‡	57% R	50% R	8% A		p=0.069*
miserable (y)	48%	40%	50%		p=6.48e-06
anxiety (y)	65%	54%	62%		p=4.96e-07
ability to confide ††	55% D	9% W	17% N		p=0.062*
health rating (fair)	25%	29%	32%		p=7.56e-18
health satisfaction (y)	85%	83%	75%		p=0.022
pain location (abdominal)	4%	3%	4%		p=2.15e-07
maternal smoking (y)	27%	32%	34%		p=0.030
smoking status (currently)	10%	11%	22%		p=4.76e-23
ever smoked (y)	62%	70%	69%		p=0.0002
IBD	35% GE	55% UC	29% CD		p=1.04e-13

Table 4.2: Table of IBD patients not-yet diagnosed p-values for Analysis of variance (ANOVA) and chi-squared (x^2).

*p-value significant before adjusted for multiple testing.

Key

† w/o=without, D=with diarrhoea.

†† A=Aspirin⁵, I=Ibuprofen⁶, P=Paracetamol⁷.

‡ R=regularly, A=always.

†† D=daily, N=never, W=once a week.

⁵medication used to reduce pain, fever, or inflammation.

⁶medication used for treating pain, fever, and inflammation.

⁷medication used to treat fever and mild to moderate pain.

4.3.3.3 IBD patients with future diagnosis

See Table 4.3 for averages of each feature in the assigned cluster, with ANOVA and x^2 p-values. See Figure 4.9 for the corresponding box plots of means.

Feature	1	3	4	6	7	adj p-value	
						ANOVA	x^2
size	362	238	114	70	115		
BMI	25.722	28.061	28.412	27.688	26.545	p=0.007	
age	55.525	57.609	59.140	57.071	55.357	p=1	
diagnosed time	-5.256	-5.168	-5.115	-5.186	-5.270	p=1	
not-yet diagnosed time	4.689	5.274	4.415	4.970	5.065	p=1	
WBC leukocyte count	6.445	7.313	8.105	10.767	5.410	p=6.13e-05	p=2.01e-71
red blood cell erythrocyte count	4.211	4.786	4.324	4.744	4.368	p=9.83e-10	p=7.42e-15
haemoglobin concentration	13.331	14.940	12.878	14.339	14.384	p=6.03e-13	p=2.16e-22
mean corpuscular haemoglobin	31.772	31.287	29.871	30.285	33.036	p=1	p=4.00e-18
platelet count	274.464	253.105	338.708	310.487	244.792	p=1	p=1.70e-21
mean platelet thrombocyte volume	9.195	9.232	8.730	9.374	8.914	p=1*	p=0.041
lymphocyte count	1.665	1.754	2.080	2.369	1.151	p=1	p=1.43e-08
monocyte count	0.434	0.531	0.506	0.730	0.383	p=0.042	p=5.13e-08
neutrophil count	4.141	4.799	5.210	7.280	3.739	p=1.48e-06	p=2.40e-36
eosinophil count	0.166	0.189	0.266	0.313	0.106	p=1	p=7.73e-09
basophil count	0.036	0.033	0.037	0.073	0.018	p=1	p=3.10e-09
alkaline phosphatase	83.544	90.358	97.972	97.393	86.206	p=0.023	p=0.980*
alanine aminotransferase	18.560	26.914	17.608	21.135	29.036	p=4.49e-09	p=9.78e-07
aspartate aminotransferase	24.760	28.389	23.863	24.657	30.944	p=2.50e-05	p=2.74e-05
urea	4.871	5.738	5.241	5.236	5.062	p=0.777*	p=0.014
creatinine	64.657	83.228	72.382	74.081	75.707	p=2.05e-12	p=2.05e-06
CRP	2.657	3.259	9.874	5.856	2.094	p=0.039	p=6.60e-20
cystatin C	0.864	0.989	0.988	0.986	0.924	p=5.22e-09	p=1.67e-05
gamma-glutamyl transferase	28.057	46.663	33.625	39.267	50.450	p=7.65e-07	p=0.141*
glycated haemoglobin HbA1c	34.729	36.509	39.669	37.097	33.343	p=1	p=0.001
IGF-1	21.125	22.308	17.648	23.944	24.045	p=0.002	
total bilirubin	8.178	9.716	6.196	7.767	11.907	p=1.11e-05	p=0.160*
triglycerides	1.421	2.333	1.930	2.074	1.764	p=4.67e-05	p=1.65e-15
urate	252.698	362.462	310.946	326.663	315.303	p=9.54e-21	p=3.08e-07
NLR	2.914	3.105	2.993	3.790	4.005	p=6.39e-06	
ethnic background (white)	98%	96%	88%	96%	98%		p=0.005
genetic sex	78% F	83% M	80% F	63% M	74% M		p=5.43e-55
age categorical †	12% a	48% c	27% d	47% c	36% b		p=0.119*
medication ††	11% I	18% A	28% P	3% R	55% O		p=0.530*
processed meat	36% (1)	30% (2)	15% (0)	13% (4)	37% (3)		p=7.05e-09
anxiety (y)	68%	56%	68%	62%	67%		p=0.986*
ability to confide ‡	53% D	62% D	19% N	13% W	15% N		p=0.211*
happy (y)	100%	93%	91%	92%	100%		p=0.597*
bowel cancer screening (y)	94%	93%	85%	96%	92%		p=0.387*

Table 4.3: Table of IBD patients diagnosed with future diagnosis p-values for Analysis of variance (ANOVA) and chi-squared (x^2).

*p-value significant before adjusted for multiple testing.

Key

† a=35-44, b=45-54, c=55-64, d=65-74.

†† A=Aspirin⁸, I=Ibuprofen⁹, P=Paracetamol¹⁰, R=Ranitidine¹¹, O=Other.

‡ D=daily, N=never, W=once a week.

⁸medication used to reduce pain, fever, or inflammation.

⁹medication used for treating pain, fever, and inflammation.

¹⁰medication used to treat fever and mild to moderate pain.

¹¹medication used to decrease stomach acid production.

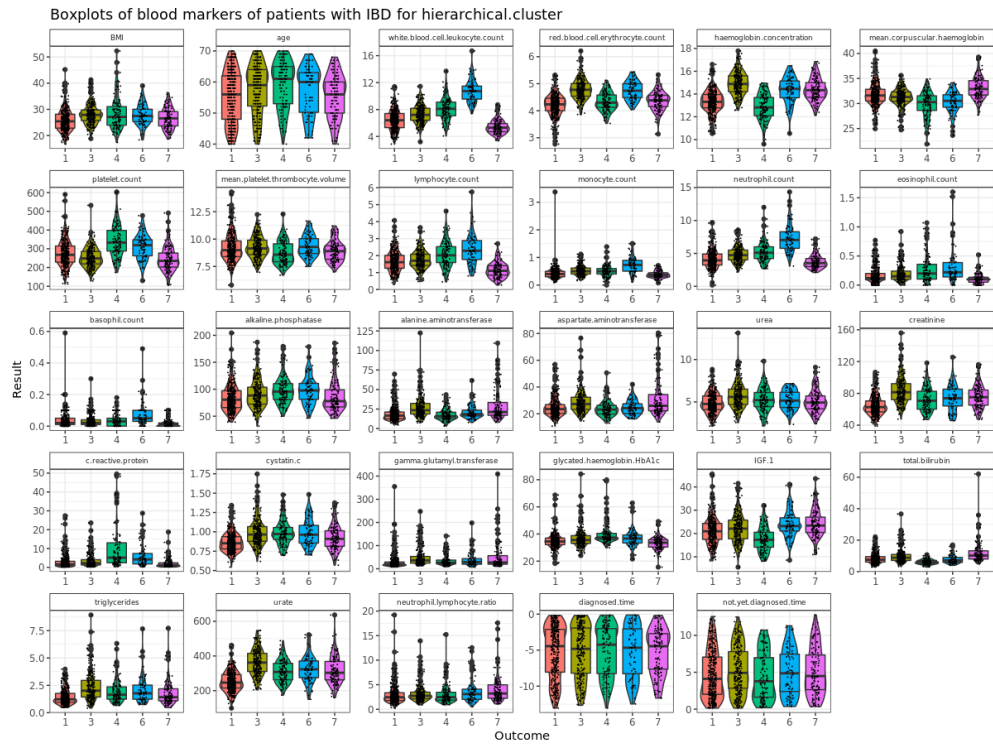


Figure 4.9: Figure of IBD-D-FD, box plots of averaged blood markers.

Below in Figure 4.10, two QoL features, are visualised: cluster 4 are an older cohort (Subfigure 4.10a) and a more genetic-ethnic mixed group (Subfigure 4.10b).

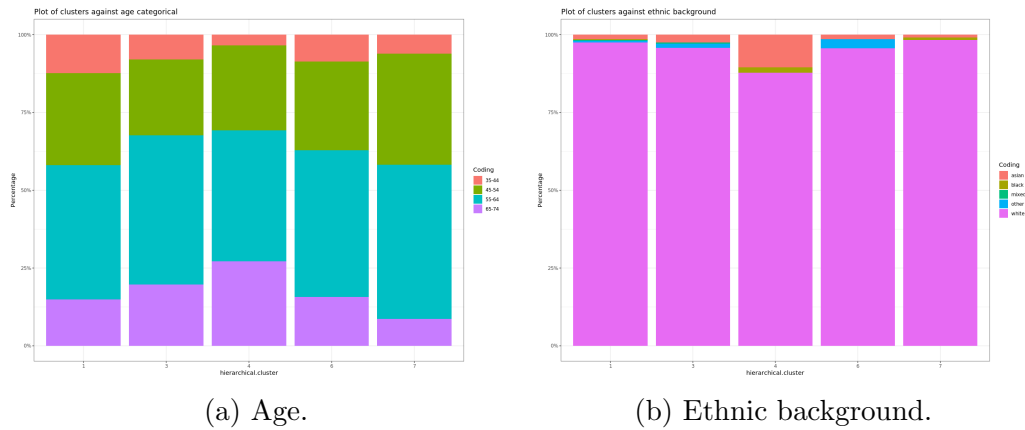


Figure 4.10: Bar plots of IBD-D-FD.

4.3.4 Supervised evaluation metrics

4.3.4.1 *Permutation variable importance*

To observe which features contribute to the clusters, a random forest feature analysis was conducted [286], using the permutation variable importance measure (PIMP) tool in R [287]. For IBD-D, see Appendix Figure D.6 (page 208), cystatin C is highly ranked. For IBD-NYD, WBC and neutrophil count are both high in feature importance - interestingly also the patient time to diagnosis showing this feature is important in cluster assignment. Finally, for IBD-D-FD, urate is the highest.

4.3.4.2 *Regression*

In this section a binomial regression was conducted to look into some significant values and important features for each cohort and cluster for predictive factors. For each cohort and their cluster of interest, an endpoint was used to conduct inner cluster analysis, see Table 4.4, including ANOVA or W and x^2 plus a binomial regression¹².

For IBD-D cluster 6 was interesting due to signs of inflammation (increased WBC, CRP, and NLR), younger average age, more diverse ethnic group, most anxious, least happy plus a poorer outlook on their health, and least smokers. For this, health rating (see average blood markers in Appendix Figure D.8, page 210) and anxiety were used as endpoints. For health rating with a PIMP analysis; creatinine, glycated haemoglobin HbA1c, and alanine aminotransferase. For anxiety: IGF-1, red blood cell erythrocyte, and age (see Appendix Figure D.7, page 209).

For IBD-NYD cluster 3 was interesting due to elevated WBC, CRP, and neutrophil count; in addition to these patients having a strong history of smoking, poorer outlook on their health, and a p-value<0.05 for their future IBD diagnosis. For this, health rating (see bar plot of health rating against IBD in Subfigure 4.11a, page 156) and IBD diagnosis (see bar plot of IBD against history of bowel cancer screenings in Subfigure 4.11b, page

¹²only values are reported with p<0.05 and features are in no particular order.

Cohort	Cluster	Endpoint	Feature	p-value	regression
IBD-D	6	health rating	WBC	p<0.05	p<0.05
IBD-D	6	health rating	basophil	p<0.05	p<0.05
IBD-D	6	health rating	neutrophil	p<0.05	p<0.05
IBD-D	6	health rating	lymphocyte		p<0.05
IBD-D	6	health rating	BMI	p<0.05	
IBD-D	6	health rating	health satisfaction	p<0.05	
IBD-D	6	health rating	medication	p<0.05	
IBD-D	6	anxiety	age	p<0.05	
IBD-D	6	anxiety	WBC		p<0.05
IBD-D	6	anxiety	lymphocyte		p<0.05
IBD-D	6	anxiety	monocyte		p<0.05
IBD-D	6	anxiety	neutrophil		p<0.05
IBD-D	6	anxiety	eosinophil		p<0.05
IBD-NYD	3	health rating	BMI	p<0.05*	
IBD-NYD	3	health rating	CRP	p<0.05*	p<0.05
IBD-NYD	3	health rating	neutrophil	p<0.05*	
IBD-NYD	3	health rating	time to diagnosis	p<0.05*	
IBD-NYD	3	health rating	ethnic background	p<0.05	
IBD-NYD	3	health rating	genetic sex	p<0.05	
IBD-NYD	3	health rating	maternal smoking	p<0.05	
IBD-NYD	3	health rating	bowel cancer screening	p<0.05	
IBD-NYD	3	health rating	IBD diagnosis	p<0.05	
IBD-NYD	3	IBD not-yet diagnosis	time to diagnosis	p<0.05*	
IBD-NYD	3	IBD not-yet diagnosis	CRP	p<0.05*	p<0.05
IBD-NYD	3	IBD not-yet diagnosis	NLR	p<0.05*	
IBD-NYD	3	IBD not-yet diagnosis	genetic sex	p<0.05	
IBD-NYD	3	IBD not-yet diagnosis	IBS	p<0.05	
IBD-NYD	3	IBD not-yet diagnosis	ability to confide	p<0.05	
IBD-NYD	3	IBD not-yet diagnosis	bowel cancer screening	p<0.05	
IBD-D-FD	6	bowel cancer screening	basophil	p=0.056	
IBD-D-FD	6	bowel cancer screening	medication	p<0.05	
IBD-D-FD	6	bowel cancer screening	IBD not-yet diagnosis	p<0.05	
IBD-D-FD	6	IBD not-yet diagnosis	alkaline phosphatase	p<0.05	p<0.05
IBD-D-FD	6	IBD not-yet diagnosis	mean corpuscular hemoglobin	p<0.05	
IBD-D-FD	6	IBD not-yet diagnosis	health rating	p<0.05	
IBD-D-FD	6	IBD not-yet diagnosis	IBD diagnosis	p<0.05	

Table 4.4: Table of binomial regression and ANOVA, W , and χ^2 p-values for particular clusters within each time cohort.

*p-value significant after adjusted for multiple testing.

156) were used as endpoints. For health rating with a PIMP analysis; NLR and eosinophil. For IBD; haemoglobin concentration and triglycerides.

Finally, IBD-NYD cluster 6 was investigated due to elevated inflammatory biomarkers (WBCs) and highest history of bowel cancer screenings, which was used as an endpoint. Whilst their IBD diagnosis p-value>0.05, it was explored for comparison purposes (see bar plot of IBD future diagnosis against health rating in Appendix Figure D.9, page 211). For bowel cancer screening with a PIMP analysis: alanine aminotransferase. For IBD; basophil, glycated haemoglobin HbA1c, and mean platelet thrombocyte volume.

4.4 Discussion

In this chapter, I have investigated the relationship between Inflammatory Bowel Disease (IBD) blood biomarkers and quantitative QoL markers in different time-split cohorts. This gave rise to potential novel insights into patients with or without IBD at baseline, allowing a deeper understanding of the differences in blood assays; complete blood count (CBC)

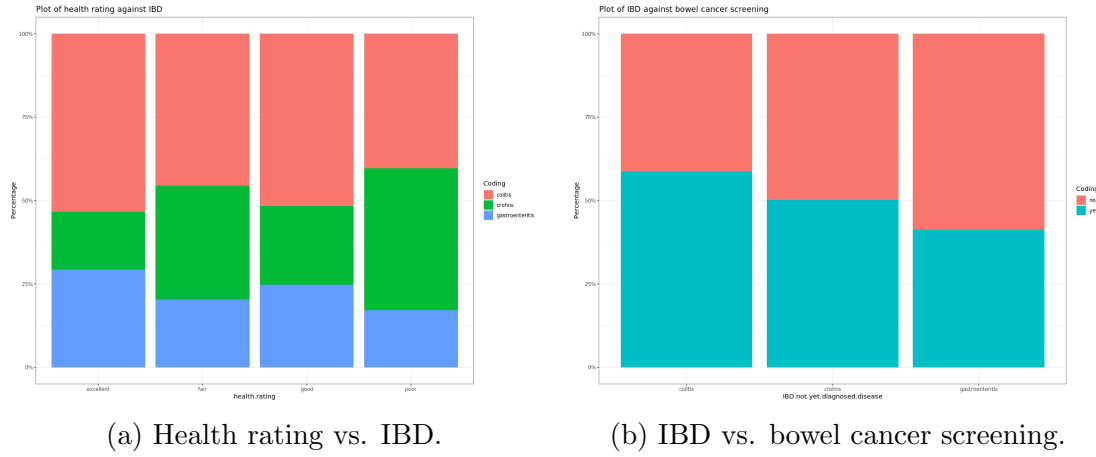


Figure 4.11: Bar plots of IBD-NYD cluster 3.

and biochemistry; and QoL mental health markers of inner IBD clusters. This allows us to observe that patients who have IBD, differ in happiness and if that relationship is corresponding to the type of IBD, how long they have experienced their disease, or blood marker of progression.

A workflow for unsupervised ML techniques of clustering has been presented and applied in order to investigate continuous blood markers and categorical QoL measurements of patients at baseline who IBD patients diagnosed (IBD-D), IBD patients not-yet diagnosed (IBD-NYD), and IBD patients diagnosed with future diagnosis (IBD-D-FD). The workflow consists of advanced techniques, used frequently, for proper exploitation of datasets. With feature elimination tests, Pearson correlation coefficient (r) revealed LDL direct (a type of cholesterol) is correlated with apolipoprotein B, both of which have been shown to link with cardiovascular diseases [289]. For both diagnosed cohorts (IBD-D and IBD-D-FD) hierarchical clustering presented the highest gap score, however, the statistic for IBD-NYD was highest for spectral clustering. In each clustering algorithm, K-means++ and Hierarchical had higher aRi agreements and silhouette coefficient sizes (supported visually).

Each time-stamped cohort characterised clusters by white blood cells (WBC), including neutrophil count, in addition to C-reactive protein (CRP). All indicators for inflammation imply that those with IBD are possibly having a flare-up, or those not-yet could be

having early signs. The majority of clusters' blood averages or relation to QoL markers reflected literature in that those who have measurements within normal measurements were happier, and those with CD had a history of smoking. This analysis combines elements from biological studies (blood markers) with clinical inaccessible data, and quantitative questionnaire feedback: an intersection of two domains for a deeper understanding of IBD, biological biomarkers, and QoL.

4.4.1 Implications

A total of 8759 patients in the UKB have chronic IBD, that is only 1.74%, and after the preprocessing (remaining 6604 patients) is only 1.31%. In the feature elimination, the r (Figure 4.3, page 145) heat map revealed LDL direct¹³ was removed due to high r with apolipoprotein B¹⁴ and cholesterol. From the PCA, (Appendix Figure D.1, page 203), apolipoprotein B and cholesterol were removed.

Using the Ul-Haq et al. [27] quartiles¹⁵ for BMI, each cohort had a significant p-value. For each cohort, all clusters are in the overweight band (25.0–29.9 kg/m²) with the exception of each cluster 1 being close to normal - interestingly reflecting the blood markers in that they reflect more normal readings [1, 84, 85, 87]. For those IBD-NYD, a patient's future diagnosis was significant, in that we see from Subfigure 4.8d (page 150) that cluster 2 had more UC and cluster 3 had more CD. Additionally, comorbidities were significant, seeing many patients having an ICD-10 code for digestive K and circulatory I.

Characterising clusters

Each time-stamped cohort, and individual cluster, is defined by various biomarkers [86].

In IBD-D, urea and cystatin C for cluster 7 is abnormally high (kidney issues); creatinine for clusters 1 and 6 are abnormally low (liver issues); CRP for 3, 6, and 7

¹³known as the “bad” cholesterol that collects in the walls of arteries.

¹⁴a protein that is involved in the metabolism of lipids.

¹⁵underweight (<18.5 kg/m²), normal weight (18.5–24.9 kg/m²), overweight (25.0–29.9 kg/m²), class I obese (30.0–34.9 kg/m²), class II obese (35.0–39.9 kg/m²) and class III obese (>=40 kg/m²) - read more here: <https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi>.

are abnormally high (inflammation); gamma-glutamyl transferase is high in 1, 3, 6, and 7 (liver issues); triglycerides high in 2, 3, and 7 (issues with arteries); and neutrophil-to-lymphocyte ratio (NLR) high in 3, 6, and 7 (inflammation). Cluster 7 seems to be characterised by their high percentage of circulatory comorbidities, use of aspirin, and have signs of inflammation. Cluster 1 and 6 seem to have liver issues. But cluster 6 also have signs of inflammation, alongside QoL-related features, e.g. high anxiety and poorest rating on health. For health rating in cluster 6, those with a poorer score had a higher WBC and neutrophil, yet a better health had a higher basophil.

In IBD-NYD, mean platelet thrombocyte volume and creatinine is abnormally high in cluster 2 (kidney issues); CRP abnormally high in cluster 3 (inflammation); gamma-glutamyl transferase and triglycerides are high in clusters 2 and 3 (liver and artery issues). Cluster 1 seems to be characterised by most health in that their BMI and blood markers are normal. Cluster 2 seems to consist of patients with artery issues through their circulatory comorbidities and used of aspirin, however these patients are mostly going to be diagnosed with UC. Finally, cluster 3 consists of an abnormally high CRP and least happy. However these patients majority get diagnosed with CD but also have a stronger history of smoking, either being current smokers or have a history of maternal smoking. For health rating in cluster 3, CRP was much higher in those with a poorer outlook in addition to a shorter diagnosis time and CD; and in IBD, CRP was higher in those with CD.

In IBD-D-FD, mean platelet thrombocyte volume was low in all clusters; neutrophil high in cluster 6 (inflammation); creatinine low in 1, 4, and 6 (liver issues); CRP high in 3, 4, and 6 (inflammation); gamma-glutamyl transferase high in 3, 4, 6, and 7 (liver issues); triglycerides high in 3, 4, 6, and 7 (artery issues); and NLR high in 3, 6, and 7 (inflammation). Cluster 1 had more normal blood markers, youngest, and happiest; whilst 4 was oldest and some markers indicating liver issues, most ethnic diverse group, and least happy. Cluster 7 were happy but most anxious. Finally, clusters 3 and 6 had signals of inflammation, yet 6 had a higher history of bowel cancer screenings and alkaline phosphatase was much lower in those with UC.

Results were more normal in IBD-D and IBD-D-FD, this could be due to patients maintaining inflammation via diet or medical interventions [1] - however IBD-D-FD lack of significant p-values could indicate that this cohort makes features noisier.

The changes in blood markers (e.g. WBC, neutrophil, and monocytes) indicate an acute inflammatory response [1, 2, 78, 83, 84, 88–91, 108] or the potential acute recurrent episode. Additionally, both CRP and NLR indicated inflammatory stress [1, 52, 80, 93–96, 108], additionally the change in platelets. However, platelets and creatinine can indicate other diseases [1]. The presence of liver issues (creatinine, gamma-glutamyl transferase) can indicate stress caused by IBD [102, 110–112]. Whilst red blood cells can indicate bleeding caused by IBD, no major indication or abnormalities were present.

Many results reflect the literature in that external factors may be contributing to the inflammation activity [1, 84, 85, 87, 97, 107, 108, 113, 114], including diet (processed meats), BMI, medication, and smoking - or how genetic males are more likely to develop UC (IBD-NYD cluster 2). We see increased levels of unhappiness in those with an inflammatory episode (IBD-D cluster 6) [21, 23, 108]. Interestingly fewer patients are miserable with UC (IBD-NYD cluster 2), seeming to contradict published literature in this area [119]. Finally, in those with more IBS (IBD-D cluster 6, IBD-NYD cluster 3), specifically with diarrhea were unhappier, reflecting works from literature [121].

4.4.2 Limitations and future endeavours

Using the UKB has limitations: the population recruited were volunteers and, therefore, may not be representative of populations due to an older age bias or more comorbidities, as brought to attention by Welsh et al. [232]. Despite the UKB aimed at an older age cohort (aged 30+, average late 50s), many patients were diagnosed ± 14 years. This may explain the low percentage of total patients as IBD is more commonly diagnosed within the ages of 15 and 40. Moreover, the numerous comorbidities within the population, specifically those patients with chronic IBD had an overweight BMI. Additionally, diagnosis in the hospital records may not be 100% correct in that some patients may present extreme

signs of IBD before diagnosis, or some comorbidity may be misdiagnosis. Splitting the data into three time-stamped cohorts also introduced bias, however the third group, IBD patients diagnosed with future diagnosis, presented noisy clusters making difficult to interpret, and so in fact splitting the data helped distinguish differences amongst those (diagnosed and not-yet diagnosed) with IBD.

A major limitation was the removal of many blood features, such as biochemistry, due to more than 10% missing. Using only 1.31% of the total UKB data for chronic IBD is also quite the limitation: more patients could have concluded in a supervised ML classification to look for predictive blood markers for IBD and QoL-related features. These are some areas for potential exploration in future projects, however, a larger cohort of patients is needed to successfully and accurately conduct prediction analysis in the clinical domain.

One major future benefit to this study would be the inclusion of a feature where patients could mention if they were having an inflammatory episode, that could be used to compare with blood and QoL markers. Or knowledge of the five fundamental phenotypes [2, 78] and a patient's experience plus blood relation. In the data itself, despite normal or abnormal (high/low) blood markers averaged across clusters, without clinician perspective or interpretation the true meaning cannot be fully exploited, and so this work should only be taken as a preliminary study which can be fully exploited and investigated with thorough clinical feedback and analysis.

Regarding the methodology, there are numerous drawbacks to the clustering techniques used. The K-means++ struggles to compute clusters on data that is heavily dense [151, 167], hierarchical clustering makes many arbitrary decisions which cannot be fixed at a later stage in addition to being sensitive to outliers [151, 167], and spectral clustering is much more computationally expensive and clustering relies on eigenvectors to be computed first [166]. All methods rely on choosing K , however I used two methods to compute K for each of the three clustering algorithms. Despite all these drawbacks of the clustering algorithms, the benefits of using multiple algorithms overcomes these issues through systematic comparisons of each clustering results.

Future work can proceed with using a final clustering technique: consensus clustering. Consensus clustering is essentially aggregating multiple algorithm results based on a similarity measure via partitions of clusters and then re-clusters the results [290]. This technique can be explored in the future to aggregate the multiple algorithms instead of separate analyses on each clustering experiment, though more work would be needed as consensus clustering methods can struggle with high number of dimensions, choosing an optimal number of clusters K , and defining distance measures [290]. Moreover, another clustering technique called, biclustering, could be applied. Of each algorithms used in this thesis, K-means, Spectral, and Hierarchical, have extended versions via biclustering, which allows clustering of data via rows against columns [291, 292]. As with each clustering algorithm, each biclustering type has different features [292] and may require their own parameter fine-tuning in addition to large computational requirements.

4.4.3 Conclusions

Clustering has demonstrated the IBD status and progression in UK patients from the UK Biobank in regards to their blood markers and QoL. The clusters have created new areas of potential investigations for a new dimension of quantitative data.

This experiment is the intersection of biomarkers from a different perspective of biological blood samples and QoL mental health features. There is a clear correlation between inflammatory episodes and a poorer outlook on health, specifically inflammatory responses, including white blood cells and CRP. There are also links with liver and circulatory issues.

As IBD can progress silently over years and includes invasive tests to diagnose [1, 52, 81, 84, 94], these results could potentially close gaps and improve patient experience [51, 52] in addition to lowering costs for health services. As patients continue through life with IBD, regular examinations are necessary and so by using preferred blood tests [52] we can improve QoL [21–23, 65, 108, 116–119].

Summary

1. Addressed the challenge of better understanding the relationship between physiological markers and quality of life (QoL) through investigating IBD complete blood counts for the potential to improve patient medical experiences:
 - Of IBD patients diagnosed, an unhappier outlook on health had a relationship with the presence of inflammatory stress.
 - Of IBD patients not-yet diagnosed, patterns emerged in those with shorter diagnosis time, Crohn's disease, smoking exposure, and negative sentiment; indicating signals of QoL and diagnosis journey.
2. Extracted a cohort of patients from the UK Biobank of those diagnosed with IBD, separated into three distinct time-split subgroups to introduce a new dimension of analysis.
3. Employed extensive clustering techniques with thorough descriptions to provide a resource for future studies to improve the reporting of ML work.
4. Work addresses the needs of patients and their preferences: blood tests can be used as a preferred tool for diagnosis or regular check-ups to look into WBC subpopulation signals, whilst also considering patient medical/smoking history and QoL features: including their opinion on their own health.
5. Whilst this study is speculative, some findings agree with existing literature, highlighting a possibility for clinical interpretation of these findings to expand the literature in future.

5 DISCUSSION

This thesis has investigated techniques into mining the patient voice for sentiment and indicators of quality of life (QoL). This work revealed the role that online forum communities play for patient experiences, the lexicon of patients, and how disease progression affects sentiment. Patients used several preferred terms for cyclopentolate (mydrilate, brand), prednisolone (pred forte & “predforte”, brand), and panuveitis (“pan uveitis”); that aren’t contained in other biomedical vocabularies. Semantic similarity revealed diarrhea and stomach cramps are more important in patient priority compared to the physician and clinician. And blood biomarkers, including inflammatory white blood cells (WBC) indicators, have a relationship with negative sentiment plus other QoL features.

Background research of the literature highlighted the utility of the patient voice and its potential for advancing clinical practice, but revealed crucial gaps in existing domain knowledge. The minimal work done to integrate the patient voice into research impedes science, despite the clear importance of QoL. Whilst biomarkers - from invasive and uncomfortable investigations - are the gold standard in medicine, a patient has preferences and so the unmet need for improving patient outcomes has been addressed in this thesis. I have explored the patient voice pertaining to two inflammatory diseases: Uveitis and Inflammatory Bowel Disease (IBD). Both immune-mediated inflammatory diseases (IMIDs) and often undifferentiated. Patients’ inflammatory condition and their voice has been researched using numerous techniques concerning the current literature challenges; including sentiment analysis, clustering, and semantic characterisations. The experiments in this thesis aimed to close the gaps by standardising ocular IMIDs, uncovering lay (patient-preferred) terms, investigating semantic similarity of clinicians and patients

for similarities and difference of the lexicon, and encountering potential biomarkers for better diagnostic techniques for IBD. This final chapter will outline how I addressed the challenges and conclude the experimental chapters.

Firstly, the development of an application ontology, Ocular Immune-Mediated Inflammatory Diseases Ontology (OcIMIDo), for ocular IMIDs has been described and advances clinical practice with a standardised vocabulary. In collaboration with domain experts, the ontology foundation was derived from a clinical consensus document, the Royal College of Ophthalmology Uveitis Clinical Dataset. OcIMIDo addresses the contradictions and structural issues in current biomedical ontologies, with extensive database cross reference (x-ref). The ontology was enhanced with patient-preferred terms leading to improved text mining and NLP tasks with more extraction and fruitful analysis, leading to a better insight into sentiment.

The second experiment explored semantic similarity between clinician and patient-derived embedding models, both an extension and another perspective into the patient voice, adding a new dimension. Whilst relevant domain ontologies, such as the Disease Ontology (DOID) exist, they did not describe the concepts in relation to inflammation in sufficient detail. Consequently an ontology for inflammation itself was vital, as inflammation is the body's first response to external threats and has a complex relationship with a wide range of diseases. The Combined Ontology for Inflammatory Diseases (COID) was built using ICD-10 as a foundation, based on anatomical structure and organises 365 inflammatory-related concepts allowing for semantic characterisations of disease subgroups for the embedding models. Creating embedding models using ClinicalBERT (CB) avoided the intense assignment of accessing clinician-generated texts; the extensive, time-consuming, and computationally expensive task of model parameter fine-tuning; and avoided thorough text preprocessing and sampling efforts. COID allowed for systematic comparisons and characterisations of the clinician and patient lexicon through analysis of embedding models. The PatientINF (PI) model, of forum conversations, was developed in addition to the combined model CB + PI, which is publicly available. The patient voice was compared

to other patient-generated texts from social media posts to evaluation patient semantics from different sources of online unstructured text. All provided the opportunity to discover disease associations not in the literature and the embedding models further automates knowledge acquisition.

Third and lastly, further sentiment work was conducted on patients with IBD to investigate the relationship between blood markers and patient QoL. These blood markers were clustered to identify groups of patients with IBD and study their characteristics. The UK Biobank (UKB) was employed here as it has synchronised records of a patient's medical history, hospital visits and diagnosis before and after recruitment - these corresponding diagnoses are past and future (post-recruitment) ICD-10 codes. These patients with chronic IBD were split into time-stamped cohorts and clustered to uncover potential new biomarkers in order to improve diagnostic methods for patients plus shorten the diagnosis time; and understand the relationship between blood markers and QoL.

5.1 Implications

The following section covers the fundamental findings from this thesis, future endeavours, and overall scientific impact.

5.1.1 Patient-preferred terms

The work in this thesis uncovered many patient-preferred terms and aligned them with their clinical counterparts. This not only bridges the gap in domain knowledge but enhances clinical vocabulary and terminology systems. I have published my work in open access journals and the ontologies are freely available, allowing for secondary research use and distribution of lay terms for clinicians to better understand patients.

Online forum discussions provide a place for patients to discuss a wide range of topics. There are geographical differences, as seen in the two ocular-related fora: exhibiting different vocabulary of medical therapies due to brands of pharmaceuticals. Analysis highlighted the patient vocabulary, including abbreviations. The appearance of these

patient-preferred terms in online text is one of the many reasons that natural language poses a problem: abbreviations and misspellings are common and text mining tasks are difficult if we don't have this lexicon known. The work in this thesis provides numerous approaches to identify the patient vocabulary and associate them with semantic resources, enabling analysis of patient voice data.

The embedding models I created in addition to exploring the preexisting CB, revealed the overwhelming prevalence of misspellings: clinicians are people too and they make spelling errors. Many ocular IMIDs were present as similar terms, despite the PI model built from a corpus of all Patient.info topics - possibly indicating that visual inflammatory conditions are frequently discussed due to impacts on QoL. My combined model, CB + PI, had higher agreement with a physician, as measured by the Pearson correlation coefficient, compared to the patient-only model implying that it retains the knowledge and semantic information from both single-domain models. From the Pearson correlation coefficient it was observed that the patient voice models consider diarrhea and stomach cramps more semantically similar, in contrast to the physician and the clinician-only CB model - showing patient priorities and their voice. These two terms not being similar in the eyes of the clinician may result in miscommunication with the patient. Many disease associations revealed by the semantic characterisations were reflected in the literature and so with future clinical interpretation, other semantic similar terms could lead to identification of new associations, or indicate misdiagnosis or conditions that present similar symptoms/signs.

There are numerous ways patients communicate with their GP, increasingly remotely via telephone or video calls, especially after/during the COVID-19 pandemic¹. As discussed, there are gaps in communication: patients not understanding jargon, or clinicians not understanding patient-preferred terms. This can hinder patient experience in many ways, clinicians not understanding a patient can delay diagnosis, which can leave patients dissatisfied and so seeking comfort/advice elsewhere. Even though a patient can join

¹March 11, 2020 - present (dated on front of this thesis).

online communities for seeking advice, support, or general conversation; which can benefit their QoL, a clinician being better aware of the patient vocabulary helps the clinician-patient relationship and improve patient outcomes, diagnosis methods, and reduces delays.

5.1.2 Sentiment analysis

Expanding the clinical domain with patient-preferred terms gives the opportunity to improve research standards. The patient lexicon is thus one insight into the patient voice, yet another is understanding how the patient is feeling.

OcIMIDo annotated with lay terms enhanced text mining and so I was able to conduct a sentiment analysis that uncovered the supportive role online forum play for the patient. It was revealed that first posts of ocular-related fora were more negative compared to replies. Sentiment analysis directly on the patient posts displayed some types of uveitis were slightly more negative, some medical therapies, complications, and more. Whilst some of this information was to the expectation of the clinical collaborator, it is valuable evidence that research should not limit analysis to clinical terminology and work highlighted aspects of the patient experience that could be better understood. For example, more patient-clinician discussions to: understand the patient and create better therapy plans to improve QoL.

My work in chapter 4 investigated the relationship between patient sentiment and biological markers. With clustering, I investigated inflammatory biomarkers of a cohort of patients with IBD to improve knowledge and advocate for blood tests: assisting with the diagnosis of IBD: less medically expensive and a less invasive experience for the patient. Clustering cohorts of patients with IBD (diagnosed at baseline against not-yet diagnosed at baseline) is an important analysis as it avoids putting a patient in “one box” and trying to understand how those with the same chronic disease are different. This level of detail into QoL of people with the same disease has potential to understand why these inner groups differ and better developing medical therapy plans or interventions best suited for the needs of the patient.

Clusters were characterised by the presence of inflammation: increase of WBC leukocyte subpopulations and C-reactive protein (CRP) adjacent to patients who rated their own health poor. In these time-stamped cohorts, the clusters with presence of inflammation, patient sentiment was more negative - showing that a flare-up of inflammation can affect a patients' QoL. Furthermore, in those who were not-yet diagnosed at baseline, there was also a relationship with time to diagnosis; and in one cluster a relationship between Crohn's disease (CD) and maternal smoking.

5.2 Future endeavours and limitations

As expressed in the literature review, many research studies primarily investigate the patient voice through questionnaires, yet introduces bias of the "Question-Behaviour Effect". I was able to avoid this through using data without clinical interference. Investigations in this thesis for sentiment analysis via NLP on forum-based or ML on quantitative-based data, should not limit future endeavours to analysis on different data, including Patient Reported Outcome Measurements (PROMs) questionnaires or other biomarkers.

Future work can build upon disease-specific PROMs focused not only on lifestyle changes but targeted on QoL from a patient perspective and disease-related influences on QoL. Work in this thesis included employing existing data for secondary use: the Royal College of Ophthalmology Uveitis Clinical Dataset, CB, and the UKB - all provided a foundation for exploration of patient voice. These resources, whilst not designed for investigations into the patient voice or sentiment analysis, contained information which allowed for additional research capabilities.

The group of domain experts who worked on OcIMIDo, and I, intend to continue development as new associations are uncovered. In addition, based on some of the work toward the end of this thesis, the development of a new ontology for QoL² is underway. Both OcIMIDo and COID can be used in semantic similarity tasks to risk-stratify patients and have application to clinical sources: EHR, genetic phenotypes in the UKB, and

²quality of life ontology <https://github.com/sap218/qolo>.

building ontology-guided predictive algorithms. Specifically, in future COID could be expanded to include biological processes and interlinking with the Gene Ontology.

The novel synonym curation of ranked terms with TF-IDF can be further enhanced with a matching algorithm to link lay terms with ontological concepts without the manual matching of a domain expert. Additionally, the TF-IDF can be explored using n-grams. There are other methods for the curation of patient-preferred terms, which previously discussed includes crowd-sourcing techniques involving both a patient and clinician - these future developments can include both quantitative and qualitative data for users of different backgrounds to label words.

Future work can build on supervised ML to release labelled data or a trained model. Supervised sentiment analysis would require a significant level of crowd-sourcing from both patients and experts: to obtain enough labels (multiple responses for an average sentiment) it would be an infeasible amount of time. Sentiment analysis can further stratified using additional context of the patient, including details such as age, general health (e.g. BMI), and if they are using online fora for themselves. This will allow for aggregation to reveal further insight into demographic relation to inflammatory diseases. More use of NLP techniques such as part-of-speech tagging and negation can be used to further understand the context of the concepts and sentiment scores.

The embedding models can be further developed with more domains. For this expansion to be done, numerous precautions are needed as model performance can depend on text size, parameter fine-tuning, and data source. Whilst CB was used due to clinician-generated texts being fairly inaccessible and benefits of publication providing analysis specifications; future work should include expansion with additional clinician-derived sources, e.g. EHR and publications. Despite work showing differences between two patient-generated texts, including *modified* BioReddit, it lacked in size to be used for expanding PI models through the repetition of terms. More patient-derived texts can be used to build on this domain and is needed to strengthen the medical lexicon that includes clinician and patient language.

As CRP is a known biomarker for IBD, more work is needed to better understand

the role WBC types and sentiment have for new potential biomarkers. Although there is evidence that smoking has a relationship with CD, works in this thesis presented a stronger link and more work is needed to better understand the relationship with maternal smoking as this could drastically change patient perspective - smoking is strongly advised against for pregnant women and so this thesis adds to the argument: telling pregnant individuals the future link of an IBD diagnosis. Finally, more investigations are needed into the relationship observed between liver complications and IBD, as many abnormal blood biochemistry markers indicated liver issues, for example future work into comorbidities.

Large sources, for example the UKB has bias due to participants being older and multimorbidities. Furthermore, a patient's sentiment can be subjective and related to external factors and so more patient responses can benefit this, e.g. if they think they are having an episode. To address this, future work should include more varied cohort of patients, e.g. age groups, various diseases (strong associations), or limited in disease, and more diverse genetic ethnic groups.

The ML workflow can be used as a foundation for future work standards: to include sufficient information and how the interlink of inflammatory diseases and sentiment (QoL) can be explored. There are future opportunities to develop a supervised ML classification for prediction, however, extensive work will be needed to ensure the process is correct, including generating a data set (clinical trial), sample size, different feature selection techniques, and robust classification algorithms.

To conclude, with these opportunities for future endeavours, the fundamental component in clinical research is clinical interpretation. Direct clinical work during the development of OcIMIDo allowed the bridge of a computational tool being developed for standardising medicine. Although some literature research was conducted to understand the disease associations highlighted by the semantic similarity results, further interpretation from a clinician is imperative to understand the clinical context behind highly ranked similarities. Finally, novel maternal smoking links with CD has been introduced in this thesis, however, without clinical clarification and extended experiments, this novel finding cannot be fully

exploited. The clustering work should be perceived as a preliminary step into a deeper understanding of sentiment and its relationship to IBD.

5.3 Conclusion and overall scientific impact

This thesis has shown that the patient voice is an invaluable resource for improving research standards and clinical domain knowledge. My work has demonstrated that ontologies are a crucial tool in science for standardising a domain of knowledge - they meet the need of better aligning computational resources with clinical and patient vocabulary.

Ocular IMIDs have been correctly standardised via OcIMIDo both structurally and conceptually, as addressed in this thesis. The obstacles presented by unstructured data has been aided with the patient voice. The ontology is enriched with patient-preferred terms as clinical synonyms, derived from online forum conversations: meeting the challenge of including the patient vocabulary in clinical research. Although the methodology included some manual curation, the computational technique is drastically shorter compared to alternative time-consuming methods. This expanded vocabulary provides the research community opportunities to analyse a wide variety of unstructured text, of different domains, for ocular IMIDs and their autoimmune aetiology. The patient vocabulary allowed for improved text mining and NLP tasks: with application to sentiment meant a more fruitful analysis and work published Pendleton et al. [242] (2021).

The patient voice was also explored using vectorisation via the development of embedding models for semantic similarity. The clinical and patient similarities and differences have shown the way each discuss inflammatory conditions and revealed patient priorities. The misspellings, abbreviations, and differences allows us to understand word associations deeper. The second challenge in this thesis was addressing the lack of open embedding models for secondary research that integrating the patient voice: I have contributed via the release of embedding model, ClinicalBERT + PatientINF, consisting of both the clinical domain and patient lexicon.

My work in sentiment and QoL (chapters 2 and 4) revealed the role patient forum

play and the relationship between a disease and blood biomarkers. First posts are more negative compared to replies; and inflammatory biomarkers have a relationship with a negative sentiment. Biomarkers and quality of life quantitation has unveiled potential diagnostic advances for shortening diagnosis time in patients with Crohn's disease: these patients were smokers, unhappier, had an abnormally high CRP, and elevated WBC. This has addressed the third challenge of understanding the relationship between physiological biomarkers and QoL to improve patient outcomes: a combination of these biomarkers can potentially avoid misdiagnosis, shorten the exhaustive diagnosis process, and improve patient experience via use of preferred investigations.

The increasingly urgent need for improved clinical care pathways and integrating the patient voice is established in this thesis. The work contributes to the recognition of patient priorities and the development of meaningful outcome measures through investigating the patient's voice. Work in this thesis introduced computational tools to avoid the exhaustive and tedious non-automatic investigations of the patient voice; yet manual investigations and clinical interpretation is still a key part of science for validation and so the future of this work extends to vital clinical evaluation.

This work sets the premise for improving standards in research with the inclusion of the patient voice. I hope the work in this thesis will: pave the way for future ontological development to include the patient lexicon; retraining of embedding models with various domains of both patient and clinical to continuously expand knowledge; and better integration of quality of life in the patient medical process adjacent to blood biomarkers. In the immediate future, this work can expand to the development of a clinical consensus document for patient outcomes.

A patient's experience with their chronic, inflammatory disease is complex. This work has shown that a patient's voice, being qualitative or quantitative, in ontologies; or combined with novel biomarkers, can be investigated for improved patient outcomes and quality of life. **The patient is the expert of their experience and their voice is indispensable to science.**

List of Figures

1.1	“Patiently waiting for advice on health, last century” by Ferguson [6].	6
1.2	Illustrating the process of acute inflammatory response in anatomy.	18
1.3	The anatomy of the human eye, highlighting where uveitis can occur.	20
1.4	Illustrating the anatomical immune response of IBD.	23
1.5	Demonstrating concepts of an ontology using anatomy as an example.	28
1.6	Visualisation of ML.	43
1.7	Visual representations of methods to gain the optimal number of clusters.	47
1.8	An example of plotting an embedding model.	57
2.1	Screenshot of a table from the RCOphth by Denniston et al. [243].	76
2.2	Graphical representation of OcIMIDo.	82
2.3	Illustration of increased text mining efforts with the addition of synonyms.	86
2.4	Box plots of TextBlob sentiment scores on the two patient fora, for a range of classes from OcIMIDo.	88
3.1	Screenshots of the ICD-10 comprehensive lists of inflammatory diseases that were included in ontology development.	98
3.2	Graphical representation of COID superclasses.	103
3.3	A t-SNE of CB.	105
3.4	Two close-up t-SNE plots of CB filtered with COID terms.	106
3.5	Full representation of CB t-SNE filtered using terms only in COID.	107
3.6	A t-SNE of PI.	108
3.7	Two close-up t-SNE plots of PI filtered with COID terms.	108
3.8	Full representation of PI t-SNE filtered using terms only in COID.	109
3.9	A t-SNE of CB + PI.	110
3.10	Two close-up t-SNE plots of CB + PI filtered with COID terms.	110
3.11	Full representation of CB + PI t-SNE filtered using terms only in COID.	111

3.12	Vectorisation representation of blepharitis.	112
3.13	Vectorisation representation of hepatitis.	113
3.14	Vectorisation representation of sinusitis.	115
3.15	Plotting ROC curve of synonyms captured from the embedding models using the 14 inflammatory terms of interest.	116
3.16	Full representation of m BR t-SNE filtered using terms only in COID.	119
3.17	Two close-up t-SNE plots of m BR filtered with COID terms.	120
3.18	A t-SNE of m BR.	120
3.19	Full representation of PI t-SNE filtered using terms only in COID.	121
3.20	Two close-up t-SNE plots of PI filtered with COID terms.	122
3.21	A t-SNE of PI.	122
3.22	Full representation of m BR + PI t-SNE filtered using terms only in COID.	123
3.23	Two close-up t-SNE plots of m BR + PI filtered with COID terms.	124
3.24	A t-SNE of m BR + PI.	124
3.25	Vectorisation representation of plantar fasciitis.	126
3.26	Vectorisation representation of ulcerative colitis.	126
4.1	Workflow of data preprocessing and ML techniques.	139
4.2	Histograms of a patient's diagnosis to recruit date.	144
4.3	r feature elimination visualisation of feature relationships for removal.	145
4.4	Hierarchical for IBD-D.	146
4.5	Gap statistic and silhouette coefficient for IBD-D-FD.	147
4.6	Figure of IBD-D, box plots of averaged blood markers.	149
4.7	Bar plots of IBD-D.	149
4.8	Bar plots of IBD-NYD.	150
4.9	Figure of IBD-D-FD, box plots of averaged blood markers.	153
4.10	Bar plots of IBD-D-FD.	153
4.11	Bar plots of IBD-NYD cluster 3.	156
A.1	An example of a DBSCAN clustering: finding patterns in arbitrary data structures. . . .	191
B.1	Bar graph illustrating similar class distribution across Olivia's Vision forum training and test sets.	195
B.2	Box plots of VADER sentiment scores on the two patient fora, for a range of classes from OcIMIDo.	196

C.1	Graphical representation of COID.	197
C.2	Vectorisation representation of vasculitis.	199
D.1	PCA feature elimination visualisation of least-contributing features for removal.	203
D.2	Variance threshold feature elimination visualisation of features for removal.	204
D.3	PCA visualisation of eigenvalues via scree and cumulative for dimension reduction.	205
D.4	Gap statistic for IBD-NYD.	206
D.5	Figure of IBD-NYD, box plots of averaged blood markers.	207
D.6	Figure of permutation variable importance measure analysis for IBD-D.	208
D.7	Figure of IBD-D anxiety as endpoint with permutation variable importance measure.	209
D.8	Figure of IBD-D box plots of averaged blood markers across health rating in cluster 6.	210
D.9	Bar plot of IBD-D-FD cluster 6 IBD not-yet diagnosis against health rating.	211

List of Tables

1.1	Structural issues in current biomedical ontologies in regards to ocular IMIDs, as highlighted by domain experts.	67
2.1	Summary of OcIMIDo sources: class count, x-ref, and synonyms.	81
2.2	Examples of constructed relationship types contained in OcIMIDo.	81
2.3	Summary of the different annotations within OcIMIDo.	83
2.4	Structural issues in current biomedical ontologies, as highlighted by domain experts, with their corrections in OcIMIDo.	83
2.5	Frequency count of various OcIMIDo classes in the two fora, with sentiment score.	87
3.1	Notation of the data sources and embedding models used in this chapter.	101
3.2	Summary of COID sources: class count, x-ref, and synonyms.	104
3.3	Description of the embedding models and data available to use for analysis.	104
3.4	The 14 inflammatory terms of interest for analysis of the embedding models.	105
3.5	Experiment one W p-values of single-domain embedding models when expanded.	116

3.6	Experiment one r table comparing normalised physician scores by Pedersen et al. [200] to the embedding models.	117
3.7	Experiment one r table comparing different seeds to present consistency of results.	118
3.8	Experiment two W p-values of single-domain embedding models when expanded.	127
3.9	Experiment two r table comparing normalised physician scores by Pedersen et al. [200] to the embedding models.	128
3.10	Experiment two r table comparing different seeds to present consistency of results.	129
4.1	Table of IBD-D.	148
4.2	Table of IBD-NYD.	151
4.3	Table of IBD-D-FD.	152
4.4	Table of binomial regression and statistical test p-values for particular clusters within each time-stamped cohort.	155
A.1	An example of normalisation and standardisation scaling.	190
B.1	Patient synonyms and class mappings in OcIMIDo.	194
D.1	Normal blood markers of genetic sexes.	202

Listings

1.1	Example of an axiom in a biomedical ontology representing uveitis.	29
2.1	Coding example of the text preprocessing steps for cleaning the forum.	77

Bibliography

- [1] Roma Pahwa, Amandeep Goyal, Pankaj Bansal, and Ishwarlal Jialal. Chronic Inflammation. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), November 2020. URL <https://www.ncbi.nlm.nih.gov/books/NBK493173/>.
- [2] Sally Hannoodde and Dian N Nasuruddin. Acute Inflammatory Response. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), November 2020. URL <https://www.ncbi.nlm.nih.gov/books/NBK556083/>.
- [3] Samera Dean, Jonathan M Mathers, Melanie Calvert, Derek G Kyte, Dolores Conroy, Annie Folkard, Sue Southworth, Philip I Murray, and Alastair K Denniston. “The patient is speaking”: discovering the patient voice in ophthalmology. *The British Journal of Ophthalmology*, 101(6):700–708, June 2017. ISSN 0007-1161, 1468-2079. doi: 10.1136/bjophthalmol-2016-309955. URL <http://dx.doi.org/10.1136/bjophthalmol-2016-309955>.
- [4] L S King. Signs and symptoms. *The Journal of the American Medical Association*, 206(5):1063–1065, October 1968. ISSN 0098-7484. doi: 10.1001/jama.1968.03150050051011. URL <http://dx.doi.org/10.1001/jama.1968.03150050051011>.
- [5] Peter N Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *American Journal of Human Genetics*, 83(5):610–615, November 2008. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2008.09.017.
- [6] T Ferguson. Online patient-helpers and physicians working together: a new partnership for high quality health care. *British Medical Journal*, 321(7269):1129–1132, November 2000. ISSN 0959-8138. doi: 10.1136/bmj.321.7269.1129. URL <http://dx.doi.org/10.1136/bmj.321.7269.1129>.
- [7] S H Kaplan, S Greenfield, and J E Ware, Jr. Assessing the effects of physician-patient interactions on the outcomes of chronic disease. *Medical Care*, 27(3 Suppl):S110–27, March 1989. ISSN 0025-7079. doi: 10.1097/00005650-198903001-00010. URL <http://dx.doi.org/10.1097/00005650-198903001-00010>.
- [8] Tessa Richards, Victor M Montori, Fiona Godlee, Peter Lapsley, and Dave Paul. Let the patient revolution begin. *BMJ*, 346:f2614, May 2013. ISSN 0959-8138, 1756-1833. doi: 10.1136/bmj.f2614. URL <http://dx.doi.org/10.1136/bmj.f2614>.
- [9] Wayne A Beach. Managing “Stable” Cancer News. *Social Psychology Quarterly*, 84(1):26–48, March 2021. ISSN 0190-2725. doi: 10.1177/0190272520976133. URL <https://doi.org/10.1177/0190272520976133>.
- [10] A J Landre-Beauvais. Semiotics and the Features of Illness. *Paris: Brosson*, 315, 1813.
- [11] Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O B Jacobsen, Daniel Danis, Jean-Philippe Gourdine, Michael Gargano, Nomi L Harris, Nicolas Matentzoglou, Julie A McMurry, David Osumi-Sutherland, Valentina Cipriani, James P Balhoff, Tom Conlin, Hannah Blau, Gareth Baynam, Richard Palmer, Dylan Gratian, Hugh Dawkins, Michael Segal, Anna C Jansen, Ahmed Muaz, Willie H Chang, Jenna Bergerson, Stanley J F Laulederkind, Zafer Yüksel, Sergi Beltran, Alexandra F Freeman, Panagiotis I Sergouniotis, Daniel Durkin, Andrea L Storm, Marc Hanauer, Michael Brudno, Susan M Bello, Murat Sincan, Kayli Rageth, Matthew T Wheeler, Renske Oegema, Halima Loughi, Maria G Della Rocca, Rachel Thompson, Francisco Castellanos, James Priest, Charlotte Cunningham-Rundles, Ayushi Hegde, Ruth C Lovering, Catherine Hajek, Annie Olry, Luigi Notarangelo, Morgan Similuk, Xingmin A Zhang, David Gómez-Andrés, Hanns Lochmüller, Hélène Dollfus, Sergio Rosenzweig, Shruti Marwaha, Ana Rath, Kathleen Sullivan, Cynthia Smith, Joshua D Milner, Dorothée Leroux, Cornelius F Boerkoel, Amy Klion, Melody C Carter, Tudor Groza, Damian Smedley, Melissa A Haendel, Chris Mungall, and Peter N Robinson. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, 47(D1):D1018–D1027, January 2019. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gky1105. URL <https://doi.org/10.1093/nar/gky1105>.
- [12] Nicole A Vasilevsky, Erin D Foster, Mark E Engelstad, Leigh Carmody, Matt Might, Chip Chambers, Hugh J S Dawkins, Janine Lewis, Maria G Della Rocca, Michelle Snyder, Cornelius F Boerkoel, Ana Rath, Sharon F Terry, Alastair Kent, Beverly Searle, Gareth Baynam, Erik Jones, Pam Gavin, Michael Bamshad, Jessica Chong, Tudor Groza, David Adams, Adam C Resnick, Allison P Heath, Chris Mungall, Ingrid A Holm, Kayli Rageth, Catherine A Brownstein, Kent Shefchek, Julie A McMurry, Peter N Robinson, Sebastian Köhler, and Melissa A Haendel. Plain-language medical vocabulary for precision diagnosis. *Nature Genetics*, 50(4):474–476, April 2018. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-018-0096-x. URL <https://doi.org/10.1038/s41588-018-0096-x>.
- [13] David Epstein. The DIY Scientist, the Olympian, and the Mutated Gene. <https://www.propublica.org/article/muscular-dystrophy-patient-olympic-medalist-same-genetic-mutation>, January 2016. URL <https://www.propublica.org/article/muscular-dystrophy-patient-olympic-medalist-same-genetic-mutation>. Accessed: 2021-10-5.
- [14] W R Centerwall and S A Centerwall. Phenylketonuria (FOLLING’s disease). The story of its discovery. *Journal of the History of Medicine and Allied Sciences*, 16:292–296, July 1961. ISSN 0022-5045, 1468-4373. URL <https://www.ncbi.nlm.nih.gov/pubmed/13691924>.
- [15] Nigel Crisp. Patient power needs to be built on strong intellectual foundations: an essay by Nigel Crisp. *British Medical Journal*, 345:e6177, September 2012. ISSN 0959-8138, 1756-1833. doi: 10.1136/bmj.e6177. URL <http://dx.doi.org/10.1136/bmj.e6177>.
- [16] M K Buller and D B Buller. Physicians’ communication style and patient satisfaction. *Journal of Health and Social Behavior*, 28(4):375–388, December 1987. ISSN 0022-1465. URL <https://www.ncbi.nlm.nih.gov/pubmed/3429807>.
- [17] Dave deBronkart. How the e-patient community helped save my life: an essay by Dave deBronkart. *British Medical Journal*, 346:f1990, April 2013. ISSN 0959-8138, 1756-1833. doi: 10.1136/bmj.f1990. URL <http://dx.doi.org/10.1136/bmj.f1990>.
- [18] Martin Fortin, Marie-France Dubois, Catherine Hudon, Hassan Soubhi, and José Almirall. Multimorbidity and quality of life: a closer look. *Health and Quality of Life Outcomes*, 5(52), August 2007. doi: 10.1186/1477-7525-5-52. URL <https://doi.org/10.1186/1477-7525-5-52>.
- [19] Annemarie A Uijen and Eloy H van de Lisdonk. Multimorbidity in primary care: prevalence and trend over the last 20 years. *The European Journal of General Practice*, 14 Suppl 1:28–32, 2008. ISSN 1381-4788, 1751-1402. doi: 10.1080/13814780802436093. URL <https://doi.org/10.1080/13814780802436093>.
- [20] Karen Barnett, Stewart W Mercer, Michael Norbury, Graham Watt, Sally Wyke, and Bruce Guthrie. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet*, 380(9836):37–43, July 2012. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(12)60240-2. URL [http://dx.doi.org/10.1016/S0140-6736\(12\)60240-2](http://dx.doi.org/10.1016/S0140-6736(12)60240-2).

- [21] P Porcelli, C Leoci, V Guerra, G J Taylor, and R M Bagby. A longitudinal study of alexithymia and psychological distress in inflammatory bowel disease. *Journal of Psychosomatic Research*, 41(6):569–573, December 1996. ISSN 0022-3999. doi: 10.1016/s0022-3999(96)00221-8. URL [http://dx.doi.org/10.1016/s0022-3999\(96\)00221-8](http://dx.doi.org/10.1016/s0022-3999(96)00221-8).
- [22] E C de Rooy, B B Toner, R G Maunder, G R Greenberg, D Baron, A H Steinhart, R McLeod, and Z Cohen. Concerns of patients with inflammatory bowel disease: results from a clinical population. *The American Journal of Gastroenterology*, 96(6):1816–1821, June 2001. ISSN 0002-9270. doi: 10.1111/j.1572-0241.2001.03877.x. URL <https://doi.org/10.1111/j.1572-0241.2001.03877.x>.
- [23] Laura T Pizzi, Christine M Weston, Neil I Goldfarb, Deborah Moretti, Nicole Cobb, Jamie B Howell, Anthony Infantolino, Anthony J Dimarino, and Sidney Cohen. Impact of chronic conditions on quality of life in patients with inflammatory bowel disease. *Inflammatory Bowel Diseases*, 12(1):47–52, January 2006. ISSN 1078-0998. doi: 10.1097/01.mib.0000191670.04605.e7. URL <http://dx.doi.org/10.1097/01.mib.0000191670.04605.e7>.
- [24] D Felce. Defining and applying the concept of quality of life. *Journal of Intellectual Disability Research*, 41:126–135, April 1997. ISSN 0964-2633. doi: 10.1111/j.1365-2788.1997.tb00689.x. URL <https://doi.org/10.1111/j.1365-2788.1997.tb00689.x>.
- [25] Kerstin Denecke and Yihan Deng. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine*, 64(1):17–27, May 2015. ISSN 0933-3657, 1873-2860. doi: 10.1016/j.artmed.2015.03.006. URL <http://dx.doi.org/10.1016/j.artmed.2015.03.006>.
- [26] Mallika Prem Senthil, Lyndell Lim, Tasanee Braithwaite, Alastair Denniston, Eva K Fenwick, Ecosse Lamoureux, Jyoti Khadka, and Konrad Pesudovs. The Impact of Adult Uveitis on Quality of Life: An Exploratory Study. *Ophthalmic Epidemiology*, 28(5):1–9, December 2020. ISSN 0928-6586, 1744-5086. doi: 10.1080/09286586.2020.1856385. URL <https://doi.org/10.1080/09286586.2020.1856385>.
- [27] Zia Ul-Haq, Daniel F Mackay, Daniel Martin, Daniel J Smith, Jason M R Gill, Barbara I Nicholl, Breda Cullen, Jonathan Evans, Beverly Roberts, Ian J Deary, John Gallacher, Matthew Hotopf, Nick Craddock, and Jill P Pell. Heaviness, health and happiness: a cross-sectional study of 163066 UK Biobank participants. *Journal of Epidemiology and Community Health*, 68(4):340–348, April 2014. ISSN 0141-7681. doi: 10.1136/jech-2013-203077. URL <http://dx.doi.org/10.1136/jech-2013-203077>.
- [28] Jen Beaumont, Jennifer Thomas, and ONS. Measuring national well-being, health, 2012. Technical report, Office for National Statistics, 2012. URL http://www.ons.gov.uk/ons/dcp171766_271762.pdf.
- [29] Tom Ferguson and Gilles Frydman. The first generation of e-patients. *British Medical Journal*, 328(7449):1148–1149, May 2004. ISSN 0959-8138, 1756-1833. doi: 10.1136/bmj.328.7449.1148. URL <http://dx.doi.org/10.1136/bmj.328.7449.1148>.
- [30] Nicole Vasilevsky, Mark Engelstad, Erin Foster, Chris Mungall, Peter N Robinson, Sebastian Köhler, and Melissa Haendel. Enhancing the Human Phenotype Ontology for Use by the Layperson. In *9th International Biocuration Conference: ICBO/BioCreative*, volume 5. F1000Research, April 2016. doi: 10.7490/f1000research.1111752.1. URL <https://pdfs.semanticscholar.org/0850/646a0c9facd051f900a345e99ae568937e02.pdf>.
- [31] Simon Eaton, Sue Roberts, and Bridget Turner. Delivering person centred care in long term conditions. *British Medical Journal*, 350:h181, February 2015. ISSN 0959-8138, 1756-1833. doi: 10.1136/bmj.h181. URL <http://dx.doi.org/10.1136/bmj.h181>.
- [32] Tessa Richards, Angela Coulter, and Paul Wicks. Time to deliver patient centred care. *British Medical Journal*, 350:h530, February 2015. ISSN 0959-8138, 1756-1833. doi: 10.1136/bmj.h530. URL <http://dx.doi.org/10.1136/bmj.h530>.
- [33] Werner Ceusters, Barry Smith, and Jim Flanagan. Ontology and Medical Terminology: Why Description Logics Are Not Enough. In *Towards an Electronic Patient Record*. Boston, MA: Medical Records Institute, 2003. URL <https://philarchive.org/archive/CEUOAM>.
- [34] Tessa Richards. Listen to patients first. *British Medical Journal*, 349:g5765, September 2014. ISSN 0959-8138, 1756-1833. doi: 10.1136/bmj.g5765. URL <http://dx.doi.org/10.1136/bmj.g5765>.
- [35] Yan Zhang, Yalin Sun, and Bo Xie. Quality of health information for consumers on the web: A systematic review of indicators, criteria, tools, and evaluation results. *Journal of the Association for Information Science and Technology*, 66(10):2071–2084, October 2015. ISSN 2330-1635, 2330-1643. doi: 10.1002/asi.23311. URL <http://dx.doi.org/10.1002/asi.23311>.
- [36] Catherine H Saunders, Curtis L Petersen, Marie-Anne Durand, Pamela J Bagley, and Glyn Elwyn. Bring on the Machines: Could Machine Learning Improve the Quality of Patient Education Materials? A Systematic Search and Rapid Review. *JCO Clinical Cancer Informatics*, 2:1–16, December 2018. ISSN 2473-4276. doi: 10.1200/CCI.18.00010. URL <http://dx.doi.org/10.1200/CCI.18.00010>.
- [37] Ahmed Allam, Peter Johannes Schulz, and Kent Nakamoto. The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: two experiments manipulating Google output. *Journal of Medical Internet Research*, 16(4):e100, April 2014. ISSN 1439-4456, 1438-8871. doi: 10.2196/jmir.2642. URL <http://dx.doi.org/10.2196/jmir.2642>.
- [38] D Spiro and F Heidrich. Lay understanding of medical terminology. *The Journal of Family Practice*, 17(2):277–279, August 1983. ISSN 0094-3509. URL <https://pubmed.ncbi.nlm.nih.gov/6875485>.
- [39] Karen L Barker, Margaret Reid, and Catherine J Minns Lowe. Divided by a lack of common language? A qualitative study exploring the use of language by health professionals treating back pain. *BMC Musculoskeletal Disorders*, 10:123, October 2009. ISSN 1471-2474. doi: 10.1186/1471-2474-10-123. URL <http://dx.doi.org/10.1186/1471-2474-10-123>.
- [40] Hedy S Wald, Catherine E Dube, and David C Anthony. Untangling the Web—the impact of Internet use on health care and the physician-patient relationship. *Patient Education and Counseling*, 68(3):218–224, November 2007. ISSN 0738-3991. doi: 10.1016/j.pec.2007.05.016. URL <http://dx.doi.org/10.1016/j.pec.2007.05.016>.
- [41] Suzy A Iverson, Kristin B Howard, and Brian K Penney. Impact of internet use on health-related behaviors and the patient-physician relationship: a survey-based study and review. *The Journal of the American Osteopathic Association*, 108(12):699–711, December 2008. ISSN 0098-6151, 1945-1997. URL <https://www.ncbi.nlm.nih.gov/pubmed/19075034>.
- [42] Erik Cambria, Amir Hussain, and Chris Eck. Bridging the Gap between Structured and Unstructured Health-Care Data through Semantics and Sentics. In *Web Science Conference*, June 2011. URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.474.6374&rep=rep1&type=pdf>.
- [43] B M Korsch, E K Gozzi, and V Francis. Gaps in doctor-patient communication. 1. Doctor-patient interaction and patient satisfaction. *Pediatrics*, 42(5):855–871, November 1968. ISSN 0031-4005. URL <https://www.ncbi.nlm.nih.gov/pubmed/5685370>.
- [44] Min Sook Park, Zhe He, Zhiwei Chen, Sanghee Oh, and Jiang Bian. Consumers’ Use of UMLS Concepts on Social Media: Diabetes-Related Textual Data Analysis in Blog and Social Q&A Sites. *JMIR Medical Informatics*, 4(4):e41, November 2016. ISSN 2291-9694. doi: 10.2196/medinform.5748. URL <http://dx.doi.org/10.2196/medinform.5748>.
- [45] Andrew R Deans, Suzanna E Lewis, Eva Huala, Salvatore S Anzaldo, Michael Ashburner, James P Balhoff, David C Blackburn, Judith A Blake, J Gordon Burleigh, Bruno Chanet, Laurel D Cooper, Mélanie Courtot, Sándor Csősz, Hong Cui, Wasila Dahdul, Sandip Das, T Alexander Dececchi, Agnes Dettai, Rui Diogo, Robert E Druzinsky, Michel Dumontier, Nico M Franz, Frank Friedrich, George V Gkoutos, Melissa Haendel, Luke J Harmon, Terry F Hayamizu, Yongqun He, Heather M Hines, Nizar Ibrahim, Laura M Jackson, Pankaj Jaiswal, Christina James-Zorn, Sebastian Köhler, Guillaume Lecointre, Hilmar Lapp, Carolyn J Lawrence, Nicolas Le Novère, John G Lundberg, James Macklin, Austin R Mast, Peter E Midford, István Mikó, Christopher J Mungall, Anika Oelrich, David Osumi-Sutherland, Helen Parkinson, Martin J Ramirez, Stefan Richter, Peter N Robinson, Alan Ruttenberg, Katja S Schulz, Erik Segerdell, Katja C Seltmann, Michael J Sharkey, Aaron D Smith, Barry Smith, Chelsea D Specht, R Burke Squires, Robert W Thacker, Anne Thessen, Jose Fernandez-Triana, Mauno Vihinen, Peter D Vize, Lars Vogt, Christine E Wall, Ramona L Walls, Monte Westerfeld, Robert A Wharton, Christian S Wirkner, James B Woolley, Matthew J Yoder, Aaron M Zorn, and Paula Mabree. Finding our way through phenotypes. *The Public Library of Science: Biology*, 13(1):e1002033, January 2015. ISSN 1544-9173, 1545-7885. doi: 10.1371/journal.pbio.1002033. URL <https://doi.org/10.1371/journal.pbio.1002033>.

- [46] Robert Hoehndorf, Paul N Schofield, and Georgios V Gkoutos. Analysis of the human diseaseome using phenotype similarity between common, genetic, and infectious diseases. *Scientific Reports*, 5:10888, June 2015. ISSN 2045-2322. doi: 10.1038/srep10888. URL <https://doi.org/10.1038/srep10888>.
- [47] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical Machine Learning in Health Care. *arXiv*, September 2020. URL <http://arxiv.org/abs/2009.10576>.
- [48] The Royal College of General Practitioners. 15-minute minimum consultations, continuity of care through 'micro-teams', and an end to isolated working: this is the future of general practice. Technical report, RCGP, May 2019. URL <https://www.rcgp.org.uk/about-us/news/2019/may/15-minute-minimum-consultations-continuity-of-care.aspx>.
- [49] L M Ong, J C de Haes, A M Hoos, and F B Lammes. Doctor-patient communication: a review of the literature. *Social Science & Medicine*, 40(7):903–918, April 1995. ISSN 0037-7856, 0277-9536. doi: 10.1016/0277-9536(94)00155-m. URL [http://dx.doi.org/10.1016/0277-9536\(94\)00155-m](http://dx.doi.org/10.1016/0277-9536(94)00155-m).
- [50] Parvathy Bowes, Fiona Stevenson, Sanjiv Ahluwalia, and Elizabeth Murray. 'I need her to be a doctor': patients' experiences of presenting health information from the internet in GP consultations. *The Journal of The Royal College of General Practitioners*, 62(604):e732–8, November 2012. ISSN 0960-1643, 1478-5242. doi: 10.3399/bjgp12X658250. URL <http://dx.doi.org/10.3399/bjgp12X658250>.
- [51] Andy McKay, Carlos Pantoja, Richard Hall, Sarah Matthews, Phil Spalding, and Rajarshi Banerjee. Patient understanding and experience of non-invasive imaging diagnostic techniques and the liver patient pathway. *Journal of Patient-Reported Outcomes*, 5(1):89, September 2021. ISSN 2509-8020. doi: 10.1186/s41687-021-00363-5. URL <https://doi.org/10.1186/s41687-021-00363-5>.
- [52] Fay Probert, Alissa Walsh, Marta Jagielowicz, Tianrong Yeo, Timothy D W Claridge, Alison Simmons, Simon Travis, and Daniel C Anthony. Plasma Nuclear Magnetic Resonance Metabolomics Discriminates Between High and Low Endoscopic Activity and Predicts Progression in a Prospective Cohort of Patients With Ulcerative Colitis. *Journal of Crohn's & Colitis*, 12(11):1326–1337, November 2018. ISSN 1873-9946, 1876-4479. doi: 10.1093/ecco-jcc/jjy101. URL <http://dx.doi.org/10.1093/ecco-jcc/jjy101>.
- [53] Kelly Young. Doctors' understanding of rheumatoid disease does not align with patients' experiences. *BMJ*, 346:f2901, May 2013. ISSN 0959-8138, 1756-1833. doi: 10.1136/bmj.f2901. URL <http://dx.doi.org/10.1136/bmj.f2901>.
- [54] Elena Davcheva. Text Mining Mental Health Forums - Learning From User Experiences. In *Information Systems*, 2018. URL <http://ecis2018.eu/wp-content/uploads/2018/09/1503-doc.pdf>.
- [55] Solène Eholié, Mike Donald Tapi Nzali, Sandra Bringay, and Clement Jonquet. MuEvo, a breast cancer Consumer Health Vocabulary built out of web forums. In *Semantic Web Applications and Tools for Life Sciences*, 2016. URL <https://hal.archives-ouvertes.fr/hal-01398253/>.
- [56] N F de Keizer, A Abu-Hanna, and J H Zwetsloot-Schonk. Understanding terminological systems. I: Terminology and typology. *Methods of Information in Medicine*, 39(1):16–21, March 2000. ISSN 0026-1270. URL <https://pubmed.ncbi.nlm.nih.gov/10786065>.
- [57] Ronald Cornet and Nicolette de Keizer. Forty years of SNOMED: a literature review. *BMC Medical Informatics and Decision Making*, 8 Suppl 1:S2, October 2008. ISSN 1472-6947. doi: 10.1186/1472-6947-8-S1-S2. URL <http://dx.doi.org/10.1186/1472-6947-8-S1-S2>.
- [58] Arwen H Pieterse, Nienke A Jager, Ellen M A Smets, and Inge Henselmans. Lay understanding of common medical terminology in oncology. *Psycho-oncology*, 22(5):1186–1191, May 2013. ISSN 1057-9249, 1099-1611. doi: 10.1002/pon.3096. URL <https://doi.org/10.1002/pon.3096>.
- [59] Sandra E Staffieri, Lisa S Kearns, Paul G Sanfilippo, Jamie E Craig, David A Mackey, and Alex W Hewitt. Crowd-sourced Ontology for Photoleukocoria: Identifying Common Internet Search Terms for a Potentially Important Pediatric Ophthalmic Sign. *Translational Vision Science & Technology*, 7(1):18, February 2018. ISSN 2164-2591. doi: 10.1167/tvst.7.1.18. URL <http://dx.doi.org/10.1167/tvst.7.1.18>.
- [60] Ivan Shun Lau, Zeljko Kraljevic, Mohammad Al-Agil, Shelley Charing, Alan Quarterman, Harold Parkes, Victoria Metaxa, Katherine Sleeman, Wei Gao, Richard J B Dobson, James T Teo, and Phil Hopkins. Natural language word embeddings as a glimpse into healthcare language and associated mortality surrounding end of life. *BMJ Health & Care Informatics*, 28(1), October 2021. ISSN 2632-1009. doi: 10.1136/bmjhci-2021-100464. URL <http://dx.doi.org/10.1136/bmjhci-2021-100464>.
- [61] Robert Leaman, Ritu Khare, and Zhiyong Lu. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, 57:28–37, October 2015. ISSN 1532-0464, 1532-0480. doi: 10.1016/j.jbi.2015.07.010. URL <http://dx.doi.org/10.1016/j.jbi.2015.07.010>.
- [62] Phebe Cramer and Tiffany Steinwert. Thin is good, fat is bad: How early does it begin? *Journal of Applied Developmental Psychology*, 19(3):429–451, July 1998. ISSN 0193-3973. doi: 10.1016/S0193-3973(99)80049-5. URL [http://dx.doi.org/10.1016/S0193-3973\(99\)80049-5](http://dx.doi.org/10.1016/S0193-3973(99)80049-5).
- [63] Bradford W Hesse, David E Nelson, Gary L Kreps, Robert T Croyle, Neeraj K Arora, Barbara K Rimer, and Kasisomayajula Viswanath. Trust and sources of health information: the impact of the Internet and its implications for health care providers: findings from the first Health Information National Trends Survey. *Archives of Internal Medicine*, 165(22):2618–2624, 2005. ISSN 0003-9926. doi: 10.1001/archinte.165.22.2618. URL <http://dx.doi.org/10.1001/archinte.165.22.2618>.
- [64] Susannah Fox and Maeve Duggan. Health Online 2013. Technical report, Pew Research Center, January 2013. URL <https://www.pewresearch.org/internet/2013/01/15/health-online-2013/>.
- [65] Benjamin Crooks, John McLaughlin, Katsuyoshi Matsuoka, Taku Kobayashi, Hajime Yamazaki, and Jimmy K Limdi. The dietary practices and beliefs of people living with inactive ulcerative colitis. *European Journal of Gastroenterology & Hepatology*, 33(3):372–379, March 2021. ISSN 0954-691X, 1473-5687. doi: 10.1097/MEG.0000000000001911. URL <https://doi.org/10.1097/MEG.0000000000001911>.
- [66] Natalia Pletneva, Alejandro Vargas, and Céilia Boyer. How do general public search online health information. Technical report, Health On the Net Foundation (HON), 2011. URL https://www.hon.ch/Survey/khresmoi_general_public_survey_results.html.
- [67] Thomas Opitz, Jérôme Aze, Sandra Bringay, Cyrille Joutard, Christian Lavergne, and Caroline Mollevi. Breast cancer and quality of life: medical information extraction from health forums. In *25th European Medical Informatics Conference (MIE)*, volume 205, pages 1070–1074, 2014. URL <https://www.ncbi.nlm.nih.gov/pubmed/25160353>.
- [68] George Gkotsis, Christoph Mueller, Richard J B Dobson, Tim J P Hubbard, and Rina Dutta. Mining Social Media Data to Study the Consequences of Dementia Diagnosis on Caregivers and Relatives. *Dementia and Geriatric Cognitive Disorders*, 49(3):295–302, August 2020. ISSN 1420-8008, 1421-9824. doi: 10.1159/000509123. URL <http://dx.doi.org/10.1159/000509123>.
- [69] Paul Wicks, Michael Massagli, Jeana Frost, Catherine Brownstein, Sally Okun, Timothy Vaughan, Richard Bradley, and James Heywood. Sharing health data for better outcomes on PatientsLikeMe. *Journal of Medical Internet Research*, 12(2):e19, June 2010. ISSN 1439-4456, 1438-8871. doi: 10.2196/jmir.1549. URL <http://dx.doi.org/10.2196/jmir.1549>.
- [70] Per E Kummervold, Deede Gammon, Svein Bergvik, Jan-Are K Johnsen, Toralf Hasvold, and Jan H Rosenving. Social support in a wired world: use of online mental health forums in Norway. *Nordic Journal of Psychiatry*, 56(1):59–65, 2002. ISSN 0803-9488. doi: 10.1080/08039480252803945. URL <http://dx.doi.org/10.1080/08039480252803945>.
- [71] Carolyn Pehora, Nisha Gajaria, Melyssa Stoute, Sonia Fracassa, Refilwe Serebale-O'Sullivan, and Clyde T Matava. Are Parents Getting it Right? A Survey of Parents' Internet Use for Children's Health Care Information. *Interactive Journal of Medical Research*, 4(2):e12, June 2015. ISSN 1929-073X. doi: 10.2196/ijmr.3790. URL <http://dx.doi.org/10.2196/ijmr.3790>.
- [72] Alexandra Balahur. Sentiment Analysis in Social Media Texts. In *Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 120–128, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL <https://aclanthology.org/W13-1617>.

- [73] A K Denniston, D Kyte, M Calvert, and J M Burr. An introduction to patient-reported outcome measures in ophthalmic research. *Eye*, 28(6):637–645, June 2014. ISSN 0950-222X, 1476-5454. doi: 10.1038/eye.2014.41. URL <http://dx.doi.org/10.1038/eye.2014.41>.
- [74] Emily Oshima Lee and Ezekiel J Emanuel. Shared decision making to improve care and reduce costs. *The New England Journal of Medicine*, 368(1):6–8, January 2013. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMp1209500. URL <http://dx.doi.org/10.1056/NEJMp1209500>.
- [75] David Wadden, Tal August, Qisheng Li, and Tim Althoff. The Effect of Moderation on Online Mental Health Conversations. In *International AAAI Conference on Web and Social Media*, volume 15, pages 751–763, May 2021. URL <https://ojs.aaai.org/index.php/ICWSM/article/view/18100>.
- [76] Tasanee Braithwaite, Melanie Calvert, Alastair Gray, Konrad Pesudovs, and Alastair K Denniston. The use of patient-reported outcome research in modern ophthalmology: impact on clinical trials and routine clinical practice. *Patient Related Outcome Measures*, 10:9–24, January 2019. ISSN 1179-271X. doi: 10.2147/PROM.S162802. URL <https://dx.doi.org/10.2147/PROM.S162802>.
- [77] Ivan Sekulić, Matej Gjurković, and Jan Šnajder. Not Just Depressed: Bipolar Disorder Prediction on Reddit. *arXiv*, November 2018. URL https://www.researchgate.net/publication/32889704_Not_Just_Depressed_Bipolar_Disorder_Prediction_on_Reddit/.
- [78] L Ferrero-Miliani, O H Nielsen, P S Andersen, and S E Girardin. Chronic inflammation: importance of NOD2 and NALP3 in interleukin-1 β generation. *Clinical and Experimental Immunology*, 147(2):227–235, February 2007. ISSN 0009-9104. doi: 10.1111/j.1365-2249.2006.03261.x. URL <https://doi.org/10.1111/j.1365-2249.2006.03261.x>.
- [79] Jacques S Beckmann and Daniel Lew. Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities. *Genome medicine*, 8(1), 2016. doi: 10.1186/s13073-016-0388-7. URL <https://doi.org/10.1186/s13073-016-0388-7>.
- [80] Isabel Bondia-Pons, Lisa Ryan, and J Alfredo Martinez. Oxidative stress and inflammation interactions in human obesity. *Journal of Physiology and Biochemistry*, 68(4):701–711, December 2012. ISSN 1138-7548, 1877-8755. doi: 10.1007/s13105-012-0154-2. URL <https://dx.doi.org/10.1007/s13105-012-0154-2>.
- [81] Andrea C Sartori, David E Vance, Larry Z Slater, and Michael Crowe. The impact of inflammation on cognitive function in older adults: implications for healthcare practice and research. *The Journal of Neuroscience Nursing*, 44(4):206–217, August 2012. ISSN 0888-0395, 1945-2810. doi: 10.1097/JNN.0b013e3182527690. URL <http://dx.doi.org/10.1097/JNN.0b013e3182527690>.
- [82] Cem Gabay. Interleukin-6 and chronic inflammation. *Arthritis Research & Therapy*, 8 Suppl 2:S3, July 2006. ISSN 1478-6354, 1478-6362. doi: 10.1186/ar1917. URL <http://dx.doi.org/10.1186/ar1917>.
- [83] Ruslan Medzhitov. Origin and physiological roles of inflammation. *Nature*, 454(7203):428–435, July 2008. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature07201. URL <https://doi.org/10.1038/nature07201>.
- [84] David Furman, Judith Campisi, Eric Verdin, Pedro Carrera-Bastos, Sasha Targ, Claudio Franceschi, Luigi Ferrucci, Derek W Gilroy, Alessio Fasano, Gary W Miller, Andrew H Miller, Alberto Mantovani, Cornelia M Weyand, Nir Barzilai, Jorg J Goronzy, Thomas A Rando, Rita B Effros, Alejandro Lucia, Nicole Kleinstreuer, and George M Slavich. Chronic inflammation in the etiology of disease across the life span. *Nature Medicine*, 25(12):1822–1832, December 2019. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-019-0675-0. URL <https://doi.org/10.1038/s41591-019-0675-0>.
- [85] George A Bray. Medical consequences of obesity. *The Journal of Clinical Endocrinology and Metabolism*, 89(6):2583–2589, June 2004. ISSN 0021-972X. doi: 10.1210/jc.2004-0535. URL <http://dx.doi.org/10.1210/jc.2004-0535>.
- [86] Kyle Strimbu and Jorge A Tavel. What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6):463–466, November 2010. ISSN 1746-630X, 1746-6318. doi: 10.1097/COH.0b013e32833ed177. URL <https://doi.org/10.1097/COH.0b013e32833ed177>.
- [87] Elizabeth A Regan, Craig P Hersh, Peter J Castaldi, Dawn L DeMeo, Edwin K Silverman, James D Crapo, and Russell P Bowler. Omics and the Search for Blood Biomarkers in Chronic Obstructive Pulmonary Disease. Insights from COPDGene. *American Journal of Respiratory Cell and Molecular Biology*, 61(2):143–149, August 2019. ISSN 1044-1549, 1535-4989. doi: 10.1165/rcmb.2018-0245PS. URL <https://doi.org/10.1165/rcmb.2018-0245PS>.
- [88] Claire Welsh, Paul Welsh, Patrick B Mark, Carlos A Celis-Morales, James Lewsey, Stuart R Gray, Donald M Lyall, Stamatina Iliodromiti, Jason M R Gill, Jill Pell, Pardeep S Jhund, and Naveed Sattar. Association of Total and Differential Leukocyte Counts With Cardiovascular Disease and Mortality in the UK Biobank. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 38(6):1415–1423, June 2018. ISSN 1079-5642, 1524-4636. doi: 10.1161/ATVBAHA.118.310945. URL <https://doi.org/10.1161/ATVBAHA.118.310945>.
- [89] Rosário Monteiro and Isabel Azevedo. Chronic inflammation in obesity and the metabolic syndrome. *Mediators of Inflammation*, 2010, July 2010. ISSN 0962-9351, 1466-1861. doi: 10.1155/2010/289645. URL <https://doi.org/10.1155/2010/289645>.
- [90] Jeroen Nijhuis, Sander S Rensen, Yanti Slaats, Francois M H van Dielen, Wim A Buurman, and Jan Willem M Greve. Neutrophil activation in morbid obesity, chronic activation of acute inflammation. *Obesity*, 17(11):2014–2018, November 2009. ISSN 1930-7381. doi: 10.1038/oby.2009.113. URL <https://doi.org/10.1038/oby.2009.113>.
- [91] Jocelyne Just, Rahele Gouvis-Echraghi, Sarah Rouve, Stephanie Wanin, David Moreau, and Isabella Annesi-Maesano. Two novel, severe asthma phenotypes identified during childhood using a clustering approach. *European Respiratory Journal*, 40(1):55–60, July 2012. ISSN 0903-1936, 1399-3003. doi: 10.1183/09031936.00123411. URL <http://dx.doi.org/10.1183/09031936.00123411>.
- [92] BioRender online software to create professional science figures. <https://biorender.com/>, 2022. URL <https://biorender.com/>. Accessed: 2022.
- [93] Jingyuan Liu, Yao Liu, Pan Xiang, Lin Pu, Haofeng Xiong, Chuansheng Li, Ming Zhang, Jianbo Tan, Yanli Xu, Rui Song, Meihua Song, Lin Wang, Wei Zhang, Bing Han, Li Yang, Xiaojing Wang, Guijin Zhou, Ting Zhang, Ben Li, Yanbin Wang, Zhihai Chen, and Xianbo Wang. Neutrophil-to-lymphocyte ratio predicts critical illness patients with 2019 coronavirus disease in the early stage. *Journal of Translational Medicine*, 18(1):206, May 2020. ISSN 1479-5876. doi: 10.1186/s12967-020-02374-0. URL <http://dx.doi.org/10.1186/s12967-020-02374-0>.
- [94] S Vermeire, G Van Assche, and P Rutgeerts. Laboratory markers in IBD: useful, magic, or unnecessary toys? *Gut*, 55(3):426–431, March 2006. ISSN 0017-5749. doi: 10.1136/gut.2005.069476. URL <http://dx.doi.org/10.1136/gut.2005.069476>.
- [95] R M Beattie, J A Walker-Smith, and S H Murch. Indications for investigation of chronic gastrointestinal symptoms. *Archives of Disease in Childhood*, 73(4):354–355, October 1995. ISSN 0003-9888, 1468-2044. doi: 10.1136/adc.73.4.354. URL <http://dx.doi.org/10.1136/adc.73.4.354>.
- [96] Séverine Vermeire, Gert Van Assche, and Paul Rutgeerts. C-reactive protein as a marker for inflammatory bowel disease. *Inflammatory Bowel Diseases*, 10(5):661–665, September 2004. ISSN 1078-0998. doi: 10.1097/00054725-200409000-00026. URL <http://dx.doi.org/10.1097/00054725-200409000-00026>.
- [97] D Adapa, Yrkm Sai, S Y Anand, S Mehaboobi, and E R Aramalla. A brief review on immune mediated diseases. *Journal of Clinical & Cellular Immunology*, 02(03), 2011. ISSN 2155-9899. doi: 10.4172/2155-9899.s11-001. URL <http://dx.doi.org/10.4172/2155-9899.s11-001>.
- [98] Haya H Al-Ani, Joanne L Sims, Oren Tomkins-Netzer, Susan Lightman, and Rachael L Niederer. Vision loss in anterior uveitis. *The British Journal of Ophthalmology*, April 2020. ISSN 0007-1161, 1468-2079. doi: 10.1136/bjophthalmol-2019-315551. URL <http://dx.doi.org/10.1136/bjophthalmol-2019-315551>.
- [99] Tasanee Braithwaite, Anuradha Subramanian, Axel Petzold, James Galloway, Nicola J Adderley, Susan P Mollan, Gordon T Plant, Krishnarajah Nirantharakumar, and Alastair K Denniston. Trends in Optic Neuritis Incidence and Prevalence in the UK and Association With Systemic and Neurologic Disease. *JAMA Neurology*, 77(12):1514–1523, December 2020. ISSN 2168-6149, 2168-6157. doi: 10.1001/jamaneurol.2020.3502. URL <http://dx.doi.org/10.1001/jamaneurol.2020.3502>.

- [100] Tasanee Braithwaite, Nicola J Adderley, Anuradha Subramanian, James Galloway, John H Kempen, Krishna Gokhale, Andrew P Cope, Andrew D Dick, Krishnarajah Nirantharakumar, and Alastair K Denniston. Epidemiology of scleritis in the United Kingdom from 1997 to 2018: Population-based analysis of 11 million patients and association between scleritis and infectious and immune-mediated inflammatory disease, March 2021. ISSN 2326-5191, 2326-5205. URL <http://dx.doi.org/10.1002/art.41709>.
- [101] D G James, A I Friedmann, and E Graham. Uveitis. A series of 368 patients. *Transactions of the Ophthalmological Societies of the United Kingdom*, 96(1):108–112, April 1976. ISSN 0078-5334. URL <https://www.ncbi.nlm.nih.gov/pubmed/1087766>.
- [102] A D Perrett, G Higgins, H H Johnston, G R Massarella, S C Truelove, and Ralph Wright. The Liver in Crohn's Disease. *QJM: An International Journal of Medicine*, 40(2):187–209, April 1971. ISSN 1460-2725. doi: 10.1093/oxfordjournals.qjmed.a067265. URL <http://dx.doi.org/10.1093/oxfordjournals.qjmed.a067265>.
- [103] Douglas A Jabs, Robert B Nussenblatt, James T Rosenbaum, and Standardization of Uveitis Nomenclature (SUN) Working Group. Standardization of uveitis nomenclature for reporting clinical data. Results of the First International Workshop. *American Journal of Ophthalmology*, 140(3):509–516, September 2005. ISSN 0002-9394. doi: 10.1016/j.ajo.2005.03.057. URL <http://dx.doi.org/10.1016/j.ajo.2005.03.057>.
- [104] M J Hogan, S J Kimura, and P Thygeson. Signs and symptoms of uveitis. I. Anterior uveitis. *American Journal of Ophthalmology*, 47(5 Pt 2):155–170, May 1959. ISSN 0002-9394. doi: 10.1016/s0002-9394(14)78239-x. URL [http://dx.doi.org/10.1016/s0002-9394\(14\)78239-x](http://dx.doi.org/10.1016/s0002-9394(14)78239-x).
- [105] B Manohar Babu and S R Rathinam. Intermediate uveitis. *Indian Journal of Ophthalmology*, 58(1):21–27, January 2010. ISSN 0301-4738, 1998-3689. doi: 10.4103/0301-4738.58469. URL <http://dx.doi.org/10.4103/0301-4738.58469>.
- [106] Andrew Dick, Annabelle Okada, and John Forrester. *Practical Manual of Intraocular Inflammation*. CRC Press, first edition, 2019. ISBN 9780367387204.
- [107] GBD 2017 Inflammatory Bowel Disease Collaborators. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet Gastroenterology & Hepatology*, 5(1):17–30, January 2020. ISSN 2468-1253. doi: 10.1016/S2468-1253(19)30333-4. URL [http://dx.doi.org/10.1016/S2468-1253\(19\)30333-4](http://dx.doi.org/10.1016/S2468-1253(19)30333-4).
- [108] Karen Kemp, Jane Griffiths, and Karina Lovell. Understanding the health and social care needs of people living with IBD: a meta-synthesis of the evidence. *World Journal of Gastroenterology*, 18(43):6240–6249, November 2012. ISSN 1007-9327, 2219-2840. doi: 10.3748/wjg.v18.i43.6240. URL <http://dx.doi.org/10.3748/wjg.v18.i43.6240>.
- [109] Bobak Moazzami, Kasra Moazzami, and Nima Rezaei. Early onset inflammatory bowel disease: manifestations, genetics and diagnosis. *The Turkish Journal of Pediatrics*, 61(5):637–647, 2019. ISSN 0041-4301. doi: 10.24953/turkjped.2019.05.001. URL <http://dx.doi.org/10.24953/turkjped.2019.05.001>.
- [110] Daniel B Graham and Ramnik J Xavier. Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature*, 578(7796):527–539, February 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2025-2. URL <https://doi.org/10.1038/s41586-020-2025-2>.
- [111] Global Guidelines Review team. Inflammatory Bowel Disease. Technical report, World Gastroenterology Organisation, 2015. URL <https://www.worldgastroenterology.org/UserFiles/file/guidelines/inflammatory-bowel-disease-english-2015.pdf>.
- [112] A D Perrett, G Higgins, H H Johnston, G R Massarella, S C Truelove, and R Wright. The liver in ulcerative colitis. *QJM: An International Journal of Medicine*, 40(158):211–238, April 1971. ISSN 0033-5622. doi: 10.1093/oxfordjournals.qjmed.a067266. URL <http://dx.doi.org/10.1093/oxfordjournals.qjmed.a067266>.
- [113] Stephan C Bischoff, Johanna Escher, Xavier Hébuterne, Stanisław Klek, Zeljko Krznaric, Stéphane Schneider, Raanan Shamir, Kalina Stardelova, Nicolette Wiersma, Anthony E Wiskin, and Alastair Forbes. ESPEN practical guideline: Clinical Nutrition in inflammatory bowel disease. *Clinical Nutrition*, 39(3):632–653, March 2020. ISSN 0261-5614, 1532-1983. doi: 10.1016/j.clnu.2019.11.002. URL <http://dx.doi.org/10.1016/j.clnu.2019.11.002>.
- [114] Justyna Kikut, Karolina Skonieczna-Żydecka, Diana Sochaczewska, Agnieszka Kordek, and Małgorzata Szczuko. Differences in Dietary Patterns of Adolescent Patients with IBD. *Nutrients*, 13(9), September 2021. ISSN 2072-6643. doi: 10.3390/nu13093119. URL <http://dx.doi.org/10.3390/nu13093119>.
- [115] May-Bente Bengtson, Camilla Solberg, Geir Aamodt, Jørgen Jahnsen, Bjørn Moum, Jostein Sauar, Morten H Vatn, and Ibsen Study Group. Clustering in time of familial IBD separates ulcerative colitis from Crohn's disease. *Inflammatory Bowel Diseases*, 15(12):1867–1874, December 2009. ISSN 1078-0998, 1536-4844. doi: 10.1002/ibd.20978. URL <http://dx.doi.org/10.1002/ibd.20978>.
- [116] Camille Zallot, Didier Quilliot, Jean-Baptiste Chevaux, Carina Peyrin-Biroulet, Rosa Maria Guéant-Rodriguez, Estelle Freling, Benjamin Collet-Fenetrier, Nicolas Williet, Olivier Ziegler, Marc-André Bigard, Jean-Louis Guéant, and Laurent Peyrin-Biroulet. Dietary beliefs and behavior among inflammatory bowel disease patients. *Inflammatory Bowel Diseases*, 19(1):66–72, January 2013. ISSN 1078-0998, 1536-4844. doi: 10.1002/ibd.22965. URL <http://dx.doi.org/10.1002/ibd.22965>.
- [117] Subrata Ghosh and Rod Mitchell. Impact of inflammatory bowel disease on quality of life: Results of the European Federation of Crohn's and Ulcerative Colitis Associations (EFCCA) patient survey. *Journal of Crohn's & Colitis*, 1(1): 10–20, September 2007. ISSN 1873-9946. doi: 10.1016/j.crohns.2007.06.005. URL <http://dx.doi.org/10.1016/j.crohns.2007.06.005>.
- [118] Lisa M Lix, Lesley A Graff, John R Walker, Ian Clara, Patricia Rawsthorne, Linda Rogala, Norine Miller, Jason Ediger, Thea Pretorius, and Charles N Bernstein. Longitudinal study of quality of life and psychological functioning for active, fluctuating, and inactive disease patterns in inflammatory bowel disease. *Inflammatory Bowel Diseases*, 14(11):1575–1584, November 2008. ISSN 1078-0998, 1536-4844. doi: 10.1002/ibd.20511. URL <http://dx.doi.org/10.1002/ibd.20511>.
- [119] W H Kim, Y S Cho, H M Yoo, I S Park, E C Park, and J G Lim. Quality of life in Korean patients with inflammatory bowel diseases: ulcerative colitis, Crohn's disease and intestinal Behçet's disease. *International Journal of Colorectal Disease*, 14(1):52–57, February 1999. ISSN 0179-1958. doi: 10.1007/s003840050183. URL <https://doi.org/10.1007/s003840050183>.
- [120] Dawn Farrell, Micol Artom, Wladyslaw Czuber-Dochan, Lars P Jelsness-Jørgensen, Christine Norton, and Eileen Savage. Interventions for fatigue in inflammatory bowel disease. *Cochrane Database of Systematic Reviews*, 4:CD012005, April 2020. ISSN 1469-493X, 1361-6137. doi: 10.1002/14651858.CD012005.pub2. URL <https://doi.org/10.1002/14651858.CD012005.pub2>.
- [121] David J Gracie, Christopher J M Williams, Ruchit Sood, Saqib Mumtaz, M Hassan Bholah, P John Hamlin, and Alexander C Ford. Negative Effects on Psychological Health and Quality of Life of Genuine Irritable Bowel Syndrome-type Symptoms in Patients With Inflammatory Bowel Disease. *Clinical Gastroenterology and Hepatology*, 15(3):376–384.e5, March 2017. ISSN 1542-3565, 1542-7714. doi: 10.1016/j.cgh.2016.05.012. URL <https://doi.org/10.1016/j.cgh.2016.05.012>.
- [122] Jacob T Bjerrum, Ole H Nielsen, Fuhua Hao, Huiru Tang, Jeremy K Nicholson, Yulan Wang, and Jørgen Olsen. Metabonomics in ulcerative colitis: diagnostics, biomarker identification, and insight into the pathophysiology. *Journal of Proteome Research*, 9(2):954–962, February 2010. ISSN 1535-3893, 1535-3907. doi: 10.1021/pr9008223. URL <https://doi.org/10.1021/pr9008223>.
- [123] Ying Zhang, Lianjie Lin, Yanbin Xu, Yan Lin, Yu Jin, and Changqing Zheng. 1H NMR-based spectroscopy detects metabolic alterations in serum of patients with early-stage ulcerative colitis. *Biochemical and Biophysical Research Communications*, 433(4):547–551, April 2013. ISSN 0006-291X, 1090-2104. doi: 10.1016/j.bbrc.2013.03.012. URL <http://dx.doi.org/10.1016/j.bbrc.2013.03.012>.
- [124] Sabina Leonelli. Bio-ontologies as Tools for Integration in Biology. *Biological Theory*, 3(1):7–11, March 2008. ISSN 1555-5542, 1555-5550. doi: 10.1162/biot.2008.3.1.7. URL <https://doi.org/10.1162/biot.2008.3.1.7>.
- [125] Robert Hoehndorf, Paul N Schofield, and Georgios V Gkoutos. The role of ontologies in biological and biomedical research: a functional perspective. *Briefings in Bioinformatics*, 16(6): 1069–1080, November 2015. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bbv011. URL <https://doi.org/10.1093/bib/bbv011>.

- [126] Thomas R Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5):907–928, November 1995. ISSN 1071-5819. doi: 10.1006/ijhc.1995.1081. URL <http://dx.doi.org/10.1006/ijhc.1995.1081>.
- [127] Russell Stuart, Norvig Peter, and Others. *Artificial intelligence: a modern approach*. Prentice Hall, USA, Upper Saddle River, NJ, 2003. URL <https://zoo.cs.yale.edu/classes/cs470/materials/aima2010.pdf>.
- [128] Robert Hoehndorf, Michel Dumontier, and Georgios V Gkoutos. Evaluation of research in biomedical ontologies. *Briefings in Bioinformatics*, 14(6):696–712, November 2013. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bbs053. URL <http://dx.doi.org/10.1093/bib/bbs053>.
- [129] Natalya F Noy and Deborah L McGuinness. *Ontology Development 101: A Guide to Creating Your First Ontology*. Stanford University, 2001. URL https://protege.stanford.edu/publications/ontology_development/ontology101.pdf.
- [130] Melissa A Haendel, Christopher G Chute, and Peter N Robinson. Classification, Ontology, and Precision Medicine. *The New England Journal of Medicine*, 379(15):1452–1462, October 2018. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMr1615014.
- [131] Panagiotis I Sergouniotis, Emmanuel Maxime, Dorothée Leroux, Annie Olry, Rachel Thompson, Ana Rath, Peter N Robinson, Hélène Dollfus, and ERN-EYE Ontology Study Group. An ontological foundation for ocular phenotypes and rare eye diseases. *Orphanet Journal of Rare Diseases*, 14(1):8, January 2019. ISSN 1750-1172. doi: 10.1186/s13023-018-0980-6.
- [132] Catia Pesquita, Daniel Faria, André O Falcão, Phillip Lord, and Francisco M Couto. Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7):e1000443, July 2009. ISSN 1553-734X, 1553-7358. doi: 10.1371/journal.pcbi.1000443. URL <http://dx.doi.org/10.1371/journal.pcbi.1000443>.
- [133] Euclid. *The Thirteen Books of Euclid's Elements*. Courier Corporation, January 1956. ISBN 9780486600888.
- [134] Georgios V Gkoutos, Paul N Schofield, and Robert Hoehndorf. The anatomy of phenotype ontologies: principles, properties and applications. *Briefings in Bioinformatics*, 19(5):1008–1021, September 2018. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bbx035. URL <https://doi.org/10.1093/bib/bbx035>.
- [135] Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. Onto2Vec: joint vector-based representation of biological entities and their ontology-based annotations. *Bioinformatics*, 34(13):i52–i60, July 2018. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/bty259. URL <http://dx.doi.org/10.1093/bioinformatics/bty259>.
- [136] Sebastian Köhler, Marcel H Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N Robinson. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *American Journal of Human Genetics*, 85(4):457–464, October 2009. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2009.09.003. URL <http://dx.doi.org/10.1016/j.ajhg.2009.09.003>.
- [137] Robert Hoehndorf, Paul N Schofield, and Georgios V Gkoutos. PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, 39(18):e119, October 2011. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkr538. URL <http://dx.doi.org/10.1093/nar/gkr538>.
- [138] Xuezhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. Human symptoms-disease network. *Nature Communications*, 5:4212, June 2014. ISSN 2041-1723. doi: 10.1038/ncomms5212. URL <https://doi.org/10.1038/ncomms5212>.
- [139] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, May 2012. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg3208. URL <https://doi.org/10.1038/nrg3208>.
- [140] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, May 2000. ISSN 1061-4036. doi: 10.1038/75556. URL <http://dx.doi.org/10.1038/75556>.
- [141] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40(Database issue):D940–6, January 2012. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkr972. URL <http://dx.doi.org/10.1093/nar/gkr972>.
- [142] Christopher J Mungall, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, and Melissa A Haendel. Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, 13(1):R5, January 2012. ISSN 1465-6906. doi: 10.1186/gb-2012-13-1-r5. URL <http://dx.doi.org/10.1186/gb-2012-13-1-r5>.
- [143] Drashti Vasant, Laetitia Chanas, James Malone, Marc Hanauer, Annie Olry, Simon Jupp, Peter N Robinson, Helen Parkinson, and Ana Rath. Ordo: An ontology connecting rare disease, epidemiology and genetic data. In *Proceedings of ISMB*, volume 30. researchgate.net, 2014.
- [144] S S Weinreich, R Mangon, J J Sikkens, M E en Teeuw, and M C Cornel. Orphanet: a European database for rare diseases. *Nederlands Tijdschrift voor Geneeskunde*, 152(9):518–519, March 2008. ISSN 0028-2162, 1876-8784. URL <https://www.ncbi.nlm.nih.gov/pubmed/18389888>.
- [145] Richard G Côté, Philip Jones, Rolf Apweiler, and Henning Hermjakob. The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, 7:97, February 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-97. URL <http://www.ebi.ac.uk/ontology-lookup/>.
- [146] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, OBI Consortium, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, November 2007. ISSN 1087-0156. doi: 10.1038/nbt1346. URL <https://doi.org/10.1038/nbt1346>.
- [147] Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne Storey, Christopher G Chute, and Mark A Musen. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37(Web Server issue):W170–3, July 2009. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkp440. URL <http://dx.doi.org/10.1093/nar/gkp440>.
- [148] Manuel Salvadores, Paul R Alexander, Mark A Musen, and Natalya F Noy. BioPortal as a Dataset of Linked Biomedical Ontologies and Terminologies in RDF. *Semantic Web*, 4(3):277–284, 2013. ISSN 1570-0844. URL <https://www.ncbi.nlm.nih.gov/pubmed/25214827>.
- [149] Christopher Ochs, Yehoshua Perl, James Geller, Sivaram Arabandi, Tania Tudorache, and Mark A Musen. An empirical analysis of ontology reuse in BioPortal. *Journal of Biomedical Informatics*, 71:165–177, July 2017. ISSN 1532-0464, 1532-0480. doi: 10.1016/j.jbi.2017.05.021. URL <http://dx.doi.org/10.1016/j.jbi.2017.05.021>.
- [150] Robert Hoehndorf, Luke Slater, Paul N Schofield, and Georgios V Gkoutos. Aber-OWL: a framework for ontology-based data access in biology. *BMC Bioinformatics*, 16:26, January 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0456-9. URL <http://dx.doi.org/10.1186/s12859-015-0456-9>.
- [151] Pieter Kubben, Michel Dumontier, and Andre Dekker. *Fundamentals of Clinical Data Science*. SpringerOpen, 2019. ISBN 9783319997124. doi: 10.1007/978-3-319-99713-1.
- [152] Charles H Calisher. Taxonomy: what's in a name? Doesn't a rose by any other name smell as sweet? *Croatian Medical Journal*, 48(2):268–270, April 2007. ISSN 0353-9504, 1332-8166. URL <https://www.ncbi.nlm.nih.gov/pubmed/17436393>.
- [153] World Health Organization (WHO). International Classification of Diseases (ICD). <http://www.who.int/classifications/icd/en/>, 2006. URL <http://www.who.int/classifications/icd/en/>. Accessed: 2020-1-6.

- [154] J D Read and T J R Benson. Comprehensive coding. *British Journal of Healthcare Computing*, 3:22–25, 1986. URL https://www.researchgate.net/profile/Tim-Benson/publication/284742447_Comprehensive_Coding/links/5666af0808ae192bbf928c09/Comprehensive-Coding.pdf.
- [155] Tim Benson. The history of the Read Codes: the inaugural James Read Memorial Lecture 2011. *Informatics in Primary Care*, 19(3):173–182, 2011. ISSN 1476-0320, 1475-9985. doi: 10.14236/jhi.v19i3.811. URL <http://dx.doi.org/10.14236/jhi.v19i3.811>.
- [156] International Health Terminology Standards Development Organisation (IHTSDO). SNOMED. <https://www.snomed.org/>, 2007. URL <https://www.snomed.org/>. Accessed: 2021-9-1.
- [157] Bakheet Aldosari, Abdullah Alanazi, and Mowafa Househ. Pitfalls of Ontology in Medicine. *Studies in Health Technology and Informatics*, 238:15–18, 2017. ISSN 0926-9630, 1879-8365. URL <https://www.ncbi.nlm.nih.gov/pubmed/28679876>.
- [158] D E Oliver, Y Shahar, E H Shortliffe, and M A Musen. Representation of change in controlled medical terminologies. *Artificial Intelligence in Medicine*, 15(1):53–76, January 1999. ISSN 0933-3657. doi: 10.1016/s0933-3657(98)00045-1. URL [http://dx.doi.org/10.1016/s0933-3657\(98\)00045-1](http://dx.doi.org/10.1016/s0933-3657(98)00045-1).
- [159] Arrigo Schieppati, Jan-Inge Henter, Erica Daina, and Anita Aperia. Why rare diseases are an important medical and social issue. *Lancet*, 371(9629):2039–2041, June 2008. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(08)60872-7. URL [https://doi.org/10.1016/S0140-6736\(08\)60872-7](https://doi.org/10.1016/S0140-6736(08)60872-7).
- [160] Alan G Palestine, Pauline T Merrill, Sophia M Saleem, Douglas A Jabs, and Jennifer E Thorne. Assessing the Precision of ICD-10 Codes for Uveitis in 2 Electronic Health Record Systems. *The Journal of the American Medical Association: Ophthalmology*, 136(10):1186–1190, October 2018. ISSN 2168-6165, 2168-6173. doi: 10.1001/jamaophthalmol.2018.3001. URL <http://dx.doi.org/10.1001/jamaophthalmol.2018.3001>.
- [161] Ling Zheng, Yehoshua Perl, Michael Halper, Sherri de Coronado, Christopher Ochs, and Hua Min. Relating Complexity and Error Rates of Ontology Concepts. *Methods of Information in Medicine*, 56(03):200–208, 2017. ISSN 0026-1270. doi: 10.3414/me16-01-0085. URL <http://dx.doi.org/10.3414/me16-01-0085>.
- [162] Steven J Steindel. A comparison between a SNOMED CT problem list and the ICD-10-CM/PCS HIPAA code sets. *Perspectives in Health Information Management*, 9:1b, January 2012. ISSN 1559-4122. URL <https://www.ncbi.nlm.nih.gov/pubmed/22548020>.
- [163] Simon de Lusignan, Tom Chan, and Simon Jones. Large complex terminologies: more coding choice, but harder to find data-reflections on introduction of SNOMED CT (Systematized Nomenclature of Medicine–Clinical Terms) as an NHS standard. *Informatics in Primary Care*, 19(1):3–5, 2011. ISSN 1476-0320, 1475-9985. doi: 10.14236/jhi.v19i1.787. URL <http://dx.doi.org/10.14236/jhi.v19i1.787>.
- [164] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–70, January 2004. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkh061. URL <http://dx.doi.org/10.1093/nar/gkh061>.
- [165] Abul Hasan, Mark Levene, and David J Weston. Natural Language Analysis of Online Health Forums. In *Advances in Intelligent Data Analysis*, pages 125–137. Springer International Publishing, 2017. doi: 10.1007/978-3-319-68765-0_11. URL http://dx.doi.org/10.1007/978-3-319-68765-0_11.
- [166] Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2, 3rd Edition*. Packt Publishing, Limited, December 2019. ISBN 9781789955750. URL <https://github.com/rasbt/python-machine-learning-book-3rd-edition>.
- [167] M Namratha and T R Prajwala. A comprehensive overview of clustering algorithms in pattern recognition. *IOSR Journal of Computer Engineering*, 4(6):23–30, 2012. URL <https://pdfs.semanticscholar.org/b0a2/0ecaf9398990bf552c0bd02e1799235a986f.pdf>.
- [168] Laura M Stevens, Bobak J Mortazavi, Rahul C Deo, Lesley Curtis, and David P Kao. Recommendations for Reporting Machine Learning Analyses in Clinical Research. *Circulation. Cardiovascular Quality and Outcomes*, 13(10):e006556, October 2020. ISSN 1941-7713, 1941-7705. doi: 10.1161/CIRCOUTCOMES.120.006556. URL <https://doi.org/10.1161/CIRCOUTCOMES.120.006556>.
- [169] Markus Svensén and Christopher M Bishop. Linear Models for Regression. In *Pattern Recognition and Machine Learning*. Springer, 2007. URL <https://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/bishop-regression.pdf>.
- [170] Sukhdev Singh Ghuman. Clustering Techniques - A Review. *International Journal of Computer Science and Mobile Computing*, 5(5):524–530, May 2016. ISSN 2320-088X. URL <https://pdfs.semanticscholar.org/dd31/3f3d07fc2672c3f184040b9a5a2ce563a173.pdf>.
- [171] Ian Jolliffe. *Principal Component Analysis*. John Wiley & Sons, Ltd, Chichester, UK, October 2005. ISBN 9780470860809, 9780470860809. doi: 10.1002/0470013192.bsa501. URL <http://dx.doi.org/10.1002/0470013192.bsa501>.
- [172] Tom Ronan, Zhijie Qi, and Kristen M Naegle. Avoiding common pitfalls when clustering biological data. *Science Signaling*, 9(432):re6, June 2016. ISSN 1937-9145, 1945-0877. doi: 10.1126/scisignal.aad1932. URL <http://dx.doi.org/10.1126/scisignal.aad1932>.
- [173] Sam Newman. *Building Microservices*. O'Reilly Media, Inc., second edition, July 2021. ISBN 9781492033974.
- [174] Peter J Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, November 1987. ISSN 0377-0427. doi: 10.1016/0377-0427(87)90125-7. URL [http://dx.doi.org/10.1016/0377-0427\(87\)90125-7](http://dx.doi.org/10.1016/0377-0427(87)90125-7).
- [175] William M Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, December 1971. ISSN 0162-1459. doi: 10.1080/01621459.1971.10482356. URL <http://dx.doi.org/10.1080/01621459.1971.10482356>.
- [176] Janus Christian Jakobsen, Christian Gluud, Jørn Wetterslev, and Per Winkel. When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. *BMC Medical Research Methodology*, 17(1):162, December 2017. ISSN 1471-2288. doi: 10.1186/s12874-017-0442-1. URL <http://dx.doi.org/10.1186/s12874-017-0442-1>.
- [177] Anwar Mulugeta, Elina Hyppönen, Mika Ala-Korpela, and Ville-Petteri Mäkinen. Metabolic subgroups and cardiometabolic multimorbidity in the UK Biobank. *medRxiv*, February 2021. doi: 10.1101/2021.02.01.21250893. URL <http://dx.doi.org/10.1101/2021.02.01.21250893>.
- [178] R B Cattell. The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, 1(2):245–276, April 1966. ISSN 0027-3171. doi: 10.1207/s15327906mbr0102_10. URL http://dx.doi.org/10.1207/s15327906mbr0102_10.
- [179] Henry F Kaiser. The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, 20(1):141–151, April 1960. ISSN 0013-1644. doi: 10.1177/001316446002000116. URL <https://doi.org/10.1177/001316446002000116>.
- [180] Hobson Lane, Hannes Hapke, and Cole Howard. *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Manning Publications, April 2019. ISBN 9781617294631. URL <https://livebook.manning.com/book/natural-language-processing-in-action/>.
- [181] Tomáš Mikolov, Wen-Tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Human Language Technologies*, pages 746–751, 2013. URL <https://aclanthology.org/N13-1090>.
- [182] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv*, January 2013. URL <http://arxiv.org/abs/1301.3781>.

- [183] Sumithra Velupillai, Hanna Suominen, Maria Liakata, Angus Roberts, Anoop D Shah, Katherine Morley, David Osborn, Joseph Hayes, Robert Stewart, Johnny Downs, Wendy Chapman, and Rina Dutta. Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. *Journal of Biomedical Informatics*, 88:11–19, December 2018. ISSN 1532-0464, 1532-0480. doi: 10.1016/j.jbi.2018.10.005. URL <http://dx.doi.org/10.1016/j.jbi.2018.10.005>.
- [184] S Trent Rosenbloom, Joshua C Denny, Hua Xu, Nancy Lorenzi, William W Stead, and Kevin B Johnson. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *JAMIA*, 18(2):181–186, March 2011. ISSN 1067-5027, 1527-974X. doi: 10.1136/jamia.2010.007237. URL <http://dx.doi.org/10.1136/jamia.2010.007237>.
- [185] Richard G Jackson, Rashmi Patel, Nishamali Jayatilleke, Anna Kolliakou, Michael Ball, Genevieve Gorrell, Angus Roberts, Richard J Dobson, and Robert Stewart. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open*, 7(1):e012012, January 2017. ISSN 2044-6055. doi: 10.1136/bmjopen-2016-012012. URL <http://dx.doi.org/10.1136/bmjopen-2016-012012>.
- [186] Jinying Chen, Emily Druhl, Balaji Polepalli Ramesh, Thomas K Houston, Cynthia A Brandt, Donna M Zulman, Varsha G Vimalananda, Samir Malkani, and Hong Yu. A Natural Language Processing System That Links Medical Terms in Electronic Health Record Notes to Lay Definitions: System Development Using Physician Reviews. *Journal of Medical Internet Research*, 20(1):e26, January 2018. ISSN 1439-4456, 1438-8871. doi: 10.2196/jmir.8669. URL <http://dx.doi.org/10.2196/jmir.8669>.
- [187] Richard Jackson, Ismail Kartoglu, Clive Stringer, Genevieve Gorrell, Angus Roberts, Xingyi Song, Honghan Wu, Asha Agrawal, Kenneth Lui, Tudor Groza, Damian Lewsley, Doug Northwood, Amos Folarin, Robert Stewart, and Richard Dobson. CogStack - experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust hospital. *BMC Medical Informatics and Decision Making*, 18(1):47, June 2018. ISSN 1472-6947. doi: 10.1186/s12911-018-0623-9. URL <https://doi.org/10.1186/s12911-018-0623-9>.
- [188] Hans-Michael Müller, Eimear E Kenny, and Paul W Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS biology*, 2(11):e309, November 2004. ISSN 1544-9173, 1545-7885. doi: 10.1371/journal.pbio.0020309. URL <http://dx.doi.org/10.1371/journal.pbio.0020309>.
- [189] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. BRAT. In *Demonstrations at the 13th Conference of the EACL*, pages 102–107. Association for Computational Linguistics, 2012. URL <https://dl.acm.org/doi/proceedings/10.5555/2380921>.
- [190] Sheng Yu, Tianrun Cai, and Tianxi Cai. NILE: Fast Natural Language Processing for Electronic Health Records. *arXiv*, November 2013. URL <http://arxiv.org/abs/1311.6063>.
- [191] Juan Miguel Cejuela, Peter McQuilton, Laura Ponting, Steven J Marygold, Raymund Stefancsik, Gillian H Millburn, Burkhard Rost, and FlyBase Consortium. tagto: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. *Database: the journal of biological databases and curation*, 2014(0):bau033, April 2014. ISSN 1758-0463. doi: 10.1093/database/bau033. URL <http://dx.doi.org/10.1093/database/bau033>.
- [192] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *System Demonstrations*, pages 55–60, 2014. URL <https://www.aclweb.org/anthology/P14-5010>.
- [193] Philipp Burckhardt and Rema Padman. Analyzing Self-Help Forums with Ontology-Based Text Mining: An Exploration in Kidney Space. In *AMIA Annual Symposium Proceedings*, volume 2015, pages 1821–1830, November 2015.
- [194] George Gkotsis, Anika Oellrich, Tim J Hubbard, and Rina Dutta. The language of mental health problems in social media. In *From Linguistic Signal to Clinical Reality*, pages 63–73. Association for Computational Linguistics, June 2016. doi: 10.18653/v1/W16-0307. URL <http://dx.doi.org/10.18653/v1/W16-0307>.
- [195] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim J P Hubbard, Richard J B Dobson, and Rina Dutta. Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific Reports*, 7: 45141, March 2017. ISSN 2045-2322. doi: 10.1038/srep45141. URL <https://doi.org/10.1038/srep45141>.
- [196] Sunghoon Lim, Conrad S Tucker, and Soundar Kumara. An unsupervised machine learning model for discovering latent infectious diseases using social media data. *Journal of Biomedical Informatics*, 66:82–94, February 2017. ISSN 1532-0464, 1532-0480. doi: 10.1016/j.jbi.2016.12.007. URL <http://dx.doi.org/10.1016/j.jbi.2016.12.007>.
- [197] Ling Zhang, Magie Hall, and Dhundy Bastola. Utilizing Twitter data for analysis of chemotherapy. *International Journal of Medical Informatics*, 120:92–100, December 2018. ISSN 1386-5056, 1872-8243. doi: 10.1016/j.ijmedinf.2018.10.002. URL <http://dx.doi.org/10.1016/j.ijmedinf.2018.10.002>.
- [198] Philip Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *arXiv*, November 1995. URL <http://arxiv.org/abs/cmp-1g/9511007>.
- [199] P Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11:95–130, July 1999. ISSN 1076-9757, 1076-9757. doi: 10.1613/jair.514. URL <http://dx.doi.org/10.1613/jair.514>.
- [200] Ted Pedersen, Serguei V S Pakhomov, Siddharth Patwardhan, and Christopher G Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299, June 2007. ISSN 1532-0464, 1532-0480. doi: 10.1016/j.jbi.2006.06.004. URL <http://dx.doi.org/10.1016/j.jbi.2006.06.004>.
- [201] Faiza Khan Khattak, Serena Jebblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, and Frank Rudzicz. A survey of word embeddings for clinical text. *Journal of Biomedical Informatics*, X, 4:100057, December 2019. ISSN 2590-177X. doi: 10.1016/j.yjbix.2019.100057. URL <http://dx.doi.org/10.1016/j.yjbix.2019.100057>.
- [202] M Nickel, K Murphy, V Tresp, and E Gabrilovich. A Review of Relational Machine Learning for Knowledge Graphs. In *Proceedings of the IEEE*, volume 104, pages 11–33. IEEE, January 2016. doi: 10.1109/JPROC.2015.2483592. URL <http://dx.doi.org/10.1109/JPROC.2015.2483592>.
- [203] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986. ISSN 0028-0836. doi: 10.1038/323533a0. URL <http://dx.doi.org/10.1038/323533a0>.
- [204] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003. ISSN 1532-4435. URL <https://www.jmlr.org/papers/volume3/tmp/bengio03a.pdf>.
- [205] Li Zhang, Jun Li, and Chao Wang. Automatic synonym extraction using Word2Vec and spectral clustering. In *2017 36th Chinese Control Conference (CCC)*, pages 5629–5632, July 2017. doi: 10.23919/ChiCC.2017.8028251. URL <http://dx.doi.org/10.23919/ChiCC.2017.8028251>.
- [206] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv*, April 2019. URL <http://arxiv.org/abs/1904.05342>.
- [207] Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'Avolio, Guergana K Savova, and Ozlem Uzuner. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5): 540–543, September 2011. ISSN 1067-5027, 1527-974X. doi: 10.1136/amiajnl-2011-000465. URL <http://dx.doi.org/10.1136/amiajnl-2011-000465>.
- [208] Yinyuan Zhang, Ricardo Henao, Zhe Gan, Yitong Li, and Lawrence Carin. Multi-Label Learning from Medical Plain Text with Convolutional Residual Models. In *Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pages 280–294. PMLR, 2018. URL <https://proceedings.mlr.press/v85/zhang18a.html>.

- [209] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 3111–3119. Curran Associates, Inc., 2013. URL <https://arxiv.org/abs/1310.4546>.
- [210] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv*, April 2020. URL <http://arxiv.org/abs/2004.10964>.
- [211] Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196. JMLR, January 2014. URL <http://www.jmlr.org/proceedings/papers/v32/le14.html>.
- [212] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv*, 5:135–146, 2017. doi: 10.48550/arxiv.1607.04606. URL <https://doi.org/10.48550/arxiv.1607.04606>.
- [213] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*, October 2018. URL <http://arxiv.org/abs/1810.04805>.
- [214] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the Dark Secrets of BERT. In *EMNLP-IJCNLP*, pages 4365–4374, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1445. URL <http://dx.doi.org/10.18653/v1/D19-1445>.
- [215] Laura Kinkade, Ahmed Allam, and Michael Krauthammer. AutoDiscern: rating the quality of online health information with hierarchical encoder attention-based neural networks. *BMC Medical Informatics and Decision Making*, 20(1):104, June 2020. ISSN 1472-6947. doi: 10.1186/s12911-020-01131-z. URL <https://doi.org/10.1186/s12911-020-01131-z>.
- [216] D Charnock, S Shepperd, G Needham, and R Gann. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *Journal of Epidemiology and Community Health*, 53(2):105–111, February 1999. ISSN 0141-7681, 0143-005X. doi: 10.1136/jech.53.2.105. URL <http://dx.doi.org/10.1136/jech.53.2.105>.
- [217] Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. What's in a Note? Unpacking Predictive Value in Clinical Note Representations. *AMIA Joint Summits on Translational Science*, 2017:26–34, May 2018. ISSN 2153-4063. URL <https://www.ncbi.nlm.nih.gov/pubmed/29888035>.
- [218] L W C Chan, T Chan, L F Cheng, and W S Mak. Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy. In *Bioinformatics and Biomedicine Workshops*, pages 467–470. IEEE, December 2010. doi: 10.1109/BIBMW.2010.5703846. URL <http://dx.doi.org/10.1109/BIBMW.2010.5703846>.
- [219] 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. *Scientific Reports*, 6:26094, May 2016. ISSN 2045-2322. doi: 10.1038/srep26094. URL <https://doi.org/10.1038/srep26094>.
- [220] Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, May 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL <https://doi.org/10.1038/sdata.2016.35>.
- [221] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv*, March 2019. URL <http://arxiv.org/abs/1903.10676>.
- [222] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B A McDermott. Publicly Available Clinical BERT Embeddings. *arXiv*, April 2019. URL <http://arxiv.org/abs/1904.03323>.
- [223] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, February 2020. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btz682. URL <http://dx.doi.org/10.1093/bioinformatics/btz682>.
- [224] Ilya Valmianski, Caleb Goodwin, Ian M Finn, Naqi Khan, and Daniel S Zisook. Evaluating robustness of language models for chief complaint extraction from patient-generated text. *arXiv*, November 2019. URL <http://arxiv.org/abs/1911.06915>.
- [225] Marco Basaldella and Nigel Collier. BioReddit: Word Embeddings for User-Generated Biomedical NLP. In *Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis*, pages 34–38, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6205. URL <http://dx.doi.org/10.18653/v1/D19-6205>.
- [226] Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. COMETA: A Corpus for Medical Entity Linking in the Social Media. In *EMNLP*, pages 3122–3137, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.253. URL <http://dx.doi.org/10.18653/v1/2020.emnlp-main.253>.
- [227] Tariq Ahmad, Michael J Pencina, Phillip J Schulte, Emily O'Brien, David J Whellan, Ileana L Piña, Dalane W Kitzman, Kerry L Lee, Christopher M O'Connor, and G Michael Felker. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. *Journal of the American College of Cardiology*, 64(17):1765–1774, October 2014. ISSN 0735-1097, 1558-3597. doi: 10.1016/j.jacc.2014.07.979. URL <http://dx.doi.org/10.1016/j.jacc.2014.07.979>.
- [228] J Tromp, W Ouwkerk, B G Demissei, S D Anker, J G Cleland, K Dickstein, G Filippatos, P van der Harst, H L Hillege, C C Lang, M Metra, L L Ng, P Ponikowski, N J Samani, D J van Veldhuisen, F Zannad, A H Zwiderman, A A Voors, and P van der Meer. Novel endotypes in heart failure: effects on guideline-directed medical therapy. *European Heart Journal*, 39(48):4269–4276, December 2018. ISSN 0195-668X, 1522-9645. doi: 10.1093/eurheartj/ehy712. URL <https://pubmed.ncbi.nlm.nih.gov/30651207/>.
- [229] Yiwang Zhou, Lu Zhao, Nina Zhou, Yi Zhao, Simeone Marino, Tuo Wang, Hanbo Sun, Arthur W Toga, and Ivo D Dinov. Predictive Big Data Analytics using the UK Biobank Data. *Scientific Reports*, 9(1):6012, April 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-41634-y. URL <https://doi.org/10.1038/s41598-019-41634-y>.
- [230] James J Ashton, Florina Borca, Enrico Mossotto, Hang T T Phan, Sarah Ennis, and R Mark Beattie. Analysis and Hierarchical Clustering of Blood Results Before Diagnosis in Pediatric Inflammatory Bowel Disease. *Inflammatory Bowel Diseases*, 26(3):469–475, February 2020. ISSN 1078-0998, 1536-4844. doi: 10.1093/ibd/izy369. URL <http://dx.doi.org/10.1093/ibd/izy369>.
- [231] Maida J Sewitch, Karen Leffondré, and Patricia L Dobkin. Clustering patients according to health perceptions: relationships to psychosocial characteristics and medication nonadherence. *Journal of Psychosomatic Research*, 56(3):323–332, March 2004. ISSN 0022-3999. doi: 10.1016/S0022-3999(03)00508-7. URL [http://dx.doi.org/10.1016/S0022-3999\(03\)00508-7](http://dx.doi.org/10.1016/S0022-3999(03)00508-7).
- [232] C E Welsh, Paul Welsh, Pardeep Jhund, Christian Delles, C Celis-Morales, J D Lewsey, S Gray, D Lyall, S Iliodromiti, J M R Gill, Naveed Sattar, and Patrick B Mark. Urinary Sodium Excretion, Blood Pressure, and Risk of Future Cardiovascular Disease and Mortality in Subjects Without Prior Cardiovascular Disease. *Hypertension*, 73(6):1202–1209, June 2019. ISSN 0194-911X, 1524-4563. doi: 10.1161/HYPERTENSIONAHA.119.12726. URL <https://doi.org/10.1161/HYPERTENSIONAHA.119.12726>.
- [233] S Velupillai, D Mowery, B R South, M Kvist, and H Dalanis. Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis. *IMIA Yearbook of Medical Informatics*, 10(1):183–193, August 2015. ISSN 0943-4747, 2364-0502. doi: 10.15265/IY-2015-009. URL <http://dx.doi.org/10.15265/IY-2015-009>.
- [234] Peter D Turney. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Machine Learning*, pages 491–502. Springer Berlin Heidelberg, 2001. doi: 10.1007/3-540-44795-4_42. URL http://dx.doi.org/10.1007/3-540-44795-4_42.
- [235] Long Ma and Yanqing Zhang. Using Word2Vec to process big text data. In *2015 IEEE International Conference on Big Data*, pages 2895–2897, October 2015. doi: 10.1109/BigData.2015.7364114. URL <http://dx.doi.org/10.1109/BigData.2015.7364114>.

- [236] Bai Xue, Chen Fu, and Zhan Shaobin. A New Clustering Model Based on Word2vec Mining on Sina Weibo Users' Tags. *International Journal of Grid Distribution Computing*, 7(3):41–48, 2014. doi: 10.14257/ijgcd.2014.7.3.05. URL <http://dx.doi.org/10.14257/ijgcd.2014.7.3.05>.
- [237] Rong Xu, Li Li, and Quanqiu Wang. Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature. *Bioinformatics*, 29(17):2186–2194, September 2013. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btt359. URL <http://dx.doi.org/10.1093/bioinformatics/btt359>.
- [238] Xuezhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. Human symptoms–disease network. *Nature Communications*, 5(1):1–10, June 2014. ISSN 2041-1723, 2041-1723. doi: 10.1038/ncomms5212. URL <http://dx.doi.org/10.1038/ncomms5212>.
- [239] John A Barry, Annie Folkard, Alastair K Denniston, Eileen Moran, and William Ayliffe. Development and validation of quality-of-life questionnaires for birdshot chorioretinopathy. *Ophthalmology*, 121(7):1488–9.e2, July 2014. ISSN 0161-6420, 1549-4713. doi: 10.1016/j.opthta.2014.01.007. URL <http://dx.doi.org/10.1016/j.opthta.2014.01.007>.
- [240] William Ollier, Tim Sprosen, and Tim Peakman. UK Biobank: from concept to reality. *Pharmacogenomics*, 6(6):639–646, September 2005. ISSN 1462-2416. doi: 10.2217/14622416.6.6.639. URL <http://dx.doi.org/10.2217/14622416.6.6.639>.
- [241] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, March 2015. ISSN 1549-1277, 1549-1676. doi: 10.1371/journal.pmed.1001779. URL <http://dx.doi.org/10.1371/journal.pmed.1001779>.
- [242] Samantha C Pendleton, Luke T Slater, Andreas Karwath, Rose M Gilbert, Nicola Davis, Konrad Pesudovs, Xiaoxuan Liu, Alastair K Denniston, Georgios V Gkoutos, and Tasanee Braithwaite. Development and application of the ocular immune-mediated inflammatory diseases ontology enhanced with synonyms from online patient support forum conversation. *Computers in Biology and Medicine*, 135:104542, August 2021. ISSN 0010-4825, 1879-0534. doi: 10.1016/j.combiomed.2021.104542. URL <http://dx.doi.org/10.1016/j.combiomed.2021.104542>.
- [243] Alastair K Denniston, Richard W Lee, Carlos Pavesio, Miles R Stanford, Philip I Murray, Annabelle Okada, H Nida Sen, and Andrew D Dick. Uveitis Dataset Clinical Dataset. Technical report, The Royal College of Ophthalmologists, October 2018. URL <https://www.rcophth.ac.uk/wp-content/uploads/2014/12/Uveitis-Data-Set-Oct-2018.pdf>.
- [244] Mark A Musen and Protégé Team. The Protégé Project: A Look Back and a Look Forward. *AI Matters*, 1(4):4–12, June 2015. ISSN 2372-3483. doi: 10.1145/2757001.2757003. URL <http://dx.doi.org/10.1145/2757001.2757003>.
- [245] Barry Smith, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan L Rector, and Cornelius Rosse. Relations in biomedical ontologies. *Genome Biology*, 6(5):R46, April 2005. ISSN 1465-6906. doi: 10.1186/gb-2005-6-5-r46. URL <https://doi.org/10.1186/gb-2005-6-5-r46>.
- [246] Barry Smith, Anand Kumar, and Thomas Bittner. Basic Formal Ontology for Bioinformatics. Technical report, IFOMIS Reports, 2005. URL <https://philpapers.org/rec/KUMIR>.
- [247] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., June 2009. ISBN 9780596555719.
- [248] G van Rossum. Python tutorial, technical report CS-R9526. Technical report, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, 1995.
- [249] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011. ISSN 1532-4435, 1533-7928. URL <http://www.jmlr.org/papers/v12/pedregosa11a>.
- [250] Steven Loria. *textblob Documentation*. Technical report, 2018. URL <https://media.readthedocs.org/pdf/textblob/latest/textblob.pdf>.
- [251] C Hutto and Eric Gilbert. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Web and Social Media*, volume 8, pages 216–225, May 2014. URL <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>.
- [252] Thomas Davidson, Dana Warmley, Michael Macy, and Ingmar Weber. Automated Hate Speech Detection and the Problem of Offensive Language. In *Web and Social Media*, volume 11. AAAI, May 2017. URL <https://ojs.aaai.org/index.php/ICWSM/article/view/14955>.
- [253] Venkateswarlu Bonta and Nandhini Kumareshtand N Janardhan. A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. *Asian Journal of Computer Science and Technology*, 8(S2):1–6, 2019. URL <https://www.researchgate.net/publication/333602124>.
- [254] StataCorp. Stata Statistical Software, 2013.
- [255] Nicolas Matentzoglou, James Malone, Chris Mungall, and Robert Stevens. MIRO: guidelines for minimum information for the reporting of an ontology. *Journal of Biomedical Semantics*, 9(1):6, January 2018. ISSN 2041-1480. doi: 10.1186/s13326-017-0172-7. URL <https://doi.org/10.1186/s13326-017-0172-7>.
- [256] Nicola Mulder, Ambroise Wonkam, and Sickle Cell Disease Ontology Working Group. Development and Implementation of the Sickle Cell Disease Ontology. *Preprints with The Lancet*, February 2019. doi: 10.2139/ssrn.3294100. URL <http://dx.doi.org/10.2139/ssrn.3294100>.
- [257] Steffen Lohmann, Vincent Link, Eduard Marbach, and Stefan Negru. WebVOWL: Web-based Visualization of Ontologies. In *Knowledge Engineering and Knowledge Management*, pages 154–158. Springer International Publishing, 2015. doi: 10.1007/978-3-319-17966-7_21. URL http://dx.doi.org/10.1007/978-3-319-17966-7_21.
- [258] Holger Stenzhorn, Stefan Schulz, Martin Boeker, and Barry Smith. Adapting Clinical Ontologies in Real-World Environments. *Journal of Universal Computer Science*, 14(22):3767–3780, 2008. ISSN 0948-695X, 0948-6968. doi: 10.3217/jucs-014-22-3767. URL <http://dx.doi.org/10.3217/jucs-014-22-3767>.
- [259] Hua Min, Hedyeh Mobahi, Katherine Irvin, Sanja Avramovic, and Janusz Wojtusiak. Predicting activities of daily living for cancer patients using an ontology-guided machine learning methodology. *Journal of Biomedical Semantics*, 8(1):39, September 2017. ISSN 2041-1480. doi: 10.1186/s13326-017-0149-6. URL <https://doi.org/10.1186/s13326-017-0149-6>.
- [260] Samantha Pendleton and Georgios Gkoutos. Jabberwocky: an ontology-aware toolkit for manipulating text. *Journal of Open Source Software*, 5(51):2168, July 2020. ISSN 2475-9066. doi: 10.21105/joss.02168. URL <http://dx.doi.org/10.21105/joss.02168>.
- [261] Ines Montani, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O'leary McCann, Maxim Samsonov, Jim Gevedi, Jim O'Regan, Duygu Altinok, György Orosz, Søren Lind Kristiansen, Roman, Explosion Bot, Leander Fiedler, Grégory Howard, Daniël de Kok, Wannaphong Phatthiyaphaibun, Lj Miranda, Yohei Tamura, Sam Bozek, murat, Edward, Mark Amery, Ryn Daniels, Björn Böing, Pradeep Kumar Tippa, and Leif Uwe Vogelsang. spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL <https://doi.org/10.5281/zenodo.1212303>.
- [262] Samantha C Pendleton. Combined Ontology for Inflammatory Diseases COID, September 2021. URL <https://doi.org/10.5281/zenodo.5524650>.

- [263] Samantha C Pendleton. PatientINF, April 2022. URL <https://huggingface.com/sap218/PatientINF>.
- [264] Aldo Gangemi, Carola Catenacci, and Massimo Battaglia. Inflammation ontology design pattern: an exercise in building a core biomedical ontology with descriptions and situations. *Studies in Health Technology and Informatics*, 102:64–80, 2004. ISSN 0926-9630. URL <https://www.ncbi.nlm.nih.gov/pubmed/15853264>.
- [265] 2021 ICD-10-CM Index > 'Inflammation, inflamed, inflammatory'. https://www.icd10data.com/ICD10CM/Index/I/Inflammation%2C_inflamed%2C_inflammatory, 2021. URL https://www.icd10data.com/ICD10CM/Index/I/Inflammation%2C_inflamed%2C_inflammatory. Accessed: 2020.
- [266] 2021 ICD-10 LOOKUP. <https://icdlist.com/?t=icd10&s=inflammation>, 2021. URL <https://icdlist.com/?t=icd10&s=inflammation>. Accessed: 2020.
- [267] Patient Platform Limited. Patient Community Forums. <https://patient.info/forums>, 1996. URL <https://patient.info/forums>. Accessed: 2020-5-16.
- [268] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9 (11), 2008. ISSN 1532-4435. URL <https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf#fbcl>.
- [269] Zhirong Yang, Yuwei Chen, and Jukka Corander. T-SNE Is Not Optimized to Reveal Clusters in Data. *arXiv*, October 2021. URL <http://arxiv.org/abs/2110.02573>.
- [270] Radim Rehurek and Petr Sojka. *Gensim—python framework for vector space modelling*. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 2011.
- [271] Leonard Richardson. *Beautiful Soup Documentation*, 2019. URL http://mde.tw/wcm2021/downloads/2019_beautifulsoup_document.pdf.
- [272] Nadezda Katricheva, Alyaxey Yaskevich, Anastasiya Lisitsina, Tamara Zhordaniya, Andrey Kutuzov, and Elizaveta Kuzmenko. Vec2graph: A Python Library for Visualizing Word Embeddings as Graphs. In *Analysis of Images, Social Networks and Texts*, pages 190–198. Springer International Publishing, 2020. doi: 10.1007/978-3-030-39575-9_20. URL http://dx.doi.org/10.1007/978-3-030-39575-9_20.
- [273] So Young Kim, Chanyang Min, Dong Jun Oh, and Hyo Geun Choi. Increased risk of appendectomy due to appendicitis after tonsillectomy in women: A longitudinal follow-up study using a national sample cohort. *Medicine*, 98(19):e15579, May 2019. ISSN 0025-7974, 1536-5964. doi: 10.1097/MD.00000000000015579. URL <https://doi.org/10.1097/MD.00000000000015579>.
- [274] J C Andreu Ballester, F Ballester, E Colomer Rubio, and M Millán Scheiding. Association between tonsillectomy, adenoidectomy, and appendicitis. *Revista Espanola de Enfermedades Digestivas*, 97(3):179–186, March 2005. ISSN 1130-0108. doi: 10.4321/s1130-01082005000300006. URL <http://dx.doi.org/10.4321/s1130-01082005000300006>.
- [275] Sang-Bae Ko. Perioperative stroke: pathophysiology and management. *Korean Journal of Anesthesiology*, 71(1):3–11, February 2018. ISSN 2005-6419. doi: 10.4097/kjae.2018.71.1.3. URL <http://dx.doi.org/10.4097/kjae.2018.71.1.3>.
- [276] Céline Kempeneers, Isabelle Paquot, Pierre Philippet, Annieta Goossens, and Vincent Bernier. Sterile folliculitis as an important diagnostic clue to Crohn's disease. *Pediatric Dermatology*, 30(6):e146–8, November 2013. ISSN 0736-8046, 1525-1470. doi: 10.1111/j.1525-1470.2012.01796.x. URL <https://doi.org/10.1111/j.1525-1470.2012.01796.x>.
- [277] P S Moraes. Recurrent vaginal candidiasis and allergic rhinitis: a common association. *American College of Allergy, Asthma, & Immunology*, 81(2):165–169, August 1998. ISSN 1081-1206. doi: 10.1016/S1081-1206(10)62804-9. URL [http://dx.doi.org/10.1016/S1081-1206\(10\)62804-9](http://dx.doi.org/10.1016/S1081-1206(10)62804-9).
- [278] Xiao Chen, Lifeng Chen, Wenjie Zeng, and Xiaofeng Zhao. Haemophilus influenzae vulvovaginitis associated with rhinitis caused by the same clone in a prepubertal girl. *The Journal of Obstetrics and Gynaecology Research*, 43(6):1080–1083, June 2017. ISSN 1341-8076, 1447-0756. doi: 10.1111/jog.13311. URL <https://doi.org/10.1111/jog.13311>.
- [279] Winter S Berry. Otitis, Sinusitis, and Mastoiditis: Ear or Facial Pain Following a Common Cold. *Introduction to Clinical Infectious Diseases*, page 37, 2018. doi: 10.1007/978-3-319-91080-2_4. URL http://dx.doi.org/10.1007/978-3-319-91080-2_4.
- [280] Rana Mady, Will Grover, and Salim Butrus. Ocular complications of inflammatory bowel disease. *The Scientific World Journal*, 2015:438402, March 2015. ISSN 1537-744X. doi: 10.1155/2015/438402. URL <https://doi.org/10.1155/2015/438402>.
- [281] Jonas P Demuro. Simultaneous acute cholecystitis and acute appendicitis treated by a single laparoscopic operation. Technical report, Hindawi, July 2012. URL <https://doi.org/10.1155/2012/575930>.
- [282] Neda Nozari. Appendicitis as a Complication of Acute Pancreatitis; A Case Report. Technical report, Govaresh, October 2016. URL <http://www.govaresh.org/index.php/dd/article/viewFile/1549/1858>.
- [283] Healthcare Cost and Utilization Project. Chronic Condition Indicator (CCI) for ICD-10-CM, October 2020. URL https://www.hcup-us.ahrq.gov/toolssoftware/chronic_icd10/chronic_icd10.jsp.
- [284] Amaia Calderón-Larrañaga, Davide L Vetrano, Graziano Onder, Luis A Gimeno-Feliu, Carlos Coscollar-Santaliestra, Angelo Carfi, Maria S Pisciotta, Sara Angleman, René J F Melis, Giola Santoni, Francesca Mangialasche, Debora Rizzuto, Anna-Karin Welmer, Roberto Bernabei, Alexandra Prados-Torres, Alessandra Marengoni, and Laura Fratiglioni. Assessing and Measuring Chronic Multimorbidity in the Older Population: A Proposal for Its Operationalization. *The Journals of Gerontology. Series A*, 72(10):1417–1423, October 2017. ISSN 1079-5006, 1758-535X. doi: 10.1093/gerona/glw233. URL <https://doi.org/10.1093/gerona/glw233>.
- [285] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, 2013. URL <https://www.R-project.org/>.
- [286] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, May 2010. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btq134. URL <http://dx.doi.org/10.1093/bioinformatics/btq134>.
- [287] Ender Celik. PIMP-Algorithm For The Permutation Variable Importance Measure, December 2015. URL <https://www.rdocumentation.org/packages/vita/versions/1.0.0/topics/PIMP>.
- [288] Wes McKinney and Others. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Semantic Scholar, 2010. URL <https://pdfs.semanticscholar.org/f6da/c1c52d3b07c993fe52513b8964f86e8fe381.pdf>.
- [289] John T Wilkins, Ron C Li, Allan Sniderman, Cheeling Chan, and Donald M Lloyd-Jones. Discordance Between Apolipoprotein B and LDL-Cholesterol in Young Adults Predicts Coronary Artery Calcification: The CARDIA Study. *Journal of the American College of Cardiology*, 67(2):193–201, January 2016. ISSN 0735-1097, 1558-3597. doi: 10.1016/j.jacc.2015.10.055. URL <http://dx.doi.org/10.1016/j.jacc.2015.10.055>.
- [290] Alexander Strehl and Joydeep Ghosh. Cluster ensembles -A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research JMLR*, 3, February 2022. ISSN 1532-4435. URL <https://www.jmlr.org/papers/volume3/strehl02a/strehl02a.pdf>.
- [291] Iven Van Mechelen, Hans-Hermann Bock, and Paul De Boeck. Two-mode clustering methods: a structured overview. *Statistical Methods in Medical Research*, 13(5):363–394, October 2004. ISSN 0962-2802. doi: 10.1191/0962280204sm373ra. URL <https://doi.org/10.1191/0962280204sm373ra>.
- [292] Beatriz Pontes, Raúl Giraldez, and Jesús S Aguilar-Ruiz. Biclustering on expression data: A review. *Journal of Biomedical Informatics*, 57:163–180, October 2015. ISSN 1532-0464. doi: 10.1016/j.jbi.2015.06.028. URL <http://dx.doi.org/10.1016/j.jbi.2015.06.028>.

- [293] Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D Chambers, Merlise Clyde, Thomas D Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P Field, Malcolm Forster, Edward I George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P Green, Anthony G Greenwald, Jarrod D Hadfield, Larry V Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J Hruschka, Kosuke Imai, Guido Imbens, John P A Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E Maxwell, Michael McCarthy, Don A Moore, Stephen L Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J Watts, Christopher Winship, Robert L Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen E Johnson. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10, January 2018. ISSN 2397-3374. doi: 10.1038/s41562-017-0189-z. URL <https://doi.org/10.1038/s41562-017-0189-z>.
- [294] Dominic Russ, John A Williams, Victor Roth Cardoso, Laura Bravo-Merodio, Samantha C Pendleton, Furqan Aziz, Animesh Acharjee, and Georgios V Gkoutos. Evaluating the detection ability of a range of epistasis detection methods on simulated data for pure and impure epistatic models. *PLoS One*, 17(2):e0263390, February 2022. ISSN 1932-6203. doi: 10.1371/journal.pone.0263390. URL <https://dx.plos.org/10.1371/journal.pone.0263390>.
- [295] Andreas Karwath, Karina V Bunting, Simrat K Gill, Otilia Tica, Samantha Pendleton, Furqan Aziz, Andrey D Barsky, Saisakul Chernbumroong, Jinming Duan, Alastair R Mobley, Victor Roth Cardoso, Luke Slater, John A Williams, Emma-Jane Bruce, Xiaoxia Wang, Marcus D Flather, Andrew J S Coats, Georgios V Gkoutos, Dipak Kotecha, and card AIC group and the Beta-blockers in Heart Failure Collaborative Group. Redefining β -blocker response in heart failure patients with sinus rhythm and atrial fibrillation: a machine learning cluster analysis. *The Lancet*, 398(103109):1427–1435, August 2021. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(21)01638-X. URL [http://dx.doi.org/10.1016/S0140-6736\(21\)01638-X](http://dx.doi.org/10.1016/S0140-6736(21)01638-X).
- [296] Furqan Aziz, Victor Roth Cardoso, Laura Bravo-Merodio, Dominic Russ, Samantha C Pendleton, John A Williams, Animesh Acharjee, and Georgios V Gkoutos. Multimorbidity prediction using link prediction. *Scientific Reports*, 11(1):16392, August 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-95802-0. URL <https://doi.org/10.1038/s41598-021-95802-0>.
- [297] John A Williams, Dominic Russ, Laura Bravo-Merodio, Victor Roth Cardoso, Samantha C Pendleton, Furqan Aziz, Animesh Acharjee, and Georgios V Gkoutos. A Causal Web between Chronotype and Metabolic Health Traits. *Genes*, 12(7), July 2021. ISSN 2073-4425. doi: 10.3390/genes12071029. URL <http://dx.doi.org/10.3390/genes12071029>.
- [298] Luke T Slater, John A Williams, Andreas Karwath, Hilary Fanning, Simon Ball, Paul Schofield, Robert Hoehndorf, and Georgios V Gkoutos. Multi-faceted semantic clustering with text-derived phenotypes. *medRxiv*, May 2021. doi: 10.1101/2021.05.26.21257830. URL <http://dx.doi.org/10.1101/2021.05.26.21257830>.

Appendices

A SUPPLEMENTARY MATERIAL

P-values

For reference a p-value of $p < 0.05$ has been used as a threshold in this thesis. Although this threshold is widely considered to be a significant threshold, recent discussions and literature have called for a stricter threshold of $p < 0.005$, see article by Benjamin et al. [293] (2018) for more information. Where possible, this thesis reports the raw p-value for readers to draw their own conclusions on significance.

Scaling

Input	Normalisation	Standardisation
0.0	0.0	-1.46
1.0	0.2	-1.88
2.0	0.4	-0.29
3.0	0.6	0.29
4.0	0.8	1.88
5.0	1.0	1.46

Table A.1: An example of normalisation and standardisation scaling.

Equations

Variance threshold

$$0.8 \times (1 - 0.8) \tag{A.1}$$

Term frequency-inverse document frequency

$$tf - idf(t, d) = tf(t, d) \times idf(t, d) \tag{A.2}$$

DBSCAN

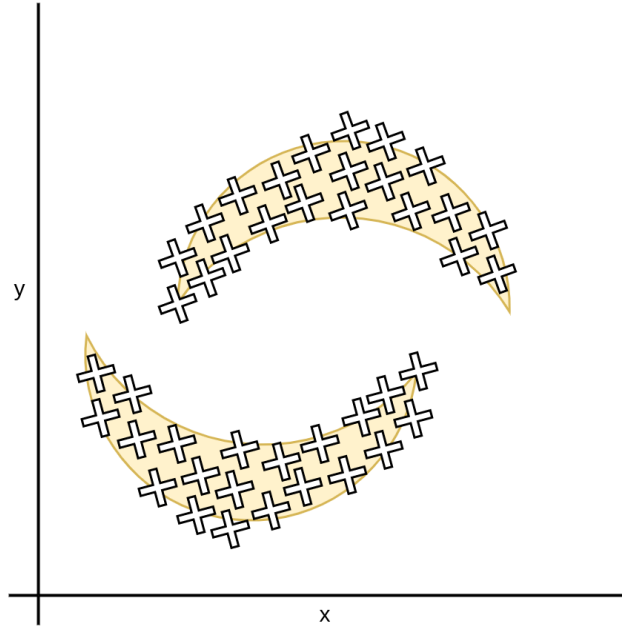


Figure A.1: An example of a DBSCAN clustering: finding patterns in arbitrary data structures.

B OCULAR INFLAMMATORY DISEASES

Synonym and class mappings

Synonym	Class
IOP	abnormal intraocular pressure
diamox	acetazolamide*
acyclovir	aciclovir
humira	adalimumab
teenager	adolescent
pain killers	analgesia
pain relief	analgesia
ACE	angiotensin converting enzyme
synechiae	anterior chamber synechiae*
the front of	anterior segment of eyeball
iritis	anterior uveitis
iopidine	apraclonidine
shunt implant	aqueous shunt
cause	associations
condition	associations
	asymmetrical*
baerveldt	baerveldt tube shunt*
cat scratch	bartonella henselae
betamethasone	betnesol
avastin	bevacizumab*
both eye	bilateral*
lumigan	bimatoprost
birdshot	birdshot retinochoroidopathy
	blind*
blurry	blurred vision
lyme	borrelia burgdorferi
alphagan	brimonidine
azopt	brinzolamide
cimzia	certolizumab
cyclosporin	ciclosporin
neoral	ciclosporin
flare	clinical symptom
cloudy	cloudy vision*
lens opacity	cloudy vision*
pink eye	conjunctivitis
corneal edema	corneal oedema
corticoid	corticosteroid
steroid	corticosteroid
cyclopental	cyclopentolate
mydrilate	cyclopentolate
maxidex	dexamethasone*
ozurdex	dexamethasone*
maxitrol	dexamethasone neomycin and polymyxin B ophthalmic
durezol	difluprednate*
trusopt	dorzolamide
effective	efficacy
enbrel	etanercept
eyeball	eye

sore	eye pain
visual field	field of vision
iluvien	flucinolone acetonide
retisert	flucinolone acetonide
fuchs	fuch's heterochromic iridocyclitis
fungal	fungus
bleeding	hemorrhage of the eye
herpes	herpes simplex virus
cold sore	herpes simplex virus 1
increased pressure	high intraocular pressure
high IOP	high intraocular pressure
A29	HLA-A29
B27	HLA-B27
B51	HLA-B51
HLA	human leucocyte antigen
advil	ibuprofen
motrin	ibuprofen
	idiopathic panuveitis*
better	improving
decreased	improving
remission	in remission off treatment
inflamed	inflammation activity
swelling	inflammation activity
swollen	inflammation activity
IBD	inflammatory bowel disease
pathology	inflammatory disorder
remicade	infliximab*
needle	injection
gradual	insidious
IU	intermediate uveitis
infusion	intravenous
IV	intravenous
drip	intravenous
vein	intravenous
eye injection	intravitreal
child	juvenile
JIA	juvenile idiopathic arthritis
JRA	juvenile idiopathic arthritis
juvenile arthritis	juvenile idiopathic arthritis
xalatan	latanoprost
	left eye*
	lensectomy*
lotemax	loteprednol
low IOP	low intraocular pressure
adult	mature
age	maturity
medication	medical therapy
medicine	medical therapy
meds	medical therapy
amethopterin	methotrexate
MTX	methotrexate
MXT	methotrexate
medrol	methylprednisolone
MS	multiple sclerosis
TB	mycobacterium tuberculosis
tuberculosis	mycobacterium tuberculosis
cellcept	mycophenolate
myfenax	mycophenolate
dilated	mydriasis
baby	neonatal
stable	no change
NSAID	non-steroidal anti-inflammatory
high pressure	ocular hypertension
low pressure	ocular hypotension
	psoriasis*
by mouth	oral
capsule	oral
pill	oral
tablet	oral
pan uveitis	panuveitis

overexposure	photophobia
light sensitive	photophobia
the back of	posterior segment of eyeball
PVD	posterior vitreous detachment
pred forte	prednisolone
predforte	prednisolone
prednisol	prednisolone
predsol	prednisolone
pre teen	prepubescent
IOL	pseudophakic
puberty	pubescent
inactive	quiescent
quiet	quiescent
lucentis	ranibizumab
relapse	recurrent
reactive arthritis	reiter's syndrome
RD	retinal detachment
RA	rheumatoid arthritis
	right eye*
vexol	rimexolone*
rituxan	rituximab
dose	route of administration
procedure	route of administration
sarcoid	sarcoidosis
sectoral laser	sectoral retinal laser photocoagulation
second line	steroid sparing immunosuppressive
dislocating	subluxed
acute	sudden
OP	surgical therapy
surgery	surgical therapy
symetric	symmetrical*
treat	therapeutic intervention
cosopt	timolol
timoptol	timolol
actemra	tocilizumab*
drop	topical to ocular surface
toxoplasmosis	toxoplasma gondii
syphilis	treponema pallidum
kenalog	triamcinolone*
mydracyl	tropicamide
surface of	tunica fibrosa of eyeball
UC	ulcerative colitis
valcyte	valganciclovir
vancocin	vancomycin
shingles	varicella zoster virus
viral	virus
eye sight	visual acuity
snellen chart	visual acuity
visual difficulty	visual impairment
	vitrectomy*
VKH	vogt koyanagi harada disease
VFEND	voriconazole
increased	worsening
worse	worsening
x ray	x-ray

Table B.1: List of the patient-preferred terms (synonyms) and matched clinical terms (class) in the Ocular Immune-Mediated Inflammatory Diseases Ontology.

*classes added directly from term frequency-inverse document frequency.

Some additional notes:

- All occurrences of “oedema” (e.g. “corneal oedema”), the domain experts saw patient use of “edema”.
- Many terms, such as “herpes simplex virus”, domain experts saw patients use shortened versions, i.e. “herpes”.

- Some synonyms without specific special characters were also included, e.g. “x ray” [x-ray] or “fuchs” [fuch’s heterochromic iridocyclitis], as request from the domain experts to ensure semantic capture of the patient-preferred term (synonym).
- Although mentioned in the Table footnote (Table B.1), classes marked with an asterisk (*) were terms derived from the term frequency–inverse document frequency for inclusion in the ontology.

Term distribution

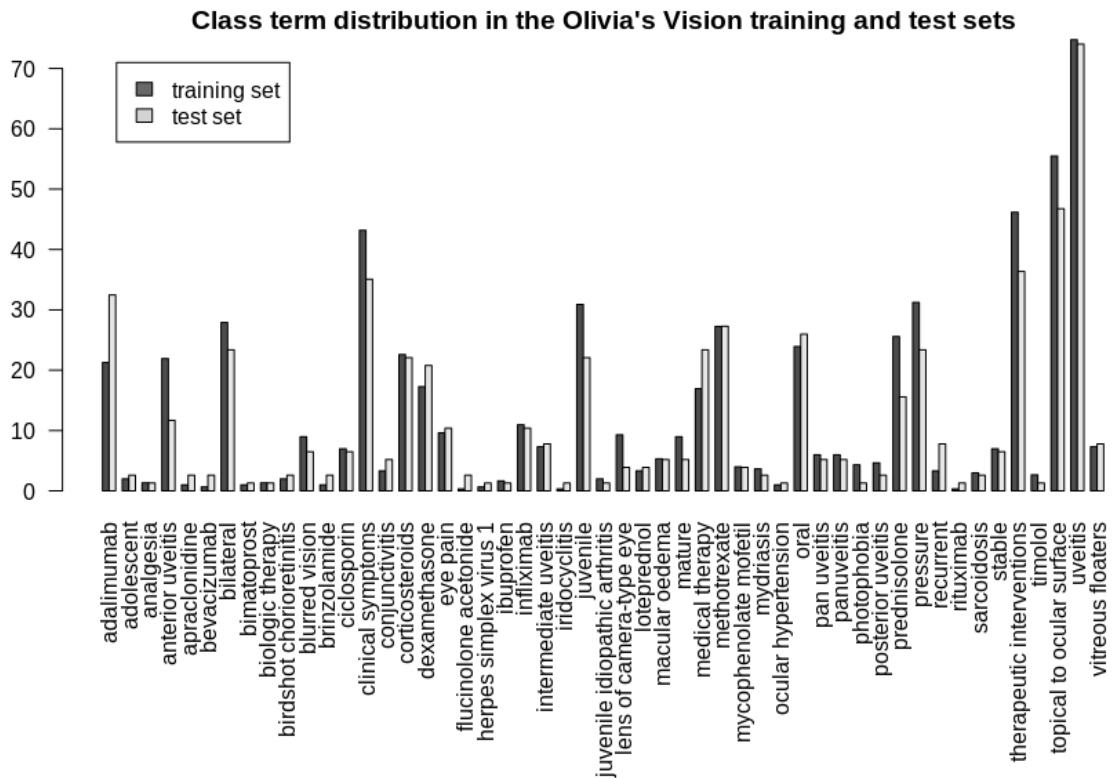


Figure B.1: Bar graph illustrating similar class distribution across Olivia’s Vision forum training and test sets.

Sentiment application

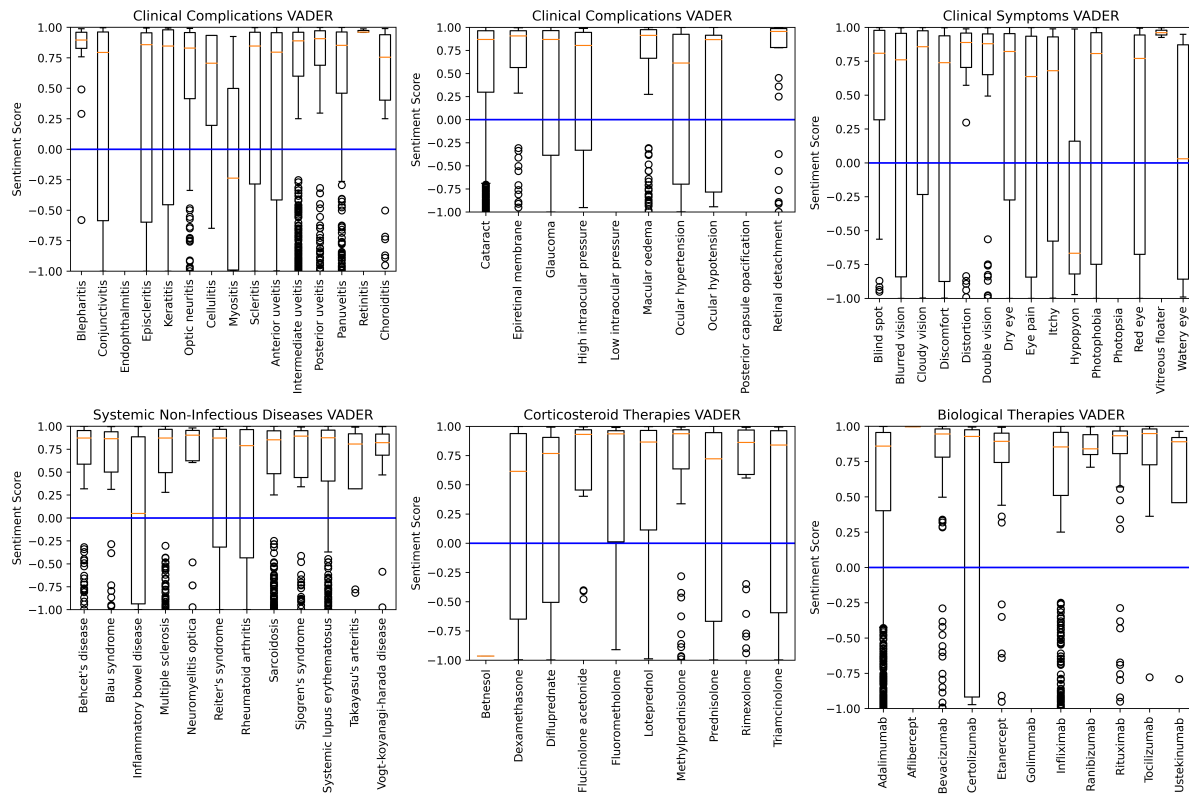


Figure B.2: Box plots of VADER natural language sentiment scores for a range of clinical phenotypes, complications, symptoms, systemic non-infectious diseases, and treatments - combining data from Olivia's Vision and Uveitis Foundation; scores (not normalised) for each class (including synonyms) ranged from -1.00 (most negative language) to +1.00 (most positive language).

C EMBEDDING MODELS

Combined Ontology for Inflammatory Diseases

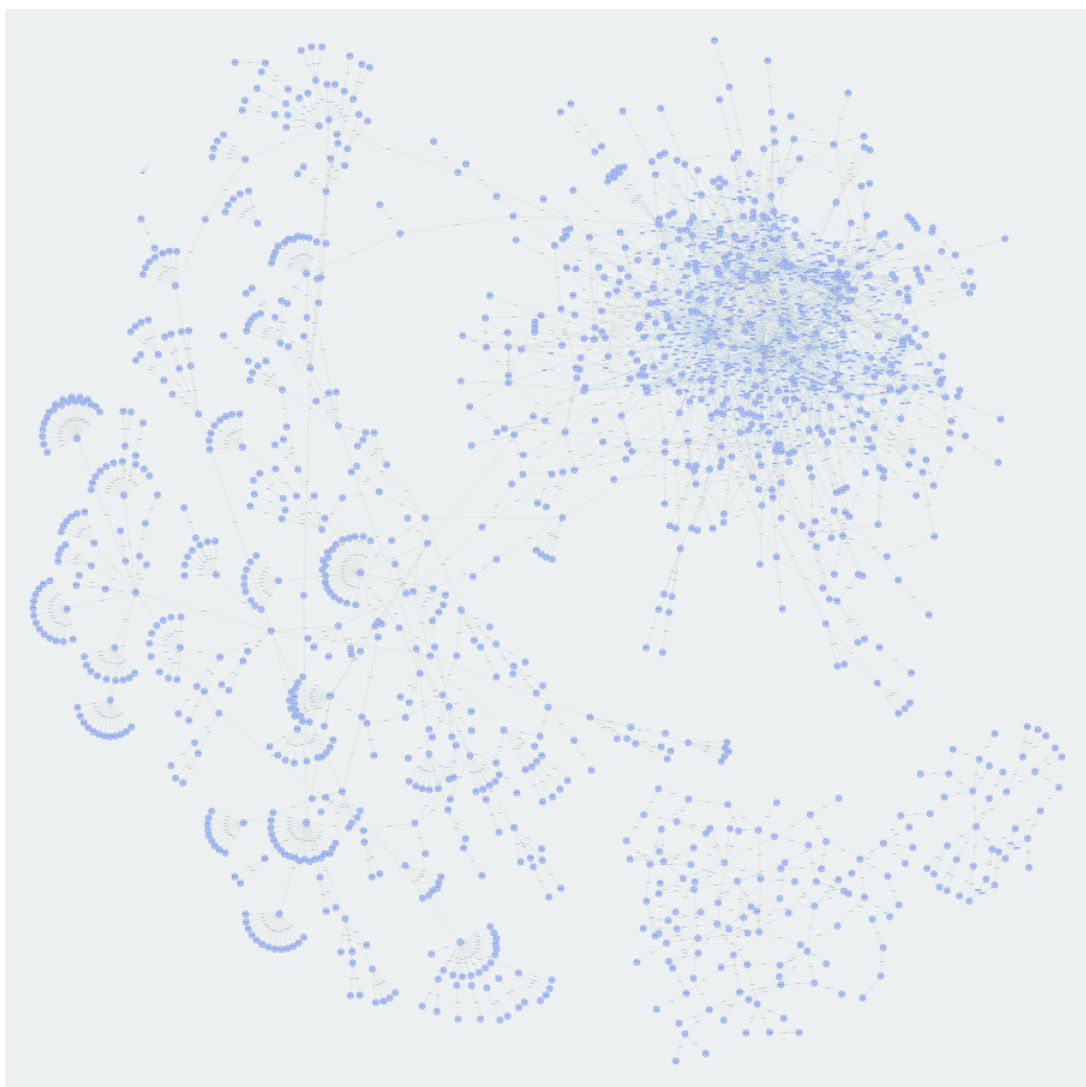


Figure C.1: Graphical representation of the Combined Ontology for Inflammatory Diseases. Each node is a class and each edge is a relationship.

Note: it should be noted that as of date-accessed 30-11-2021 WebVOWL was no longer available, however, as of date-accessed 01-03-2022 it became public again.

Semantic similarities

Semantic similarities in experiment one

Atrophic vaginitis

In CB the most similar terms are sialadenitis, “duondenitis” (misspelling of duodenitis: inflammation of duodenum lining, first part of the intestines), vulvovaginitis, pyelonephritis (urinary tract infection), adenitis, and more inflammatory conditions. In PI the most similar terms are vaginitis, “belanitis” (misspelling of balanitis), “adnitis”, peritonitis (inflammation of stomach lining), lymphoedema (swelling of body tissues), and more inflammatory conditions. With very few overlapping of CB. In CB + PI, the most similar terms are vulvovaginitis, vaginitis, and atrophic. Additionally, chorioretinitis and scleritis, are both visual inflammatory disorders.

Cellulitis

In CB the most similar terms are misspellings: “cellulititis”, “celllulitis”, “celulitis”, and “cellullitis”. Followed by misspellings of osteomyelitis: “osteomyelolitis”, “osteomyelelitis”, and “osteomyilitis”. Then phlebitis (inflammation of a vein) and panniculitis (inflammation of subcutaneous tissue). In PI the most similar terms are “cellulitus”, cellular, uveitis, myelitis with misspelling “mylitis”, and osteomyelitis. In CB + PI, the most similar terms are mostly misspellings of cellulitis: “cellulititis”, “cellulitis”, “celulitis”, “cellulitus”, “cellulitic”, “cellullitis”, “cellulits”, and more.

Cystitis

In CB the most similar terms are “cystititis” and misspellings of cholecystitis: “cholecystitis”, “choleycystitis”, “cholycystitis”, “choleystitis”, and “chlolecystitis”. Followed by mastitis, and “pylonephritis” and its correction pyelonephritis. In PI the most similar terms are cholecystitis, “cystitus”, “captitis” (misspelling of capitis, ringworm of scalp), “burstitis” (misspelling of bursitis), and proctitis. In CB + PI, the most similar terms are misspellings: “cystititis”, “cystitus”, and “cystits” (not in either CB or PI). Followed by misspellings of cholecystitis: “cholecystitis”, “choleycystitis”, “choleycystitis”, “choleystitis”, “chlolecystitis”, and more. With majority overlapping with CB.

Folliculitis

In CB the most similar terms are “panniculitis” and its correction panniculitis, “ticulitis” (shortened of diverticulitis), and misspellings of cellulitis: “cellullitis”, “celulitis”, and “celllulitis”, and finally “collitis” (misspelling of colitis) and more inflammatory conditions. In PI the most similar terms are “folliculitus”, itis, “vaculitis” (misspelling of vasculitis), diverticulitis, and cheilitis. Followed by eye inflammatory disorders: blepharitis, scleritis, and iritis. In CB + PI, the most similar terms are “folliculitus”, “panniculitis” and its correction panniculitis, “follicular” (misspelling of follicular), and “ticulitis”.

Vasculitis

In CB the most similar terms are vasculitic, angiitis, “pruitis”, typhlitis (inflammation of the cecum, the beginning of the large intestine), and omphalitis (infection of the umbilical stump). In PI the most similar terms are misspellings: “vasculitus”, “vasculittis”, “vascultis”, “vasulitis”, “vaculitis”, and “vasculities”. In CB + PI, the most similar terms are also misspellings from majority PI but also CB: “vasculittis”, “vasculitus”, “vasculitic”, “vasculits”, “vascultis”, “vasculitides”, “vasculidites”, and “vasculities”. Appendix Figure C.2 illustrates the semantic networks.

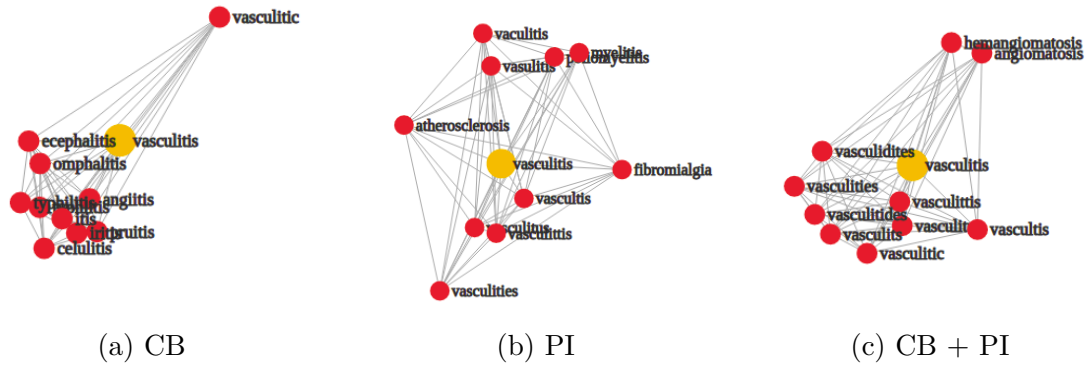


Figure C.2: Vectorisation representation of vasculitis.

Semantic similarities in experiment two

Arthritis

In m BR the most similar terms are osteoarthritis, arthritic, spondylitis, uveitis, cystitis, and osteomyelitis. In PI the most similar terms are osteoarthritis, urethritis, scleritis, blepharitis, arachnoiditis, and osteomyelitis. In m BR + PI the most similar terms are arthritic, osteoarthritis, urethritis, scleritis, uveitis, blepharitis, and osteomyelitis.

Blepharitis

In m BR the most similar terms are uveitis, costochondritis, cystitis, colitis, spondylitis, and gastritis (inflammation of the stomach lining). In PI the most similar terms are scleritis, meningitis, proctitis, osteomyelitis, urethritis, UTIs, and colitis. In m BR + PI the most similar terms are scleritis, proctitis, meningitis, uveitis, cystitis, and pancolitis.

Cellulitis

In m BR the most similar terms are osteomyelitis, spondylitis, necrosis (death of body tissue), uveitis, colitis, and prostatitis. In PI the most similar terms are vasculitis, osteomyelitis, colitis, polymyositis (a disease that causes muscles to become inflamed), folliculitis, and spondylitis. In m BR + PI the most similar terms are vasculitis, pancolitis, diverticulitis, osteomyelitis, colitis, and polymyositis.

Cystitis

In m BR the most similar terms are uveitis, spondylitis, rhinitis, osteomyelitis, colitis, bronchitis, and UTIs. In PI the most similar terms are hepatitis, proctitis, colitis, UTIs, pancreatitis, scleritis, and urethritis. In m BR + PI the most similar terms are uveitis, pancolitis, hepatitis, colitis, UTIs, proctitis, and tonsillitis.

Folliculitis

In m BR the most similar terms are osteomyelitis, colitis, diverticulitis, prostatitis, uveitis, and pancolitis. In PI the most similar terms are osteomyelitis, tonsillitis, diverticulitis, vasculitis, colitis, and cellulitis. In m BR + PI the most similar terms are diverticulitis, tonsillitis, vasculitis, colitis, uveitis, and osteomyelitis.

Hepatitis

In $_m$ BR the most similar terms are dermatitis, prostatitis, uveitis, colitis, and pancolitis. In PI the most similar terms are cystitis, proctitis, pancreatitis, UTIs, and colitis. In $_m$ BR + PI the most similar terms are cystitis, proctitis, pancolitis, uveitis, and colitis.

Scleritis and episcleritis

In $_m$ BR the most similar terms are uveitis, rhinitis, spondylitis, osteomyelitis, and cystitis. In PI the most similar terms are meningitis, osteomyelitis, blepharitis, urethritis, and proctitis. In $_m$ BR + PI the most similar terms are uveitis, meningitis, blepharitis, proctitis, and urethritis.

Sinusitis

In $_m$ BR the most similar terms are bronchitis, bursitis, uveitis, cystitis, and colitis. In PI the most similar terms are sinus, cystitis, bronchitis, UTIs, and polymyositis. In $_m$ BR + PI the most similar terms are sinus, uveitis, “sinusinfection” (sinus infection), cystitis, and sinuses.

Vasculitis

In $_m$ BR the most similar terms are uveitis, osteomyelitis, colitis, cellulitis, spondylitis, pancolitis, and diverticulitis. In PI the most similar terms are osteomyelitis, cellulitis, colitis, scleritis, spondylitis, diverticulitis, and folliculitis. In $_m$ BR + PI the most similar terms are colitis, pancolitis, osteomyelitis, diverticulitis, cellulitis, uveitis, and spondylitis.

D CLUSTERING IBD PATIENTS FROM THE UK BIOBANK

Normal blood markers

Blood marker	Genetic male normal range		Genetic female normal range	
	min	max	min	max
White blood cell leukocyte count	3.53	9.57	3.53	9.57
Red blood cell erythrocyte count	3.96	5.5	3.96	5.5
Haemoglobin concentration	12.14	16.27	12.14	16.27
Mean corpuscular haemoglobin	25.69	32.95	25.69	32.95
Platelet count	169.06	397.1	169.06	397.1
Mean platelet thrombocyte volume	7.54	11.24	7.54	11.24
Lymphocyte count	0.65	4.25	0.65	4.25
Monocyte count	0.17	1.21	0.17	1.21
Neutrophil count	1.47	7.06	1.47	7.06
Eosinophil count	0.03	0.77	0.03	0.77
Basophil count	0.01	0.13	0.01	0.13
Reticulocyte count	0.02	0.11	0.02	0.11
Alkaline phosphatase	30	130	30	130
Alanine aminotransferase	0	55	0	55
Apolipoprotein B	0.5	1.55	0.5	1.55
Aspartate aminotransferase	5	34	5	34
Urea	2.5	7.8	2.5	7.8
Cholesterol	0	4	0	4
Creatinine	64	104	49	90
C-reactive protein	0	5	0	5
Cystatin C	0.62	1.15	0.62	1.15
Gamma-glutamyl transferase	11	59	8	33
Glycated haemoglobin HbA1c	20	42	20	42
Total bilirubin	0	21	0	21
Triglycerides	0	1.7	0	1.7
Urate	388.96	671.84	203.32	583.44
Vitamin D	50	300	50	300

Table D.1: List of normal ranges* of blood markers for the different genetic sexes.
*from the UK Biobank.

Methods

Feature elimination

PCA

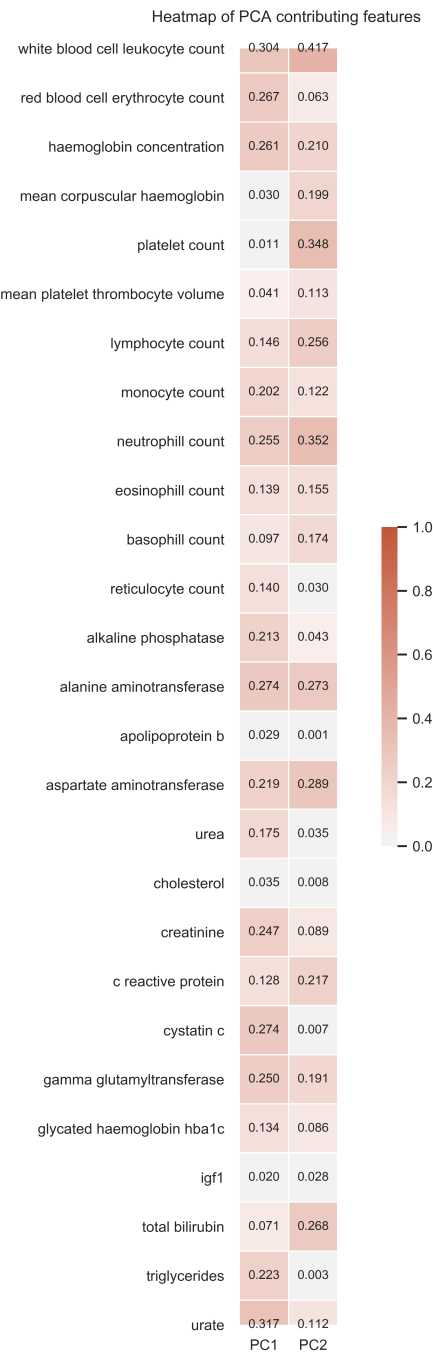


Figure D.1: PCA feature elimination visualisation of least-contributing features for removal.

Variance threshold

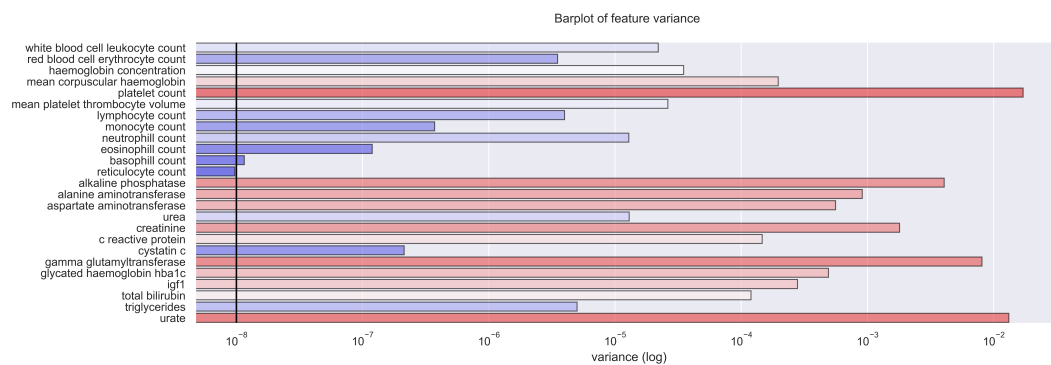


Figure D.2: Variance threshold feature elimination visualisation of features for removal.

Dimension reduction

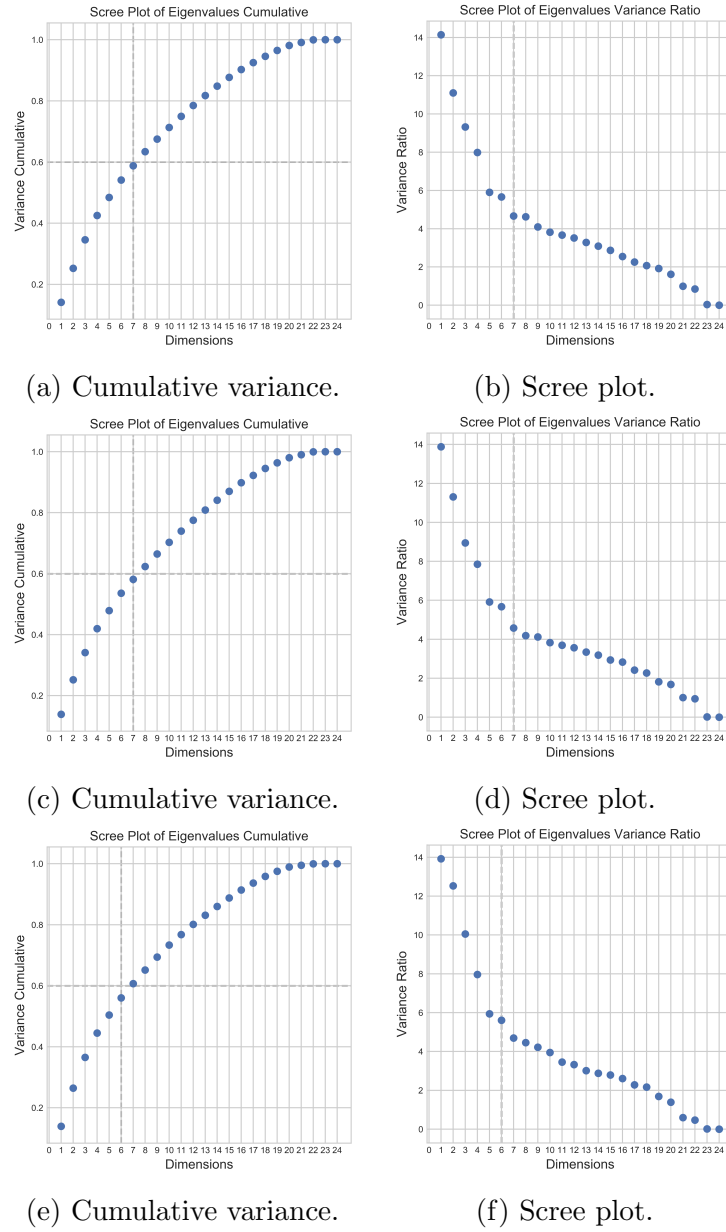
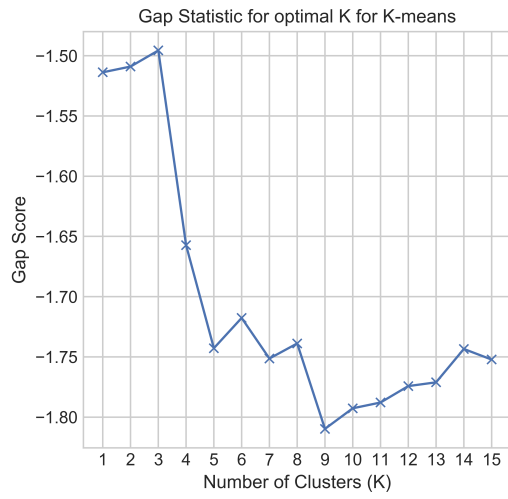


Figure D.3: PCA visualisation of eigenvalues via scree and cumulative.

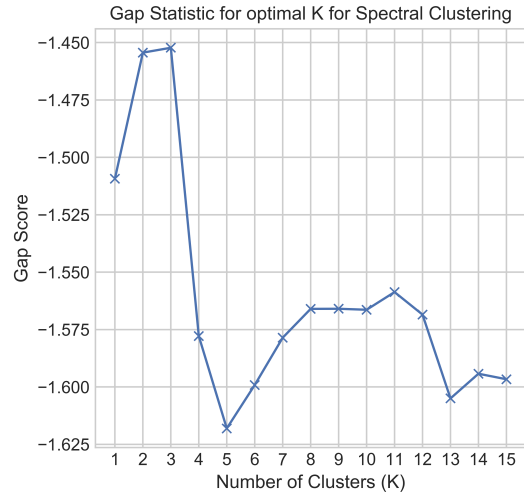
Key: (a) & (b) IBD patients diagnosed; (c) & (d) IBD patients not-yet diagnosed; (e) & (f) IBD patients diagnosed with future diagnosis.

Clustering

Gap statistic for IBD patients not-yet diagnosed



(a) K-means++.



(b) Spectral.

Figure D.4: Gap statistic for IBD-NYD.

Clustering box plots of not-yet diagnosed

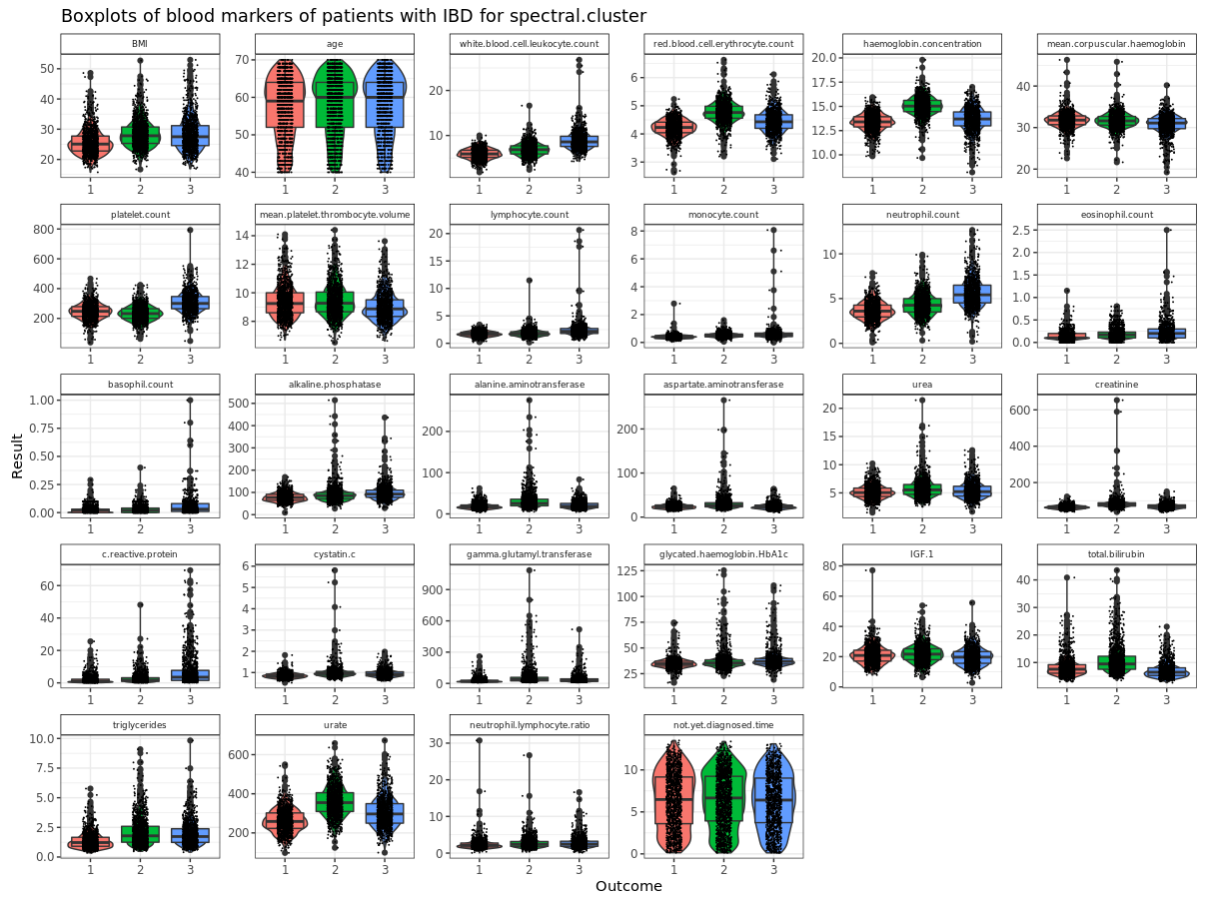


Figure D.5: Figure of IBD-NYD, box plots of averaged blood markers.

Permutation variable importance for IBD patients diagnosed

PIMP

- The mean decrease accuracy expresses how much accuracy the model losses by excluding each variable - in that if accuracy decreases, the feature removed was deemed important.
- The mean decrease in Gini coefficient is a measure of how each variable contributes (distribution) to the similarity of the decision tree nodes and leaves (outcome).

The higher the value of both indicates the higher the importance of the feature for the random forest model.

cl.rf

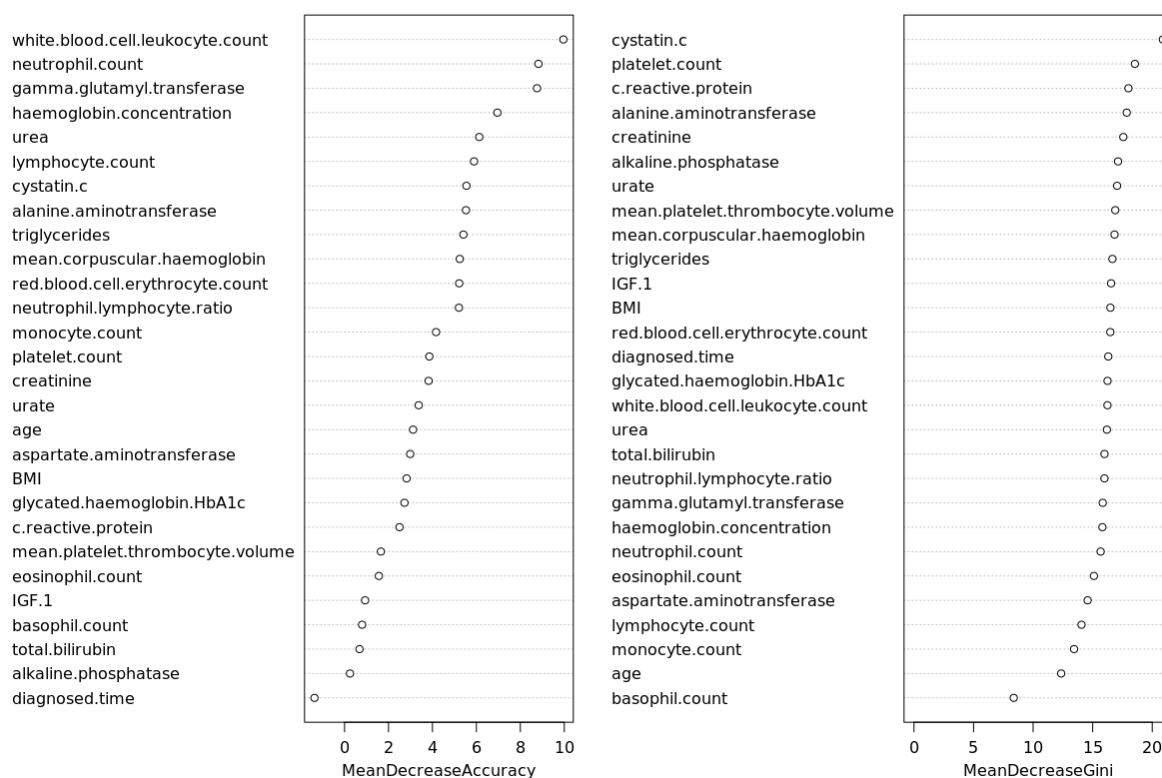


Figure D.6: Figure of permutation variable importance measure analysis for IBD-D.

Regression

Diagnosed

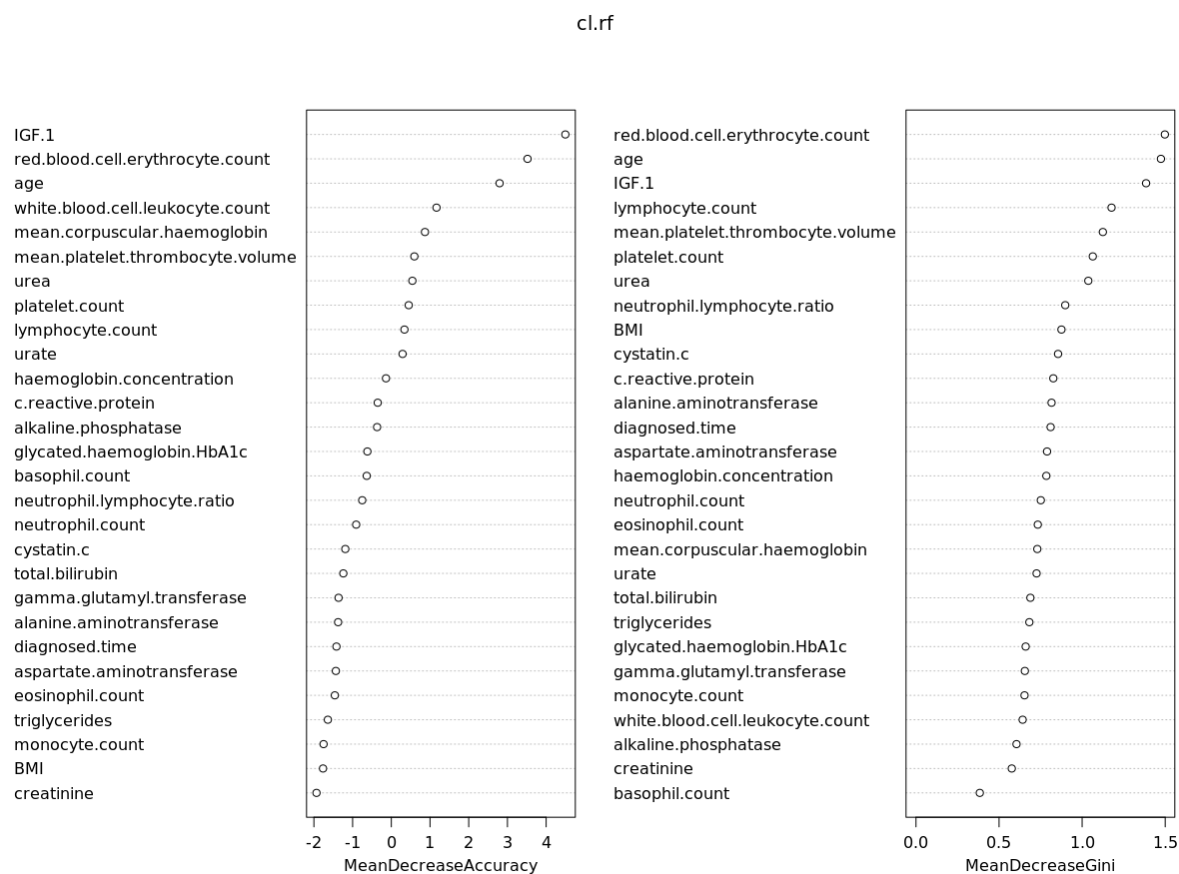


Figure D.7: Figure of IBD-D anxiety as endpoint with permutation variable importance measure.

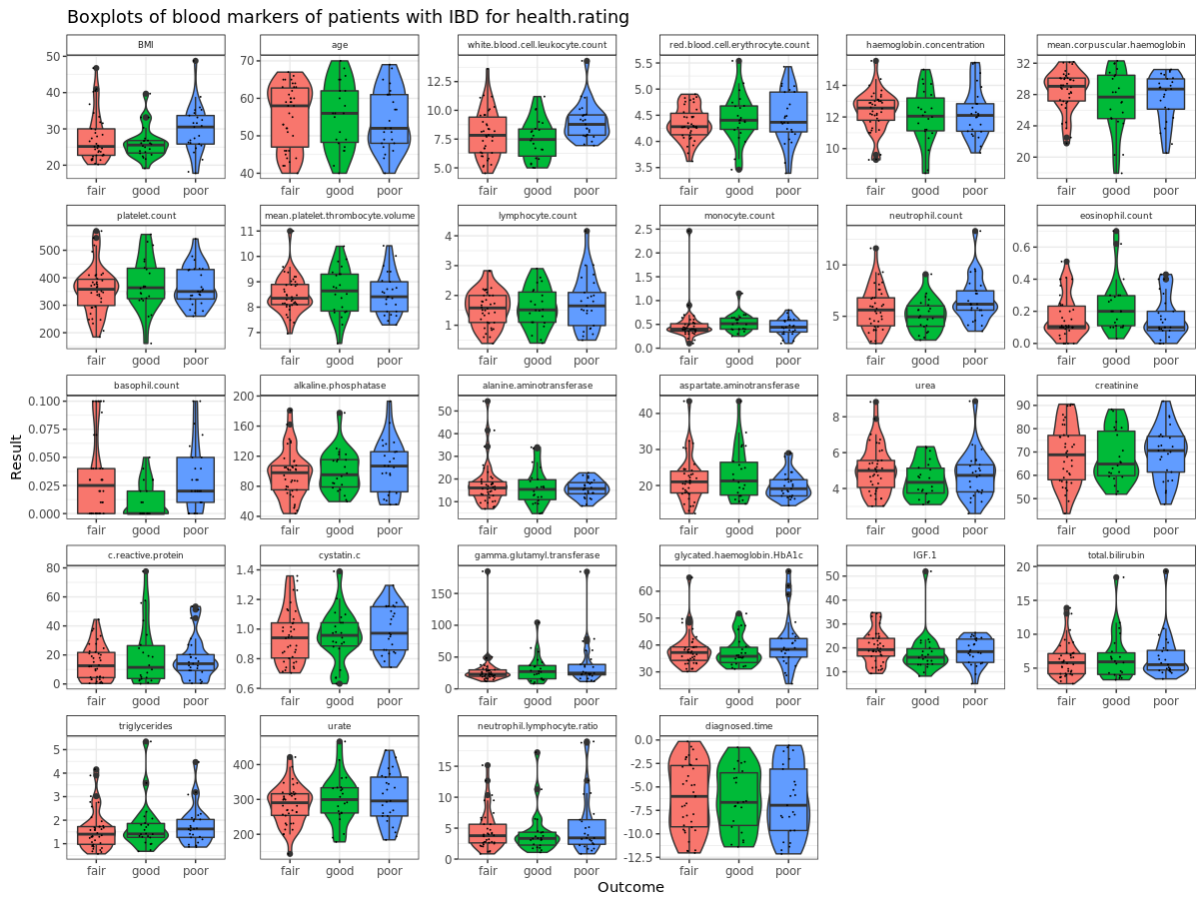


Figure D.8: Figure of IBD-D box plots of averaged blood markers across health rating in cluster 6.

Diagnosed with a future diagnosis

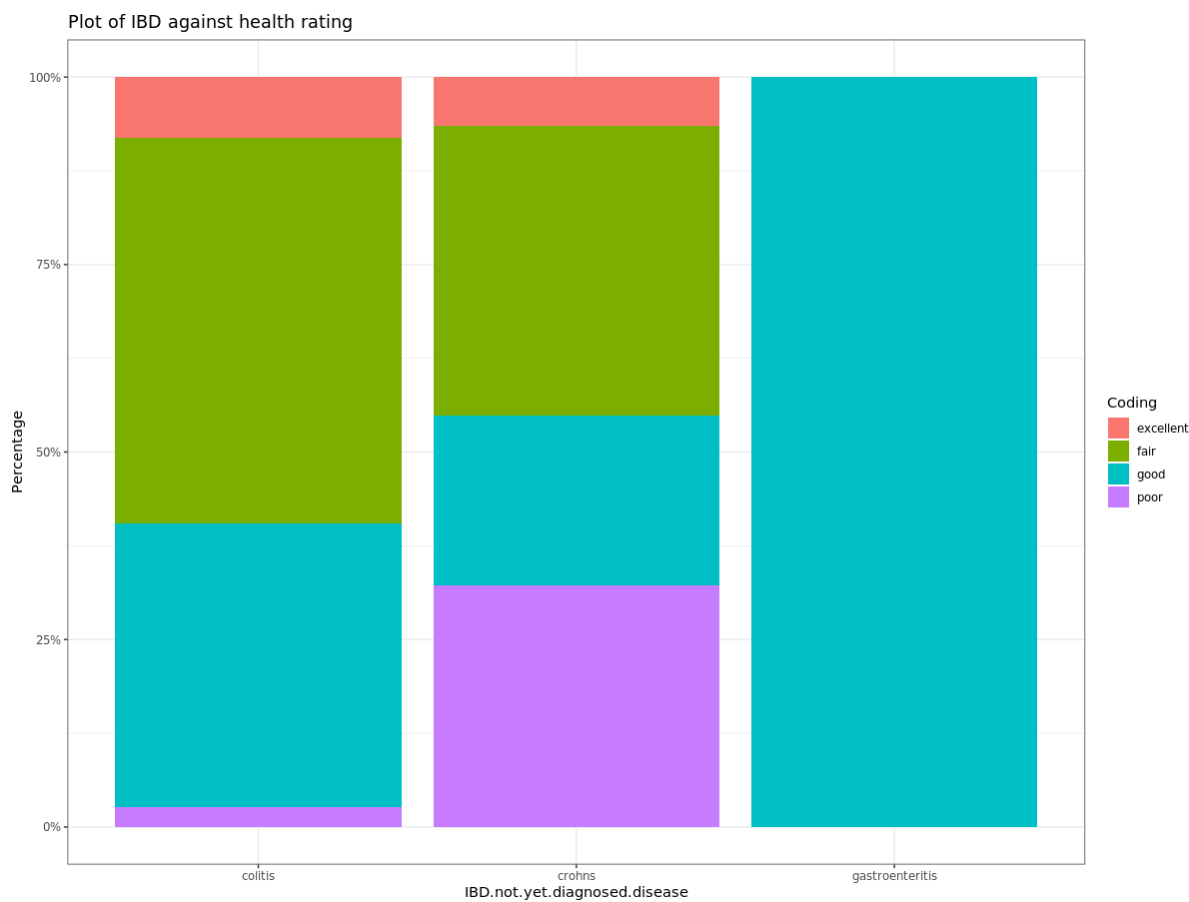


Figure D.9: Bar plot of IBD-D-FD cluster 6 IBD not-yet diagnosis against health rating.

PUBLICATIONS

Below is a list of all publications - and open source resources for secondary research use - produced over the time of this thesis and research, sorted by date.

- *PatientINF*. HuggingFace, 2022 Pendleton [263].
- *Evaluating the detection ability of a range of epistasis detection methods on simulated data for pure and impure epistatic models*. PLOS One, 2022 Russ et al. [294].
- *Combined Ontology for Inflammatory Diseases*. Zenodo, 2021 Pendleton [262].
- *Redefining β -blocker response in heart failure patients with sinus rhythm and atrial fibrillation: a machine learning cluster analysis*. The Lancet, 2021 Karwath et al. [295].
- *Multimorbidity prediction using link prediction*. Scientific Reports, 2021 Aziz et al. [296].
- *A causal web between chronotype and metabolic health traits*. Genes, 2021 Williams et al. [297].
- *Klarigi: Explanations for Semantic Groupings*. bioRxiv, 2021 Slater et al. [298].
- *Development and application of the ocular immune-mediated inflammatory diseases ontology enhanced with synonyms from online patient support forum conversation*. Computers in Biology and Medicine, 2021 Pendleton et al. [242].
- *Jabberwocky: an ontology-aware toolkit for manipulating text*. Journal of Open Source Software 2020, Pendleton and Gkoutos [260].

Important Term Definitions

adenitis	inflammation of a gland
appendicitis	inflammation of the appendix
arachnoiditis	inflammation of the arachnoid, membrane that surrounds nerves of the spinal cord
arthritis	inflammation of the joints
atrophic vaginitis	inflammation of the vaginal walls
autoimmune	the body's immune system attacking itself, a subset of immune-mediated diseases
balanitis	inflammation of the penis
biomarker	biological marker, measuring biological processes
blepharitis	inflammation of eyelids
bronchitis	infection of the main airways of the lungs
bursitis	inflammation of joint
cellulitis	inflammation of deeper skin layers
cholecystitis	inflammation of the gallbladder
chorioretinitis	inflammation of the choroid lining of the retina, a type of retinitis, a type of posterior uveitis: inflammation of the uvea in the eye
clinical letter	summary of a patient's visit to hospital and primary form of communication for clinicians
complete blood count	a set of medical tests to investigate a patient's blood: looking at their health
costochondritis	inflammation of the cartilage that joins ribs to breastbone
cystitis	inflammation of the urinary tract
dermatitis	general term that describes a common skin irritation
diverticulitis	inflammation of pouches in intestines
endometriosis	non-inflammatory tissue outside the uterus causing pain
folliculitis	inflammation of hair follicles
hepatitis	inflammation of the liver
hidradenitis	skin condition that causes abscesses and scarring
iritis	type of anterior uveitis: inflammation of the uvea in the eye

layman	language in a simplistic form, that everyone can understand without need of domain knowledge
mastitis	inflammation of breast tissue
myelitis	inflammation of the spinal cord
osteomyelitis	inflammation occurring in the bone
pancolitis	inflammation of the entire colon, a type of ulcerative colitis
pancreatitis	inflammation of the pancreas
panniculitis	inflammation of subcutaneous tissue
phenotype	observable trait, can be defined as height or presence of a disease
plantar fasciitis	inflammation of the plantar fascia
proctitis	inflammation of the rectum
pruritus	itchy skin
rhinitis	inflammation caused by allergen
scleritis and episcleritis	inflammation of the white of the eye
sialadenitis	salivary gland infection
sinusitis	inflammation of the sinuses
spondylitis	inflammation of spinal bones
tendonitis	inflammation or irritation of a tendon
tonsillitis	inflammation of the tonsils
urethritis	inflammation of the urethra
urinary tract infection	infection of urinary tract, abbreviated to UTI
vasculitis	inflammation of the blood vessels
vulvovaginitis	inflammation of both the vagina and vulva