Using Novel Data Types for Big Data Research in Epilepsy: Patient Records, Clinic Letters and Genetic Mutation

Lacey, Arron S.

Using Novel Data Types for Big Data Research in Epilepsy: Patient Records, Clinic Letters and Genetic Mutation

Submitted to Swansea University in fulfilment of the requirements for the Degree of Doctor of Philosophy

Arron Lacey

BSc MSc MRes [redacted]

Data Science

Swansea Neurology Research Group Swansea University Medical School

Swansea University

February 2019

Declaration

I, *Arron Lacey*, confirm this thesis has not been submitted towards a previous degree or other qualification, and is intended for submission of a Doctor of Philosophy, awarded by Swansea University.

Signed: (candidate) Date:

I, *Arron Lacey*, confirm that all of the work presented in this thesis is my own unless otherwise indicated. I have provided footnotes in the text where I have received assistance and have indicated permissions for presenting work which is not my own.

Signed: (candidate) Date:

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loans after expiry of a bar on access approved by the Swansea University.

Signed: (candidate) Date:

Abstract

Introduction: The aims of this thesis was to explore novel data types in healthcare that could enhance epidemiology studies in epilepsy and to develop novel methods of analysing routinely collected linked healthcare data, unstructured free text in hospital clinic letters and genetic variation.

Method: The SAIL Databank was used to source linked healthcare data for people with epilepsy across Wales to study the effects of epilepsy and social deprivation, coding of epilepsy in GP records and the educational attainment of children born to mothers with epilepsy. Hospital clinic letters from Morriston Hospital in Swansea were analysed using Natural Language Processing techniques to extract rich clinic data not typically recorded as part of routinely collected data. An automated pipeline was developed to predict the pathogenicity of Single Nucleotide Polymorphisms to prioritize potential disease-causing genetic variation in epilepsy for further in-vitro analysis.

Results: Incidence and prevalence of epilepsy was found to be strongly correlated with increased social deprivation, however a 10 year retrospective follow-up study found that there was no increase in deprivation following a diagnosis of epilepsy, pointing to deprivation contributing to social causation of epilepsy rather than epilepsy causing social drift. An algorithm was developed to accurately source epilepsy patients from GP records. Sodium Valproate was found to reduce educational attainment in 7 year olds by 12%. A Natural Language Processing pipeline was developed to extract rich epilepsy information from clinic letters. A pipeline was created to predict pathogencity of epilepsy SNPs that performed better than commonly used software.

Conclusion: This thesis presents novel studies in epilepsy using population level healthcare data, unstructured clinic letters and genetic variation. New methods were developed that have the potential to be applied to other disease areas and used to link different data types into routinely collected healthcare records to enhance further research.

Contents

List of Figures

List of Tables

Acronyms

1000g 1000 Genomes

AED Anti-Epileptic Drug

ALF Anonymous Linking Field

Bash Bourne Again SHell

BLAST basic local alignment search tool

BLOSUM BLOcks SUbstitution Matrix

CADD Combined Annotation Dependent Depletion

CBZ carbamazepine

CI confidence interval

CIPHER Centre for the Improvement of Population Health through E-records Research

Clinvar Clinical Variation database

CSV Comma Separated Variable

CUI Concept Unique Identifier

DALY Disability Adjusted Life Year

dbSNP Single Nucleotide Polymorphism atabase

DNA deoxyribonucleic acid

DSSP Dictionary of Secondary Structure in Proteins

EEG electroencephalogram

epi4k Epilepsy 4000 genomes project

ExAc Exome Aggregation Consortium

GATE General Architecture for Text Engineering

gnomAD Genome Aggregation Database

GUI Graphical User Interface

GWAS Genome Wide Association study

HGMD Human Gene Mutation Database

HPC High Performance Computer

ICD International Classification of Diseases

IGRP Information Governance Review Panel

ILAE International League Against Epilepsy

indel insertions/deletion

IQ Intelligence Quotient

JAPE Java Annotations Pattern Engine

LEV levetiracetam

LTG lamotrigine

LSOA Lower Super Output Area

MAF Minor Allele Frequency

MESH Medical Subject Headings

MRI Magnetic Resonance Image

MRCONSO UMLS Concept file

MRREL UMLS Relationship file

NGS Next Generation Sequencing

NHS National Health Service

NWIS NHS Wales Informatics Service

OR odds ratio

PEDW Patient Episode Database for Wales

PHT phenytoin

POS Part Of Speech

PSIBLAST Position Specific Iterated BLAST

PSIPRED PSI Blast Prediction for secondary structure

R R programming language

READ READ clinical coding system

SAIL Secure Anonymised Information Data Linkage

SD standard deviation

SIFT Sorting Intolerant From Tolerant

SNOMED SNOMED clinical coding system

SNP Single Nucleotide Polymorphism

SQL Standard Query Language

SUDEP Sudden Unexplained Death in Epilepsy

SVM Support Vector Machine

TPM topiramate

UCSC University of California, Santa Cruz

UK United Kingdom

UMLS Unified Medical Language System

Uniprot the Universal Protein Resource

UNIX UNIX operating system

VCF variant call format

VEP Variant Effect Predictor

VPA sodium valproate

WES Whole Exome Sequencing

WERN Wales Epilepsy Research Network

WGS whole genome sequencing

WIMD Welsh index of multiple deprivation

Acknowledgements

I would like to thank my supervisors Professor Mark Rees and Dr Seokyung Chung for providing me with their guidance and support. Their experience was vital in helping shape this thesis and it would have not been completed without their help. I would also like to thank Dr Owen Pickrell for providing me with invaluable experience and expertise that was essential towards carrying out the studies in this thesis. I would also like to thank Beata Fonferko-Shadrach for her keen eye for detail and help with the analysis of clinic letters and GP records.

I would like to thank the Swansea Neurology Research Group for their support: Dr Rhys Thomas, Dr Cathy White and Professor Mike Kerr, Dr Anna Derrick, Adam Higgins and Sarah Dawes. I would also like to thank my colleagues at the SAIL Databank for both their professional and personal support (cakes), particularly Ashley Akbari and Daniel Thayer for line managing me, Simon Thompson for his technical support and Professor Ronan Lyons and Professor David Ford for the opportunities they have given me.

Finally I would like to thank my family, especially my wife Sophie Lacey for her unending encouragement and understanding during completion of this thesis as well as my son Dylan for providing many welcome distractions from both study and sleep over the last 18 months.

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Papers related to Thesis

- Lacey, A.S., Pickrell, W.O., Thomas, R.H., Kerr, M.P., White, C.P. and Rees, M.I., 2018. Educational attainment of children born to mothers with epilepsy. J Neurol Neurosurg Psychiatry, pp.jnnp-2017. [3]

- Pickrell, W.O., Lacey, A.S., Bodger, O.G., Demmler, J.C., Thomas, R.H., Lyons, R.A., Smith, P.E., Rees, M.I. and Kerr, M.P., 2015. Epilepsy and deprivation, a data linkage study. Epilepsia, 56(4), pp.585-591. [4]

- Fonferko-Shadrach, B., Lacey, A.S., White, C.P., Powell, H.R., Sawhney, I.M., Lyons, R.A., Smith, P.E., Kerr, M.P., Rees, M.I. and Pickrell, W.O., 2017. Validating epilepsy diagnoses in routinely collected data. Seizure-European Journal of Epilepsy, 52, pp.195-198.[5]

Chapter 1

Introduction

1.1 Summary of Thesis Themes

This thesis documents research in epilepsy across 3 themes: epidemiology and big data, natural language processing of clinic letters and predicting pathogenicity of single nucleotide polymorphisms. The main aim was to describe how different methods and data types across these themes can be brought together to enhance the opportunity for epilepsy based research. This chapter introduces relevant studies related to the three themes of research to support the motivations for this thesis, and chapter two describes a comprehensive overview of the methods used to carry the work in this thesis.

Chapter 3 documents the results for 3 longitudinal epidemiological studies in epilepsy using the SAIL Databank. The SAIL Databank is a research platform for conducting population level healthcare studies in Wales and specialises in anonymous linked "Big Data" across various healthcare services across Wales. The three studies in chapter 3 explore the effects of epilepsy on social deprivation, a validation of GP recorded epilepsy diagnoses and the effects of exposure to antiepileptic drugs in the womb and the impact it has on educational attainment in 7 year old children.

Chapter 4 aims to explore how data in unstructured clinic letters can be included in epidemiology studies by using Natural Language Processing techniques. Only a proportion of unstructured data such as that in clinic letters, discharge reports or radiology and examination reports get entered into structure databases and audited for research purposes, with many rich patient data often missing and not available for research. Chapter 4 presents a study using NLP techniques to extract rich patient information from 200 clinic letters from the Morriston hospital epilepsy clinic.

Chapter 5 explores various techniques to predict pathogencity of a particular type of genetic mutation called singe nucleotide polymorphisms, or SNPs. There are many SNPs that have been documented as the cause of various types of epilepsy, but with over 3 million variants in a person's genome it is difficult to predict the impact of these mutations in terms of likelihood of developing a disease. Bioinformatics pipelines aim to reduce the search space within the human genome to focus on a very small set of variants for further study. Part of these pipelines involves functional analysis and there are many programs that specialize in predicting pathogenicity of SNPs, in which the accuracy of these programs can differ in different disease areas. Chapter 5 aims to incorporate the knowledge from existing systems to build a pipeline that accurately predict the pathogenicity of epilepsy SNPs.

1.2 Epilepsy

Epilepsy is a disease characterized by unprovoked seizures that can be distinctly different from other types of seizures such as febrile seizures that occur mainly in children during a fever and dissociative seizures that occur for psychological reasons. It effects 1% of the population (600,000 individuals in the UK) [6] [7] and it has been estimated that over 50 million people worldwide have epilepsy [8]. The International League Against Epilepsy (ILAE) define epilepsy as any one of the following [9]:

1. Two unprovoked seizures occurring more than 24 hours apart
2. One unprovoked (or reflex) seizure and a probability of further seizures similar to the general recurrence risk (at least 60%) after two unprovoked seizures, occurring over the next 10 years
3. A diagnosis of an epilepsy syndrome

Epileptic seizures are treated with anti-epileptic drugs (AEDs) in which patients may require a combination of AEDs to help control their seizures, and some may not respond to AEDs at all, known as refractory epilepsy.

1.2.1 Epilepsy types

There are various epilepsy types that can be defined in various ways:

- Type of seizure
- Age at which seizure began
- Causes of seizure

- The part of the brain involved during a seizure
- Severity and duration of seizures
- EEG electroencephalogram patterns
- Brain imaging
- Mode of inheritance
- Other disorders in additions to seizures
- Patterns of seizures during the day (at day or night)

While there are many *syndromes* within epilepsy such as Juvenile Myclonic Epilepsy, Dravet Syndrome and Lennox-Gaut syndrome, epilepsy seizures are also used as a diagnostic tool in clinical practice and combined with various other factors seizure type will underpin an epilepsy syndrome. The use of seizures as a diagnostic tool is useful in terms of choosing AED treatment, where different seizure types have well defined AED regimes. Seizure categories in epilepsy are broadly defined by two types, that being generalized or focal seizures.

1.2.2 Generalized seizures

Generalized seizures originate rapidly from bilaterally distributed brain networks i.e. generalized seizures affect the entire brain [10]. The ILAE recognize generalised seizures as described in table 1.1. [11]

Table 1.1: Generalized seizures recognized by the ILAE.

Name	Description
Tonic-clonic seizures	Initial phase of stiffness (tonic) followed by jerking (clonic) and a loss of consciousness. Gradual recovery with minute/hours of post ictal confusion
Clonic seizures	Similar to tonic clonic seizures without stiffness
Typical absence seizures	Sudden, brief (generally <10s) periods of loss of awareness with behavioural arrest (staring episodes) with rapid recovery, occasional eye movements and automatisms
Atypical absence seizures	Longer than typical absences and frequently associated with myoclonic or atonic attacks. Start and finish more gradually, focal features more prominent, and more retained awareness, than typical absences
Myoclonic absence seizures	Very brief (<1 sec) 'electric-shock' muscle contractions with sudden onset and cessation. Single muscle to generalised jerking. Consciousness generally not impaired
Tonic seizures	Sustained muscular contraction lasting <1 minutes with rapid recovery
Eyelid myoclonia	Quick upward jerk of the eyelids lasting around 3 seconds
Myoclonus	Spasmodic jerks or twitches in various muscles (positive) or brief laspes in concentration (negative)
Atonic seizures	Sudden loss in muscle strength causing the patient to drop to floor. Sometimes called "drop attacks"

1.2.3 Focal seizures

Focal seizures originate from one hemisphere in the brain in which some types may cause absences and loss of consciousness. Additionally some types of focal seizures can spread to the entire brain which are called secondary generalised seizures. The ILAE recognizes focal seizure types as described in table 1.2.

Table 1.2: Focal seizures recognized by the ILAE

Name	Description
Focal sensory seizures	Brief disturbance in taste, touch, smell or sight usually lasting no more than 2 minutes
Focal motor seizures	A seizure with localized motor activity. There may be spasm or clonus (jerking) of one muscle or a muscle group and this may remain localized or it may subsequently spread to adjacent muscles
Frontal lobe	Frequently occurring during sleep. Brief, rapid onset and cessation. Prominent motor features, sometimes with posturing and head version. Frequent bizarre automatisms / behaviours and vocalisation
Temporal lobe	"Generally longer in duration than frontal lobe seizures. Variety of sensory disturbances including psychic (d´ej'a vu , jamaisvu, fear), gustatory and olfactory hallucinations. Sensation of epigastric disturbance. Oro-facial automatisms (e.g. chewing, sucking) or fidgety hand movements. Frequently altered awareness. Auditory features with lateral temporal lobe involvement"
Gelastic seizures	A rare type of seizure that involves a sudden burst of energy, usually in the form of laughing or crying.
Hemiclonic seizures	Entirely 1-sided, unilateral, clonic convulsions
Secondarily generalized seizures	Focal seizures evolving into generalized seizures, most often with tonic-clonic convulsions. The partial seizures, which were once limited to one hemisphere of the brain, progress to encompass the entire brain bilaterally

1.2.4 Causes of epilepsy

The causes of epilepsy can be broadly defined as either *symptomatic* where there is a physically identifiable change in structure of the brain, or *genetic* where there is no apparent change in structure of the brain and is therefore assumed to be caused by a genetic mutation inherited from a person's parents. Both types of epilepsy each account for around 50% of all epilepsies respectively. Symptomatic epilepsies are usually caused by injury to the brain through birth trauma, neurodegenerative diseases, brain neoplasms, cerebrovascular disease and brain malformations. For symptomatic epilepsy to be ruled out in the presence of epileptic seizures i.e. genetic epilepsy, there must be no evidence of structural changes in the brain to be detected. There exists a grey area over what constitutes symptomatic epilepsy or genetic epilepsy in that some conditions cause various deficiencies in the supply of glucose to the brain caused by a known genetic mutation, in which no structural changes are present in the brain, yet the patient is classified as having epilepsy.

1.2.5 Genetics of epilepsy

For the 50% of epilepsies that are caused by genetic mutation, the incidence of genetic epilepsy passed on to first degree relatives have been shown to be up to 4 times that in the population than those that do not have a first degree relative with epilepsy [12]. Mutations found in a small amount of genes have been found to cause Idiopathic Generalized Epilepsy (IGE). Multiple family studies and twin studies have found that IGE has a common gene origin, but it is likely that some forms of epilepsy have multi-gene modes of inheritance [13] [14] [15]. There is evidence for different sets of genes producing different epilepsy syndromes such as Juvenile Myoclonic Epilepsy (JME) [16] [17]. Currently, mutations on the SCN1A voltage-gated sodium channel gene account for the largest amount of IGE syndromes [18] [19] with over 150 mutations attributed to infantile and childhood onset epilepsy. In general, seizure-related syndromes are accounted for by mutations across multiple genes that code for ion channel proteins, where examples of such proteins are given in table 1.3.

Table 1.3: A list of ion channel domains and proteins and how mutations correlate to epilepsy phenotypes.

Channel Mutations in Epilepsy			
Ion Channel	**Gene**	**Phenotype**	**Inheritance**
Acetylcholine receptor	CHRNA2	ADNFLE	Single Gene
	CHRNA4	ADNFLE	Single Gene
	CHRNB2	ADNFLE	Single Gene
Calcium	CACNA1A	CPS,GTCS	Single Gene
	CACNA1H	CAE,IGE	Complex
	CACNB4	IGE	Complex
Chloride	CLCN2	IGE	Single Gene
GABA receptor	GABRG2	CAE/GEFS+/FS	Single Gene
	GABRA1	JME,CAE	Single Gene
Potassium	KCNQ2	BFNC1	Single Gene
	KCNQ1	BFNC1	Single Gene
	KCND2	mTLE2	Single Gene
Sodium	SCN1A	GEFS/SMEI	Single Gene
	SCN2A	BFNIC	Single Gene
	SCN1B	GEFS+	Single Gene

Despite some clear single gene relationships for various epilepsy syndromes there is also evidence for single gene overlap, and thus the relationship between known genes in epilepsy appears to be more complex[14]. Many factors determine development of the disease outside of an observed mutation such as mode of inheritance and

gene expression, so presence of a mutation that is known to cause epilepsy in one person, may not cause epilepsy in another. Recruitment of families with an extensive history of a disease is the first challenge in furthering our understanding of the complex relationships of genes and disease. Sourcing cohorts of patients comprising of such families is a lengthy and expensive process - family history needs to be determined as accurately as possible and blood samples need to be taken to analyse each persons' DNA. Processing and analysing DNA is also incredibly expensive, where most whole genome sequencing is not done within the research department that will analyse the resulting genome. Whole genome sequencing typically gets outsourced to dedicated laboratories at a cost per genome. Next Generation Sequencing (NGS) has revolutionized the process of sequencing a persons whole genome and is now the leading method that supports whole genome/exome based research.

1.2.6 Epidemiology of epilepsy

Epilepsy prevalence has been measured in various studies [20]-[21] where it ranges between 0.3-0.8% in developed countries and 0.43-1.4% in developing countries. In chapter 3 of this thesis a study of epilepsy prevalence in Wales is presented that estimated the prevalence of epilepsy to be 0.77% of the population, where higher prevalence is found in more deprived areas (1.13%) than less deprived areas (0.49%), a trend which is also seen in the incidence of new cases of epilepsy.

Incidence of epilepsy is typically highest in children and the elderly, with lower incidence between the age of 18-65, and incidence is double in men over 65 than women over 65 [22]. The majority of incident epilepsy in children are due to genetic factors, where the incidence of epilepsy later in life is due to symptomatic factors as the incidence of neurodegenerative diseases, brain neoplasms and stroke also increase with age. Figure 1.1 presents the "U-shape" curve that describes epilepsy incidence across all ages:

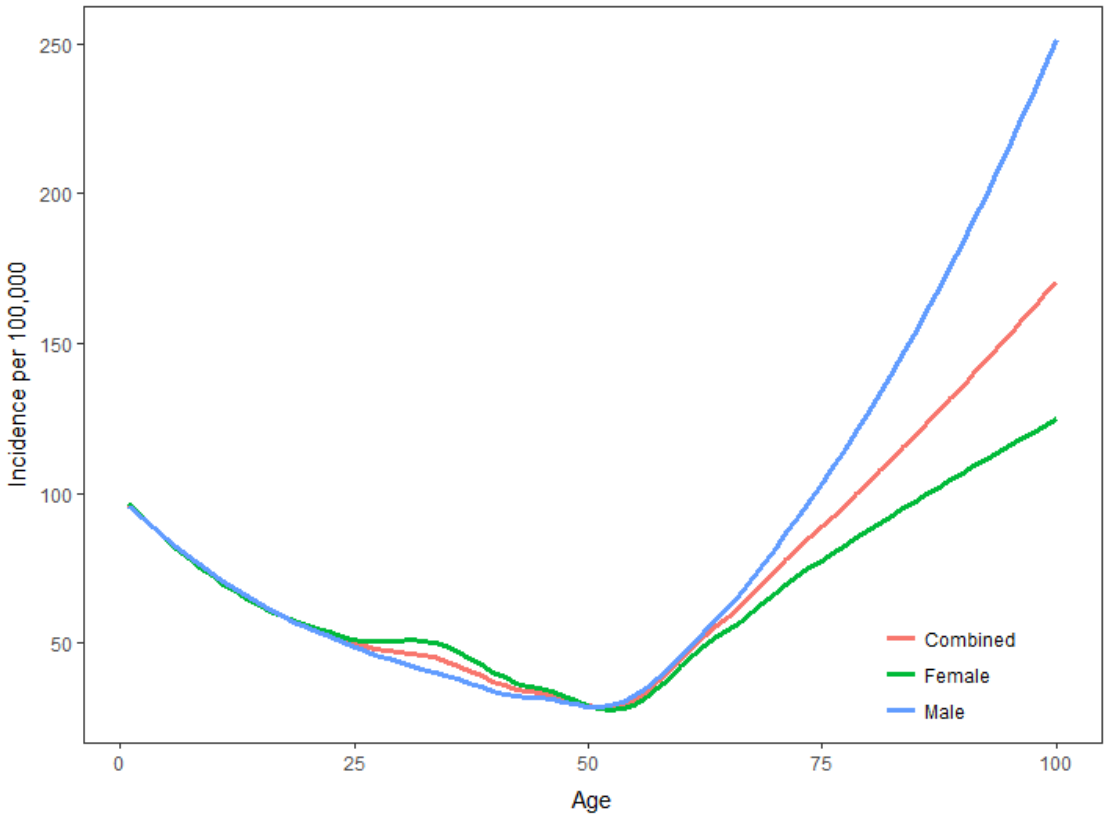

Figure 1.1: Incidence of epilepsy from a study in Iceland showing incidence per 100,000 stratified by age and sex.

1.2.7 Anti-epileptic drugs

Anti-epileptic drugs (AEDs) are a group of drugs that aim to suppress unprovoked seizures by suppressing the rapid firing of neurons in the brain during a seizure by binding to specific receptors in the brain and inhibiting voltage dependant sodium currents [23]. AEDs also aim to prevent the spread of seizures in it's early phase to other parts of the brain [24]. Around 50% of patients treated with AEDs have a 25-50% reduction in seizures, with other patients with more modest reduction [25]. While AEDs can be effective, around half of epilepsy patients experience adverse effects from a first line AED [26] [26]. Prescribing trends in anti-epileptic drugs have changed in recent years due to evidence of some such side effects of some AEDs.

A study using Welsh GP data held accessed via the SAIL Databank showed that newer AEDs such as Lamotrigine have been prescribed as a first line AED with increasing frequency over a ten year a period between 2000 and 2010, and older AEDs such as Sodium Valproate have seen a reduction in prescribing in women of child bearing age, probably due to evidence suggesting valproate can produce cognitive dysfunction if exposed to children *inutero* [1] [27]. Figure 1.2 shows prescribing trends of first line AEDs prescriptions in Wales.

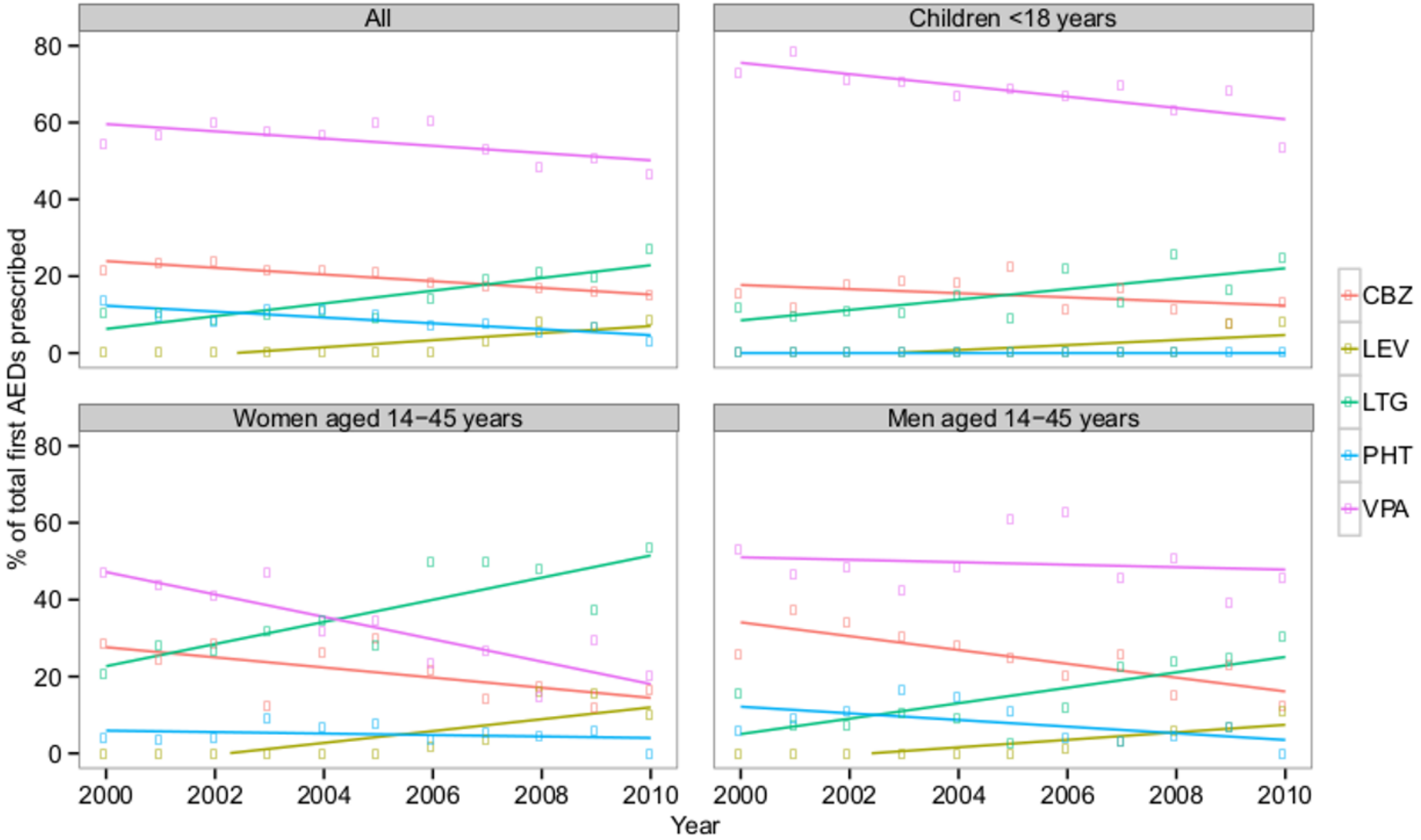

Figure 1.2: Prescribing trends of first line AEDs in Wales between 2000-2010. Newer drugs such as Lamotrigine have been adopted as per the SANAD study guidelines, where a decline in valproate prescriptions to women of child bearing age is also observed [1]

Choice of AEDs to treat seizures doesn't just include seizure control of the individual. Well documented side effects of AEDs include behavioural issues, decline in cognitive function, migraines and psychiatric disorders [28]-[29]. Various AEDs have been studied in relation to weight gain and loss, in which some drugs such as Levetiracetam and Sodium Valproate, while some drugs have been shown to cause weight loss, complicating the issue of AED prescriptions in patients with conditions such as diabetes mellitus [30]- [31].

AED choice is also important when prescribing in pregnant women. Various studies have associated *inutero* exposure to sodium valproate with a variety of effects on offspring that include reduced IQ, decline in motor and language skills as well as general decline in cognitive abilities [27]-[32]. Recently a Danish study found that children exposed to valproate *intuero* perform worse then their peers in national tests [33]. Chapter 4 in this thesis presents a study using Welsh Key Stage 1 education tests and AED prescribing data in pregnant women that found *inutero* exposure to sodium valproate and AEDs in combination are associated with decreased educational attainment in children aged 7 [3].

1.2.8 Burden and impact of epilepsy

The Global Disease Burden Study has estimated that epilepsy contributes to 1% of all days lost due to ill health and that on average epilepsy forms 0.5% of total disease burden as measured by the Disability Adjusted Life-Year (DALY) score. [34]. Epilepsy places a huge burden on those who suffer unprovoked seizures on a daily basis, a burden which is also shared by the relatives and peers of someone with epilepsy [35].

Epilepsy is a condition associated with a range of co-morbid conditions. Around 40% of adolescents with epilepsy also have an additional neurological condition and 1 in 4 persons with epilepsy of any age has a learning disability [36]. Behavioural issues are prevalent in children with epilepsy exhibited both in school and at home [37] [38]. Children with epilepsy are a stigmatized group and are twice as likely to be bullied at school than their peers [39], and a qualitative study of children with refractory epilepsy viewed seizures as a barrier to a normal life [40].

Seizures disrupt short term information storage, especially nocturnal seizures when most memory consolidation takes place. The physical impact of seizures on the brain is associated with memory loss in people with epilepsy ranging from concentration issues to chronic forgetfulness [41]. Auras occurring before and during a seizure and other lapses of concentration contribute to poor recall of which 1 in 4 people cannot recall experiencing auras or lapses in concentration [42] and people with epilepsy

fail to document around 50% of recent seizures. Poor memory loss therefore also contributes to difficulties in learning along with increased risk of being born with learning difficulties.

Around 80% of epilepsy is prevalent in developing countries in which there exists not only a barrier to effective care and seizure treatment, but there is an undeniable trend in epilepsy, social deprivation and social stigma [43] [44]. People with epilepsy are likely to experience prejudice and discrimination in all walks of life as well as being at high risk of abuse and violence [45] [46]. Young people with epilepsy are often discouraged from pursuing their chosen career path [35] and face discrimination in life ranging from diminished access to various insurance schemes and employment opportunities [47]. One survey of employers found that 16% felt that they didn't have jobs for someone with epilepsy and 21% considered employing someone with epilepsy as "a major issue" [48].

Complimentary to the fact that epilepsy is more apparent in developing nations, lower socio-economic status is a risk factor for epilepsy in adults [49] and the various social struggles people with epilepsy face is a strong argument for epilepsy causing social drift. Multiple studies have associated social deprivation with epilepsy in both new cases of and existing epilepsies [50] [51]. There are two main hypotheses for social deprivation in epilepsy; social causation and social drift. Social causation in epilepsy could be explained by factors associated with both deprivation and causes of epilepsy, namely such as perinatal hypoxic injury, head trauma, and cerebrovascular disease [21] [52] [53]. Social drift is hypothesised to explain some of the high deprivation seen in people with epilepsy for various reasons related to social stigma and discrimination in employment. Chapter 4 presents a study of social deprivation in Welsh patients with epilepsy and in both newly diagnosed patients and patients with existing epilepsy.

1.3 Big data and patient records as a resource for research

In this section, patient records and large linked datasets are presented as a method for researching the epidemiology and burden of epilepsy for large cohorts. The SAIL Databank is presented as such a resource that is utilized in this thesis to conduct population level epilepsy research in Wales.

1.3.1 Electronic Healthcare Records

Up to 1990 in the UK, patient records were largely paper based. Large scale Electronic Healthcare Record (EHR) linkage as a research method has grown as a direct result of embedding computer systems in primary care to make the transition from paper based records to electronic records. In 1987, two GPs from Egton, Yorkshire Dr Peter Sowerby and Dr David Stables formed the Egton Medical Information Systems Group (EMIS) to develop software that could capture paper based records into electronic format. The result was a commercial rollout of the EMIS software in 1990, and EMIS, as well as many other clinical audit systems that followed, enabled GPs to capture routinely collected information at point of care and have instant access to those records at a later time. As thousands of GP practices started to employ such systems, the resulting records were quickly used to monitor trends and performance of GP practices [54] and measure outcomes in patient care. After years of patient interactions with GPs being recorded electronically, large scale retrospective epidemiological were made possible because access to patient records were much faster compared to sourcing paper based records.

Using the first EHRs as a form of research was a success mainly because the data collected by GPs to inform patient care was mutually beneficial to inform public health. GPs not only recorded patient details important to building a picture of patient care such as diagnoses and medication, they recorded patient details using clinical coding systems to summarise their interactions. In 1990 the READ clinical coding system had matured for 8 years and was considered for use in computerised coding systems [55]. Developed from the early 1980's onwards and still being developed today, Dr James Read built a clinical coding system consisting of 250,000 codes that could be used to describe the details of a patient's medical notes. While clinical coding systems such as ICD have existed for over a hundred years [56], the READ code system was the first in the UK that could classify disease, symptoms, prescriptions and referrals in one heirachy. The real advantage of using computerised records and embedding coding systems such as READ used in GP practices, and ICD-10 used in secondary care is how EHRs can be queried using computer languages such as SQL. Patient information can be rapidly accessed and partitioned by these coding structures to create cohorts of certain disease or medication retrospectively. This is by far more efficient and reliable than having to process free text such as medical notes manually, a problem still being solved today through Natural Language Processing - discussed later on in this chapter. The electronic patient record, when combined with clinical coding systems are a potential data source for fast, large scale research.

However, a fundamental understanding of the purpose and context of why and

how EHRs are recorded must be taken into account if intended for use in research. Patient records are designed to be collected at point of contact which usually means a consultant or GP. The records are therefore a reflection of how a patient should be cared given their presentation at point of contact, and are certainly not designed for research purposes. For example, a record of a prescription does not necessarily mean the patient has adhered to a treatment plan, or a diagnosis code used by a GP could indicate a diagnosis subject to a specialist referral rather than a definite diagnosis. Even when administrative staff take more of a role in entering details of patient records, the ability to generate factually correct patient records relies on communication with consultants or specialized training to translate consultants finding into clinical codes. While technology moved forward the ability to create patient records more efficiently, the sources of error remain the same as when paper records were used. For research purposes this broadly means that any conclusions are limited by the quality of data entered into patient records, or to put more crudely - garbage in, garbage out.

Aside from data entry being influenced by how patient records will be used in-house by medical professionals, incorrect data is ingrained in patient records. Human error is a factor in any data entry tasks, but data entry in a live healthcare setting is arguably more difficult than most data entry tasks. Clinical coding is a fast growing profession within secondary care that requires a strict set of exams to qualify as a clinical coder. Their job is to sift through consultants, surgeons, junior doctors and pharmacists notes to build a patient profile and turn them into discrete episodes of care described by ICD-10 and OPCS-4 codes. This process is detective-like by nature, often having to piece together conflicting medical opinions, sifting through short-hand patient notes and assigning a subset of the 16,000 codes ICD-10 codes to describe disease and morbidity, and OPCS-4 codes to describe operations [57]. The potential for error is large without the high standard of training and continuous communication with the various healthcare professionals that are responsible for treating the patients. In contrast to the use of clinical coders, GPs are expected to enter data into patient records at point of care. While clinical audit systems such as EMIS aim to help GPs accomplish this task, it is incredibly difficult for GPs to have a working knowledge of the 250,000 code list in the READ code system while entering data and caring for patients in an average consultation time of just 11 minutes. Under these pressures it is easy to imagine why GPs may be forced to cut corners or omit certain aspects of coding. This can lead to systematic error in coding such as using codes that do not accurately describe the patient or limiting their use of READ codes to a very small subset regardless of what the patient presents with.

The electronic patient record has however become the cornerstone informing medical

practice through either research or immediate feedback of data at point of care. The explosion of research based on information in EHRs has no doubt furthered the case for more emphasis to be placed on accuracy and maintenance of EHRs and making EHRs as robust as possible. There have been incentives such as the Quality of Outcomes Frameworks (QOF) that pay GPs to use a wide variety of READ codes in clinical practice in which the effectiveness of well coded EHRs post-QOF showed reduction in mortality, hospital admissions and the improvements in the management of chronic conditions such as diabetes [58] [59].

Consequently, the demand for a patient record for both healthcare and research has evolved beyond what is recorded in primary and secondary care. Various national health registers and audits ranging from the Office of National Statistics Deaths and Births register, Congenital Anomaly Register and Information Service (CARIS) [60], Welsh Cancer and Intelligence and Surveillance Unit (WCISU) and various biobank datasets from clinical trials all feed into the patient record. Social care datasets and tertiary health programs provide useful measures of patient care outside of first and second line services, as well as administrative and demographic data such as the Welsh Demographic Service to explore geographical and social deprivation effects on health. Perhaps the most exciting addition to the patient record is genetic data due to the potential for deeper understanding inherited disease and the opportunity to develop personalised medicines. The patient record is beginning to include information that is not even found in database format or collected via traditional auditing methods - namely free texts such discharge letters and consultant reports that contain far richer data than any of the datasets previously mentioned, if it can be processed. Any dataset that can feed into the patient record is beneficial, especially for research, a prospect which becomes eve more powerful when such datasets are successfully integrated together.

1.3.2 The SAIL Databank

The linkage of big data is a corner stone of public health research. The potential to mine patient profiles from national datasets produces novel research that directly impacts policy. While randomized control trials are the benchmark for studying health interventions and drug use, retrospective, longitudinal studies produced from linked data is much cheaper, faster and statistically more powerful. There are many established data centers that take advantage of linking routinely collected data, but the largest government funded initiative for using data linkage in health research in the UK is the Farr Institute of Health Informatics Research. The Farr Institute was a collaboration of 21 academic and health institutions in the UK, where the four main data centers are co-ordinated through The Health e-Research Center at the

University of Manchester, University College London, University of Dundee and the SAIL Databank at Swansea University. Since then the Health Data Research (HDR UK) initiative was set up to continue innovation in big data linkage for research, of which new centers such as the Sanger Institute are now included as a means to bring genetic data to the patient record.

The SAIL Databank is a repository for national health datasets in Wales, hosted at the HDR UK cite at Swansea University, that contains linked anonymised health records at a patient level [61] [62]. Developed in 2006, the SAIL databank aimed to take advantage of emerging technologies to capture patient level data and provide a platform to link datasets on a large scale. National datasets such as primary and secondary care, mortality and birth records, geographic and socio-economic status all have routinely collected data that date back 25 years and when linked together produce research potential greater than the sum of it's parts. Figure 1.3 shows the different types of data held in the SAIL databank:

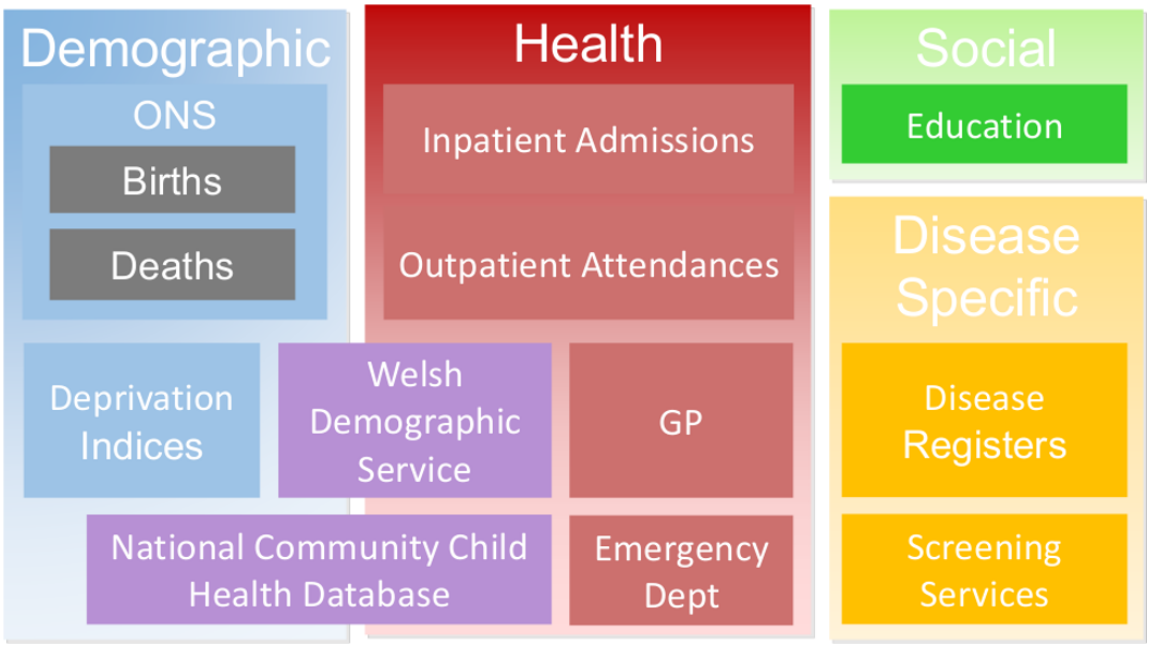

Figure 1.3: The core SAIL datasets. Each dataset can be linked anonymously via an encrypted NHS number

1.3.3 Anonymous patient records

Datasets held in SAIL are anonymised using a split file procedure. Each data provider splits their dataset into 2 parts - one containing all demographic data which is sent to a trusted third party (TTP), and the other containing only clinical data that is sent directly to SAIL. An internal system ID that bears no relation to the patient ID is the only field shared between the split files. The TTP, in this case the National Welsh

Information Service and SAIL each encrypt the split datasets, and are then combined using a shared encryption key. The result is a completely anonymised dataset that can be linked to all other datasets within SAIL.

Users can assess the repository through the SAIL gateway - a remote server with a Graphical User Interface (GUI) front-end. The SAIL databank is powered by an IBM Blue-C supercomputer which provides extremely fast database queries, capable of completing queries on databases of millions of records within seconds. The data is warehoused and made available as a repository of relational databases, in which the data can be queried using IBM DB2 Structured Query Language (DB2 SQL). As data is anonymised there is no ethical approval needed to query data. However an independent Information Governance Research Panel (IGRP) consisting of multi-disciplinary professionals in the field of health care and health care research exists, to which project studies plans are scrutinized to ensure the research question is valid and answerable using SAIL data, as well as ensuring that no individuals can be identified. Once a project is approved, data can be requested out of the SAIL gateway. An internal team of researchers view all outputs to ensure no sensitive data leaves the SAIL gateway.

1.3.4 SAIL studies

The SAIL databank has been used in a diverse range of healthcare studies. The main type of studies conducted are retrospective longitudinal studies that take advantage of millions of person-years of data across multiple health datasets, although follow up studies from patient recruitment have also been carried out.

Child Health and Births

The Wales Electronic Cohort for Children (WECC) is an e-cohort of children in Wales set up to study a range of social and environmental determinants and outcomes of child health. The WECC cohort is the largest e-cohort for children in health (804,290 children, 375,025 mothers between 1998-2008) and was built from routinely collected data in the SAIL Databank [63]. Several studies have used the WECC cohort to research health and social outcomes in children. One study found a 4 fold increase in hospital admissions for children born at a gestational age of 33 weeks (41.5 per 100 child years) compared to gestational age of 40 weeks (9.8 per 100 child years) [64]. The impact of skull fractures and inter-cranial injury was associated with poorer academic performance in Key Stage 1 assessments compared to a control group, and a higher risk of hospital admission was observed in children with a mental health disorder or living with parents that had an alcohol misuse problem recorded in GP

records [65] [66]

The Congenital Anomaly Register and Information Service for Wales (CARIS) dataset was linked to ONS birth records in the SAIL Databank to estimate the prevalence of Turner Syndrome to 1 in 4901 female births [67]. An example of bespoke datasets being linked to the SAIL Databank is a study using data from the Singleton Hospital Maternity Ward at Swansea University that were linked to primary and secondary care datasets to study the relationship of BMI during pregnancy and health utilization. Based on data of 484 pregnancies it was found that healthcare costs during pregnancy was 37% higher in obese women compared to those with normal weight. Demographic data was linked to Key Stage 1 education data showing that a clear trend in reduction of educational attainment was seen with increasing numbers of house moves, even in children that moved prior to the Key Stage 1 assessment period (< 5 years of age) [68].

Mental Health

A cohort ascertainment study using GP records in the SAIL Databank specified sets of READ codes to define anxiety and depression. Using results of the Caerphilly Health and Social Needs Survey (CHSNS), combinations of depression and anxiety diagnoses, medication and symptoms showed that high positive predictive value could be achieved, but it is likely that depression and anxiety are under reported in GP records[69]. A further analysis of GP recording of depression showed that diagnoses recorded in GP settings have declined while antidepressant prescribing has increased in adolescents, indicating GP coding habits change over time and highlights the importance of understanding reference data for epidemiology studies from routinely collected healthcare data [70].

The Suicide Information Database Cymru (SID-Cymru) was set up using mortality and secondary care data in the SAIL databank to identify 2664 cases of suicide in Wales between 2003-2011 [71]. The SID-Cymru dataset was used to obtain suicides following alcohol related emergency admissions to hospital which showed that women were at double the risk of suicide than men and that 10% of suicides took place within 4 weeks of admission [72].

Multiple Sclerosis

The UK MS Register was set up to obtain rich patient and clinically reported information on patients with MS. Patients can upload their medical data via web forms and social media, to which they have consented for this information to be used in healthcare research. All data is hosted by the SAIL Databank and is available

to be linked to existing datasets in SAIL [73]. Responses from the web-portal were used to assess how MS patients fair on the Hospital Anxiety and Depression Scale (HADS), where HADS measures anxiety and depression using a scoring system (0-7 neither reported, 8-10 mild, 11-14 moderate and 15-21 severe). The results of 4178 respondents showed that the median HADS score was 15.7, with over half of respondents score ≥ 8 for depression and just under half scoring $\geq$ for anxiety [74]. A follow study correlated increasing prevalence of anxiety and depression with increasing physical disability where anxiety or depression was reported in 38% of people with mild physical disability, 66.7% with moderate physical disability and 71% with high disability [75]. Patients in the MS Register also answered a survey to determine their generic health status from an EQ-5D in which people with MS scored 22% less than the UK mean of 82% [76]

Diabetes

A study of ONS birth records in the SAIL Databank identified 1250 pregnancies where the mother had existing diabetes and 1358 gestational diabetes in which contrary to hypotheses of "obesity programming" in children born to mothers with diabetes, little evidence of this effect was found unless the mother was also found to be in the highest weight tertile during pregnancy [77]. 1577 children between the ages 0-15 with type-1 diabetes from the Brecon Group Register were linked to hospital admissions in SAIL. The study found a 480% incidence of hospital admissions in which the incidence rate decreases 15% with each increasing 5 year age band [78].

HbA1c measurements recorded in GP records were compared before and after an incident stroke in patients with existing type-2 diabetes. 1741 diabetes patients were identified having HbA1c measurements before or after an incident stroke and were age and sex matched 1:4 to a control group of patients with diabetes that had not had a stroke. On average there was a 7.5% decrease in HbA1c measurement after the incident stroke, indicating increased monitoring post-stroke in patients with diabetes may result in better glycemic control [79].

Cardiovascular Disease

A study linking hospital records for patients admitted for acute myocardial infarction (AMI) (n = 30,633), stroke (37,888) and sub arachnoid haemorrhage (SAH) (1753) to ONS Death certificates in the SAIL Databank explored the effects of social deprivation on 30 day mortality following admission. Baseline 30 day mortality rates for AMI, stroke and SAH were 14.3%,21,4% and 35.6% respectively, however a 24%,24% and 32% increase in mortality was observed when comparing the lowest deprived quintile to the highest [80]. Statin use in patients presenting with incident acute coronary

syndrome (ACS) to the Cathlab Unit in Morriston hospital was studied to investigate if guidelines were being taken up to prescribe statins post ACS. 80% of patients were prescribed statins with simvastatin most common, however only 38% were prescribed a high dose, leading to the conclusion that statin use post-ACS is under utilized in Wales [81].

Cathlab data for patients with aortic stenosis was used to show that Transcatheter Aortic Valve Implantation (TAVI) was more effective than being medically managed in terms of prospective health utilisation and mortality over a 2 year period. Mortality rates were half that in the TAVI group (19.2% vs 41.7%) and experienced less hospital stay length (0.86 vs 1.84% person days per year) and costs within primary and secondary care was half that than the medically managed group (£6059 vs £11001) [82]. In contrast to many studies indicating worse health and social outcomes in people from deprived areas, including many SAIL studies, a study of patients with coronary artery disease (CAD) found no health utilization or treatment inequality across Welsh Index of Multiple Deprivation deciles [83].

Epilepsy

An algorithm using GP records was developed to identify cases with epilepsy using a combination of AEDs and epilepsy diagnosis codes. This algorithm was used to study prescribing trends of first line AEDs in people with epilepsy between 2000-2010. The study showed a sharp decrease in Sodium Valproate prescribing to women of child bearing age, and that recent guidelines from the MHRA to prescribe lamotrigine as a first line AED prescription had been taken up in Wales. [1]. The effects of sodium valproate, lamotrigine, levetiracetam, topiramate and carbamazepine on weight change were explored in a cohort 1423 epilepsy patients. Significant weight gain (+1% body weight) in levetiracetam and significant weight loss (-2.62% body weight) in topiramate was found, where the other 3 AEDs showed no significant change in weight [84].

A collaborative study between Manchester University and Swansea University found that people with epilepsy are twice as likely to die from suicide than people without epilepsy, are 3 times more likely to die accidentally, are 5 times more likely to die of accidental medication poisoning and are 3.5 times more likely to die of intentional medication poisoning [85]. Emergence admissions of patients with epilepsy who had attended ED for reasons specific to epilepsy were studied in which social deprivation and living alone were identified as risk factors for ED attendance and of these patients, and psychiatric co-morbidities and learning disabilities than epilepsy patients who had not attended ED for epilepsy specific reasons [73].

1.4 Natural Language Processing: Using clinic letters as a data source for research

Natural Language Processing (NLP) is a multi disciplinary field of linguistics and computer science that aims to construct computer algorithms that can automatically parse unstructured text into more manageable forms. Typically these algorithms are aimed at unstructured texts reflective of human language such as clinic letters, and a suitable task for NLP might be to automatically extract letters where a diagnosis of a certain disease is written in the text. At present, only patient records stored in structured databases are immediately accessible for healthcare research and epidemiology. For decades researchers have been able to take advantage of the codified format of these datasets to make quick gains in epidemiological research. ICD-10, READ and SNOMED-CT codes can quickly be used to manipulate cohorts of patients with structured query languages, but other forms of medical information are slowly starting be adopted into a big data patient record. While huge efforts go into producing an organised patient record at point of care, a large amount of patient information is still recorded only by free text. These include consultant notes, discharge letters, GP correspondence, radiology reports and even structured databases may contain so fields that store free text. Free text in is seen as a rich source of patient information not found in the structured patient records. but remains a challenge to bundle the information into a database format that lends itself to data manipulation. While there is no agreed way to process free text correctly, a number of scientific disciplines have come together to address this problem.

Natural Language Processing (NLP) is a discipline that aims to process free text into easily accessible information, such as a summary or database. NLP draws upon advances in statistical theory, machine learning, artificial intelligence and computer science to create programs or models that understand the nuances of human (natural) language.

Early NLP techniques developed in the 1960's and 1970's used rule sets and pattern matching techniques to infer meaning from text [86],[87],[88]. These early works focussed on creating extensive hand crafted rules that used the relationships between text units such as nouns, verbs and adjectives to extract structured items of information. Using complicated human rule sets to parse language relies solely on human knowledge of text to predict patterns in advance that would capture items of interest and the context they are found in. The scale of this problem has required more sophisticated approaches to be developed. The advent of machine learning in the 1980's provided ways to parse text not through fixed sentences, but through teaching

a computer to learn the role of each word in a sentence [89]. Part Of Speech (POS) tagging abandons highly conceptual human rule sets declared to a computer algorithm prior to analysis, and analyses the relationships of verbs, adjectives and nouns in relation to a dictionary of words of interest i.e. medical terminology. Machine learning allows a ground truth such as phrases known to confirm disease or symptom, be used to train an algorithm to recognize the patterns between each word in a sentence and words of interest. This machine learning approach does not rely on complex rules sets, but rather learns the language used to describe cohort characteristics. Structured concepts such SNOMED terminology can be "mined" out of free text, including information that would go unseen or undefined by prior rule sets. In the methods chapter various NLP techniques are described and tested to define characteristics about patients with epilepsy from clinical notes.

1.4.1 Part of Speech Tagging

Part of Speech (POS) Tagging involves assigning word classes such as verbs, nouns, adjectives as well as more complex classes such as qualifiers, prepositions and adverbs to tokens in text. This cannot be achieved by a simple lookup because words can be assigned as different word classes based on context. For example the word haemorrhage in the phrase "there is a chance she will haemorrhage" is classed as a verb, but used in the phrase "she has had a haemorrhage" it is classed as a noun. Assigning the correct word class for each token is crucial to NLP tasks such as information extraction where word classes can be built into rules or machine learning processes as basic building blocks that help identify concepts within text.

The development of POS tagging has relied on analysing large corpora of many of documents such as the Brown Corpus [90] and the Cobuild project [91] so that common data sources can be used to both develop and benchmark POS tagging algorithms. The Brown Corpus consisted of 1 million words of English prose from randomly selected scientific publications and was used to develop custom tagsets to encapsulate detailed tags that extend beyond basic word classes. The Brown Corpus was manually annotated over many years and served as a target tagset for computerized algorithms. The first attempt to develop a computerised algorithm from the Brown Corpus used human logic such as an article followed by a noun can occur i.e. "Dr House", but in general an article followed by a verb does not occur. This approach yielded an initial accuracy of 70% [92].

Both larger corpora and machine learning approaches were adopted to increase POS Tagging accuracy. Hidden Markov Models - a probability state classifier was tested on the Lancaster-Oslo-Bergen Corpus [93] of British English. Hidden Markov

Models were able to take bi-grams, tri-grams and n-grams as input from the manually annotated POS tags to calculate the probability of a TAG for each element of an n-gram. This was a popular method of POS-tagging which used the scalability of dynamic programming to produce fast and accurate taggers [94], [95], [96]. Some of the most widely used POS taggers are rule-based. The Brill tagger [97] uses a set of rules that recursively updates tags during repeated phases. An initial phase is run generating most likely tags, in which set of conditions are imposed to correct each tag. This process is repeated until a threshold is met in terms of the proportion of tags corrected. The Penn Treebank POS tagset project annotated POS tags over a corpus consisting of 4.5 million American-English words from the Brown Corpus and the Wall Street Journal using a combination of Church's PARTs method [94] and manually correcting any errors, in which this method was measured to be twice as fast as a fully manual annotation method [98]. While most POS tag studies up to this point published an automated algorithm for POS tagging, the original Penn Treebank paper does not describe such a method but provides a widely used benchmark corpus for training POS taggers (machine learning, rule based or hybrid) for unseen samples of text. Table 1.4 shows the PENN treebank tags that are commonly used to tag texts:

Table 1.4: A table of PENN treebank POS tags. `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

Tag	Description	Tag	Description
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

1.4.2 Shallow parsing

Shallow parsing or " text chunking" is the process of grouping tokens into n-grams such as sentences, phrases as well as further sub-units commonly known as chunks. The sentence "The radiologist was able to observe the tumours visible against the background" can be split up into five chunks "The radiologist" "was able to observe" "the tumours" "visible" "against the background", in which one chunk usually provides context to other chunks in close proximity. Chunks are defined as sub-units of text that contain a "potential governor" - a handle placed at the rightmost part of a chunk [99]. The potential governors in the above phrase would be "radiologist", "observe", "tumours", "visible", "background".

The Brill POS tagger algorithm and PENN treebank tagsets were used as input to develop a rules based noun phrase chunker that defined noun phrase chunks such as "she has epilepsy" and "her seizures are frequent" [100]. Further algorithms were introduced to classify verb phrases (VP), prepositional phrase (PP), adjective phrases (ADJP) and adverb phrases (ADVP) [101]. The Conll-2000 shared task [2] defined 13 different chunk types as targets for classifiers to learn by providing a fully annotated version of the Penn Treebank Corpus. These chunk types are shown in table 1.5.

Table 1.5: The chunk types defined as part of the CONLL-2000 shared task [2]

Chunk Type	Definition	Examples
NP	Noun phrase - phrases beginning with a noun	"Mr Jones", "He was", "a year"
VP	Verb phrase - phrases beginning with a verb	"may want to increase", "could be a", "broke the"
PP	Prepositional phrase - a phrase to place context to nouns	"at night", "because of ", "due to"
ADVP	Adverb phrase - pre or post modifier to a verb or verb phrase	"very well", "overdosed earlier", "quickly"
SBAR	Subordinated clause - conjunction between other phrase types	"so that", "even if", "until"
ADJP	Adjective phrase - phrases beginning with an adjective	"upset with her seziures", "prolapsed disk"
PRT	Particles - verb/adverb attached to non inflected words	"look up", "on and off", "in and out", "get out"
CONJP	Conjunction phrase - multiword conjuntions to list additional phrases	"as well as", "not only", "but also"
INTJ	Interjection - phrase containing an ubrupt remark	"oh", "alas!", "good grief!"
LST	List marker - denotes a list	"firstly...", "1.", "lastly", "a,b,c"

Eleven algorithms were submitted to the Conll-2000 shared tasks and the most accurate algorithm achieved an F-score of 93.4% [102]. Shallow parsing algorithms have usually focussed on employing machine learning and statistical learners such as Hidden Markov Models, Support Vector Machines, Naive Bayes and Conditional Random Fields [103], [104], [96], [105].

1.4.3 Named Entity Recognition

Named Entity Recognition (NER) is the process of tagging specific words or phrases and labelling them into entities and such as persons, addresses, identification numbers, diseases, symptoms or medication. NER tasks usually involve mapping entities to a user specified dictionary. NLP tasks focussing on healthcare typically map entities to medical ontologies such as as MetaMap, The Unified Medical Language System, SNOMED-CT and ICD-10 [106],[107]. NER has advanced significantly since scientific events and competitions were set up in the 90's, with the Messaging Understanding Conference (MUC-6) [108] set up specifically to bring together groups researching

NER. Many other annual conferences such as HUB-4 [109], the Information Retrieval and Extraction Exercise (IREX) conference for Japanese NER [110] and the Automatic Content Extraction (ACE) [111] conference have since been set up to maintain a focus on NER and are still running today.

The main challenges for NER tasks are:

- **Word ordering** - the order of words within a phrase can change the meaning of entities to be labelled. For example the phrase "she had a blood pressure check" identifies that this person has had a blood pressure measurement, but "to check her blood pressure" implies this is something that should happen in the future and not be assumed this person has had a blood pressure measurement at the current time.
- **Inflexions** - suffixes and prefixes can change to indicate a different meaning for a word. For example "big/biggest", "vomit/vomited", "informed/uninformed" etc.
- **Homopgraphs** - the same words can have different meaning. The word "fine" can mean something is normal or it can describe a procedure or wound. Research in word-sense disambiguation is dedicated to addressing homographs.
- **Synonymy** - the opposite of homographs in that multiple entities can mean the same concept i.e. "Focal seizures/ partial seizures".
- **Negation** - certain trigger words such as "not/no/never/unlikely" indicate that an entity has not been found, and therefore should not be attributed to a positive finding in text. It is common to tag entities with negation status.
- **Word relationships** - and extension of negation. The surrounding words around an entity dictate it's context. For example the phrase "if the results are positive, prescribe Lamotrigine" suggest a prescription of Lamotrigine is only positive given the context of a result prior to the word Lamotrigine.
- **Temporal qualifiers** - describing a temporal feature attached to an entity involves measuring the proximity and order of trigger words relating to past, present or future tenses.

1.4.4 NLP tools and software

Table 1.6: A list of common NLP software

Toolkit	Description
Stanford Natural Language Processing Toolkit [112]	Open source Stanford CoreNLP toolkit. Contains standard NLP applications (POS taggers, chunker, NER). Developed by Stanford University.
NLTK [113]	Natural Language Toolkit developed in Python
Apache UIMA [114]	Open source Java based Unstructured Information Management Application developed by the Apache Software Foundation
Apache OpenNLP [115]	Open source NLP toolkit adopted by the Apache Software Foundation and developed by the open source community
Apache cTakes [116]	Open source healthcare information extraction system developed by the Mayo Clinic
GATE [117]	The General Architecture for Text Engineering. An open source NLP architecture developed at Sheffield University. Contains a variety of standard NLP applications as well as user contributed plugins. Has a rich GUI and can be run in embedded systems.
IBM WATSON Content Analytics [118]	A proprietary NLP product produced by IBM. Makes use of the UMIA framework and has a rich GUI.
Spacy [119]	A python library that supports many NLP tasks including deep learning and pre-annotated corpora
Open NLP (R package) [120]	An R package that interfaces to the Apache OpenNLP tools written
TM (R package) [121]	A text mining framework written in R. Contains many NLP applications
Apache UIMA RUTA [122]	A UIMA framework for executing rule based scripts for NLP applications
Gensim [123]	A python library for vector space modelling of large text corpora. Developed at Mararyk University
Word2Vec [124]	An algorithm to produce words embeddings for topic modelling. Developed at Google trained on a Google News corpus.
GloVe [125]	An algorithm developed at Stanford University to produce word embeddings

1.4.5 Validating NLP algorithms

Any algorithm proposed for an NLP task can be measured against a human annotator, or multiple annotators. Annual shared tasks such as CoNLL specify a problem statement as well as providing or using large, annotated corpora to compare against algorithm submissions. Comparison between a human annotator and an NLP algorithm on a binary classification results in true positives (target class labelled as target class by algorithm), false positives (non-target class labelled as target class), true negatives (non-target class not identified as target class) and false negatives (target class not identified as target class). By assigning each item identified by both human annotator and an algorithm as a true positive, true negative, false positive or false negative, the overall accuracy can be measured in various ways. In NLP tasks precision, recall and F1-score [126] are widely used and are defined as as:

$$Precision = \frac{TP}{TP * FP}$$

$$Recall = \frac{TP}{TP * FN}$$

$$F1score = \frac{Precision * Recall}{Precision + Recall}$$

Recall is a measure of the proportion of all possible target classes identified by the algorithm, where precision is a measure of the proportion of classes identified by the algorithm are true. The F1-score is the mean of precision and recall and is usually reported as the overall accuracy.

NLP algorithms can be compared against multiple annotators by scoring algorithms against the agreement of multiple annotators, as measured by Cohen's Kappa statistic. A Kappa-like statistic was presented by Galton in 1892 as a method of identifying fingerprints using human raters where a match was identified if a certain percentage of raters could agree that a unseen sample matched that of a known fingerprint. Cohen's Kappa statistic was formally introduced in 1960 [127] and is defined as:

$$\kappa = 1 - \frac{p_o}{p_e}$$

where p_o is the probability of an observed class by multiple raters and p_e is the random chance of observing all possible classes. κ therefore represents the class agreement by multiple raters normalized by the probability of all classes where the class agreement

is 100% when $\kappa = 1$. Agreement measures such as Cohen's Kappa statistic are employed for inter-annotator agreement for NLP tasks [98]. This is useful to estimate the relevant difficulty of a task in which some tasks may yield low inter-annotator agreement and therefore sets a lower expectation for an algorithm's ability to perform the task.

Human annotation is time consuming and has been listed as one of the major challenges in NLP for healthcare applications. Annotation tools such as BRAT[128] ehost [129] provide a method of rapidly annotating documents in which data files (xml, custom output) store all the annotations in each document. These output files are designed to be read into NLP applications to directly compare annotations picked up by humans annotators and those identified by the algorithm. Thus allows for computation of accuracy measures to be automated and therefore many NLP models can be validated automatically.

1.4.6 NLP clinical information applications

There have been extensive studies focussing on creating clinical extraction NLP systems for specific disease areas. The NLP system developed as part of the Linguistic String Project (LSP) was one of the first systems to be used for clinical information extraction [130],[131]. Developed in 1987, a qualitative study first described a system comparing human annotated notes with an automated extraction system in radiology reports, reports and discharge summaries [132]. The Medical Language Extraction and Encoding System (MEDLEE) was developed to detect disease mentions from radiology reports and was scored against physicians' interpretations of 230 radiology reports, achieving 87% recall and 78% precision across all disease mentions [133],[134]. Medlee has also been extended to detect breast cancer from mammogram reports, as well as forming the basis of the GENIES that extracts molecular pathways from journal articles [135],[136].

The SymText NLP application [137] was used to detect bacterial pneumonia from chest X-ray scans. The majority vote of 3 physicians was used to manually score 292 X-ray reports from the LDS hospital in Utah and compare the annotations to extracted annotations using SymText. The SymText system uses a syntactical rule-based approach combined with a Naive Bayes Classifier to extract 76 different radiographic findings and 89 different diseases from chest x-ray reports [138]. Pneumonia concepts were split into 4 categories: acute bacterial pneumonia, infiltrate pneumonia, aspiration pneumonia and support pneumonia. Physician average precision and recall for acute bacterial pneumonia compared favourably with that of three physicians and outperformed annotations by lay-persons.

The i2b2 project used cTAKES and HITex (Health Information Text Extraction) to extract Crohn's disease, Ulcerative Colitis, multiple sclerosis (MS), and Rheumatoid arthritis [139]. A recent study on patients with known MS identified from electronic healthcare records used NLP techniques to accurately extract attributes specific to MS, namely: Expanded Disability Status Scale, Timed 25 Foot Walk, MS subtype and age of onset [140]. A study used clinic letters, available at www.mtsamples.com, to determine whether sentences containing disease and procedure information were attributable to a family member using the BioMedICUS NLP system and variety of phenotype data was extracted from 300 randomly chosen journal titles [141],[142]

There have also been several epilepsy based NLP studies and applications developed. The rule based Epilepsy Data Extraction and Annotation (EpiDEA) system was developed to extract epilepsy information from epilepsy monitoring unit discharge summaries. Categories of information included EEG pattern, past medications and current medication extracted from 104 discharge summaries from Cleveland hospital [143]. The rule-based Phenotype Extraction in Epilepsy (PEEP) pipeline was developed to extract epileptogenic zone, seizure semiology, lateralising sign, interictal and ictal EEG patterns from epilepsy monitoring discharge summaries as Cleveland hospital [144]. A machine based learning NLP pipeline was also developed to identify a rare epilepsy syndrome from discharge summaries and EEG reports [145]

Medication extraction has also been an area of interest for NLP research. The Medication Information Extraction system (MedEx) was developed to extract prescription information, including drug name, dosage, strength and frequency. On a validation set of 50 discharge summaries and 25 clinic notes MedEx was able to achieve an F-score of 93.2% and 90% respectively. CPRD prescription data was used to validate 220 prescriptions from anonymised GP records, in which a rule based system was able to achieve 91% accuracy [146] A rule-based NLP application was developed and applied to the NHS Scotland Prescribing Information System (PIS) [147]. On a validation set of 15,593 prescriptions the system achieved 94.7% accuracy when extracting full prescription information and was able to generate structured outputs for 92.3% of 458,227,687 dosage instructions in the PIS.

There has been a particular focus on mapping extracted terms from NLP applications to existing clinical ontologies. This is particularly important for linking extracted terms to routinely collected data that use ontologies such as SNOMED-CT and ICD-10. A study using 23 citations in the Annals of Internal Medicine and the expert opinion of three physicians described a system that mapped 89% of terms identified by the physicians to MeSH terms [148]. Various studies have focussed on mapping extracted terms to UMLS concepts for it's ability to map to many other ontologies such

as SNOMED-CT, READ and ICD-10. A study evaluated the use of UMLS concepts as look up terms for congestive heart failure, chronic obstructive pulmonary disease, acute bacterial pneumonia, neoplasm, pleural effusion without congestive heart failure compared to manually curated lists by clinicians. This study found that using UMLS concepts improved retrieval of terms over than of clinician specified terms[149],[106] and further studies have used a variety of patient information sources such as patient charts, MEDLINE citations and surgical notes [150],[151],[152]. The MetaMap project used UMLS codes to retrieve medical terms from Medline citations and compared the results to the use of NLM MeSH terms. By using UMLS concepts they reported a 14% improvement on using MeSH terms and is now one of the most widely used algorithms in the NLP community to patient information to UMLS codes. SNOMED-CT terms are widely used in patient records and several studies have proposed methods to map free text to SNOMED-CT. Several studies have used veterans' patient records as a source to validate varying accuracy across a variety of conditions including acute renal failure, venous thromboembolism, pneumonia, myocardial infarction, sepsis (82%,38%,59%,64% and 89% accuracy respectively).

1.4.7 Genetic Mutation

Single Nucleotide Polymorphisms (SNPs) are the most prevalent type of genetic mutation in the human genome, with most genes having multiple non-synonymous SNPs and accounting for around 90 percent of genetic variation [153], [154]. SNPs appear in both the coding and non-coding region of the genome in which both have been associated with disease phenotypes. The human genome consists of around 3.2 billion pairs of DNA, with SNPs appearing every 1000-2000 base pairs (between 2-3 million per genome) [155] and on average one person will have 250-300 potentially damaging SNPs that are directly or indirectly associated with disease [156],[157]. SNPs fall into three broad categories.there Nonsynonymous (or missense) SNPs represent a mutation in nucleotide base triplets that cause a change in the amino acid the nucleotides code for. Synonymous SNPs, while containing mutations in nucleotides, do not cause a change in the resulting amino acid. This is due to the fact that there are multiple combinations of nucleotide triplets that code for the same amino acid. Frameshift SNPs involves a deletion or insertion in the amino acid sequence. Multiple insertions and deletions can also occur, but SNPs are estimated to contribute to over 90% of all known genetic mutation [154] of which 50% of SNPs have shown to be common (20% of the population have a given SNP) [158].

SNPs can fall anywhere in the genome meaning they are found in regions of DNA that do not code for protein sequences (99% of the human genome) and in coding regions that do code for proteins, known as the exome. Non coding regions have controversially

been labelled in the past as "junk DNA", where it was assumed that mutations in non-coding regions of the human genome would not contribute to disease, however the ENCODE project showed that over 80% of non-coding regions serve some purpose such as promotors, enhancers and silencers - all important roles in gene regulation [159]. Despite SNPs in non-coding regions shown to cause diseases such as pancreatic agenesis [160], juvenile idiopathic arthritis [161] and auto-immune conditions [162], the majority of genetic mutation research, especially in non-synonymous SNPs has been focussed on coding regions in the exome as this demographic has been shown to be responsible for around 50% of human inherited disease [163].

1.5 Pathogenicity of SNPs

Pathogenicity in terms of genetic mutation can be described as the ability of a genetic mutation to cause disease. The American College of Medical Genetics and Genomics stated that pathogenic mutations can be determined on two types of evidence: a variant has been previously reported as the cause of disease, or the variant has not been reported as the cause of disease but is expected to be declared as such in future [164]. These categories are broad and problematic for the definition of pathogenicity as illustrated by a study of 402 published severe disease mutations showing that 27% of these were either common or lacked direct evidence for pathogenicity [165]. Further guidelines such as those proposed by The US National Human Genome Research Institute have suggested 5 categories of mutation associated with causation of disease [166]:

- Pathogenic: contributes mechanistically to disease, but is not necessarily fully penetrant (i.e., may not be sufficient in isolation to cause disease).
- Implicated: possesses evidence consistent with a pathogenic role, with a defined level of confidence.
- Associated: significantly enriched in disease cases compared to matched controls.
- Damaging: alters the normal levels or biochemical function of a gene or gene product.
- Deleterious: reduces the reproductive fitness of carriers, and would thus be targeted by purifying natural selection.

These definitions are far more reflective of that fact that genetic mutation and it's association with disease can range from single-gene, single-mutation causation of disease (high penetrance), such as in cystic fibrosis [167] or Huntington's disease [168] than can be discovered through familial inheritance studies through to common-disease,

common-variant associations where multiple mutations common in the population sum to causative effects (low penetrance) such as Alzheimer's disease [169] that require large-scale GWAS studies to identify disease risk. The ClinVar dataset curates a record of all variants that can be defined as pathogenic, likely pathogenic, likely benign, benign and uncertain following review of aggregated submissions and publications [170]. Other databases such as dbSNP [171] and Uniprot/Swissprot also record pathogenicity/benign status where the entry point in terms of association to disease meets at least the minimum requirement as set out previously in [166], that is, a variant deemed as pathogenic at least contributes to a disease phenotype.

1.6 Whole genome sequencing and the need for SNP prediction paradigms

Frederick Sanger introduced dideoxynucleotide sequencing, or Sanger sequencing of DNA [172] were manual techniques harnessed to sequence individual genes and cells. More sophisticated techniques such as shotgun sequencing were introduced in which bacterial genomes such the influenza genome (2 million DNA base pairs) was able to be sequenced [173], with the first eukaryote genomes, a strain of yeast - Saccharomyces cerevisiae (12 million base pairs) and a type of worm - nematode elegans (100 million base pairs) were sequenced [174],[175]. Next Generation Sequencing technologies were developed to allow high-throughput sequencing of large genomes such as the common fruitfly - Drosophila melanogaster (135 million base pairs) [176] and finally in 2001 the human genome (3 billion base pairs), costing an estimated £750 million US dollars [177]. The cost of whole genome sequencing has vastly decreased since Craig J Venter sequenced his genome in 2001 as part of the Human Genome Project from £750 million US dollars to £1000 US dollars today. This remarkable decrease in cost is owed to advances in sequencing technology over the past 15 years, and to highlight this, Figure 1.4 shows the decrease in whole genome sequencing when compared to what the decrease might be if following Moore's Law.

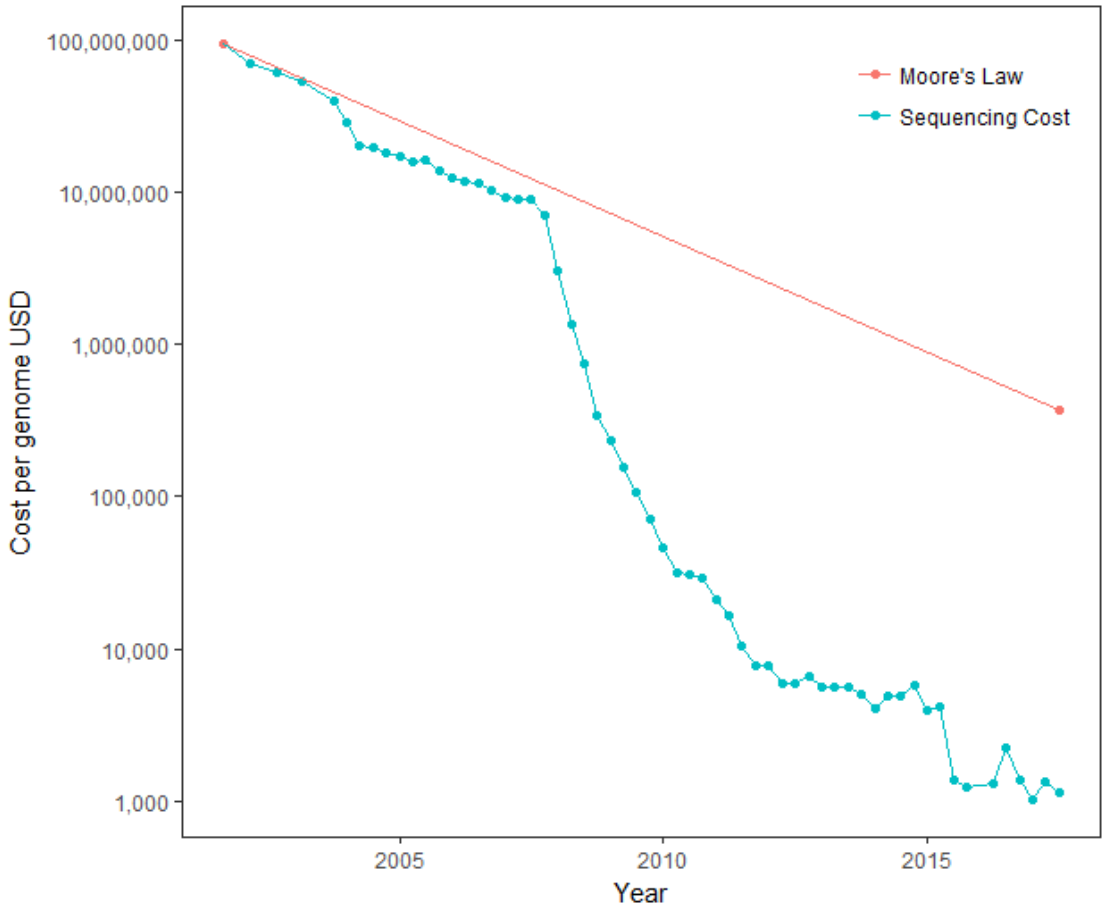

Figure 1.4: Decrease in whole genome sequencing since the Human Genome Project when compared to the expected rate of decrease following Moore's Law

The relatively low cost of whole genome sequencing has enabled whole genome / exome cohort projects such as The 1000 Genomes Project [178], The 100,000 Genomes Project [179], UK Biobank [180], ExAC [181] and aggregation services such as the Genome Aggregation Database (gnomAD) that contains 15,496 genomes and 123,136 exomes for unrelated individuals. The number of variants processed from whole genome sequencing outstrips the ability for genetic research to comprehensively study each and every variant through traditional laboratory techniques [182]. Typically whole genome sequencing one person will generate 3,000,000 variants and whole exome sequencing will generate 30-50,000 variants and it is therefore vital to find ways of focussing on variants likely to impact disease in advance of more thorough analyses [183].

1.7 Common features used in prediction of pathogenicity

There are a variety of hypotheses that help when considering the pathogencity of SNPs. A study of 561 disease causing SNPs and from the SWISSPROT database that contained details of 2D and 3D structure found that over 70% of these SNPs were found in structurally important regions such as binding sites and sites with low solvent accessibility [184]. The allele frequency on non-synonymous SNPs that caused disease was found to be lower than that of non-synonymous SNPs, indicating that nature selects against pathogenic SNPs and that disease causing SNPs may be found at sites that are conservative [153]. The BLOSUM62 matrix was developed to produce a system that scored all possible 210 amino acid substitution of the 20 standard amino acids based on an alignment of 500 protein sequences (BLOSUM62 matrix shown in Figure 1.5).

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Figure 1.5: The BLOSUM 62 matrix where higher scores indicate higher frequency of substitution. Each amino acid substitution is scored in accordance to it's frequency, where lower frequency substitutions are said to be conserved and potentially selected against by natural selection.

On average 5 of the possible 19 substitutions for a given amino acid has found to be non-conservative and is a potential predictor for SNPs associated with disease [153]. A study comparing the distributions of 1169 disease associated non-synonymous SNPs and 741 neutral SNPs found a significant difference when plotted against BLOSUM62 scores inferring that lower frequency substitutions are associated with pathogenicity [185]. The Sifting Intolerant From Tolerant predictor[186],[187] was

developed from the hypothesis that position specific information of where a SNP is found might be a predictor of disease/neutral status given that the plasticity of proteins can change across sequences [188], [189],[190]. To test this, the SIFT study took datasets that collected variants found in the Laci, HIV-1 and T4 lysozyme proteins [191],[192],[193],[194] and performed a multiple sequence alignment to other proteins in their respective families using PSI-BLAST [195] and compared disease/non-disease SNPs in these proteins to the position of the multiple sequence alignment. The results showed that the accuracy of SIFT was 66%, 86% and 45% for the SNPs in each protein respectively and that using position specific information from multiple sequence alignment was able to correctly identify the disease status of 14% more SNPs overall when compared to using BLOSUM62 scores. The study showed that position specific information was however more important in some proteins than others, in particular the low prediction accuracy of the T4 lysozyme proteins.

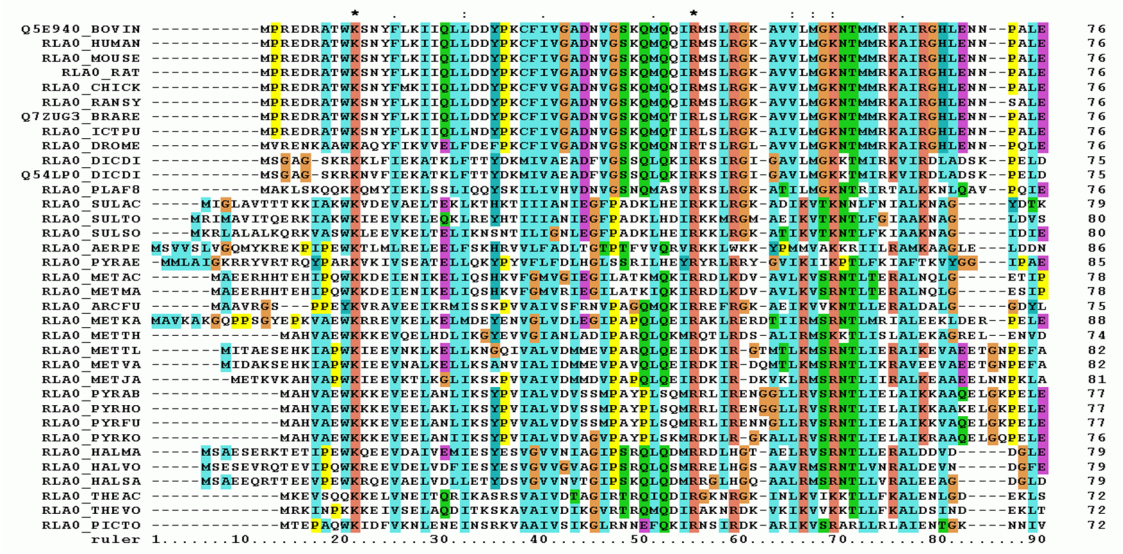

Figure 1.6: A multiple sequence alignment of transmembrane proteins from different species. Conserved regions are in red where alignment of different proteins shows no difference in amino acids across all proteins in this position. Conserved regions are therefore hypothesized not to tolerate geneitc variation and are deemed hotspots for pathogenic mutations.

The PolyPhen-2 prediction server was developed with a range of features that included multiple sequence alignments and molecular function information [196]. Prominent features used by PolyPhen was difference position specific scores for both wild type and mutant based on multiple sequences alignments [197], predictions of functional region such as transmembranes [198], coils [199] and peptide signals [200] as well as known 2D and 3D structure from DSSP and PDB [201],[202]. The study used a Naive Bayesian classifier, a type of machine learning classifier, to take these features as input and classify into three categories: benign, possibly damaging and probably damaging.

The HumDiv (3,155 damaging and 6,231 benign SNPs) and HumVar (13,032 damaging and 8,946 benign SNPs) datasets, datasets annotated with damaging/benign status from Uniprot, were used for training/testing of the Naive Bayes classifier using a 5-fold cross validation technique where training and testing sub-sets were randomly sampled. The PolyPhen method was able to achieve 93% and 72% true positive rates for damaging SNP detection using a 20% false positive rate on the HumVar and HumDiv datsets respectively [203].

1.7.1 Machine learning classification for SNP pathogenicity

Machine learning classification is a staple of SNP prediction techniques. They are used effectively when a large number of SNPs in both benign and pathogenic classes are used to train a model with features hypothesized to be useful in discriminating between pathogenic and benign SNPs. Machine learning techniques use the interaction of every feature against all other features to make a prediction, meaning that even features not thought to have a strong influence on classification may be useful to increase accuracy by a small amount overall, or increase accuracy substantially in subsets of samples that do not adhere to common hypotheses such as position specific sequence information or structural features.

The PolyPhen server uses a Naive Bayes classifier, but many other machine learning techniques have been used in SNP prediction as shown in Table 1.7.

Table 1.7: Commonly used SNP prediction programs that utilize machine learning

Algorithm	Citation	Machine Learning Classifier
DANN	[204]	Convolutional Neural Network
MetaLR	[205]	Logistic regression
MutationAssessor	[206]	SNP prediction application
REVEL	[207]	Random Forest
FATHMM	[208]	Hidden Markov Model
PolyPhen	[196]	Naïve Bayes
CADD	[209]	Support Vector Machine
MetaSVM	[205]	Support Vector Machine
VEST3	[210]	Random Forest
PHD-SNP	[211]	Gradient Boosting Learner
PolyPhred	[203]	Decision Tree

1.7.2 SNP datasets

A number of databases provide pathogenic/benign status. There are many public databases that record SNPs from population studies such as the National Center for Biotechnology Information (NCBI) SNP database, dbSNP [171], ClinVar [212] and the Human Genome Variation database, HGVBase [213]. Other databases focus on curating SNPs known to be associated with causing disease (rather than pathogenicity) such as the Online Mendelian Inheritance in Man (OMIM) [214], the Genetic Association Database (GAD) [215] and The Human Gene Mutation Database (HGMD) [216]. One of the most widely used datasets along with HGMD (FATHMM, REVEL, CADD, DANN) is the humvar dataset `https://www.uniprot.org/docs/humsavar` curated by Uniprot, which as of June 2018 contains over 70,000 SNPs with known status and has been used to train classifiers such as PolyPhen, VEST3, MetalR, MetaSVM and PHD-SNP.

Currently the REVEL classifier has the highest published accuracy when compared with other existing scores in two independent test sets, Clinvar and SWISSVAR, achieving AUC measures of 0.83 and 0.95 respectively. REVEL is perhaps unique in that it only uses a relatively small set of 13 features, none of which are derived from protein features such as secondary structure or multiple sequence alignments, but instead it uses the output of 13 SNPs prediction classifiers and combines the results using a Random Forest classifier [207]

1.7.3 Bioinformatics software and annotation programs to obtain SNP features

As well as obtaining SNPs with known pathogenicity status, it is important for machine learning classifiers to obtain many features such as multiple sequence alignments and region annotation. There are various bioinformatics pipelines that can help source these features. There exists many multiple sequence alignment programs that can be used to obtain conservation measures at each position of a sequence. These include BLAST [217], PSIBLAST [195], MUSCLE [218], CLUSTAL [219], UBLAST [220] and HMMER [221]. In addition to multiple sequence aligners for conserved regions of sequences, there are various tools exist such as GERP (Genome Evolutionary Rate Profiling) and phyloP [222], phastcons [223] and SiPHy (Site Specific Phylogenetic analysis) [224] that compute conservation as a function of evolutionary selection over all positions of protein sequences.

Region annotation is useful to detect if a SNP falls with certain functional regions such as transmembranes or binding sites. While databases such as DSSP and PDB

exist that catalogue such features, there are many programs that compute or predict such features. The PSIPRED server can be used to predict if positions in an amino acid sequence are likely to be in a coil, helix or sheet category of secondary structure [225], and various programs can predict functional regions such as binding sites [226], transmembrane regions [198] and protein-protein interaction sites [227].

There are also annotation aggregation services that take a SNP as input (usually rsID, BED or HGVS format) and query large pre-computed databases to gather a vast range of features that include multiple sequence alignments, regional annotations, scores from many prediction software (dbNFSP [228]) and allele frequency measures computed as part of 1000 genomes, exAC and gnomAD. Three commonly used aggregation tools are Variant Effect Predictor [229], snpEFF [230] and Annovar [231]. Such tools are available as web services and standalone programs and particularly useful in SNP annotation because their databases storing many SNP features are kept up to date, making them a one-stop shop for SNP annotation.

1.8 Chapter Summary

This thesis is summarised as follows:

- Epilepsy is prevalent in 1% of the population and can be caused by inherited genetic mutation of acquired through lesions in the brain due to injury. Epilepsy can be presented in many sub-types and seizure types of which many different types are attributed to inherited genetic mutation in certain genes. Severity of epilepsy can range in terms of seizure frequency, and various treatment regimes using anti-consultants exist to control seizures

- The burden of epilepsy can be measured in other ways than seizure frequency. Patients living with epilepsy are known to have lower socio-economic status and those with poorly controlled epilepsy perform poorer in education than their peers.

- Epilepsy research in epidemiology is restricted to using routinely collected healthcare records that often contain rich data desired for impactful studies in epilepsy. Large data banks such as the SAIL Databank can provide data linkage across multiple health and social care datasets, however this still does not solve the problem of data shortage for people with epilepsy

- Natural Language Processing is an emerging field in healthcare that has benefited from machine learning and the explosion of big datasets. Natural Language Processing offers the potential to collect data form clinic letters that does not

end up in routinely collected datasets.

- The emergence of Next Generation Sequencing has enabled genetic research to be conducted on a much larger scale and help with genetic discoveries. The sheer volume of data recorded poses a problem in how to make sense of millions of variant data per individual. Consequently prediction paradigms are necessary to prioritise which variants require thorough genetic analysis to make genetic research cost effective.
- Various prediction techniques have been developed with the aid of machine learning and adopted for genetic research. Such programs exist as part of larger pipelines that filter pathogenic variants from benign variants, however many lack the specificity to prevent large amounts of benign variants being needlessly analysed in molecular assays.
- Linked data of routinely collected data, clinic letters and genetics will play an important role in the future of healthcare research, and efforts must be made to facilitate the linkage of emerging data types at scale.

1.9 Summary of aims and objectives

A summary of the aims and objectives are:

1. Using routinely collected healthcare data stored in the SAIL databank to identify persons with epilepsy
2. Link primary care records of persons with epilepsy to other datasets within the SAIL databank to study societal burden of epilepsy and the effects of antiepileptic drugs
3. Explore the possibilities of incorporating clinical free text into existing patient records by using NLP techniques. Target information will consist of data that is difficult to obtain or non-existent within the SAIL databank
4. Investigate various methods of determining or predicting pathogenicty of SNPs, in particular to study SNPs found within epilepsy associated genes.
5. Create an algorithm that accurately predicts pathogenicity for missense SNPs in epilepsy to aid SNP prioritisation for downstream structure/functional analysis

Chapter 2

Methods

The following chapter describes the methods and materials used to mine health trends in data from the SAIL databank, extract information from unstructured clinic letters using NLP building a pipeline for assessing the impact of SNPs, with a focus on how the outputs of NLP and SNP analysis could potentially enrich data audited from routinely collected healthcare records. The SAIL databank holds a large all-Wales database of routinely collected electronic healthcare records that can be potentially be enriched by linking bespoke datasets, such as clinic letters (unstructured text) and whole genome/exome (NGS) data. Patient records at the SAIL databank anonymous - specifically by use of an encrypted NHS number, so that these records can be linked anonymously across different datasets i.e. GP records and hospital admissions. The NLP methods described in this chapter and chapter 4 are intended to be used to extract information from clinic letters using an automated computer algorithm in which the results could be linked to the SAIL databank and provide rich patient information that doesn't exist in SAIL. Similarly, the SNP pipeline described in this chapter and chapter 5 is designed to take in SNP information, filter it on various criteria and output the data in a format that can be linked to patient data in the SAIL Databank. Much of the code underpinning the methods is written in SQL, R, perl, bash for calling open source bioinformatics programs, as well as an array of UNIX/GNU programs such as *sed* and *awk*. Complete code is found at `www.github.com/https://github.com/arronlacey/Epilepsy-GATE-app` (Chapter 3).

2.1 The SAIL Databank

The SAIL databank holds many routinely collected electronic healthcare databases, where a person's records can be identified and linked between datasets by an encrypted

NHS number called an ALF (Anonymous Linking Field). Figure 2.1 shows the anonymisation and linkage process that ensure SAIL uses encrypted NHS numbers or Anonymous Linking Fields (ALFs) to perform linkage.

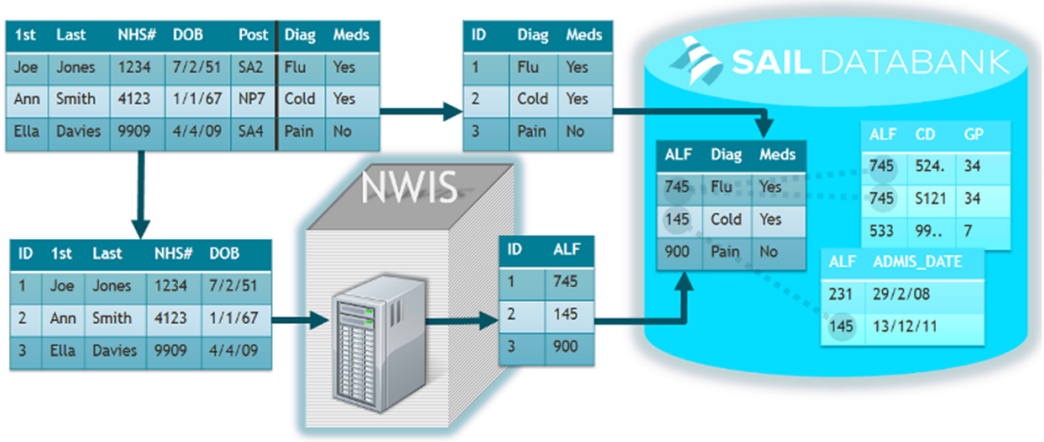

Figure 2.1: SAIL Databank split file procedure. Data is split at source into identifiable data and clinic data. The identifiable data sent to The NHS Wales Information Service where each NHS number is encrypted before being sent to the SAIL Databank. The clinic data is sent directly to the SAIL Databank, and is joined to the encrypted identifiable data by an internal system ID that is present in both datasets.

2.1.1 Ethics and Governance

All proposed SAIL projects undergo a review process by an IGRP (Information Governance Review Panel) for approval. Researchers must complete an IGRP form that details the scope of the project and what datasets are required. Approved projects must be deemed feasible and present no risk to patient identification. While all NHS numbers are encrypted in SAIL, it may be possible for Clinicians, without taking data out of the SAIL databank to identify a patient with a rare combination of information in their medical records e.g. a 35 year old man living in the ABMU health board area that has Becker's muscular dystropy, focal seizures with secondary generalisation and being prescribed multiple anti-epileptic drugs for multiple seizures per day. SAIL does not allow individual level data to be taken out of the gateway and must be summarized so that it conforms to small number disclosure rules of never reporting on groups with less than 5 persons. Data must be requested out and reviewed by the SAIL Analyst Team to enforce these rules.

2.1.2 Assessing the burden of disease using the SAIL Databank

The work of this thesis focuses largely on how to identify disease phenotypes, their effects on quality of life and how to treat and manage the disease. The Wales

Epilepsy Research Network (WERN) at Swansea University has conducted world class research into epilepsy genetics, and has both clinical and research expertise in the field of epilepsy. The SAIL databank presents an ideal opportunity to further our understanding of the burden of epilepsy in health care. Where exome analysis can identify the exact cause and mechanisms of epilepsy, patient records within SAIL can answer questions on social outcomes, health utilisation and drug efficacy for patients with epilepsy. Being a member of WERN and integrating into the wide variety of researchers has helped form novel and important questions for this thesis.

2.1.3 Forming Research Questions

In 2008 a group survey conducted on behalf of the DUETs (Database of Uncertainty about the Effects of Treatments) and James Lind Alliance brought together patients, clinicians, patient carers and researchers to address questions regarding treatment of conditions, including epilepsy `www.library.nhs.uk/duets`. In addition, all the Neurology Consultants across Wales were asked to contribute research questions that SAIL may or could answer. Many of the items highlighted form the research basis for this thesis involving SAIL data, with an emphasis on:

1. Research into better treatments and seizure control
2. Research into ensuring current treatments are as effective as possible
3. Research into stigma associated with epilepsy
4. Research into epilepsy and other medical conditions
5. Research into patient information

These questions set many challenges when using data in the SAIL databank and were considered to prioritise the studies described in chapter 3 of this thesis. The first problem is to correctly identify epilepsy phenotypes within routinely collected primary care data, and the second is to obtain treatment and social status of those with epilepsy. A scoping exercise was carried out to identify datasets that would contain the required data build an epilepsy patient profile within SAIL data. Four core datasets were identified within the SAIL databank to obtain epilepsy metrics - General Practice dataset, Secondary care dataset (PEDW), the Welsh Demographic Service dataset and the Office of National Statistics (ONS) Deaths dataset.

2.2 SAIL Datasets

The following section describes the core datasets in SAIL, including those used in this thesis.

2.2.1 GP dataset

Each GP practice in Wales uses a clinical information system to capture patient records. The Primary Care IM& T Programme developed a piece of software called Audit+ that is provided free of charge to all Welsh GPs. It was designed to facilitate the capture and transfer of GP data to external sources. A SAIL module is built into Audit+ that automatically two data extracts: file 1 contains demographic data and file 2 contains clinical data, where an internal system ID is shared between both files. Both files are securely transferred to SAIL and are anonymised via the encryption process outlined in [61]. Each GP practice in Wales has been invited to participate, where only those consenting have their data transferred into SAIL. Over length of study of this thesis, SAIL GP participation has increased from 40% of Welsh GP practices to 75%.

The dataset contains two unique identifiers: patient ID (encrypted NHS Number) and GP ID (also encrypted). Demographic data available includes week of birth, sex and dates of the beginning/end of registration with a given GP practice. Clinical information includes reason for attendance via version 2 READ codes, date of attendance and any laboratory result such as blood pressure. There are over 300,000 READ codes that can be used to define diseases, drug prescriptions, symptoms, referrals to specialists and laboratory results. A subset of READ codes were used to define:

1. Epilepsy diagnosis
2. Anti-epileptic drugs (AEDs)

Using a combination of these two categories, a diagnosis of epilepsy in the SAIL dataset is defined as a patient that has a diagnosis code for epilepsy followed by a repeat anti-epileptic drug prescription within 6 months. A repeat AED prescription is used as a pre-cautionary measure to prevent suspected, unconfirmed epilepsies being used in any epilepsy cohorts generated. The reasoning behind this is that epilepsy and treatment plans are confirmed by a specialist, not in a GP setting. It is possible a GP may use a diagnosis code for epilepsy to denote suspected epilepsy as well as a code to refer to a specialist. The repeat prescription is indicative of that patient being seen by a specialist and prescribed a AEDs beyond a potential trial period.

Defining epilepsy in the GP dataset held within SAIL is useful because it can be used to build up a patient profile of epilepsy at first point of health care in Wales, as well as link to other datasets.

2.2.2 Secondary Care dataset

The SAIL databank receives an annual extract of all Welsh inpatient hospital data. This dataset is processed using the same split file procedure where the National Wales Information System (NWIS) acts as a TTP and processes the extract useable within SAIL called PEDW. PEDW contains all hospital admissions in Wales from 1998 and uses the ICD-10 coding system to record admissions to inpatients. Each hospital employs teams of clinical coders to records the reason for admission as determined from consultation of doctors' notes, as well as any operations and costs associated with the admission. Other variables included within PEDW are date of admission, length of stay and consultation speciality. Date recorded in PEDW is designed to reflect the care pathway of an admission to secondary care, and with careful use of ICD-10 code selection, it is possible to identify patients that are admitted to secondary care and the reason for being admitted. In particular there are ICD-10 codes for seizures and status epilepticus, as well as other common outcomes associated with epilepsy such as seizures and stroke.

2.2.3 Welsh Demographic Service

The Welsh Demographic Service (WDS) was introduced in 2009 to manage administrative and demographic data for NHS patients in Wales. It contains address information as recorded by a patients' GP, which is mapped to Lower Super Output Area (LSOA), Middle Super Output (MSOA) and Local Health Board (LHB) by the address postcode. These are geographical units with LSOA's accounting for between 5-10 postcodes and can be mapped to The Welsh Index of Multiple Deprivation (WIMD) to measure social deprivation. The Welsh Government's official measure of deprivation is the Welsh Index of Multiple Deprivation (WIMD) `http://wales.gov.uk/topics/statistics/theme/wimd/?lang=en` and is readily available in the SAIL databank for any person registered with a Welsh GP. It comprises of 8 domains:

1. Income
2. Employment
3. Health
4. Education

5. Access to Services
6. Community Safety
7. Physical Environment
8. Housing

where a deprivation score can be obtained for each domain or combined into one score. The WIMD is therefore one way of measuring social deprivation to some degree in those living with epilepsy. An explanation of how the WIMD score is calculated is given in Figure 2.2.

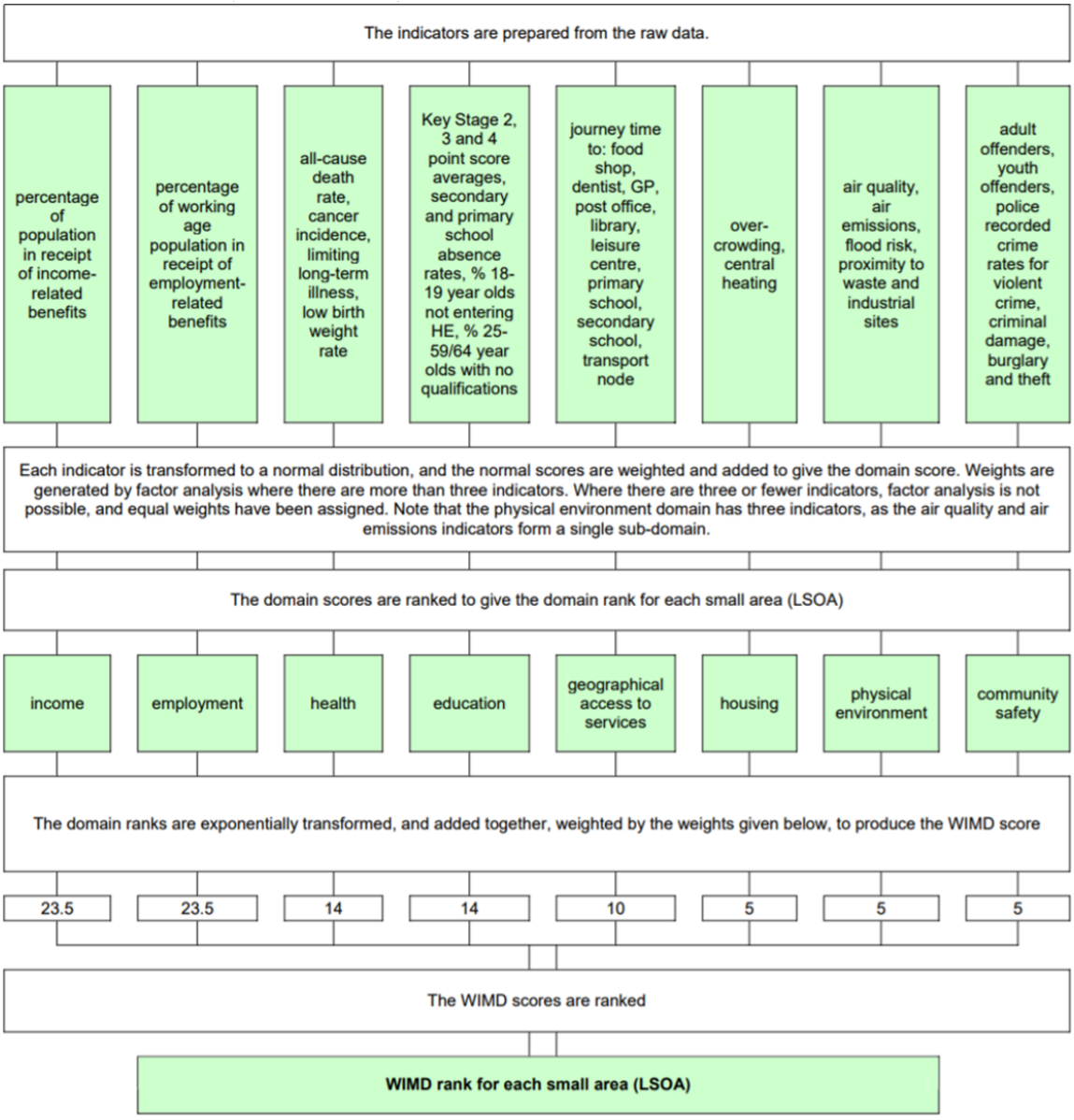

Figure 2.2: Flow chart of how the WIMD score is calculated from 8 different domains. Taken from `http://webarchive.nationalarchives.gov.uk/20150505155421/http://gov.wales/docs/statistics/2011/110831wimd11summaryen.pdf`

2.2.4 ONS deaths

SAIL contains both annual and monthly extracts from the ONS database. This dataset contains date of death, primary and secondary causes of death, location of death (LSOA), place of death (hospital,home etc) and age of death. ONS deaths contains all deaths in Wales from 2003 onwards. Causes of mortality are not well understood in epilepsy related deaths, and death certificate information alone can only provide so much evidence. For example Sudden Unexpected Death in Epilepsy (SUDEP), an uncommon outcome of epilepsy rarely gets recorded in death certificates due to the cause of death not being clear, and so the amount of people dying from SUDEP is thought to be underestimated in the epilepsy population. In some cases it is thought to be preventable, however the risk factors are not well understood and there is currently no genetic explanation for SUDEP, and as such a research priority in epilepsy. The work in this thesis hypothesizes that linking death certificate information from ONS into GP and secondary care records can provide further insights into risk factors associated with SUDEP.

2.3 Data Linkage

2.3.1 Structured Query Language

SQL (Structured Query Language) queries use set theory to join multiple datasets together and aggregate individual records into groupings for statistical analysis. Unique identifiers are present in all datasets within SAIL, this is usually the encrypted NHS number of a patient called an Anonymous Linking Field, ALF. For example, consider the two tables and SQL code used to join them by the ALF:

```
##      ALF_E GNDR_CD DRUG    GP_DATE GP_PRACTICE
## 1 20000001       M  LTG 2001-01-01         SW1
## 2 20000001       M  VPA 2001-04-02         SW1
## 3 20000002       F  CBZ 2001-04-10         NEA
## 4 20000003       M  CBZ 2001-11-04         NEA
## 5 20000004       F  VPA 2001-01-31        CAF4
## 6 20000005       F  LTG 1999-04-06        POW2
```

```
##      ALF_E  HOSP_DATE HOSP_ID
## 1 20000001 2001-05-10     7AE
## 2 20000003 2001-11-11     ONA
## 3 20000004 2001-02-14     9DN
## 4 20000005 2007-03-04     G4K
## 5 20000006 2009-03-15     7AE
```

```
select distinct gp.ALF_E, gp.GNDR_CD, gp.DRUG, gp.GP_DATE,
hosp.HOSP_ADMIS, hosp.HOSP_DATE

FROM
SAILGP gp inner join SAILHOSP hosp

on gp.ALF_E = hosp.ALF_e
where hosp.HOSP_DATE between gp.GP_DATE and gp.GP_DATE + 3 months
```

Figure 2.3: A simple SQL script to return encrypted patient identifiers along with the gender and date of birth of the patient where the patient must be female and born after the 1st of January 1990.

The first table is GP data containing prescriptions for anti-epileptic drugs, and the second table is hospital admissions for seizures. It is possible to find persons who have been admitted to hospital for a seizure within 3 months of a drug prescription. The code in figure 2.3 demonstrates how to join the GP data and hospital data together, and the results are shown in figure 2.3.1.

```
##      ALF_E GNDR_CD DRUG    GP_DATE HOSP_ADMIS  HOSP_DATE
## 1 20000001       M  LTG 2001-01-01          N       NULL
## 2 20000001       M  VPA 2001-04-02          Y 2001-05-10
## 3 20000002       F  CBZ 2001-04-10          N       NULL
## 4 20000003       M  CBZ 2001-11-04          Y 2001-11-11
## 5 20000004       F  VPA 2001-01-31          Y 2001-02-14
```

Once a basic data linkage is established amongst SAIL datasets the aim is to filter the linked datasets down to a "final" dataset ready for statistical analysis.

2.3.2 Quality checking routinely collected data

It is important to note that routine healthcare datasets were not designed for research purposes, rather collected at point of care for costing purposes or to supplement decision making processes. With this in mind, any routinely collected healthcare dataset requires careful consideration of how to interpret the data and understand the limitations. The majority of the data is entered manually by trained professionals, but there is still the potential for human error in data entry in which it is impossible for some error to be rectified by retrospectively cleaning the data.

Data cleansing refers to removing or transforming data so that is is meaningful when used in statistical analysis. For example, blood measurements such as cholesterol can be expressed in millimoles per litre (**mmol/L**) or milligrams per decilitre **mg/dL**.

In the UK it is usually the case that the former is used to record cholesterol, however this is not always the case and the units are not attached. Limits of normality must be defined to determine if a reading is likely to expressed in certain units or not by looking at the value recorded. While there is an obvious difference in the values recorded within different units it is possible for two extreme measurements of each of the scale of different units to have a similar value. Some values may even appear outside of any limits of normality of any units of measurement and might require from the study.

Cleansing on measurement values are relatively more straightforward than making assumptions on how other items of data are recorded. Two examples of this are the dates GP records are documented and the reason for admission to secondary care. In the first example you assume that a recording of a prescription in a GP database was made on that date appearing in the records. However this could have been entered by administrative staff retrospectively, or even entered by both the GP and staff where it appears that the same prescription was made on two separate days. Many diagnosis codes for epilepsy appear before electronic records were even integrated into healthcare systems, suggesting that historical information is entered by GPs where the date appearing in the records is an approximation of when someone was first diagnosed with epilepsy.

In the second example, ICD-10 codes are used to record episodes of care in secondary care. For each episode 14 ICD-10 codes may be entered *to record the details of care required*. This can lead to some interesting codes being used at point of admission as it may be in the interest of hospital staff to include background information about patients to tailor their care through an inpatient stay. Chronic conditions such as diabetes or asthma often get recorded even if they are unrelated to the admission. A distinction has to be made because codes that are entered to detail care required does not directly translate to reason for admission. These 14 diagnostic positions are ordered by priority of care, however this leads to many of the high priority positions being taken up by generic symptoms such as "chest pain" or "out of breath", and for complicated episodes of care there could be multiple valid reasons for admission. There is clearly no "one-size fits all" approach into which positions are used as a proxy for reason for admission as the actual reason for admission is not known. This rules out an opportunity to conduct a sensitivity analysis to determine how many positions are included.

Finally, some other obvious errors appear in large datasets that are common and well known, but nevertheless have to be taken into account. These are errors such as men appearing to be pregnant (incorrect gender code assigned), patients seeing their

GP after they have died (incorrect date of death) and people still alive longer than the known human life span (incorrect date of birth). A particularly painful error to identify in large datasets are when NHS numbers are incorrectly entered at point of care. For example there are records in SAIL that suggest one person has seen two different GPs in different health boards on the same day, where this behaviour can be traced back for numerous years. While it isn't impossible for this to be happen, it is most likely that the multiple persons are sharing an NHS number due to a mistake at the point of registration. In this instance cross checking with other datasets may help determine which person the NHS number truly belongs to, and therefore which set of records to exclude.

All of these cleansing considerations were taken into account for the various studies in this thesis when using the SAIL databank.

2.3.3 Statistical Analysis

The SAIL gateway contains various software packages to analyse linked datasets in SAIL. Studies in this thesis use the open source R statistical software language to produce statistics and figures. SAIL has strict guidelines on what data is allowed to be taken out of the secure gateway and be included in publications. The results of statistical analysis must not contain data that could potentially identify a person, even from anonymous data. Therefore number of persons in group outputs may not be reported below 5 persons as persons in this group could potentially be identified through linking many datasets together and building a detailed patient profile. Individual level data is therefore restricted from being reported as part of any statistical analysis.

2.4 Natural Language Processing

To construct NLP algorithms and validate their ability to extract important items of text from clinic letters, clinic letters were sourced from Morriston Hospital [1]. Patients with known epilepsy were sourced form clinic letters held in the Swansea Epilepsy Database that stores patient data from the Epilepsy unit in Morriston hospital, and patients without epilepsy were sourced from general neurology clinics. All letters were de-identified by replacing identifiable information with fake information. These letters were then made available to aid constructing an automated NLP algorithm to extract epilepsy specific information from the clinic letters. The clinic letter in Figure 2.4 is representative of the clinic letters used in this study.

Re: **Mstr. John Red** **D.O.B: 01/10/2001**
5 Lone lane, Camden, London, SW6 7AA

Diagnosis: Juvenile absence epilepsy
Possible non-epileptic attacks

Medication: Epilim Chrono 250mg am, 250mg nocte (16mg/kg/day)

Follow Up: 1 year

John was reviewed in the Epilepsy Clinic today. John had an EEG last which showed abnormal spike and wave activity which was supportive of JME and now has regular appointments with the specialist nurse. At the onset he was having around 3-4 seizures per day and was put on Sodium Valproate 250mg bd which helped somewhat. We had asked him to keep a diary of his seizures, as well as what his mother said was "jerking" on his right side during sleep. We discussed the possibility of these being non-epileptic attacks, but because they are infrequent it is hard to determine the source of these.

John had a further EEG last September and upon review he was weaned off Sodium Valproate and prescribed Epilim Chrono 500mg bd. His tonic clonic seziures appear to be well controlled, and has remained seizure free since December 2014.

I have arranged for John to be seen again in a year. If he continues to remain seizure free, we will discharge him back to his GP. Dr Jones will arrange to meet with John to discuss lifestyle issues.

Yours sincerely,
Dr Smith.

Figure 2.4: A example clinic letter. The letter contains real patient data, but all identifiable information has been anonymized

2.4.1 Software

The open source GATE v8.4.1 (General Architecture for Text Engineering) `https://gate.ac.uk/` framework was used to construct an algorithm to extract epilepsy specific information from clinic letters, as other details such as clinic date, patient NHS number and date of birth. GATE allows users to build rule sets by combining custom gazetteers (user defined dictionaries) with mechanisms form specifying word

[1] performed by Ms Beata Fonferko-Shadrach

ordering by writing JAPE (Java Annotation Pattern Engine) scripts. GATE also provides plugins to perform common NLP tasks such as POS (Part-of-Speech) tagging, NER (Named Entity Recognition) and phrase identification, while also allowing the user to put together such plugins as modules in a pipeline. A GATE pipelines are constructed with a provided GUI (Graphical User Interface).

2.4.2 Part of Speech tagging

POS tagging formed the basis of information extraction for this study. Words in text are classified into grammatical units such as verbs, adjectives and nouns. The following phrase:

> Mary has focal epilepsy and has been taking Lamotrigine for five years

can be tagged in the following way:

> Mary—NNP has—VBZ focal—JJ epilepsy—NN and—CC has—VBZ been—VBN taking—VBG Lamotrigine—NNP for—IN five—CD years—NNS

where tags are in red. The ANNIE POS Tagger is used to POS tag clinic letters used in the epilepsy NLP algorithm, in which the possible tags are given in table 2.1.

Table 2.1: ANNIE POS tags and their descriptions

POS tag	Description
CC	coordinating conjunction: 'and', 'but', 'nor', 'or', 'yet', plus, minus, less, times (multiplication), over (division). Also 'for' (because) and 'so' (i.e., 'so that').
CD	cardinal number
DT	determiner: Articles including 'a', 'an', 'every', 'no', 'the', 'another', 'any', 'some', 'those'.
EX	existential 'there': Unstressed 'there' that triggers inversion of the inflected verb and the logical subject; 'There was a party in progress'.
FW	foreign word
IN	preposition or subordinating conjunction
JJ	adjective: Hyphenated compounds that are used as modifiers; happy-go-lucky
JJR	adjective comparative: Adjectives with the comparative ending 'er' and a comparative meaning. Sometimes 'more' and 'less'.
JJS	adjective superlative: Adjectives with the superlative ending 'est' (and 'worst'). Sometimes 'most' and 'least'.
LS	list item marker: Numbers and letters used as identifiers of items in a list.
MD	modal: All verbs that don't take an 's' ending in the third person singular present: 'can', 'could', 'dare', 'may', 'might', 'must', 'ought', 'shall', 'should', 'will', 'would'.
NN	noun singular or mass
NNP	proper noun singular: All words in names usually are capitalized but titles might not be.
NNPS	proper noun plural: All words in names usually are capitalized but titles might not be.
NNS	noun plural
NP	proper noun singular
NPS	proper noun plural
PDT	predeterminer: Determiner like elements preceding an article or possessive pronoun such as 'all/PDT his marbles', 'quite/PDT a mess'.
POS	possessive ending: Nouns ending in "s' or "'.
PP	personal pronoun
PRP	possessive pronoun,such as 'my', 'your', 'his', 'his', 'its', 'one's', 'our', and 'their'.
RB	adverb: most words ending in 'ly'. Also 'quite', 'too', 'very', 'enough', 'indeed', 'not', 'n't', and 'never'.
RBR	adverb comparative: adverbs ending with 'er' with a comparative meaning.
RBS	adverb superlative
RP	particle: Mostly monosyllabic words that also double as directional adverbs.
STAART	start state marker (used internally)
SYM	symbol: technical symbols or expressions that aren't English words.
TO	literal "to"
UH	interjection: Such as 'my', 'oh', 'please', 'uh', 'well', 'yes'.
VBD	verb past tense: includes conditional form of the verb 'to be'; 'If I were/VBD rich...'.
VBG	verb gerund or present participle
VBN	verb past participle
VBP	verb 3rd person singular present
VB	verb base form: subsumes imperatives, initiatives and subjunctives.
VBZ	verb 3rd person singular present
WDT	'wh' determiner
WPH	possessive 'wh' pronoun: includes 'whose'
WP	'wh' pronoun: includes 'what', 'who', and 'whom'.
WRB	'wh' adverb: includes 'how', 'where', 'why'. Includes 'when' when used in a temporal sense.

2.4.3 Gazetteers

The GATE framework makes extensive use of dictionaries, or gazetteers to tag words with higher level information than simple grammatical units. The Bio-YODIE plugin for GATE was used to find biomedical references in the clinic letters. Bio-YODIE

Table 2.2: UMLS representation of a subset of epilepsy terms. The CUI (Concept Unique Identifier) is a code assigned to biomedical concepts. The source column represents the original coding system the term exists in, and the SCUI column is the source code used within a particular system. For example "Epilepsy" exists in both ICD10 and READ coding systems, but map to the same CUI in UMLS despite having unrelated SCUIs.

CUI	Term	SCUI	Source
C0014544	Epilepsy	G40	ICD10
C0014544	Epilepsy	F25	READV2
C0014549	Tonic-Clonic Epilepsy	F2510	READV2
C0014558	Other Epilepsy	G40.8	ICD10
C0477371	Progressive myoclonic epilepsy	F1321	READV2
C1827284	Intractable occipital lobe epilepsy	425054007	SNOMEDCT_US
C1827878	Refractory localization-related epilepsy	422724001	SNOMEDCT_US
C1827284	Refractory occipital lobe epilepsy	425054007	SNOMEDCT_US
C1827691	Intractable frontal lobe epilepsy	425237009	SNOMEDCT_US
C1827974	Refractory parietal lobe epilepsy	425349008	SNOMEDCT_US
C1827974	Intractable parietal lobe epilepsy	425349008	SNOMEDCT_US

which the UMLS (Unified Medical Language System) - 600 healthcare coding systems such as ICD-10, READ and SNOMED CT combined into a unified coding system - to form the basis of a gazetteer that maps any text found in a document to a UMLS code.

The Bio-YODIE plugin scans all text in a document, and where it finds a match to a term the text is tagged and assigned a CUI code (mapping shown in Table 2.2). Custom gazetteers were also defined to incorporate information such as certainty levels ("likely","probably","doubtful"....) or hypothetical modifiers ("to see","we may","to determine"...) to help define the context to which UMLS mappings are found.

2.4.4 Context Algorithm

The Context algorithm developed by Harkema et al [232] was used to determine if terms are negated e.g. "Mary does not have epilepsy" or if they are attributed to someone other than the patient, such as a family member. The Context algorithm also tags symptoms in terms of their temporal context such as historical or hypothetical. The GATE plugin implements the Context algorithm through multiple gazetteers that contain trigger words for various contexts, such as pre and post negation terms, temporal triggers and triggers for family members. These triggers are related to biomedical terms found in the text by writing rules in the JAPE scripting language.

2.4.5 JAPE rules

Implementing POS tagging and gazetteer mapping provides information ranging from simple grammatical units to meaningful biomedical concepts and contextual terms. The JAPE scripting language was used to weave together sequences of tagged terms to form rules. After tagging, a phrase containing prescription information might look like:

He	**is**	**taking**	**Lamotrigine**	**250**	**mg**	**twice**	**a**	**day**
PRP	VBZ	VBG	NNP	CD	NN	RB	DT	NN
person	current		DRUG	number	unit	word-num		temporal
patient			C0064636	quantity		quantity		calendar

where each word has been assigned multiple tags by mapping to various user defined gazetteers. Further context can be built by combining words, for example "250" and "mg" is a unit of measurement that could be annotated and used for downstream processes.

He	**is**	**taking**	**Lamotrigine**	**250**	**mg**	**twice**	**a**	**day**
PRP	VBZ	VBG	NNP	CD	NN	RB	DT	NN
person	current		DRUG	number	unit	word-num		temporal
patient			C0064636	quantity		quantity		calendar
				Measurement		**Frequency**		

A JAPE script that could create this rule is shown would be:

In general, larger rules are built by layering smaller rules. For example, if Measurement and Frequency have previously been defined by JAPE rules, those annotations can be used in further JAPE rules.

He	**is**	**taking**	**Lamotrigine**	**250**	**mg**	**twice**	**a**	**day**
PRP	VBZ	VBG	NNP	CD	NN	RB	DT	NN
person	current		DRUG	number	unit	word-num		temporal
patient			C0064636	quantity		quantity		calendar
				Measurement		**Frequency**		
			Prescription					

```
# this is a comment that is ignored when running the script

Phase: Measurement
# phase to be run in larger pipeline
Input: Number Unit
# Input type i.e. read in gazetteers for Number and Unit
Options: control=appelt

/*
 * Find measurements from combining numbers and units
 * i.e. 250 mg
 */
Rule: find_measurement
# when Number and Unit appear consecutively
(
  ({Number}):num
  ({Unit}):unit
):match
-->
# create new annotation called "Measurement" containing the following information
:match.Measurement = { Rule = findMeasurement,
                       Quantity = :num.String,
                       Unit = :unit.String # Unit i.e. mg
                     }
```

Figure 2.5: A JAPE script to annotate measurements. Gazetteers are used as input to the JAPE script, and depending on the order of words tagged by a gazetteer, a rule can be triggered and create a "Measurement" annotation.

```
# this is a comment that is ignored when running the script

Phase: Prescription
# phase to be run in larger pipeline
Input: Drug Measurment Frequency
# Input type i.e. read in gazetteers for Number and Unit
Options: control=appelt
/*
 * Find prescriptions from combining numbers and units
 * i.e. Lamotrigine 250 mg once per day
 */
Rule: find_prescription 1
# combination of annotations for prescription
(
  ({Drug}):drug
  ({Measurement}):measure
  # Frequency is optional - denoted by ?
  ({Frequency}?):frequency
):match
-->
# create new annotation called "Prescription" containing the following information
:match.Measurement = { Rule = findPrescription, # rule reference
                       Drug = :drug.Name # Drug name
                       Quantity = :measure.Quantity,
                       Unit = :measure.unit,
                       Num_Dose = :frequency.num,
                       Frequency.period = :frequency.calendar
                     }
```

Figure 2.6: A JAPE script to annotate prescriptions. Previous annotations writtin in JAPE rules can be directly used as input to build larger rules.

The are additional operators than can be used such as the **?** operator used in figure 2.6.

Table 2.3: List of JAPE operators than can be applied to any annotation

Operator	Description
?	optional
*	zero or more
+	one of more
!	any other than specified annotation
[x]	exact length of annotation
[x,y]	range length of annotation
|	OR
,	AND
==	exact match
!=	not equal to
==∼	partial match via regex
!=∼	not equal to regex
contains	annotation contains specified annotations
!contains	annotation doesn't contain other specified anotations
within	annotation exists with specified annotation
!within	annotation does not exist within specifided annotation

2.5 Predicting functional impact of Single Nucleotide Polymorphisms

2.5.1 Pipeline to determine the effect of SNPs

The SNP pipeline aims to determine the effect of SNPs and their impact on presentation of a disease phenotype, in this case if the SNP is implicated in disease or not. This largely revolves around collecting and processing data that builds a final dataset of protein features for any given SNP, often called annotation. Machine learning methods are applied on the final datasets to predict whether a SNP has the potential to be pathogenic and candidates for disease causality.

2.5.2 Data sources

The prediction of functional impact of SNPs largely relied on obtaining biological annotations from publicly available reference data. The main type of data required are meta-data for a given SNP, such as the protein affected, the region that the SNP

is expressed and the conservation score at the position of the SNP. The following list of databases were used to source SNP features:

- **University of California Santa Cruz (UCSC) Genome Browser** - provides protein level reference data from chromosomal SNP co-ordinates. Datasets used were snp144, kgXref and knownGene. - `https://genome.ucsc.edu/`.
- **Uniprot** - Stores various reference data on proteins, in particular the FASTA sequence files - `http://www.uniprot.org`
- **Humvar** - a set of SNPs used to train the machine learning algorithms in the pipeline - `http://www.uniprot.org/docs/humsavar`

2.5.3 Obtaining Protein Features

Figure 2.7 describes the various steps in the pipeline to collect the necessary data to make a prediction for a SNP. Given a list of SNPs in chromosomal format each SNP goes through the following pipeline architecture. At each point various cleaning and data wrangling are performed and fed into downstream programs, each accumulating protein features that are useful with regards to protein annotation, but also important for predicting the effects of SNP in terms of disease.

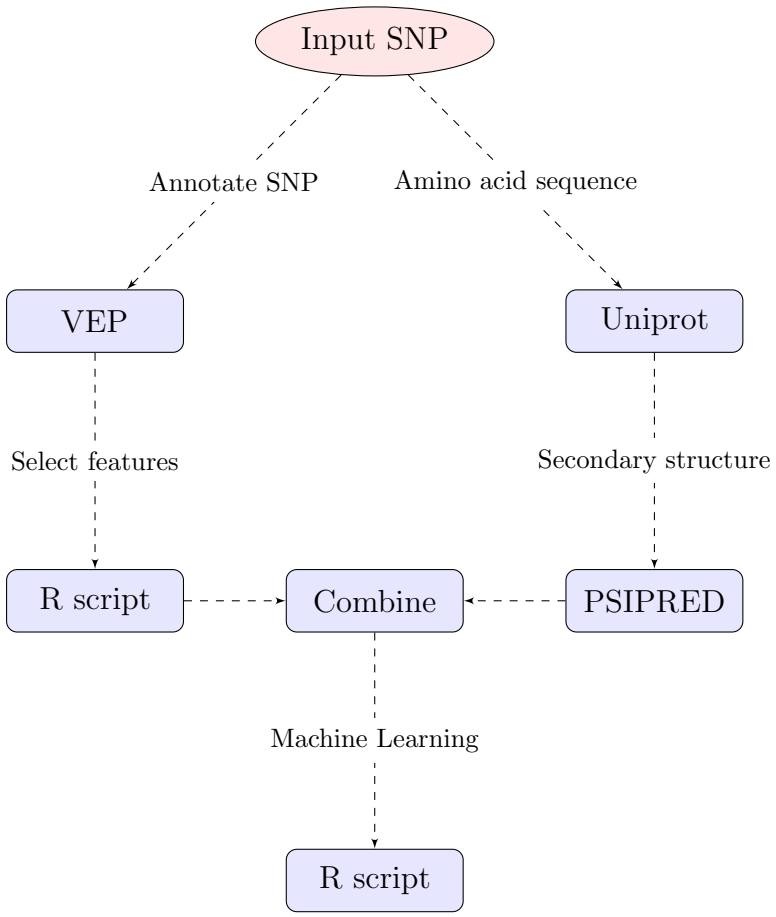

Figure 2.7: A flow chart of the pipeline. Purple nodes are databases and green nodes are processes. The user can specify SNPs in chromosomal format as input to the pipeline. The end result is the SNP data with protein level data that includes indexes generated by downstream programs and database annotation.

SNP format

SNPs are expressed as chromosomal co-ordinates when called from NGS pipelines. However many protein annotation, prediction and filtering programs require SNPs to be expressed in protein co-ordinates, with various meta-data also included such as gene name and protein ID. The first step of the pipeline is to make this conversion. There are various methods to do this however the conversion was performed on many thousands of SNPs without human interaction. The University of California Santa Cruz (UCSC) genome browser hosts an online database that can be queried with MYSQL scripts. With the chromosome number, position, wild type and mutation nucleotides it was possible to search for the protein co-ordinates and gene information using the code in figure 2.8.

An important piece of data for this pipeline are the raw amino acid sequences in the form of FASTA files. The protein IDs in the previous script can be used in conjunction

```
mysql --user=genome --host=genome-mysql.cse.ucsc.edu -A hg19 -D hg19 -e
"select distinct substr(S1.chrom,4,2) as chrom,
S1.chromStart,
S1.chromStart,
substr(S1.observed,1,1) as wild,substr(S1.observed,3,1) as snp
X.geneSymbol,X.spID,S1.class,
S1.bitfields,S1.name
from snp144 as S1, knownGene as K, kgXref as X
where X.geneSymbol = '"$1"'
and K.chrom = S1.chrom and X.kgID = K.name
and S1.chromStart >= K.cdsStart
and S1.chromStart < K.cdsEnd and S1.class = '"$2"';"
```

Figure 2.8: MYSQL script to retrieve protein descriptors from chromosomal position.

```
// #!/bin/bash

// download fasta seqs given file of uniprot ids

// file name is first parameter of command
file=$1

// protein IDs are contained in first column
ids=($(cat ${file} | awk '{print $1}'))

// loop through IDs and get fasta using uniprot API
for i in "${!ids[@]}" ; do
      curl -sS "http://www.uniprot.org/uniprot/"${uniqids[i]}".fasta"
      >> $file.out.fasta ;
done
```

Figure 2.9: FASTA file retrieval using the Uniprot API. SNPs are substituted into each sequence and passed onto downstream programs

with the Uniprot web interface to retrieve FASTA files. These can essentially be accessed via a unique uniprot URL i.e. `http://www.uniprot.org/uniprot/P23415.fasta`. These can either be typed into an internet browser or programmatically retrieved using webscraping tools such as GNU **cURL**. Figure 2.9 shows how **cURL** can be implemented via a bash script to download fasta sequences from the Uniprot API.

This again allows automation and no human interaction - providing that the protein IDs are passed on from previous programs. Using the Uniprot web interface also does not require storage of FASTA files on a local computer, where each FASTA file is downloaded in a few seconds. Later on in the pipeline a GNU AWK program is used to programmatically substitute the mutant SNP in place of the wild type. This is particularly useful for structural modelling where higher dimensional structures are built from raw amino acid sequences.

Protein annotation with Variant Effect Predictor

Variant Effect Predictor (VEP) `https://www.ensembl.org/Tools/VEP` was used to annotate SNPs. VEP accepts SNPs in chromosomal format, bed format or dbSNP rsID and can scan 35 publicly available datasets ranging from SNPs reported in national GWAS studies, functional prediction scores, binding site locations and regional sub-sequences within proteins. The following features (dataset name) are collected using VEP and used for functional prediction in the pipeline:

- Conserved genomic elements (phastCons,siPhy,GERP)
- Binding sites
- Cytogenic band
- Variants disrupting microRNAs
- Variants disrupting binding sites
- Reported structural variants
- SNP predictions from existing tools
- 1000 genomes frequency annotations
- ExAC frequency annotations
- gnomAD frequency annotations

The last 3 items are used to filter SNPs - frequency based filtering from 1000 Genomes and ExAC allow common SNPs, rare SNPs and unseen SNPs to be separated, where SNPs found in DbSNP contain any published links to clinically observed pathogenicty.

Existing prediction programs

The scores of existing programs for a given SNP is used for both comparison against the algorithm developed as part of this thesis, but also used as input features to the classification process. Many of these programs provide bespoke protein features used with each algorithm and vary between higher sensitivity and specificity and so are useful to include as input features. The Variant Effect Predictor algorithm was used to annotate SNPs against the dbNSFP `https://sites.google.com/site/jpopgen/dbNSFP` and table 2.4 lists all of the existing SNP prediction software used as features in the SNP pipeline.

Table 2.4: List of popular SNP prediction software

Name	Website
Polyphen2	http://genetics.bwh.harvard.edu/pph2
SIFT	http://sift.jcvi.org/
FATHMM	http://fathmm.biocompute.org.uk/
Provean	http://provean.jcvi.org/index.php
MetaSNP	http://snps.biofold.org/meta-snp/
LRT	http://www.genetics.wustl.edu/jflab/index.html
MutationTaster	http://www.mutationtaster.org/
MutationAssessor	http://www.mutationtaster.org/
FATHMM	http://fathmm.biocompute.org.uk/
MetaSVM	http://wglab.org/research
MetaLR	http://wglab.org/research
VEST -	http://karchinlab.org/apps/appVest.html
CADD	http://cadd.gs.washington.edu/
GERP++	http://mendel.stanford.edu/SidowLab/downloads/gerp/
DANN	https://cbcl.ics.uci.edu/public_data/DANN/
fitCons	http://compgen.cshl.edu/fitCons/
Phylop	http://ccg.vital-it.ch/mga/hg19/phylop/phylop.html
SiPhy	http://portals.broadinstitute.org/genome_bio/siphy/index.html
REVEL	https://sites.google.com/site/revelgenomics/

Predicting secondary structure using PSIPRED

There are many secondary structure prediction servers that accurately predict which amino acids in a sequence belong to certain types of secondary structural folds, namely beta sheets, alpha coils and helices. The Critical Assessment of protein Structure Prediction (CASP) [233] is a community driven initiative to enhance knowledge of structural prediction. Held bi-annually from 1994, many open source applications have been developed and tested for comparison, where the PSIPRED structural prediction server [225] has regularly featured as one of the most accurate programs of predicting secondary structure from amino acid sequences. PSIPRED is used in the pipeline to predict the difference in secondary structure change between wild type amino acid sequences and sequences with a SNP substituted in. Due to the relatively long processing time (between 15-30 minutes per sequence), the part of the pipeline where PSIPRED is run is hosted on the HPC Wales cluster. This reduces running time to a few minutes per sequence, where sequences can also be run in parallel across the many compute nodes within the HPC Wales cluster.

PSIPRED uses PSIBLAST alignments of proteins with know secondary structure

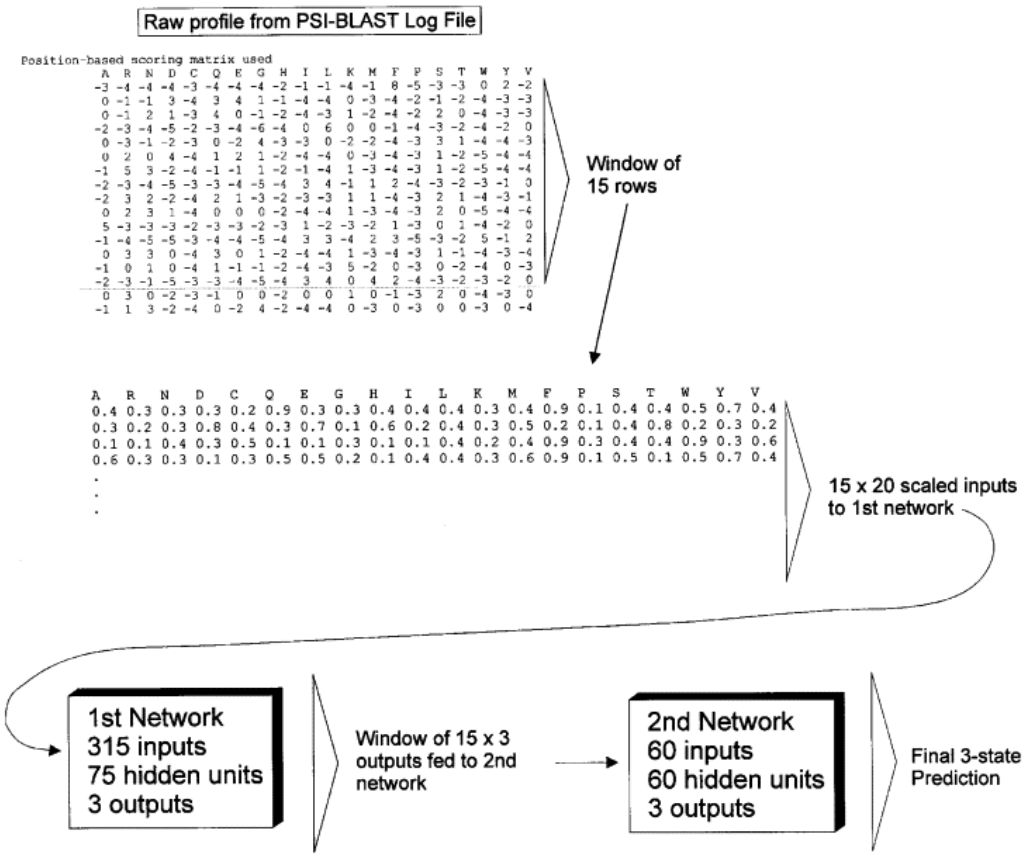

Figure 2.10: PSIPRED algorithm: multiple sequence alignments of known protein structures are built up from an input sequence. Position specific scoring matrix is used to train a neural network to predict the secondary structure of novel proteins.

and then trained with a two-step artificial neural network [234]. For a given protein sequence, PSIBLAST efficiently finds similar sequences via string matching and aligns all the sequences where a scoring matrix is calculated to determine the similarity of each position of each sequence. These position specific scores are used as input to the two stage neural network, along with the ground truth of known secondary structure. The most likely predicted secondary structure fold at each point along the multiple sequence alignment is then assigned by the output of the neural network. It is important to note that the predicted secondary structure at each amino acid in the sequence is determined by which proteins are assigned to the multiple sequence alignment by PSIBLAST. Because PSIBLAST searches on sub-sequences of proteins before building up the alignment, SNPs have the potential to source a small set of proteins not found in an alignment built from the wild type. This small set of proteins can make a difference (albeit small) in the final prediction.

PSIPRED requires amino acid protein sequences in FASTA file format. To compare the output of wild type protein sequences and sequences containing a SNP, the AWK script in figure 2.12 was written to automatically substitute in the SNP to the wild

POS	AA	SS	%COIL	%HELIX	%SHEET	AA	SS	%COIL	%HELIX	%SHEET	DCOIL	DHELIX	DSHEET
350	R	H	0.067	0.931	0.027	R	H	0.06	0.938	0.024	0.000049	0.000049	0.000009
351	L	H	0.078	0.911	0.052	L	H	0.075	0.902	0.044	0.000009	0.000081	0.000064
352	R	H	0.125	0.91	0.02	R	H	0.111	0.933	0.018	0.000196	0.000529	0.000004
353	R	H	0.098	0.912	0.02	R	H	0.077	0.943	0.017	0.000441	0.000961	0.000009
354	R	H	0.133	0.857	0.037	R	H	0.206	0.746	0.035	0.005329	0.012321	0.000004
355	Q	H	0.137	0.887	0.037	R	H	0.171	0.812	0.047	0.001156	0.005625	0.0001
356	K	H	0.261	0.755	0.026	K	H	0.297	0.667	0.04	0.001296	0.007744	0.000196
357	R	H	0.367	0.699	0.013	R	H	0.437	0.527	0.021	0.0049	0.029584	0.000064
358	Q	C	0.51	0.49	0.028	Q	C	0.586	0.344	0.037	0.005776	0.021316	0.000081
359	N	C	0.698	0.275	0.028	N	C	0.721	0.251	0.025	0.000529	0.000576	0.000009
360	K	C	0.725	0.245	0.044	K	C	0.711	0.252	0.056	0.000196	0.000049	0.000144

Figure 2.11: PSIPRED output comparing the predicted secondary structure of a wild type GLRA2 sequence with the same sequence having a clinically benign SNP at position 355 of the sequence. The secondary structure prediction is normalised between coil, helix and sheet, where the absolute difference between the wild type and SNP are calculated in the three rightmost columns. The red line indicates the SNP, where other lines are neighbouring amino acids and predictions. It can be seen that while the predicted secondary structure doesn't change, the amino acids closer to the SNP have a larger change in the raw score than those further out from the SNP.

```
BEGIN { FS="_" }
/^>/ {
id=$1;p=$2; wild=$3;subs=$4; c=$NF; next
}
{
s=1
e=length($0)
print id"_"p"_"wild"_"subs">\n"substr($0,s,p-1) c substr($0,p+1,e)
}
```

Figure 2.12: An AWK script to substitute a SNP in place of a wild type amino acid within a FASTA sequence. The script takes in 4 parameters that can be read from a text file in the form of 4 columns (protein ID, position of the SNP, amino acid of the wild type, P, and the amino acid of the SNP. These are then used tho subsitute the wild type in the position of the SNP with the SNP of the amino acid whilst preserving the original protein flanking either side of the SNP)

type sequence. The fasta header is also modified in this step to contain SNP position, wild type and reference for easy comparison between the wild type and SNP secondary structure output files.

2.5.4 Predicting SNP Impact Using Machine Learning

The R programming language was used to train various machine learning models to classify a SNP as either pathogenic or benign using protein features. A ground truth dataset was established by obtaining protein from the Clinvar `ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/` database for epilepsy SNPs and the Humvar `https://www.uniprot.org/docs/humsavar` database for disease non-specific datasets that contain information on a SNP known to cause disease or is benign. These datasets were then annotated using the methods described previously..

The following machine learning algorithms shown in Table 2.5 were used and tested

on both datasets:

Table 2.5: Machine learning algorithms used in SNP prediction in chapter 5

Classifier	R package	CRAN link
Random Forest	randomForest	https://cran.r-project.org/web/packages/randomForest/randomForest.pdf
C45 Decision Tree	rpart	https://cran.r-project.org/web/packages/rpart/rpart.pdf
Support Vector Machine	e1071	https://cran.r-project.org/web/packages/e1071/index.html
Logistic Regression	base r	NA
Artifical Neural Networks	nnet	https://cran.r-project.org/web/packages/nnet/nnet.pdf
Naïve Bayes	naivebayes	https://cran.r-project.org/web/packages/naive-bayes

2.5.5 Training and testing

The machine learning process was split into 2 phases: training and testing. The training phase used 75% of the humvar data to train each classifier with the remaining 25% used as an unseen test set to predict pathogenicty status. A sampling method called cross validation was used where the 75-25% split is selected randomly, in this case 5 times so that 5 unique models are generated on 5 unique training-test sets.

2.5.6 Receiver Operator Curves

Receiver Operator Curves (ROC) are used as a means to validate the accuracy of each classifier. Figure 5.11 shows a ROC curve for multiple classifiers:

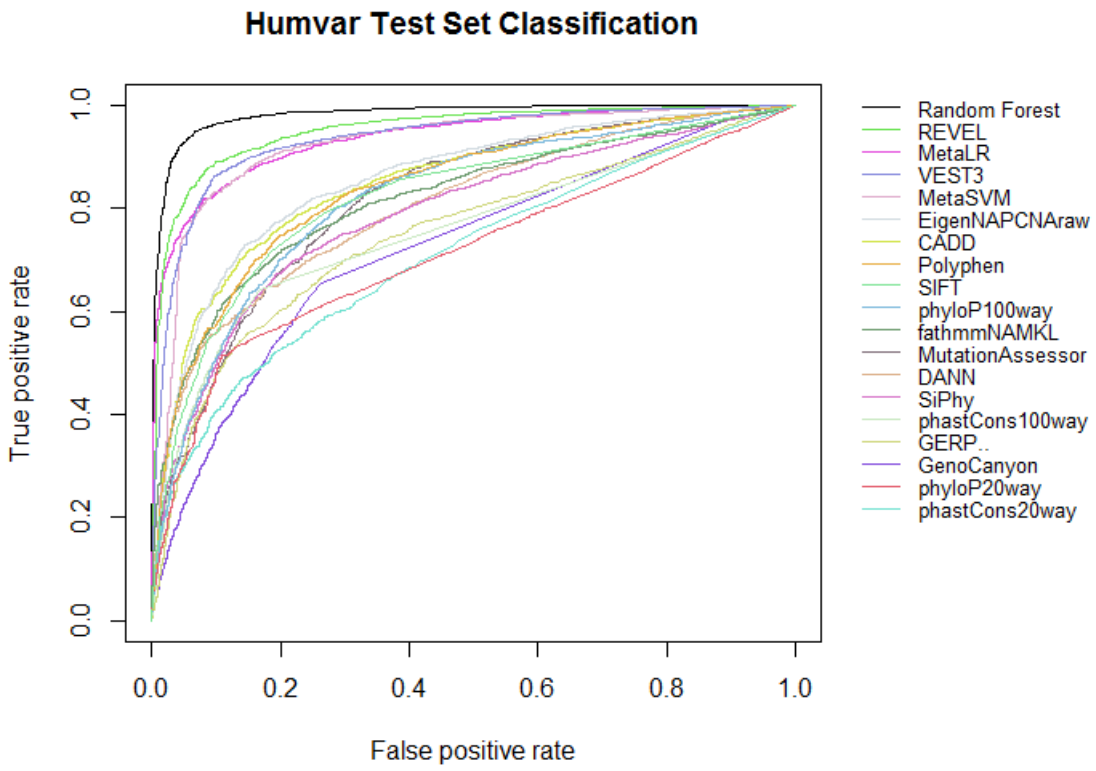

Figure 2.13: ROC curve comparing the classifier from this thesis (black) to scores from other classifiers when predicting disease/benign status on the humvar test set

ROC curves are useful in that each point in the curve plots the sensitivity vs specificity of an algorithm at discrete scoring thresholds. The result is a curve in which the ideal sensitivity vs specificity point can be read off (i.e. 95% sensitivity or 95% specificty) and reproduced for unseen samples with the associated threshold value. Two other interesting properties of ROC curves are the Area Under the Curve measure - a measure of accuracy of the classifier, and the two diagonal lines indicating the line of chance (bottom left to top right) and the line of accuracy (bottom right to top left). Classifiers can easily be compared to the line of chance to see how better it performs than random choice, and the section of the ROC curve that intersects with the line of accuracy is the point denoting the highest accuracy achievable by the algorithm. A ROC curve approaching the top left of the graph indicates a perfect score.

2.6 Chapter Summary

- The SAIL Databank was used to carry out retrospective longitudinal studies in people with epilepsy
- SQL queries were used extensively to link datasets together within the SAIL

databank and extract data in a format ready for statistical analysis

- The R programming was used to carry out statistical analysis
- 240 epilepsy clinic letters from Morriston hospital were used to conduct an NLP study for extracting epilepsy specific information from clinic letters
- The open source GATE application was used as a framework for NLP development. The main concepts behind algorithm development was the inclusion of dictionaries used for tagging terms of interest, such as UMLS codes and writing JAPE scripts that declare rule sets to produce annotations based on tagged terms. Standard NLP applications are also used as part of GATE such as tokenization, POS tagging and chunking
- An epilepsy clinician reviewed 200 test letters and these were compared against the algorithm
- Two SNP datasets were used to test a bespoke SNP classifier for pathogenic/benign status. The Humvar data contains over 70,000 disease non-specific SNPs that were used to trian the classifier, and the Clinvar dataset was used to obtain epilepsy SNPs
- Various open source bioinformatics programs were used for SNP annotation data. The Varaint Effect Predictor was used to annotate SNPs with conservation and frequency data as well as existing SNP prediction scores. PSIPRED was used to calculated the difference in secondary structure prediction between wild type sequences and mutation sequences
- Machine learning was used to test various classifiers, of which this was programmed in R. These were compared against existing scoring software obtained from dbNSFP annotations from VEP

Chapter 3

Results: Analysing Routinely Collected Healthcare Records for Epilepsy Research

The first of the three results chapters show how epilepsy and its impact on health and social factors can be studied using routinely- collected health records in the SAIL databank. Using Welsh GP records an algorithm was created to determine people with robustly-confirmed epilepsy. By comparing GP records in SAIL to patient details in the paediatric neurology department in Morriston Hospital , the algorithm is able to categorise 85% of patients as having epilepsy while excluding nearly all cases where a lack of clinical evidence confirms the absence of an epilepsy diagnosis . Antiepileptic drugs (AEDs) are studied in terms of how GPs prescribing habits have changed over time, as well adverse effects of AEDs including weight gain and cognitive decline in children born to mothers prescribed AEDs during pregnancy. The relationship of social deprivation and epilepsy is explored where people diagnosed with epilepsy tend to come from areas of higher social deprivation. This chapter presents the strengths and limitations of studying the impact of epilepsy using routinely-collected data. Each study formed the basis of a published paper, in which footnotes are used to account for any specific work undertaken by co-authors.

3.1 Prevalence, Incidence and the Social Deprivation Profile of Epilepsy in Wales

The aim of this retrospective study was to determine if prevalence and incidence of epilepsy is due to social drift or social causation by using GP records and demographic

data between 2004-2010. The first objective was to identify people with epilepsy within SAIL. The GP dataset in SAIL (see 3.1) contains READ codes pertaining an interaction with a GP that was recorded in a patient's GP record. Various codes can be entered such as diagnoses, prescriptions, symptoms, laboratory tests and medical advice. This dataset was used to query codes for both diagnosis codes for epilepsy as well as anti epileptic drug prescriptions (AED). In discussion with clinicians within the Swansea Neurology Research Group, an appropriate method discussed to extract epilepsy cases was to use a combination of repeat AED and epilepsy diagnosis codes. This would pick people with unresolved epilepsy while also excluding people with AED prescriptions exclusively for mental health disorders and pain management. The use of a repeat AED prescription was also included to exclude uses of diagnosis codes used as a way to recorded suspected diagnoses that require follow up.

Table 3.1: SAILWGPV.EVENT_ALF_E is a table in the SAIL Databank that stores GP patient records.

Field name	Description
PRAC_CD_E	Encrypted General Practice code
ALF_E	Anonymous linking field representing an encrypted NHS number
WOB	Week of Birth - defaults to Monday of week of birth
GDNR_CD	Gender code 1=male, 2=female, 9=unknown
LOCAL_NUM	Local number identifier - a unique patient number
EVENT_CD_VRS	Determines code type such as READ v2, SNOMED etc
EVENT_CD	Recorded clinical information during the event
EVENT_VAL	The value associated with the recorded event
EVENT_DT	Date of the event
EPISODE	Denotes if event is due to ongoing care or first recording of diagnosis
SEQUENCE	The number of records for a specific event

The EVENT_CD column was queried with a list of READ codes that defined AED prescriptions and epilepsy diagnosis, and the WOB column was used to define week of birth for splitting patients into age bands, particularly between adults and children as a child's social deprivation is unlikely to be effected by an epilepsy diagnosis.

3.1.1 Defining Epilepsy in the SAIL Databank

The Quality of Outcomes Framework (QOF) `https://www.nice.org.uk/standards-and-indicators/qofindicators` aims to encourage GPs to keep patient records as complete as possible. By providing a paid incentive to use certain READ codes to record patient details, finding patients with diseases such as epilepsy should be

possible by querying GP records. The usage of READ codes used to record details of patients with epilepsy was explored to determine if it is possible to select people with known epilepsy.

Patients receiving drug treatment for epilepsy in the last six months		
Epilepsy	F25%	excluding F2501; F2511; F2516; F256.%; F258.–F25A.; F25y4
Progressive myoclonic epilepsy	F1321	
Traumatic epilepsy	SC200	
Patient receiving AEDs	dn%	
Patient receiving AEDs	d26%	
Epilepsy Exceptions*		
Patient unsuitable	9h61	
Informed dissent	9h62	
Patients age 16 and over on drug treatment for epilepsy who have a record of seizure frequency in the previous 15 months (4 points 90%)		
Fit frequency	6675	
Seizure-free >12 months	667F	
No seizures on treatment	667P	
1–12 seizures a year	667Q	
2–4 seizures a month	667R	
1–7 seizures a week	667S	
Daily seizures	667T	
Many seizures a day	667V	
Patients age 16 and over on drug treatment for epilepsy who have been seizure-free for last 12 months recorded in last 15 months		
Seizure-free >12 months	667F	
Epilepsy resolved.	21260	Note:This is a five-byte code that can only be used in five-byte systems)

Figure 3.1: Table 1 of 2 defining QOF codes for recording information on epilepsy in patient records. Table taken from `https://www.epilepsy.org.uk/sites/epilepsy/files/primary-care-resource/A18-Tool.pdf`

Epileptic seizure-free exception	
Patient on maximal tolerated anticonvulsant therapy	8BL3
Epilepsy medication review	8BIF
Epilepsy drug side-effects	6677
Epilepsy treatment changed	6678
Epilepsy treatment started	6679
Epilepsy treatment stopped	667A
Pregnancy advice	
Contraception advice read code	6110
Pre-conception advice read code	671J0
Pregnancy advice read code	67AF
Pregnancy advice – exception codes	
Contraceptive counselling inappropriate or declined	8IAg.; 8IB2.
Pre-conception advice inappropriate or declined	8IAh.; 8IB3.
Pregnancy advice inappropriate or declined	8IAi.; 8IB4.
Hysterectomy and equivalent	685H; 685I; 685K; 9O8Y; 7E05.%; 7E040; 7E042; 7E043; 7E046; 7E049; 7E04B; 7E04G; 7L0A.%; 26L3
Sterilisation	7E10.%; 7E111; 7E113; 7E115; 7E15.%; 7E160; 7E162; 7E1C; 7E1C3; 7E1D0; 159A; ZV25x; ZV252

Figure 3.2: Table 2 of 2 defining QOF codes for recording information on epilepsy in patient records.Table taken from `https://www.epilepsy.org.uk/sites/epilepsy/files/primary-care-resource/A18-Tool.pdf`

Using the READ codes tables in figures 3.1 and 3.2 an algorithm based on the presence of an epilepsy diagnosis code and a repeat AED prescription was used to determine if a person is known to be living with epilepsy on a given day by querying their GP records. Figure 3.3 depicts a timeline of how a combination of AED prescriptions and epilepsy diagnosis code entered into a GP record and are used to identify epilepsy, and figure 3.4 provides SQL code to implement the process.

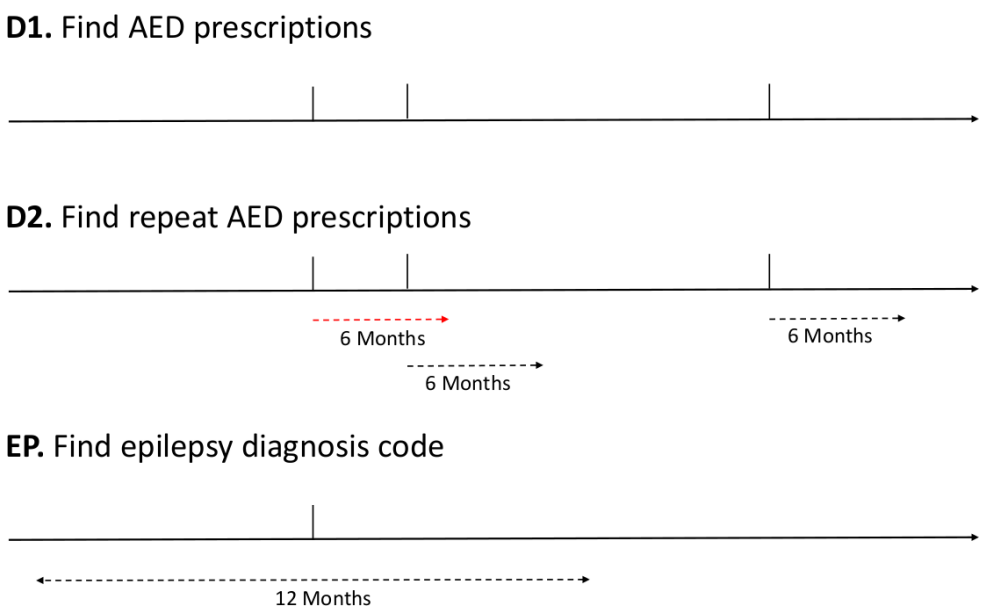

Figure 3.3: Visual explanation of the algorithm used to capture epilepsy diagnoses. All AED prescriptions are first found using GP records (D1), in which AED prescriptions pairs within 6 months after the initial prescription are classed as a repeat AED prescription (D2). Epilepsy diagnosis codes appearing in GP records 12 months either side of the first prescription of each of the AED pairs are queried, and where there is a match a person is classified by the algorithm as having an epilepsy diagnosis at the time of the first AED in the pair.

The SQL query in figure 3.4 was looped through the years 2004-2010, where for those who had a new diagnosis in a given year, all previous years were checked for absence of a diagnosis (incident cases). Those with both new and known epilepsy from previous years that satisfy the extraction criteria in later years contribute to the prevalent population of people with a diagnosis of epilepsy.

The SQL query in 3.4 was also used to determine prevalence and incidence of epilepsy, where prevalence is defined as the number of people with known epilepsy divided by the number of people in the population, and incidence is defined as the number of new cases of epilepsy in a given year divided by the number of people in the population. Annual prevalence between 2004 and 2010 was calculated by identifying the number of people with known epilepsy living in Wales on the 1st of January of a given year and dividing by the total number of people living in Wales on the same day. Annual incidence of epilepsy was also calculated by identifying all newly diagnosed patients in a given year, divided by the total number of persons registered as living in Wales in the same year.

```
-- first AED D1
SELECT DISTINCT D1.ALF_E, MIN(D1.EVENT_DT) AS FIRST_AED FROM
(SELECT DISTINCT ALF_E, EVENT_DT FROM SAILWLGPV.EVENT_ALF
WHERE EVENT_CD LIKE 'dn\%'
OR EVENT_CD LIKE 'do\%'
-- READ codes beginning with dn/do are AEDs
AND EVENT_DT BETWEEN '2000-01-01' AND '2000-31-12'
-- find first AED prescription in a given year
) \textbf{D1}

INNER JOIN

-- repeat AED D2
(SELECT DISTINCT ALF_E, EVENT_DT FROM SAILWLGPV.EVENT_ALF
WHERE EVENT_CD LIKE 'dn\%'
OR EVENT_CD LIKE 'do\%'
AND EVENT_DT BETWEEN '2000-01-01' AND '2001-06-01'
-- find potential repeat AEDs up to 6 months after first AED
) \textbf{D2}
ON D1.ALF_E = D2.ALF_E
-- match ALF_E in tables D1 and D2
AND D2.EVENT_DT BETWEEN D1.EVENT_DT AND D1.EVENT_DT + 6 MONTHS
-- search up to 6 months after first AED

INNER JOIN
-- epilepsy diagnosis code
(SELECT DISTINCT ALF_E, EVENT_DT FROM SAILWLGPV.EVENT_ALF
WHERE EVENT_CD LIKE 'F25\%'
-- READ codes beginning with F25 is an epilepsy diagnosis
AND EVENT_DT BETWEEN '1999-06-01' AND '2001-06-01'
-- allow 6 months before and after possible repeat
-- AED windows (i.e. suspected epilepsy)
) \textbf{EP}
ON D1.ALF_E = EP.ALF_E   # match ALF_E in tables D1 and EP
AND EP.EVENT_DT BETWEEN D1.EVENT_DT - 6 MONTHS AND D1.EVENT_DT + 6 MONTHS
-- search up to 6 months before/after first AED
}
```

Figure 3.4: An SQL query to extract epilepsy cases using a combination od diagnosis and AED READ codes. In this example, the first known confirmation that a person has epilepsy in 2000 is given by the first AED prescription in their GP record, while also having a repeat AED within 6 months as well as an epilepsy diagnosis up to 6 months either side of the first identified repeat prescription window

3.1.2 Social deprivation

The Welsh Demographic Service dataset was used to obtain social deprivation, which contains WIMD score that are assigned to 1896 geographical Lower Super-Output Areas (LSOAs) in Wales, each with around 1,500 people. Each LSOA is ranked from most deprived to least deprived according to its corresponding WIMD score and then grouped into deciles, with decile 1 being the most deprived and decile 10 the least deprived. For this specific study, WIMD 2011 deciles were used to measure an individuals' social deprivation where for each person in the study it was possible to query their demographic records to determine what LSOA code they live within on any day of the year, in which the 1st of January was chosen for each study year. Tables 3.2 and 3.3 were used to obtain a person's address and link it to their WIMD score by linking the LSOA_CD between both tables:

Table 3.2: The SAILWDSDV.AR_PERS table in the SAIL Databank holds individuals address, provided as the address when registering with a GP. Each address is also assigned to a Lower Super Output Area (LSOA) which is a geographical area comprising of around 1500 individuals.

Field name	Description
PERS_ID_E	Encrypted Person Identifier
RALF_E	Encrypted Residential address sourced from address given at GP registration
LSOA_CD	Lower Super Output Area code - approx 1500 person per area
FROM_DT	Day resident started living at address
TO_DT	Day resident stopped living at address

Table 3.3: The SAILREFRV.WIMD2008_OVERALL_INDEX table in the SAIL Databank contains a link between an LSOA code and various Welsh Index of Multiple Deprivation measures.

Field name	Description
LSOA_CD	Lower Super Output Area code ∼1500 people
LSOA_DESC	Name of geographic area
SCORE	Raw score comprising of 8 indicators
RANK	LSOA rank based on raw score. 1=most deprived
DECILE	Decile LSOA belongs to. 1=most deprived
QUNTILE	Quintile LSOA belongs to. 1=most deprived

For each ALF_E in the study it was possible to determine the WIMD Decile at the

```
select distinct arp.alf_e, w.wimd2008_QUINTILE
-- Get ALF_E from Welsh Demographic Service dataset
from SAILWDSDV.AR_PERS
inner join SAILWDSDV.AR._PERS arp
on arp.pers_id_e = argp.pers_i d_e
-- Get address data
join SAILWDSDV.AR_PERS_add AR
on arp.pers_id_e = ar.pers_id_e
-- Get WIMD quintile at 1st January 2004 i.e. census date
join sailx031v. LSOA_refr w
on w.lsoa_cd = ar. lsoa_cd
and '2004-01-01' between AR.from_dt and AR.to_dt
 -- ensure ALF_e is in contibuting SAIL gp practices
join SAILWLGPV.PATIENT_ALF_CLEANS ED GP
on ARGP.prac_cd_e = GP.prac_cd_E
}
```

Figure 3.5: An SQL query to link an ALF_E to their address in the Welsh Demographic dataset in SAIL, and how to linke the WIMD quintile to the address on a given day i.e. 1st January 2004

start of each year on the study window using the SQL query in Figure 3.5 and was used to compare the prevalence and incidence epilepsy across WIMD quintiles:

Sex and age were included as covariates where age groups were categorized as 0-5; 6-12; 13-21; 22-45; 25-45; 46-64, and 65 years or over in relation to their age in the study year. Figure 3.6 shows a summary flowchart of how the cohort was selected and Table 3.4 compares the study population in 2010 with that of the Welsh population.

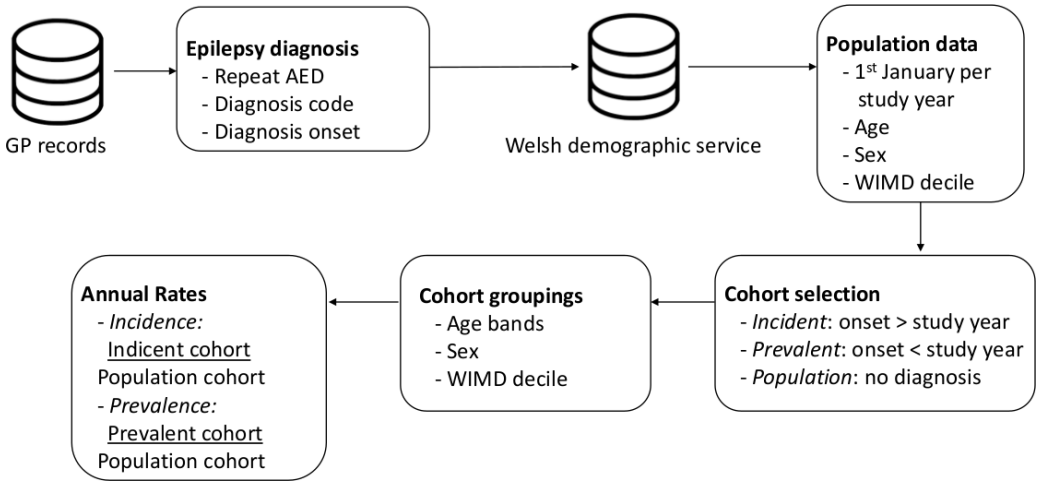

Figure 3.6: Flow chart of cohort selection. GP records were used to identify people with epilepsy (and therefore those that did not have epilepsy). The Welsh Demographic Service dataset was then used to sample age bands, sex and WIMD deciles on the first of January in each study year. For every unique combination of covariates, incidence and prevalence was calculated.

Table 3.4: Study population characteristics in 2010 as compared to the Welsh population (measured by the 2011 WIMD data). Table taken from [4]

		Study population in 2010	**Wales population**
Total Number		**1,178,558**	**3,169,594**
Sex	**Male**	588,476 (49.9%)	1,582,144 (49.9%)
	Female	590,082 (50.1%)	1,587,446 (50.1%)
Age (years)	**0-5**	73,716 (6.3%)	206,148 (6.5%)
	06-12	86,809 (7.4%)	235,681 (7.4%)
	13-21	142,333 (12.1%)	367,981 (11.6%)
	22-45	374,090 (31.7%)	999,254 (31.5%)
	46-64	290,612 (24.7%)	793,247 (25.0%)
	>64	210,998 (17.9%)	567,282 (17.9%)
Deprivation (WIMD decile)	**1**	109,703 (9.3%)	318,275 (10.0%)
	2	122,291 (10.4%)	315,689 (10.0%)
	3	91,478 (7.8%)	315,983 (10.0%)
	4	124,033 (10.5%)	317,000 (10.0%)
	5	121,894 (10.3%)	313,995 (9.9%)
	6	127,573 (10.8%)	325,662 (10.3%)
	7	101,077 (8.6%)	309,675 (9.8%)
	8	113,994 (9.7%)	324,457 (10.2%)
	9	124,752 (10.6%)	307,093 (9.7%)
	10	141,763 (12.0%)	321,954 (10.2%)

Over the study period, the mean epilepsy prevalence was 0.77% (95% CI 0.76 to 0.79%) and there were 2,390 incident cases of epilepsy, giving a mean incidence rate of 29.5/100,000 per year (95% CI 28.3 to 30.7). A breakdown of prevalence and incidence for each year is given in Table 3.5. Given that the sensitivity of the epilepsy algorithm in the validation study was 84% and the prevalence of epilepsy in the UK in 2011 was reported to be 0.97% by the Joint Epilepsy Council (`http://www.epilepsyscotland.org.uk/pdf/Joint_Epilepsy_Council_Prevalence_and_Incidence_September_11_%283%29.pdf`), the mean prevalence of 0.77% seems reasonable.

Table 3.5: Breakdown of epilepsy prevalence and epilepsy incidence by WIMD decile. Table tak

Deprivation (WIMD decile)	Mean epilepsy prevalence 2004-2010			Epilepsy incidence 2004-2010		
	Number of cases	Mean population	Mean prevalence (%)	Number of cases	Patient years at risk	Incid
1	1211	107,464	1.13	304	752,250	40.41
2	1164	119,990	0.97	305	839,931	36.31
3	845	89,671	0.94	228	627,696	36.32
4	969	121,421	0.8	270	849,944	31.77
5	989	119,345	0.83	257	835,416	30.76
6	911	125,983	0.72	247	881,878	28.01
7	712	99,640	0.71	186	697,479	26.67
8	742	111,456	0.67	197	780,195	25.25
9	734	122,153	0.6	208	855,069	24.33
10	684	140,053	0.49	188	980,374	19.18

Figure 3.7 shows that the mean prevalence of epilepsy is double in most deprived (1.13%) compared to least deprived (0.49%), and the mean incidence of epilepsy is also double in most deprived (40.41 per 100,000) compared to least deprived (19.18 per 100,000), identifying a strong trend that epilepsy is associated with increased social deprivation.

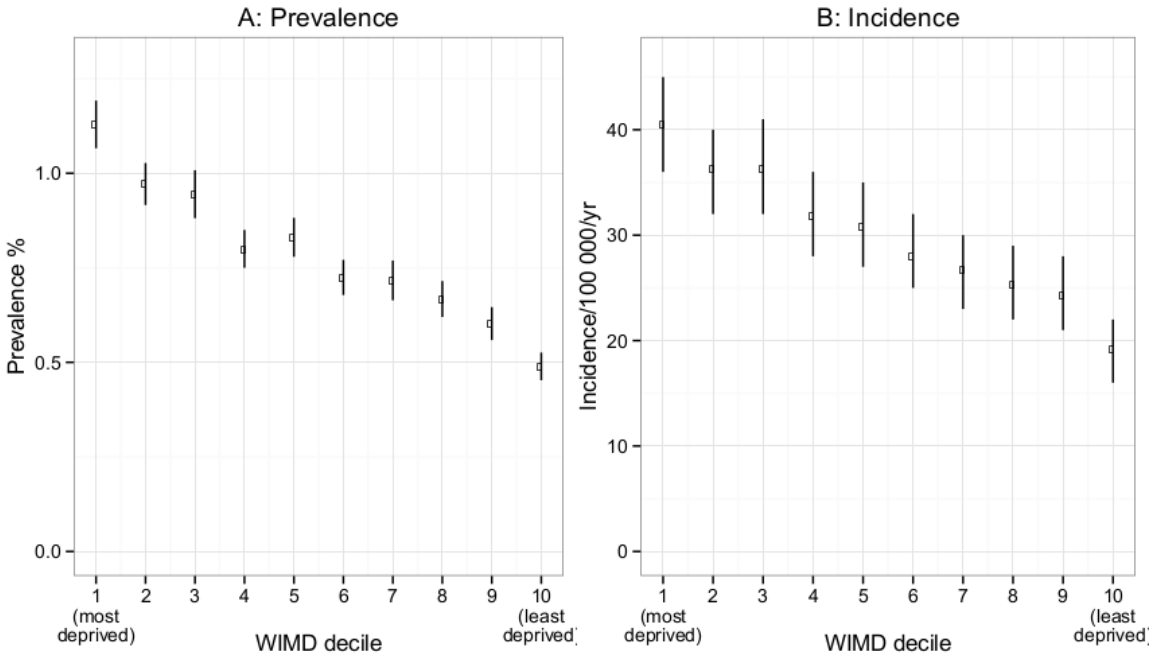

Figure 3.7: Plots of (A) epilepsy prevalence and (B) epilepsy incidence by WIMD (deprivation) decile. Error bars indicate 95% confidence intervals. Figure taken from [4]

The mean prevalence and incidence was calculated for each WIMD decile together with confidence intervals using binomial and Poisson models, respectively. The LSOA WIMD decile data for prevalence and incidence of epilepsy aligned with a 2001 LSOA shape file from the Office of National Statistics `https://data.gov.uk/dataset/fa883558-22fb-4a1a-8529-cffdee47d500/lower-layer-super-output-area-lsoa-boundaries` to produce a geographical representation of deprivation, epilepsy prevalence, and epilepsy incidence in Figure 3.8 [1]. Geographical areas were excluded where GP information was not available for at least 5% of the population of that area.

It is possible to see a correlation of deprived areas (dark blue) with areas that have a high prevalence and incidence of epilepsy. Due to densely populated urban areas skewing deprivation on an LSOA level, it can be seen in enlarged portions of the map (Swansea, Cardiff and Newport) that while it is possible to see the correlation

[1]With assistance from Dr Joanne Demmler

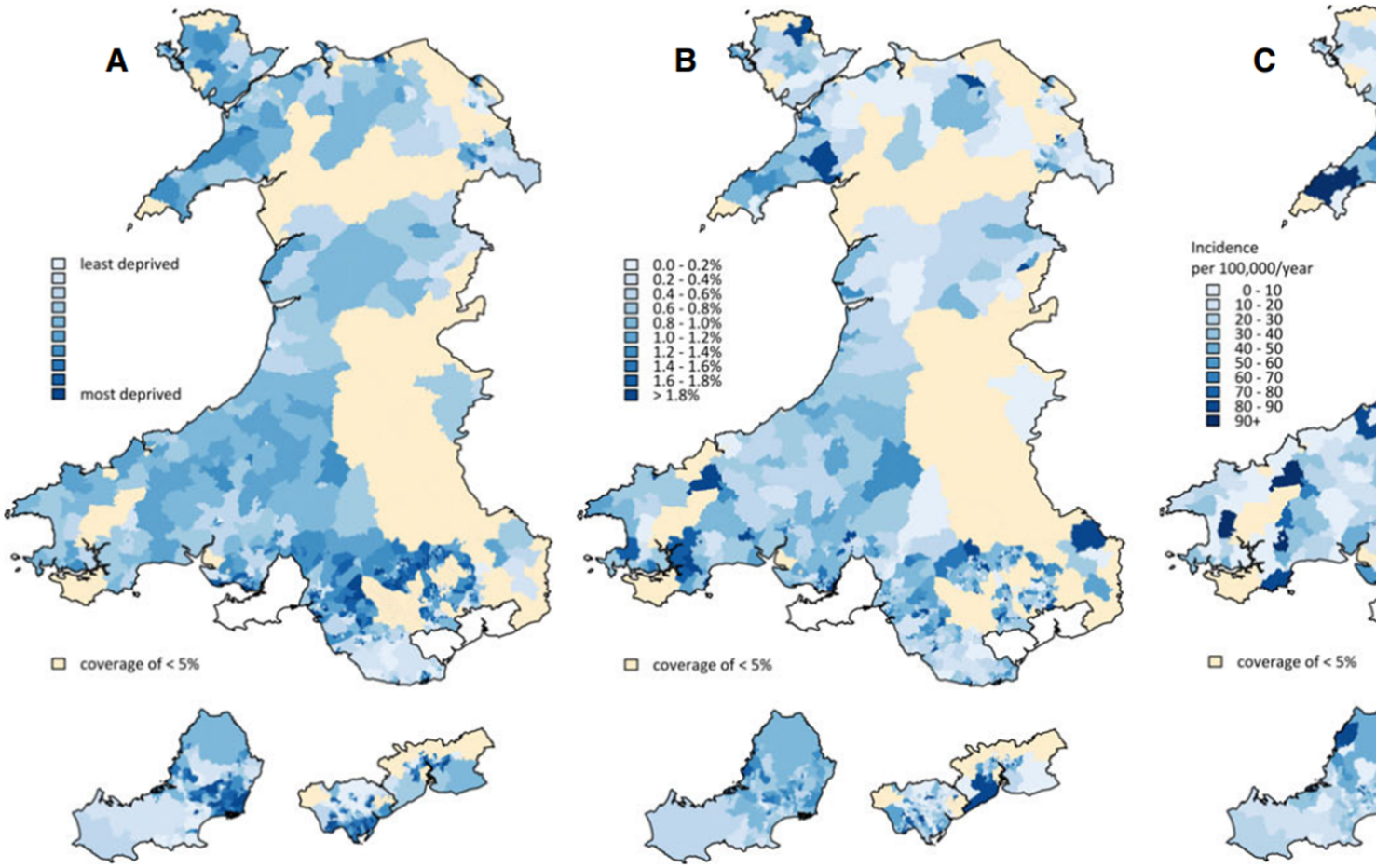

Figure 3.8: Maps of Wales showing each LSOA (areas with population of around 1,500); Yellow areas represent v of the population) and are not shown. (A) Deprivation measured by WIMD decile, (B) epilepsy prevalence, Enlarged areas represent the densely populated areas of the cities of Swansea, Cardiff, and Newport (left to rig

between social deprivation, prevalence and incidence of epilepsy, the correlation is not as clear when viewing Wales as a whole. Odds and incident rate ratios are presented in Table 3.6 for WIMD deciles, sex and age bands, where a WIMD decile of 1, males, and age band 0-5 are used as a reference where ORs and IRs were calculated using multiple logistic regression and Poisson regression models respectively [2]. It can be seen that even after adjusting for WIMD deciles, there is still a significant effect of epilepsy prevalence and incidence.

Table 3.6: Variable Adjusted epilepsy prevalence odds ratio Adjusted epilepsy incidence rate ratio The odds and incidence rate ratios for deprivation (second row of the table) are given per WIMD decile when compared to the population in decile 1, for example, the odds ratio of epilepsy prevalence in WIMD decile 3 = 0.922 x 0.94 when compared to the population in decile 1. Table taken from [4]

Variable	Adjusted prevalence odds ratio	Adjusted incidence rate ratio
Deprivation (per WIMD decile)	0.922 (0.920 to 0.925; p <0.001)	0.936 (0.923 to 0.950; p <0.001)
Sex		
Male	1.0 (ref)	1.0 (ref)
Female	0.981 (0.966 to 0.997; p = 0.018)	0.853 (0.787 to 0.924; p <0.001)
Age (years)		
0-5	1.0 (ref)	1.0 (ref)
6-12	2.572 (2.372 to 2.792; p <0.001)	0.999 (0.828 to 1.207; p = 0.993)
13-21	3.419 (3.169 to 3.694; p <0.001)	0.950 (0.799 to 1.134; p = 0.565)
22-45	5.570 (5.183 to 5.994; p <0.001)	0.573 (0.488 to 0.676; p <0.001)
46-64	6.371 (5.928 to 6.859; p <0.001)	0.567 (0.479 to 0.673; p <0.001)
>64	6.778 (6.304 to 7.300; p <0.001)	1.098 (0.935 to 1.296; p = 0.261)

3.1.3 Follow up cohort

A cohort of adults aged older than 18 years with a new diagnosis of epilepsy between January 1, 2000 and December 31, 2002 was selected as a follow up cohort to measure any difference in WIMD decile 10 years after their diagnosis. Only adults were selected, as a child's deprivation status is determined by their parents' deprivation status and children move with their parents. For each person in this cohort who remained within the study population, a comparison of WIMD decile between time of diagnosis and either 10 years after diagnosis or time of death was used to test the hypothesis that social drift plays a role in increased deprivation for people with epilepsy.

[2]Statistical Analysis performed by Dr Owen Pickrell

Table 3.7: Summary of follow up statistics for 10 year follow up cohort. Table taken fro

	Number	Mean length of follow-up in years (SD)	Mean length of follow-up in years (SD)	(Mean Decile
All	582	52.42 (20.2)	7.9 (3.3)	0.04 (p
Alive	352	42.96 (16.4)	10 (0.0)	-0.02 (
Younger than 41 years at diagnosis	172	28.85 (6.4)	10 (0.0)	0.05 (p
41 years or older at diagnosis	180	56.44 (10.7)	10 (0.0)	-0.08 (
Died	230	66.91 (16.7)	4.7 (3.3)	0.13 (p

613 new cases of epilepsy were identified in adults between January 1, 2000 and December 31, 2002. Thirty-one patients (5%) had moved out of the study population. Of the remaining 582 cases: 352 (60%) remained alive and were followed for 10 years; 230 (40%) died and were followed for a mean of 4.7 years (standard deviation [SD] 3.3 years). A Wilcoxon signed-rank test was used in the cohort study to test the null hypothesis that there was no significant change in WIMD decile following diagnosis. Table 3.7 summarizes the cohort population and figure 3.9 shows a graph of the change in WIMD decile.

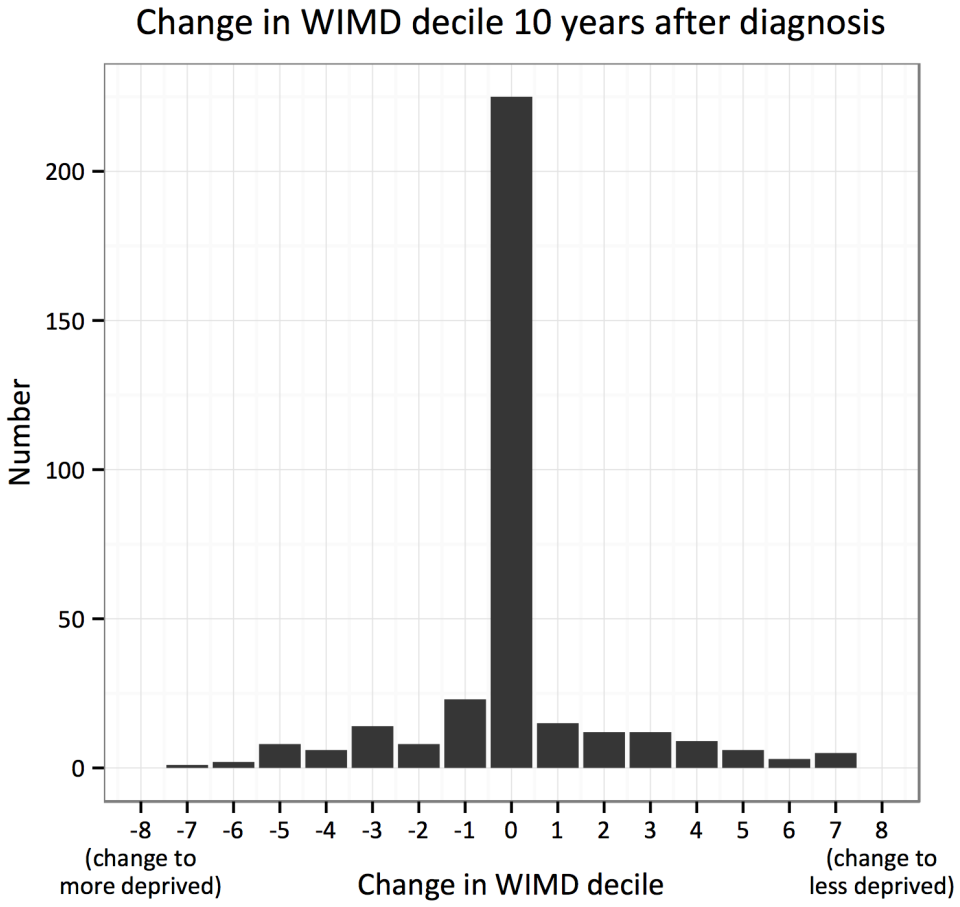

Figure 3.9: Changes in WIMD decile over 10 years for with incident epilepsy diagnosed between January 1, 2000 and December 31, 2002. Figure taken from [4]

3.2 Validating epilepsy status from electronic healthcare records

The previous study used an SQL algorithm to determine an epilepsy diagnosis from GP records and compare those identified as having epilepsy with patients in the Cardiff Epilepsy database. However, it was not possible to measure the specificity of the algorithm due to a lack of a comparison cohort of people that definitely did not have epilepsy. This study aimed to validate the accuracy of algorithms using GP records to identify people with epilepsy from anonymised, linked, routinely-collected Welsh healthcare data contained within the SAIL databank.

3.2.1 Study population

To validate epilepsy status through the use of READ codes in the SAIL GP records, a "gold standard" dataset of patients with known epilepsy was sourced using the Swansea Epilepsy Database within Morriston Hospital [3]. A comparison cohort of patients without epilepsy was sourced from general neurology clinics in Morriston Hospital. There were 918 patients from the Swansea Epilepsy Database with known epilepsy (283 (29%) generalised epilepsy, 510 (53%) focal epilepsy, 125 (13%) unclassifiable epilepsy and 42 (4%) with an uncertain diagnosis), of which 100 adults and 50 children were randomly selected to form the validation set of known epilepsy. A further 300 letters from general neurology clinic letters were manually reviewed to exclude those with known epilepsy, and 100 adults and 50 children were randomly selected to form the validation set of non-epilepsy patients. The 300 person cohort was then linked to their corresponding GP records in the SAIL databank.

3.2.2 Algorithm validation

Three different algorithms were tested to identify people with epilepsy within the SAIL Databank. Using a READ codes within GP records, diagnosis codes for epilepsy as well as AEDS were used in the following way

- A) individuals with an epilepsy diagnosis code and two consecutive anti-epileptic drug (AED) prescription codes within 12 months of diagnosis
- B) individuals with an epilepsy diagnosis code only
- C) individuals with two consecutive AED prescription codes only.

For a full list of READ codes used to define epilepsy please see the code list in

[3]Data sourced by Beata Fonferko-Shadrach

the appendix item 1. 145 of the 150 reference cases with epilepsy (97%) and 143 of the 150 reference cases without epilepsy (95%) were registered with a SAIL GP. True positive (TP) cases had a hospital diagnosis of epilepsy and were identified within SAIL as having epilepsy; true negative (TN) cases did not have epilepsy as confirmed by hospital records and were not identified as having epilepsy within SAIL; false positive (FP) cases did not have epilepsy as confirmed by hospital records and were identified as having epilepsy within SAIL; and false negative (FN) cases had a hospital diagnosis of epilepsy and were not identified as having epilepsy within SAIL. Positive predictive value (PPV) was defined as TP/(TP+FP); sensitivity TP/(TP + FN); specificity TN/(TN+FN) and false positive rate (FPR) as FN/(FN+TN). Youden's index (J) was then calculated using sensitivity plus specificity, as a measure of the accuracy of the algorithms. J ranges from -1 to 1 (J=1 for a perfect test)[235]. Confidence limits were calculated using the exact binomial method. The sensitivity, specificity, positive predictive value, false positive rate and accuracy of each of the three algorithms are shown in Table 3.10.

Patients within SAIL identified as having epilepsy			Hospital neurology service diagnosis of epilepsy		Positive predictive value (95% CI)	Sensitivity (95% CI)	False positive rate (95% CI)
Algorithm Used							
			Yes	No			
A - Epilepsy diagnosis & AED	All patients	Yes	122	2	98% (94–100)	84% (77–90)	1% (0–5)
		No	23	141			
	Adults	Yes	84	2	98% (92–100)	87% (78–93)	2% (0–7)
		No	13	96			
	Children	Yes	38	0	100% (91–100)	79% (65–90)	0% (0–8)
		No	10	45			
B - Epilepsy diagnosis only	All patients	Yes	125	5	96% (91–99)	86% (80–91)	3% (1–8)
		No	20	138			
	Adults	Yes	85	2	98% (92–100)	88% (80–93)	2% (0–7)
		No	12	96			
	Children	Yes	40	3	93% (81–99)	83% (70–93)	7% (1–18)
		No	8	42			
C - AED only	All patients	Yes	133	39	77% (70–83)	92% (86–96)	27% (20–35)
		No	12	104			
	Adults	Yes	91	38	71% (63–78)	94% (87–98)	39% (30–49)
		No	6	60			
	Children	Yes	42	1	98% (94–100)	88% (75–95)	2% (0–12)
		No	6	44			

Figure 3.10: The accuracy of algorithms A,B and C in being able to determine epilepsy status from GP reco

The results show that anonymised GP records can be used to accurately identify patients with epilepsy diagnosed by a hospital specialist in the UK. Since maximizing specificity was most important, while aiming to keep sensitivity as high as possible, algorithm A (specificity 99%, sensitivity 84%) is best suited to identify epilepsy from GP records, where algorithm C could be exclusively used for children(specificity 98%, sensitivity 98%). Algorithm B shows that using a diagnosis code for epilepsy alone also achieves a high level of accuracy in adults (specificity 98%, sensitivity 88%), which is a 1% increase in specificity over algorithm A, but when combining adult and children it has lower specificity than algorithms A and C respectively. These results compare well to other epilepsy validation studies conducted in Australian, Italian and American healthcare systems that report similar accuracy (specificity 100%, 99.8%, 94% and sensitivity 85.9%,81%,82%) [236–238], and this study is the first epilepsy validation study using gold standard patient records accuracy in the UK.

There is a clear difference in epilepsy reporting in GP records between adults and children. It appears that GPs record a diagnosis code for a lower proportion of children than adults, resulting in only 79% sensitivity for children using algorithm A, but this is likely due to the many years required to determine a clear diagnosis of epilepsy in children. There is a large difference in specificity between adults and children for algorithm C (61%-98%) where it is likely conditions other than epilepsy in adults (e.g. migraine, mental health disorders and neuropathic pain) are classified as having epilepsy by the algorithm, but AEDs are rarely prescribed for anything other than epilepsy in children in the UK [239].

There was little difference in performance between algorithms A and B. Algorithm A (epilepsy + a repeat prescription) had slightly higher specificity than algorithm B (epilepsy diagnosis only) and algorithm A had slightly higher sensitivity, but their overall accuracy was comparable as seen by their Youden's Index measurement. From this study, it seems that GP diagnosis codes for epilepsy could be used on their own to identify people with epilepsy form GP records in the UK, which can be explained in that an epilepsy diagnosis should be made in secondary care by a specialist in the UK, and then recorded by GPs in GP records [240].

3.3 Educational attainment of children born to mothers with epilepsy

There are currently many AEDs used for seizure control, but some methods of seizure control during pregnancy can have effects on the unborn child. Valproate is the most effective drug for treating genetic generalized epilepsy,[26] but recent prospective psychometric studies have demonstrated cognitive impairment and neurodevelopmental disorders in 30-40% of children exposed to valproate *inutero*,[241, 242] as well as a significant decrease in intelligence quotient(IQ)[27, 243]. Women with epilepsy who have satisfactory control with valproate and are planning a family therefore have a difficult decision to make. In the United Kingdom the Medicines and Healthcare Products Regulatory Agency (MHRA), issued stringent guidance for all clinicians prescribing valproate to women of child-bearing potential in 2015. An International League Against Epilepsy (ILAE) task force made seven recommendations, the first of which is where possible, valproate should be avoided in women of childbearing potential. Women with epilepsy who are taking AEDs are presently advised to continue them throughout pregnancy, primarily because of the risks of convulsive seizures to mother and her unborn child.

To be able to counsel mothers adequately about the risks of uncontrolled seizures during pregnancy and cognitive outcomes for their children, it is important to know whether the psychometric differences demonstrated in research conditions translate to children in the community. This study was conducted to investigate the effect of AED exposure *inutero* on the educational attainment of children born to mothers with epilepsy using anonymised, routinely-collected healthcare records and the results of a standard national educational assessment.

3.3.1 Cohort selection

The Child Health dataset in the SAIL databank was used to select encrypted identifiers for children as well as a linked ID to the mother. Gestational age, maternal age were also extracted as covariates used for control matching between mothers with epilepsy at time of birth and those without. All fields in the Child Health dataset are given in table 3.8.

Table 3.8: SAILCHDV.CHILD is a table in the SAIL Databank that contains birth records and relates each child's NHS number to their mother.

Field Name	Description
CHILD_ID_E	Internal child ID
ALF_E	Encrypted NHS number of child
MAT_ALF_E	Encrypted NHS number of mother
MAT_WOB	Week of Birth of mother
WOB	Week of Birth of child
BIRTH_WEIGHT	Weight of the child at birth.
BIRTH_WEIGHT_CAT	Derived variable. Classes for birth weights.
BIRTH_TM	Time of birth
GNDR_CD	Sex of child
APGAR_1	APGAR score taken at 1 minute
APGAR_2	APGAR score taken at 5 minutes
GEST_AGE	Best estimate of gestation at time of delivery
TOT_BIRTH_NUM	Number of deliveries for multiple births
BIRTH_ORDER	Order of deliveries for multiple births, by ALF_E
MAT_AGE	Age of mother in years at delivery
PROV_SITE_CD	Hospital provider site code
STILLBIRTH_FLG	Stillbirth flagged only in case of stilbirth
DOD	Date of death of child
LHB_CD	Local healthboard code
DEL_CD	Delivery code indicating type of delivery
LABOUR_ONSET_CD	Method of labour onset
MOTHER_CARE_CD	Type of maternity care allocated for mother
BREASTFEED_BIRTH_FLG	Breastfeeding at birth
BREASTFEED_8_WKS_FLG	Breastfeeding at 8 weeks
LSOA_CD	Lower Super Output Area containing mother's address

For each birth record, the mother's data was linked to their social deprivation as described in the social deprivation study earlier in this chapter. WIMD quintiles at the time birth were used as an additional covariate during the control matching procedure. These data were then linked to the GP dataset in SAIL using algorithm A from the epilepsy validation study (epilepsy diagnosis + repeat AED prescription) to determine if the mother had known epilepsy during the pregnancy. A control group was created (with 4:1 matching) matched for maternal age, week of gestational age, and WIMD decile at the time of birth between mothers who had known epilepsy during pregnancy and those that did not have epilepsy.

3.3.2 Education dataset

Educational attainment data for Key Stage 1 from the Department for Children Education, Lifelong Learning and Skills (DCELLS) dataset was available in the SAIL databank between the years 2003-2008. The DCELS dataset contains attainment for children in mathematics, language (English or Welsh) and science in which each subject is awarded a level between 1 (lowest) and 3 (highest). In certain circumstances children may obtain an unclassified or working towards meaning that they do not achieve the required grade to pass the year. The core subject indicator (CSI) is defined as the proportion of children achieving a minimum standard in all three KS1 subjects, that being a level 2 or higher in each subject. Given that KS1 results (taken at the age of 7) were only available within SAIL for the years 2003-2008, SAIL GP records were queried for women with epilepsy who gave birth between 1996 and 2001. Table 3.9 shows all field available within the Key Stage 1 dataset:

Table 3.9: SAILDCELV.PRE16_KS1 is a table in the SAIL Databank contains all-Wales education data between 2003-2008 for Key Stage 1. Three subjects (Maths,Science,Engish/Welsh) as well as a Core Subject Indicator are provided to indicate the level of attainment per child

Field name	Description
BATCH_NUM	Batch number
LEA	Local education Authority code
ESTAB_E	Encrypted educational establishment code
PUPIL_IRN_E	Internal pupil reference number
CSI	Core Subject Indicator
EN1	English teacher assessment 1
EN2	English teacher assessment 2
EN3	English teacher assessment 3
ENSUB	English teacher assessment subject level
MA1	Maths teacher assessment 1
MA2	Maths teacher assessment 2
MA3	Maths teacher assessment 3
MASUB	Maths teacher assessment subject level
SC1	Science teacher assessment 1
SC2	Science teacher assessment 2
SC3	Science teacher assessment 3
SC4	Science teacher assessment 4
SCSUB	Science teacher assessment subject level
CY1	Welsh teacher assessment 1
CY2	Welsh teacher assessment 2
CY3	Welsh teacher assessment 3
CYSUB	Welsh teacher assessment subject level
YEAR	Census year
URN	Internal school reference number
NEWBES	Pupils from non English/Welsh education system

Each child who was born to mothers with known epilepsy during pregnancy was then linked to their education data and compared to those children not born to mothers with epilepsy. Figure 3.11 shows each step of the cohort ascertainment and linkage:

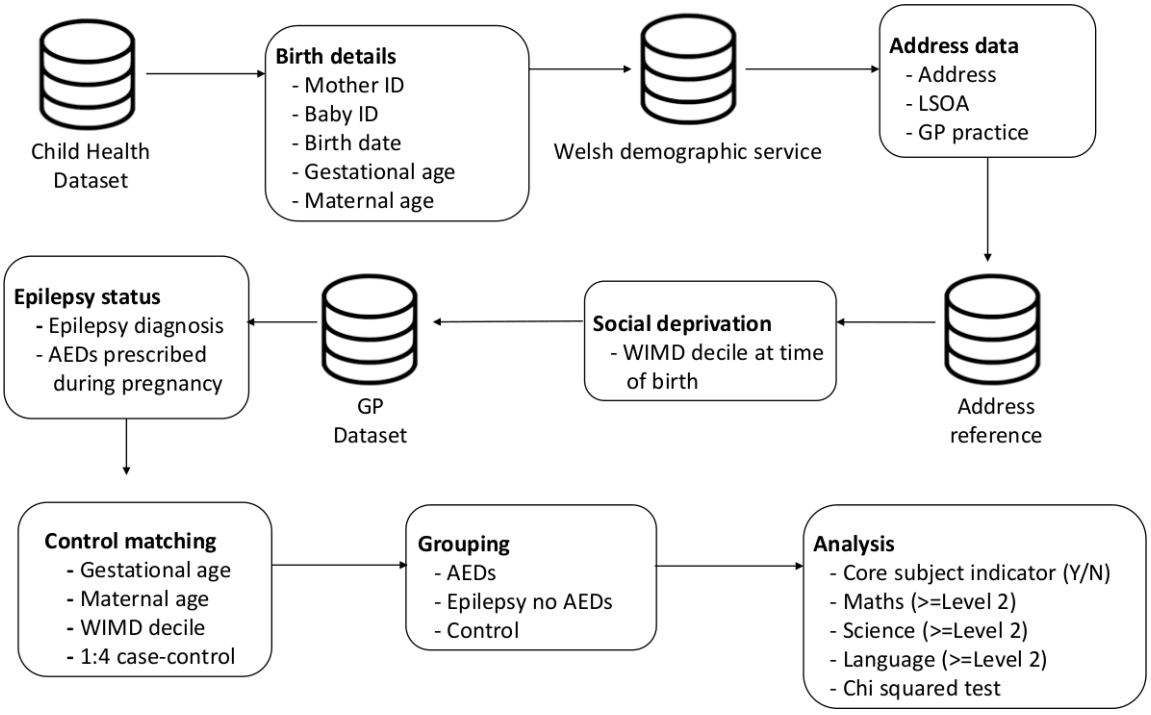

Figure 3.11: Flow chart of cohort ascertainment. 4 datasets were queried: General Practice, ONS Births, Welsh Demographic Service and Welsh Education Dataset.

3.3.3 Results

A total of 440 children were identified with KS1 results available between 2003 and 2008 who had mothers with epilepsy diagnosed before their pregnancy, and the mothers were stratified into five groups based on AED prescription during pregnancy (carbamazepine, lamotrigine, sodium valproate, multiple AEDs or no AEDs prescription) - see table 3.12. Only prescription information was available, but it is not expected that adherence differs across different AED prescriptions.

The proportion of children in each group achieving at least a level 2 in each subject is shown in figure 3.13.

	COHORT		N	Male (%)	N children w/ epilepsy	Mean birth weight / kg (sd, p-value*)	Mean maternal age / years (sd, p-value*)	Mean ges / weeks (sd, p-valu
Mothers with epilepsy	Carbamazepine		84	43 (51)	2	3.17 (0.69, 0.03)	28.51 (5.44, 0.02)	38.71
	Lamotrigine		24	10 (42)	1	3.38 (0.62, 0.68)	24.75 (5.29, 0.09)	38.72
	No drug		178	95 (53)	9	3.28 (0.54, 0.25)	26.39 (5.53, 0.59)	39.31
	Sodium valproate		115	58 (50)	7	3.32 (0.53, 0.86)	25.75 (5.62, 0.10)	39.11
	Polytherapy		39	20 (51)	2	3.16 (0.69, 0.14)	27.84 (6.03, 0.21)	38.24
		With valproate	20	12(60)	0	3.23(0.57,0.3)	26.95(5.58,0.8)	38.
		With other	19	8(42)	2	3.08(0.81,0.13)	28.78(6.49,0.16)	37.7
Control group			1,756	889 (51)	0	3.33 (0.55,1.00)	26.43 (5.49, 1.00)	39.09
All Wales			159,849	82,049 (51)	0	3.37 (0.58, 0.02)	27.42, (5.81, < 0.001)	39.27,(2

Figure 3.12: Descriptive statistics of the study cohort. The control group comprised of a 1:4 match on matern Welsh Index of Multiple Deprivation (WIMD) quintile. WIMD quintiles are a measure of deprivation (see me the most deprived and quintile 5 being the least deprived. sd=standard deviation. *p-values are for comparisons the "Mothers with Epilepsy" group with the control group. Table taken from [3]

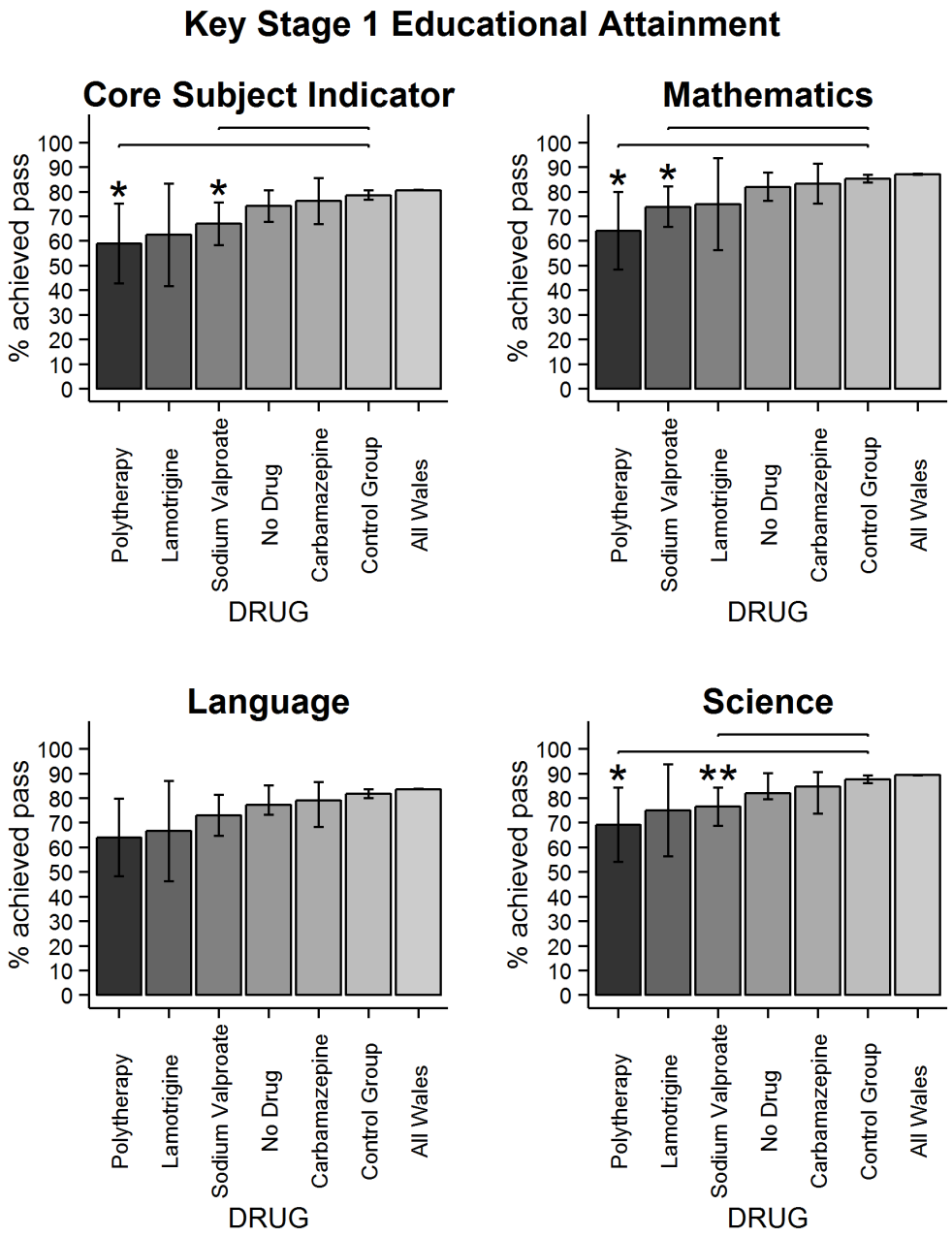

Figure 3.13: Key Stage 1 results stratified by subject and study groups. Each group was compared to the matched control group. Significant differences in attainment (* p ¡ 0.05, ** p ¡ 0.005) between each group and the matched control are shown. The p-values have been Bonferroni corrected for multiple testing (see Methods section). The All Wales group is shown as a regional comparator only and not used to test for significance. Figure taken from [3]

These results show that children born to mothers with epilepsy being prescribed sodium valproate during pregnancy have a significantly lower level of achievement in KS1 tests across all indicators, with fewer children achieving the minimum standard when compared to the matched control group by (CSI = -12.7% less than the control group, mathematics = -12.1%, language = -10.4%, science = -12.2%). Also fewer children born to mothers with epilepsy being prescribed multiple AEDs during

pregnancy achieved the national standard in KS1 tests when compared to the matched control group by (CSI = -20.7% less than the control group, mathematics = -21.9%, language = -19.3%, science = -19.4%). There was no significant decrease in attainment in children born to mothers with epilepsy that were not prescribed an AED during pregnancy according to their GP records. Excluding children with epilepsy and mothers who were recorded as smoking during pregnancy did not change the significance of these results.

3.4 Chapter summary

This chapter has presented various epidemiological studies in epilepsy using linked, anonymized healthcare data in the SAIL Databank. A study of GP coding of epilepsy was undertaken by comparing epilepsy diagnoses in GP records held in the SAIL databank to a gold standard dataset in Morriston hospital of patients with and without an epilepsy. The results showed that using a repeat AED prescription as well as an epilepsy diagnosis code is important in capturing epilepsy patients with high specificity, while also maintaining good sensitivity. Using this algorithm the incidence and prevalence of epilepsy in Wales was linked to social deprivation using the Welsh Index of Multiple Deprivation and showed that in more deprived areas there is both a higher incidence and prevalence of epilepsy. In a follow up study of newly diagnosed epilepsy patients, there appeared to be no increase in social deprivation leading to the conclusion that social deprivation in epilepsy is due to social causation rather than social drift. Finally, the effects of AEDs prescribed during pregnancy on children's Key Stage 1 educational attainment was explored. Children born to mothers that were prescribed sodium valproate or multiple AEDs in combination perform worse than a control group and has important cognitive outcomes for pharmaco-exposed children.

Chapter 4

Using Natural Language Processing techniques to extract clinical information from unstructured text

This chapter aims to explore the potential of extracting rich information from epilepsy clinic letters using NLP techniques. The motivation for extracting information from clinic letters when routinely collected information is already available for research purposes, is that the information available is often limited in detail. For example, in all of the studies presented in the previous chapter they each lack specific epilepsy and seizure type, dosage details for prescribed drugs as well as finer details such as results from EEG and MRI scans. These data are available in other sources of information, namely free texts in healthcare settings. Manually reading through clinic letters to obtain rich information is time consuming, and so an automated method would be desirable to extract this data. The results in this chapter present an NLP method to extract rich epilepsy information from clinic letters stored in Morriston hospital and the Swansea Epilepsy Database.

4.1 Clinic letters

The Swansea Epilepsy database was used to source patients with epilepsy that had clinic letters written by epilepsy specialists at Morriston hospital. Such letters contain very detailed information regarding a patient's epilepsy such as seizure type, seizure frequency and results of examinations and investigations, and even contains information where a patient experiences symptoms similar to epilepsy that is in fact

due to non-epileptic attack disorder. Permission to use clinic letters from the Swansea Epilepsy database was given under the condition all patient details were psuedo anonymized, where validation of any algorithm was undertaken by a clinician. 240 clinic letters were manually de-identified hospital clinic letters and used to build and test the algorithm [1]. 40 letters were used for training purposes to build rule sets, and a validation set of 200 letters to test the accuracy of the algorithm. The validation set contained letters originating from various outpatient clinics (145 adult epilepsy, 37 paediatric epilepsy, and 18 general neurology), from first and follow-up appointments, and written by eight different clinicians.

4.2 A rule based NLP approach to extract epilepsy information from clinic letters

Two approaches were considered when developing the NLP pipeline - machine learning and manually constructing rule sets. Machine learning based approaches require vast amounts of training data for the algorithm to achieve high accuracy where rule sets can take advantage of human knowledge when constructing rules. Given that only 240 letters were available for this study, this limitation was considered when deciding between a rule based and machine learning based NLP approach to analyse these letters. Due to the very large training datasets required for machine learning purposes (tens of thousands), a rule based approach in which human knowledge could be quickly built into logical rules and processed by a computer program was favoured. Therefore a rule based NLP approach was used to build an epilepsy clinical extraction pipeline that could capture data within epilepsy clinic letters.

4.2.1 The General Architecture for Text Engineering

The General Architecture for Text Engineering (GATE) framework was used to build a rule based NLP pipeline. Two open source applications freely available and configured for GATE were used - the biomedical named entity linking pipeline (Bio-YODIE plugin) and the South London and Maudsley medication application (SLaM). The main focus was to map clinical terms found in text to the Unified Medical Language System (UMLS) concepts so that a structured dataset could be constructed, much like the datasets that exist within the SAIL Databank. The ANNIE pipeline was used for basic POS tagging and sentence boundary detection, and the JAPE scripting language was used to program various rule sets using items of information tagged using Bio-YODIE, SLaM, ANNIE as well as custom dictionaries defined to supplement

[1] De-identification performed by Beata Fonferko-Shadrach

these plugins. Figure 4.1 shows a flow chart of the various pipeline components:

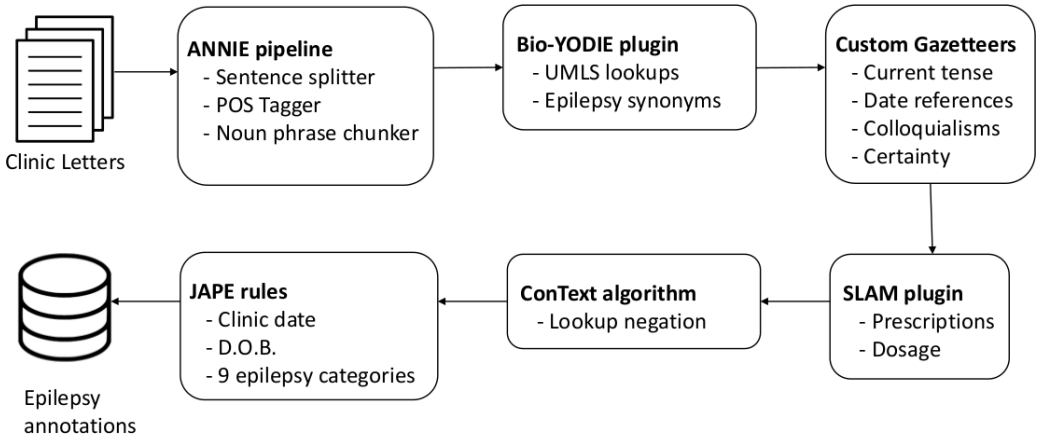

Figure 4.1: Overview of the GATE pipeline and the various components used to generate annotations

A list of predefined categories specified by a neurologist [2] formed the basis of the important information to be extracted from the clinic letters:

[2]List provided by Dr Owen Pickrell

Table 4.1: Definitions of each category intended to be extracted

Category	Details
Clinic date	The date the patient visited the clinic.
Date of Birth	The patient's date of birth.
Epilepsy diagnosis	Items of information which confirmed a diagnosis of epilepsy e.g. "this lady has a diagnosis of focal epilepsy" or "... has recurrent unprovoked generalised tonic-clonic seizures". Frequently there is diagnostic uncertainty in epilepsy clinic letters e.g. "this lady probably has frontal-lobe epilepsy" or "I am uncertain whether the blackouts are epileptic"; and so we defined five levels of certainty (1=no diagnosis, 2=unlikely, 3=uncertain, 4=likely, 5=definite) to each information item associated with an epilepsy diagnosis. We specified that the epilepsy diagnosis must be attributable to the patient (e.g. not a family member); and did not include items of information that described epilepsy clinic attendance, or a discussion about epilepsy in general, as confirmation of an epilepsy diagnosis.
Epilepsy type	Whether the patient had focal or generalised epilepsy or an epilepsy syndrome where epilepsy type could be inferred. For example generalised epilepsy if the letter confirmed juvenile myoclonic epilepsy. This was based on the UMLS CUI extracted with the epilepsy diagnosis information. We only used explicit mentions of epilepsy types or syndromes within the clinic letters, and did not use other information, such as seizure type or investigation results, to infer epilepsy type.
Seizure type	Specific seizure types e.g. "focal motor seizures" or "absence seizures". Seizures types were categorized into focal seizures or generalised seizures at the validation stage.
Seizure frequency	The number of seizures in a specific time period e.g. "two seizures per day", "seven seizures in a year", or "seizure free since last seen in clinic."
Medication	An identifiable drug name with a quantity and frequency e.g. "Lamotrigine 250mg bd".
Investigations	The type of investigation and classification of results (normal or abnormal). UMLS CUI codes were used to assign a normal / abnormal value to investigation results, using the simplified abnormal outcomes gazetteers. We categorised the investigation results into CT, MRI, and EEG results at the validation stage.

4.2.2 Defining rules

The following sections demonstrates the specifics of how the rules for each category were built using the components in 4.1.

```
Phase: Dates
# accept TIMEX3 and Lookup2 annotations
# Lookup2 are user defined Lookups desgined to signify keywords
Input:TIMEX3 Lookup2
Options: control=appelt

Rule: ClinicDate

( #includes words such as "clinic", "hospital", "outpatient" etc
  ({Lookup2.majorType == "organization", Lookup2.minorType == "health_term"})
  # followed by a TIMEX3 annotation where date is explicit i.e. full date
  ({TIMEX3.foundByRule == "date_r1b-explicit"}|
  {TIMEX3.foundByRule == "date_r0h-explicit"})

):match
-->
# create new annotation "ClinicDate"
:match.ClinicDate = {rule = ClinicDate1, value = :match.TIMEX3.timexValue}
```

Figure 4.2: JAPE script to obtain clinic datae given an input of TIMEX3 and customs annotations relating to clinic visits (LOOKUP2)

4.2.3 Clinic date and date of birth

Within each letter there were various dates pertaining to different items of information such as referring to previous clinic visits, date of scans and prescriptions, as well as clinic date, date of birth and date the letter was typed up by administrative staff. The TIMEX plugin available within GATE was used to extract dates written in a variety of ways (01/01/2001 or 1st of January 2001 etc.) that were also found within the context of words/phrases suggesting clinic visits defined by custom gazetteers. The JAPE script in figure 4.2 was used to extract clinic dates.

Similarly date of birth was captured by combining TIMEX3 annotations that specify full dates i.e. day/month/year with strings such as "D.O.B", "DOB:" and "Date of birth".

4.2.4 Epilepsy diagnosis, epilepsy type and seizure type

Rules were built to capture phrases related to an epilepsy diagnosis attributed to a patient. Some sample phrases, where only the first two phrases would be considered to have a confirmation of epilepsy using the algorithm developed in this chapter are:

I suspect he has generalized epilepsy
She was diagnosed with focal epilepsy
She doesn't have epilepsy, but has non-epileptic attacks
I saw this gentleman regarding epilepsy

The first step was to identify words within phrases that indicate a mention of

```
select distinct CUI,STR,SAB,CODE from mrconso where STR = "epilepsy";
+----------+----------+-------------+-----------+
| CUI      | STR      | SAB         | CODE      |
+----------+----------+-------------+-----------+
| C0014544 | Epilepsy | ICD10       | G40       |
| C0014544 | Epilepsy | MTH         | NOCODE    |
| C0014544 | Epilepsy | SNOMEDCT_US | 267698007 |
| C0014544 | Epilepsy | SNOMEDCT_US | 84757009  |
+----------+----------+-------------+-----------+
4 rows in set (0.00 sec)
```

Figure 4.3: MySQL script used to query the UMLS RRF files. The table MRCONSO contains a list of all CUIs that encompass various different coding systems such as ICD 10 and SNOMED CT. By searching the STR column for the word epilepsy, the corresponding CUI is given

epilepsy. The Bio-YODIE plugin was used to map any term found in a document to a medical concept as part of the UMLS ontology, defined as a Lookup. To look specifically at epilepsy concepts, a gazetteer of epilepsy terms was built by specifying CUI codes relating to epilepsy and used to filter all Lookups within a document. Using an installation of the UMLS Metathesaurus Rich Release Format (RRF) files `https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html` the following MySQL scripts were used to query the UMLS relationship datasets for sub codes of epilepsy:

The mysql script in figure 4.3 finds all possible CUIs for the description "epilepsy" of which there is one CUI (C0014544) that unifies other existing coding systems that also describe epilepsy. This CUI was then used in script 4.4 to find child codes, or sub codes of epilepsy using the relationship file MRREL:

The script example in 4.4 shows an example list, limited to 15 items (out 2532 in total) of UMLS concepts and CUIs that are children of the epilepsy CUI by linking UMLS concepts in the MRCONSO table to the MRREL relationship table. These CUIs were used to build a Flexible Gazetteer which functions as a filter for annotations produced by the Bio-YODIE plugin. Once a subset of epilepsy terms found by Bio-YODIE has been produced, the terms found require further context to form a diagnosis. Each Bio-YODIE Lookup annotation has multiple attributes, of which the following were used to write a diagnosis annotator:

- **Negation** (Context plugin) - if the term has a negative context i.e. does not have epilepsy
- **TUI/Unique Identifier Type** (UMLS) - each concept within UMLS is

```
select distinct a.CUI1, a.CUI2, b.STR as child from mrrel a #relationship file
-- join relation dataset to mrconso
inner join mrconso b
-- on the parent CUI = the child CUI
on a.CUI2 = b.CUI
-- join child CUIs to mrconso
inner join mrconso c
-- on child CUI = child CUI
on a.CUI1 = c.CUI
-- specify parent as epilepsy
where c.CUI like "C0014544%" and REL = "CHD"
-- get first 15 rows only
limit 15;
+----------+----------+-----------------------------------------------------+
| CUI1     | CUI2     | child                                               |
+----------+----------+-----------------------------------------------------+
| C0014544 | C0014544 | Epilepsy                                            |
| C0014544 | C0014544 | Epilepsy NOS                                        |
| C0014544 | C0014544 | Epilepsy, NOS                                       |
| C0014544 | C0270850 | Idiopathic generalized epilepsy                     |
| C0014544 | C0270850 | Idiopathic generalised epilepsy                     |
| C0014544 | C0270850 | Idiopathic generalized epilepsy, NOS                |
| C0014544 | C0494475 | Tonic-clonic seizures                               |
| C0014544 | C0494475 | Tonic-clonic seizure                                |
| C0014544 | C0494475 | Tonic - clonic seizures                             |
| C0014544 | C0477371 | Other epilepsy                                      |
| C0014544 | C0477370 | Other generalized epilepsy and epileptic syndromes  |
| C0014544 | C0014553 | Absence Epilepsy                                    |
| C0014544 | C0494474 | Special epileptic syndromes                         |
| C0014544 | C2584947 | Anoxic epileptic seizure                            |
| C0014544 | C2919602 | Witnessed epileptic seizure                         |
+----------+----------+-----------------------------------------------------+
```

Figure 4.4: MySQL script used to query the UMLS RRF files and find all child codes of epilepsy

attributed to a semantic type e.g. "Disease or Syndrome", "Procedure", "Sign or Symptom" or "Clinical Drug" with each assigned a TUI code.

- **Experiencer** (Context plugin) - if the term is referenced to the primary person within the text, i.e. the patient or other such as family members

There are many components other than finding the word "epilepsy" that determine if an epilepsy diagnosis has been confirmed. Custom gazetteers were created to search for terms such as "Diagnosis:" found in structured elements of clinic letters, and custom gazetteers were also created to specify 5 levels of certainty (5 being most certain) of a term to differentiate phrases such as "it is doubtful that she has epilepsy" and "this is probably a case of complex partial seizures" [3]. Table 4.2 shows a list of terms and their certainty levels that was used to attach a certainty level to any Lookup:

[3]With assistance from Dr Owen Pickrell

Table 4.2: A gazetteer of terms used to determine 5 levels of certainty attached to an epilepsy diagnosis. A confirmed diagnosis must have a value of 4 or 5.

Term	Level	Term	Level	Term	Level
ruled out	1	possibility of	3	suspected	4
unlikely	2	?	3	suggestive	4
doubtful	2	uncertain	3	treated as	4
doesn't	1	might	3	treating this as	4
doubt	2	potential	3	probably	4
unsure	2	potentially	3	suspicion	4
unclear	2	further clarification	3	I think	4
not convinced	2	further investigation	3	impression is	4
remote	2	to be confirmed	3	sounds like	4
improbable	2	to be sure	3	sound like	4
not likely	2	to see if	3	suspect	4
??	2	could be	3	suspicious	4
remote possibility	2	to see whether	3	certain	5
unusual	2	likely	4	definite	5
possible	3	probable	4	are dealing with	5

The JAPE script in Figure 4.5 is one example of how potential diagnoses was captured.

The UMLS CUI codes (767 in total) from running the mysql query in Figure 4.4 were used to filter out non-epilepsy related Lookups. Figure 4.6 shows a screenshot from GATE of the features attributed to a Lookup. These features were used to determine if a Lookup was negated, its certainty level, type and it's UMLS CUI code and can be used for further downstream annotations such as only including Lookups with a certainty level greater than 3 as a confirmed diagnosis.

```
Phase: Diagnosis
Input:  Lookup Sentence
Options: control=all

Rule: getDiagnosis
(
  ({Lookup.PREF == "Diagnosis"} | {Lookup.PREF == "Diagnosed"} |
  {Lookup.label == "suffers"})
    (
    {Lookup.STY == "Disease or Syndrome"} |
    {Lookup.STY == "Sign or Symptom", Lookup.PREF != "Fit NOS"} |
    {Lookup.STY == "Mental or Behavioral Dysfunction"} |
    {Lookup.STY == "Congenital Abnormality"} |
    ({Lookup.STY == "Diagnostic Procedure"} |
     {Lookup.Temporality == historical})?
  )* # allow for further/nested diagnoses within a phrase
  (
  {Lookup.STY == "Disease or Syndrome"} |
  {Lookup.STY == "Sign or Symptom", Lookup.PREF != "Fit NOS"} |
  {Lookup.STY == "Mental or Behavioral Dysfunction"} |
  {Lookup.STY == "Congenital Abnormality"} |
  ({Lookup.STY == "Diagnostic Procedure"} |
    {Lookup.Temporality == historical})?)?
  ):item
):label
-->
:item.Diagnosis = { rule = "getDiagnosis", PREF = :item.Lookup.PREF,
                    CUI = :item.Lookup.inst,
                    STY = :item.Lookup.STY, Negation = :item.Lookup.Negation,
                    Experiencer = :item.Lookup.Experiencer,
                    Temporality = :item.Lookup.Temporality,
                    # store certainty for later i.e. >4 = diagnosis
                    Certainty = :item.Lookup.Certainty}
```

Figure 4.5: JAPE script to extract diagnosis using various contexts such negation, semantic types and certainty terms

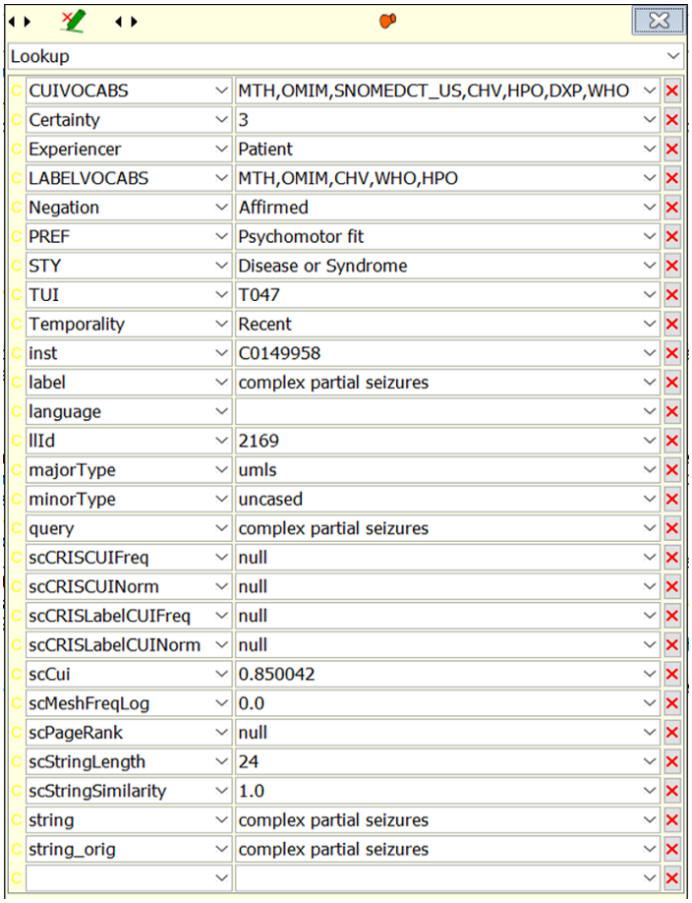

Figure 4.6: A Lookup for the phrase "possible complex partial seizures". The "Certainty" features was added through development of custom gazetteers and JAPE rules. The rest of the features come as default from the BIO-Yodie plugin in GATE, and the "Negation" feature was produced by modifying the Context plugin in GATE to add more stop words.

4.2.5 Seizure frequency

Seizure frequency was annotated by extracting the following items of information within text: mention of a seizure (subject to negation and certainty), number or range of seizures and the time period over which the seizures occurred. Some example phrases, of which the first four provide a measure of the number of seizures over a period of time are:

She is having 5-10 seizures per week.
He describes what are probably focal seizures. These happen at least once per day.
Since last April he has had 5 seizures.
He has had more than 20 episodes since his last visit.
She was diagnosed with epilepsy after having 5 seizures.

The approach taken was that seizure frequency can be split into three parts: mention

```
Phase: PartialDate
# Token as input ensures any other annotation type stops rule from firing in between
# input annotations i.e. strictly day of month, followed by month, followed by year.
Input: DayDate Month Numeric Token
Options: control=appelt

Rule: partialDate
(
  #gazetteer for "1st", "2nd", "3rd" etc
  ({DayDate})?
  # gazetteer for months
  {Month}
  #any number, logically this will always be a year
  ({Numeric})?
):match
-->

:match.PartialDate = {rule = partialDate,
                       # record the day of month
                       day=:match.DayDate.value,
                       # record the month
                       month=:match.Month.month,
                       # record the year
                       year=:match.Numeric.value
                       }
```

Figure 4.7: JAPE script to define all possible ways of specifying a date including partial "April 25th" and full "April 25th 1992"

of a seizure, a time period, and number of seizures. Gazetteers and JAPE rules were written to reflect these three components. Initially seizures were filtered from all Lookups identified by the BIO-Yodie plugin, but during the development of the algorithm it was found that seizure mentions aren't often specified formally e.g. "2-4 complex partial seizures per day" but rather colloquially by both patient and clinician e.g. "20 episodes since his last visit" (see example 4 above) or "15 events every morning". A custom gazetteer of terms was created to reflect this, but to preserve specificity and to distinguish "episodes" as seizures from other episodes such as episodes of depression or anxiety, a JAPE rule was written to only associate colloquial terms as seizures where a formal seizure type, such as complex partial seizure, is mentioned elsewhere in the letter.

Two approaches were taken to annotate time periods. The first was to create a JAPE rule for inferring implicit time references such as "Since April she has had around 20 seizures" where given the clinic date, a time period could be calculated. This involved capturing calendar references that span from month names i.e. April, to full date references i.e. 1st April or 1st April 2005. The Jape script in Figure 4.7 show how all of these forms are captured in a single rule using the Kleene operator for optional arguments.

The second approach was to annotate explicit mentions of a time period such as "10

per week". Custom gazetteers for temporal terms such as "since","in the last" ,"per week", "a day" were created with attributes for each. Table 4.3 show a sample of the gazetteer used to define explicit time periods:

Table 4.3: A sample of phrases used to define explicit time periods.

Item	Number	Time Unit	When
a year	1	year	
a year	1	year	
b.d	2	day	
b.d.	2	day	
b.i.d.	2	day	
bd	2	day	
in the evening	1	day	evening
every morning	1	day	morning
in the morning	1	day	morning
daily	1	week	
every day	1	week	
in a day	1	day	
in a single week	1	day	
in a week	1	week	
per day	1	day	
o.d.	1	day	
at night	1	day	night
a month	1	month	

The JAPE script in 4.7 and the gazetteer in Table 4.3 were used to reference various points in time and combined with seizures mentions to annotate seizure frequency as well the individual components to calculate the frequency:

4.2.6 Medication

The SLaM (South London and Maudsley) medication application for GATE [244] was used and modified to annotate documents with prescription information that include drug name, tablet size, unit of measurement and frequency. The SLaM application comes with custom gazetteers for drugs derived from BNF code lists, units and frequency terms which are then used as input to various JAPE rules. The drugs annotated with the SLaM application did not contain any code reference such as UMLS or READ, so the BNF gazetteer was swapped with the UMLS gazetteer

```
Phase: SeizureFrequency
Input: NumberRange timePeriods Lookup3 Sentence startList Split
Options: control=all

# define one of may rules in JAPE script
# i.e. seizureFrequency0a, seizureFrequency0b ...
Rule: seizureFrequency0a
(
    # could be a number or range pertaining to seizure quantity
    ({NumberRange}):X1
    {Lookup3.Negation == Affirmed}
    i.e. terms such as "since", "during"
    {timePeriods.period==yes}
    # number of days/month defining a time period
    ({NumberRange}):X2
):match
-->
:match.SeizureFrequency = { SeizureType=:match.Lookup3.PREF,
                             rule = seizureFrequency0a,
                             PREF = "Fit Frequency",
                             CUI = "C0149775",
                             seizureNum = :X1.NumberRange.value,
                             seizureLower = :X1.NumberRange.N1,
                             seizureUpper = :X1.NumberRange.N2,
                             timeNum = :X2.NumberRange.value,
                             timeLower = :X2.NumberRange.N1,
                             timeUpper = :X2.NumberRange.N2,
                             period = :match.timePeriods.time-unit
}
```

Figure 4.8: JAPE script to extract certain ways of experessing seizure frequency

used by BIO-Yodie and the JAPE scripts that came with the SLaM application were modified to accept Bio-YODIE annotations.

Other JAPE rules were also modified and supplemented with custom gazetteers to capture further details about prescriptions, such as if a prescription mention was historical, or if a prescription were to be made pending further follow up. This made it possible to select current prescriptions only. The JAPE script in figure 4.9 shows how custom gazetteers for words such as "pending", "may", "try", "previously" etc. were used to pad out prescriptions with further context:

During development of the algorithm, a common way of expressing directions to take a prescription multiple times a day was found to include times of day:

$$\underbrace{\text{Lamotrigine}}_{\textbf{Drug}}\ \underbrace{\text{250mg in the morning,}}_{\textbf{Direction 1}}\ \underbrace{\text{200mg at night}}_{\textbf{Direction 2}}$$

Where both directions must be captured to sum to a daily dose of 550mg of Lamotrigine per day. The JAPE script in figure 4.10 uses Kleene operators to accommodate multiple directions when appearing consecutively without a Lookup (usually a drug given this pattern) in between them:

```
Phase: DrugStatus
Input: Measurement DoseFrequency Lookup Numeric ContextPrescription NewLine
Options: control=appelt

Rule: doseMatch0
(
    # optional (?) gazetteer of terms such as "descreased","continue","try"
    ({ContextPrescription})?
    # drug tagged in Bio-YODIE or measurement
    # i.e. Lamotrigine 200mg or 200mg Lamotrigine
    ({Lookup.STY=="Pharmacologic Substance"} | {Measurement})
    # Repeat again i.e. if drg was picked up in first line
    # logically measurement should be picked up in second, vice-a-versa
    ({Lookup.STY=="Pharmacologic Substance"} | {Measurement})
    # twice a day, once in the morning etc
    {DoseFrequency}
    # another optional context i.e. prescribe <Presription> if.....
    ({ContextPrescription})?
):match
-->
:match.Prescription0 = { rule = doseMatch0,  drug = :match.Lookup.PREF,
                         CUI = :match.Lookup.inst,
                         dose-val = :match.Measurement.quantity,
                         dose-unit =:match.Measurement.units,
                         dose-frequency = :match.DoseFrequency.frequency,
                         time-unit = :match.DoseFrequency.time-unit,
                         dose-interval = :match.DoseFrequency.interval,
                         Context = :match.ContextPrescription.context}
```

Figure 4.9: JAPE script to extract prescriptions

4.2.7 Investigations - CT, MRI and EEG scans

Two attempts were made to capture details of CT, MRI and EEG scans. The first attempt used CUI subsets in a similar way to how Bio-YODIE annotations were filter for epilepsy specific annotations, however terms relating to scan result were too specific to map directly to language used within clinic letters. Table 4.4 shows a sample of how UMLS concepts related to EEGs are highly specific in terms of string matching to phrases within a text:

```
1
2  Phase: LazyPrescription
3  Input: Measurement Token Lookup
4  # Pick up all possible mentions, rather than longest match.
5  Options: control=all
6
7
8   # JAPE rules for when multiple doses are repeated off for just one drug
9   # maximum three does per prescription are captured
10  # don't need to worry about b.d., twice a day etc
11
12 Rule: lazyMatch
13 Priority: 100
14 (
15     # Get drug mention
16     ({Lookup.STY=="Pharmacologic Substance"} | {Lookup.STY=="Clinical Drug"})
17     # Get first measurement (quantity and unit)
18     # Don't need how many times per day, explicitly says "in the morning"
19     ({Measurement}):m1
20     # Allow a toke that isn't a Lookup...i.e. a comma or "and"
21     ({Token,!Lookup})?
22     # Get second measurement (quantity and unit)
23     # Don't need how many times per day, explicitly says "at night"
24     ({Measurement}):m2
25     ({Token,!Lookup})?
26     # Optional further dose
27     ({Measurement}):m3
28
29 ):match
30 -->
31 :match.Prescription = { rule = lazyMatch,
32                          drug = :match.Lookup.PREF,
33                          CUI = :match.Lookup.inst,
34                          dose-val1 = :m1.Measurement.quantity,
35                          dose-val2 = :m2.Measurement.quantity,
36                          dose-val3 = :m3.Measurement.quantity,
37                          dose-unit1 =:m1.Measurement.units,
38                          dose-unit2 =:m2.Measurement.units,
39                          does-unit3=:m3.Measurement.units,
40                          # hardcoded i.e. each unique directions will
41                          # be assgned once per day
42                          dose-frequency = "1", time-unit = "day"}
```

Figure 4.10: JAPE script to extract multiple prescription directions.

Table 4.4: A sample of UMLS terms and the important information within each. UMLS terms such as these shown are difficult to map directly to text within clinical texts, where the terms of interest from within each UMLS term are much easier to map.

UMLS term	Term of interest
EEG shows generalized, bilateral, synchronous, symmetrical discharge	symmetrical discharge
EEG with hyperventilation-induced focal epileptiform discharges	focal epileptiform discharges
EEG with hyperventilation-induced generalized epileptiform discharges	generalized epileptiform discharges
MRI shows leukoencephalopathy with cavitation	leukoencephalopathy + cavitation
MRI shows congenital abnormalities of the posterior fossa	congenital abnormalities
MRI shows short, thick corpus callosum	abnormal corpus callosum
Diffuse cerebral atrophy on CT and MRI	cerebral atrophy
Aplasia of posterior semicircular canal on CT scan	aplasia
Low density white matter on CT scan	low density white matter

In table 4.4 the UMLS terms in column 1 only get mapped to terms in text if the exact term is found, but due to the length and specificity of some terms there was low sensitivity in picking up investigations. Therefore custom gazetteers that use smaller terms categorised as normal (CUI:C0560017) or abnormal (CUI:C0151611) were produced to maximize to maximize sensitivity of investigation outcomes that preserve what is important i.e. normal or abnormal results [4]. Table 4.5 shows a list of terms used to identify possible investigation findings:

[4]with assistance from Dr Owen Pickrell

Table 4.5: A list of custom terms used to indicate a possible EEG finding

Custom term	Derived UMLS concept	Custom term	Derived UMLS concept
abnormal	C0151611	normal	C0560017
abnormal EEG	C0151611	normal EEG	C0560017
abnormalities	C0151611	photosensitive	C0151611
abnormality	C0151611	photosensitive	C0151611
burst suppression	C0151611	photosensitivity	C0151611
clear	C0560017	polyspike	C0151611
did not capture any events	C0560017	poly-spike	C0151611
dysrhythmic	C0151611	polyspike and wave	C0151611
EEG normal	C0560017	right side slowing	C0151611
epileptic	C0151611	sharp	C0151611
epileptiform	C0151611	spike	C0151611
epileptogenic	C0151611	spike and wave	C0151611
failed to alter	C0560017	spikes	C0151611
focal slowing	C0151611	spike-wave	C0151611
focus	C0151611	temporal slowing	C0151611
generalised slowing	C0151611	unremarkable	C0560017
irregular	C0151611	unstable	C0151611
left side slowing	C0151611		

Once terms were identified within the text, the JAPE rule in figure 4.11 was written to associate them to investigation names using strings such as "EEG", "MRI" and "CT" as long as they were found after the investigation name and within the same paragraph:

4.2.8 Validation of Algorithm

After developing the pipeline, 200 unseen letters were used to validate accuracy against a clinician [5]. For each category of information, the scope of what was expected to be extracted or not extracted was discussed with the clinician. The clinician then annotated every letter for each category, including multiple mentions from the same category. Separately, the pipeline was run against the 200 test letters and annotated

[5]performed by Dr Owen Pickrell

```
Phase: Investigations
# Investigation are Investigation types i.e. EEG
# InvestigationFinding are outcomes i.e. lesion
# p is paragraph to ensure
Input: Investigation p InvestigationFinding
Options: control=brill

Rule: getInvestigationsOutcomes
(
    # get investigation type i.e. EEG
    ({Investigation}):invest
    # followed by outcomes (+ means one or more)
    (({InvestigationFinding}):outcome)+
):match
-->
:match.Investigations =
{ rule = getInvestigationsOutcome1,
  #store type
  INVESTIGATION = :invest.Investigation@string,
  #store outcome
  Outcome = :outcome.InvestigationFinding@string,
  # store outcome CUI
  CUI = :outcome.InvestigationFinding.CUI,
  #store negation status
  Negation = :outcome.InvestigationFinding.Negation}
```

Figure 4.11: JAPE script to extract investigation outcomes

for each category and the results of the clinician and pipeline were reviewed, where all disagreements were manually reviewed.

Precision, recall and F1-score were used as measures to determine the accuracy of the pipeline and are defined as:

$$Precision = \frac{TP}{TP * FP}$$

$$Recall = \frac{TP}{TP * FN}$$

$$F1score = \frac{Precision * Recall}{Precision + Recall}$$

True positives were defined as both the pipeline and clinician identifying a positive finding such as confirming an epilepsy diagnosis, false positives were defined as the pipeline identifying a positive finding where the clinician did not, and false negatives were defined as the algorithm failing to identify a positive finding where the clinician was able to. To resolve any mention where there was a disagreement, either the disagreement remained after manual review, or in small proportion of cases where the clinician had made a mistake, the clinicians record was corrected. Results are given in table 4.6 where a "per item" score is based on every possible mention within the categories, and a "per letter" score assumes that identification of one true positive

within a given category is a true positive for that category as a whole i.e. if there are 3 mentions of an epilepsy diagnosis, if the pipeline was able to identify just one, the per letter score for epilepsy diagnosis would be a true positive.

Table 4.6: Precision, recall and F1-score are calculated across 9 epilepsy specific categories as well as clinic da approaches have been considered - the first measures the algorithm's accuracy for every mention (N=1925) acros in the manual review by the clinician, and the second approach aggregates results from multiple mentions per there are multiple true mentions regarding confirmation of epilepsy in a single letter, we assign a single true posit picks up at least one of these mentions, with the same logic used to determine false positives, true negatives an

Variables	Per item performance				Per letter performance	
	N items	Precision %	Recall %	F1 score %	N letters	Precision %
Clinic Date	191	98.9	97.4	98.2	186	100
Date of Birth	201	100	98	99	199	100
Epilepsy Confirmed	383	88.1	99	88.5	150	94.1
Epilepsy Type	89	89.9	79.8	84.5	70	91
Focal Seizures	145	96.2	69.7	80.8	69	96.7
Generalised Seizures	151	88.8	52.3	65.8	76	89.7
Seizure Frequency	153	86.3	53.6	66.1	119	92.2
Medication	316	96.1	94	95	157	98.6
CT Scan	17	55.6	58.8	57.1	16	76.9
MRI Scan	109	82.4	68.8	75	66	86.7
EEG	170	81.5	75.3	78.3	79	86.6
All	1925	90.6	80.8	85.4	1187	96.6

The pipeline obtained an overall precision, recall and F1-score of 91%, 81% and 85% on a per item basis, where high scores were obtained in prescription (F1=95%), confirmation of an epilepsy diagnosis (93%), epilepsy type (84%) and presence of focal seizures (81%). The algorithm was less accurate in identifying CT (57%), MRI (75%) and EEG results (78%), seizure frequency (66%) and generalised seizure terms (66%) given the complexity and high variance in expressing these concepts in clinic letters. The pipeline achieved even higher overall scores for precision, recall, and F1-score (96%, 87%, 91%) on a per letter basis, in which given how the final data is to be used for further research purposes, a decision for each category can be made accurately across measures such as epilepsy diagnosis, epilepsy type, seizure type and prescriptions.

4.3 Chapter Summary

A gold standard dataset of de-identified clinic letters was used to build and test an NLP pipeline, that was found to accurately extract novel information about epilepsy when compared to manual review by a clinician. The use of UMLS terminologies, in particular the ability to map findings to CUI codes can be powerful in curating structured datasets that can be linked to other routinely collected data such as GP and hospital patient records, where these data can be processed programmatically rather than via manual review. Some categories such as diagnosis of epilepsy, epilepsy type and prescriptions can be extracted with high accuracy, but some concepts such as EEG/MRI/CT investigations and seizure frequency remained difficult to extract and further improvements are necessary for further research purposes. However, the information that the pipeline can extract well would improve the richness of data such as GP records that are held within the SAIL databank.

While gold standard letters have been used, the pipeline was tested on a relatively small number of letters sourced from one Health-board with a limited number of writing styles and letter structures, therefore the generalizability of the pipeline may be limited and would benefit from a larger test set. The pipeline does not however rely on the structure of clinic letters and is designed to use free text without relying on dedicated sections in the letters. A "per item" and "per letter" score was calculated to validate both the accuracy of the pipeline, but also to validate how information within letters can be used practically for further research by summarising all items in a clinic letter and giving a decision boundary on the category as a whole.

Chapter 5

Predicting functional impact of Single Nucleotide Polymorphisms

The aim of this chapter was to investigate the best method for indicating that a Single Nucleotide Polymorphism (SNP) is a pathogenic consideration for epilepsy/neurology phenotypes. Given that an exome contains 30-40 thousands SNPs it is important to prioritise those that may contribute to disease so that they can be studied in downstream functional validation and link to observable clinical outcomes. Various machine -learning techniques were explored to identify the most accurate method of classifying disease (pathogenic) and benign SNPs. These techniques were compared to existing prediction software when tested on a disease non-specific dataset of SNPs as well as epilepsy specific SNPs. The *humvar* dataset was identified from an extensive literature review to be a commonly-used training dataset for machine learning classification of SNPs. It contains over 70,000 pathogenic and benign SNPs obtained from published studies. These SNPs were used to train various machine learning algorithms which were then compared to existing prediction algorithms widely-used in the literature. A set of SNPs found in genes associated with epilepsy were also scored using the trained algorithm. All code used in this chapter can be found at the following Github repository `https://github.com/arronlacey/PhD-Chapter5`.

5.1 Features

An automated feature extraction pipeline was built so that for each mutation in the *humvar* dataset, the pipeline obtains 30 protein features which are all used to train

and test various machine learning algorithms. The pipeline consists of bash and R scripts to pull data from multiple websites, public databases and derive further features with downstream processing. Figure 5.1 shows a flowchart of the main processes in the feature extraction pipeline:

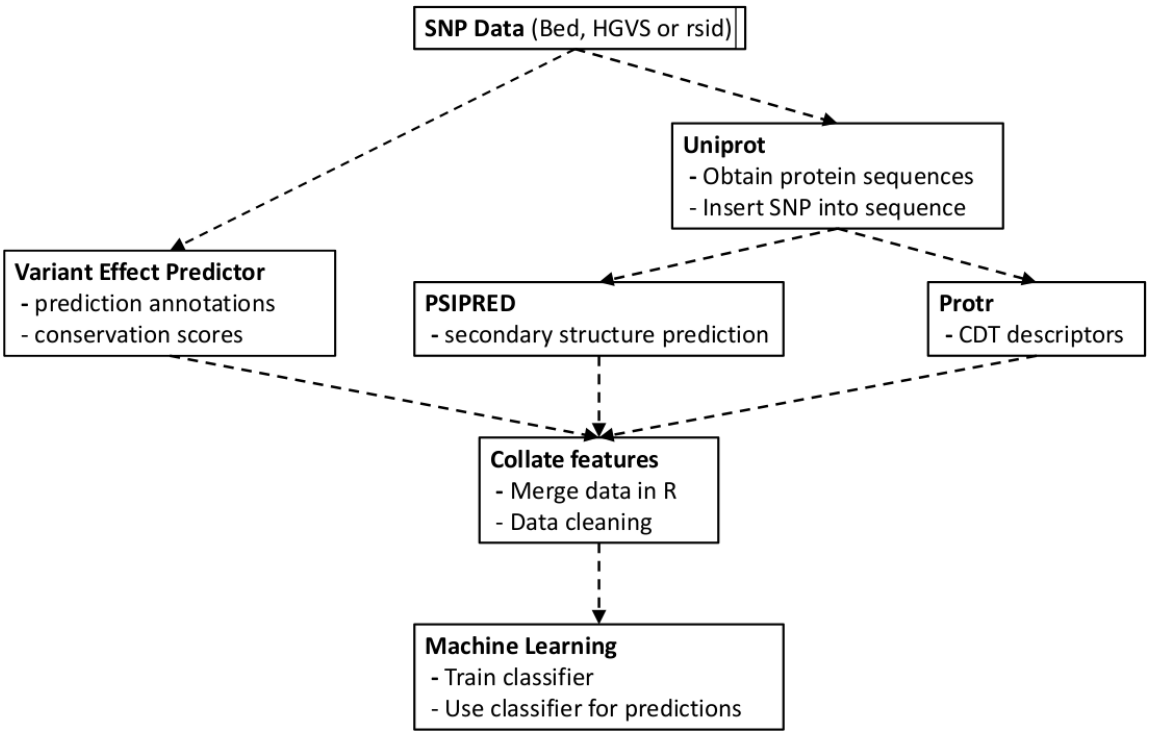

Figure 5.1: Pipeline of SNP data collection. The data is used to train a classifier that can be used to predict disease/benign status of a SNP.

The pipeline was built to accommodate chromosomal or amino acid co-ordinates in BED and HGVS formats, as well as rsid format. Sections 5.1.1 - 5.1.3 document how each process in the feature extraction pipeline was built.

5.1.1 Variant Effect Predictor

The humvar dataset rsids were used as input to the Variant Effect Predictor (VEP) annotating system to obtain conservation scores and existing prediction software scores for each SNP. Table 5.1 describes the 38 features collected to be used as part of the training data for machine learning.

Table 5.1: Features obtained from VEP

Feature	Description
DANN_score	Deep learning SNP prediction score [204]
GM12878_fitCons_score	Fitness conservation score for lymphoblastiod cells [245]
GM12878_fitCons_score_rankscore	Rankscore for lymphoblastoid cells [245]
GenoCanyon_score_rankscore	Prediction score for non-coding function regions [246]
H1NAhESC_fitCons_score_rankscore	Fitness conservation score for human embryonic stem cells [245]
HUVEC_fitCons_score_rankscore	Fitness conservation score for umbilical vein epithelial cells [245]
MetaLR_score	SNP prediction score using 9 existing SNP software, trained with logistic regression [205]
MutationAssessor_score_rankscore	SNP prediction application for cancer variants [206]
REVEL_score	SNP prediction aggregation score using 8 SNP prediction software for predicting pathogenicity of rare variants [207]
fathmmNAMKL_coding_score	SNP prediction score using Hidden Markov Models [208]
integrated_fitCons_score	Combined fitcons score
integrated_fitCons_score_rankscore	Combined fitcons rank score
PolyPhen_score	SNP prediction score using a Naïve Bayes classifier [196]
SIFT_score	SNP prediction score using protein conservation methods [187]
CADD_raw_rankscore	Combined Annotation–Dependent Depletion SNP prediction score [209]
DANN_rankscore	DANN converted rankscore [204]
EigenNAPCNAraw_rankscore	SNP prediction score using unsuperivsed learning [247]
FATHMM_converted_rankscore	FATHMM converted rankscore
GERP.._RS_rankscore	Conservation score in humans based on 1,092 genomes [248]
MetaLR_rankscore	MetaLR rankscore
MetaSVM_rankscore	SNP prediction score using 9 existing SNP software, trained with SVM [205]
MutationTaster_converted_rankscore	Mutation taster rankscore
PROVEAN_score	SNP prediction in non-coding regions [249]
REVEL_rankscore	REVEL rankscore
SiPhy_29way_logOdds_rankscore	SIte-specific PHYlogenetic analysis [224]
VEST3_rankscore	Variant Effect Scoring Tool [210]
fathmmNAMKL_coding_rankscore	FATHMM coding rankscore
phastCons100way_vertebrate_rankscore	Evolutionary conservation (ranked) scores in vertabrae [223]
phastCons20way_mammalian_rankscore	Evolutionary conservation (ranked) scores in mammals [223]
phyloP100way_vertebrate_rankscore	Score predicting non-neutral substitution rates in vertebrae [222]
phyloP20way_mammalian_rankscore	Score predicting non-neutral substitution rates in mammals [222]
Reliability_index	Reliability index as calculated by SNAP2
GERP.._NR	Conservation score in humans based on 1,092 genomes [248]
SiPhy_29way_logOdds	SiPhy log odds score
phastCons20way_mammalian	Evolutionary conservation scores in mammals [223]
gnomAD	Genome aggregation SNP frequency in population combined exomes and genomes
gnomAD_exomes	Genome aggregation SNP frequency in population in exomes
gnomAD_genomes	Genome aggregation SNP frequency in population in genomes

Table 5.2: Amino acid attributes with three group classification. Each classification is given by a unique set of amino acids. Table reproduced with permission from `https://cran.r-project.org/web/packages/protr/vignettes/protr.html`

Atrribute	Group 1	Group 2	Group 3
Hydrophobicity	Polar	Neutral	Hydrophobicity
	R, K, E, D, Q, N	G, A, S, T, P, H, Y	C, L, V, I, M, F, W
Normalized van der Waals Volume	0-2.78	2.95-4.0	4.03-8.08
	G, A, S, T, P, D, C	N, V, E, Q, I, L	M, H, K, F, R, Y, W
Polarity	4.9-6.2	8.0-9.2	10.4-13.0
	L, I, F, W, C, M, V, Y	P, A, T, G, S	H, Q, R, K, N, E, D
Polarizability	0-1.08	0.128-0.186	0.219-0.409
	G, A, S, D, T	C, P, N, V, E, Q, I, L	K, M, H, F, R, Y, W
Charge	Positive	Neutral	Negative
	K, R	A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V	D, E
Secondary Structure	Helix	Strand	Coil
	E, A, L, M, Q, K, R, H	V, I, Y, C, W, F, T	G, N, P, S, D
Solvent Accessibility	Buried	Exposed	Intermediate
	A, L, F, C, G, I, V, W	R, K, Q, E, N, D	M, S, P, T, H, Y

5.1.2 CTD Descriptors

CDT (Composition/Transition/Distribution) descriptors were used to assign physiochemical attributes to each SNP [250]. The protr R package contains a function that calculates the global distribution of amino acid attributes classed into 3 categories as a percentage of all amino acids in a given sequence. These attributes and categories are shown in Table 5.2

The protr R function was modified to output the attributes and categories for every amino acid in a sequence so that data on both a wild type and a SNP could be obtained and the differences compared. The modified function is shown in Figure 5.2

```r
extractCTDCraw = function(x) {
  k<-as.numeric(x[,2])
  prot<-as.character(x[,1]) # protein identifier
  pos<-as.numeric(x[,2]) # SNP position
  wild<-as.character(x[,3]) # wild type amino acid
  sub<-as.character(x[,4]) # sub is the mutation

  group1 = list(
    'hydrophobicity'  = c('R', 'K', 'E', 'D', 'Q', 'N'),
    'normwaalsvolume' = c('G', 'A', 'S', 'T', 'P', 'D', 'C'),
    'polarity'        = c('L', 'I', 'F', 'W', 'C', 'M', 'V', 'Y'),
    'polarizability'  = c('G', 'A', 'S', 'D', 'T'),
    'charge'          = c('K', 'R'),
    'secondarystruct' = c('E', 'A', 'L', 'M', 'Q', 'K', 'R', 'H'),
    'solventaccess'   = c('A', 'L', 'F', 'C', 'G', 'I', 'V', 'W'))
  group2 = list(
    'hydrophobicity'  = c('G', 'A', 'S', 'T', 'P', 'H', 'Y'),
    'normwaalsvolume' = c('N', 'V', 'E', 'Q', 'I', 'L'),
    'polarity'        = c('P', 'A', 'T', 'G', 'S'),
    'polarizability'  = c('C', 'P', 'N', 'V', 'E', 'Q', 'I', 'L'),
    'charge'          = c('A', 'N', 'C', 'Q', 'G', 'H', 'I', 'L',
                          'M', 'F', 'P', 'S', 'T', 'W', 'Y', 'V'),
    'secondarystruct' = c('V', 'I', 'Y', 'C', 'W', 'F', 'T'),
    'solventaccess'   = c('R', 'K', 'Q', 'E', 'N', 'D'))
  group3 = list(
    'hydrophobicity'  = c('C', 'L', 'V', 'I', 'M', 'F', 'W'),
    'normwaalsvolume' = c('M', 'H', 'K', 'F', 'R', 'Y', 'W'),
    'polarity'        = c('H', 'Q', 'R', 'K', 'N', 'E', 'D'),
    'polarizability'  = c('K', 'M', 'H', 'F', 'R', 'Y', 'W'),
    'charge'          = c('D', 'E'),
    'secondarystruct' = c('G', 'N', 'P', 'S', 'D'),
    'solventaccess'   = c('M', 'S', 'P', 'T', 'H', 'Y'))
  xSplitted = substr(x[1,5],pos,pos)

  # Get groups for each property & each amino acid
  g1 = lapply(group1, function(g) which(xSplitted %in% g))
  names(g1) = paste(names(g1), '1.', sep = '.')
  g2 = lapply(group2, function(g) which(xSplitted %in% g))
  names(g2) = paste(names(g2), '2.', sep = '.')
  g3 = lapply(group3, function(g) which(xSplitted %in% g))
  names(g3) = paste(names(g3), '3.', sep = '.')
}
```

Figure 5.2: R function modified from the protr package to obtain CTD amino acid groups and categories for each position of a protein sequence

5.1.3 Secondary structure prediction

The pipeline uses the open source program PSIPRED to predict the secondary structure state (coil, helix, sheet) at each amino acid position for both the wild type and the variant protein sequences. For each SNP the protein sequence was downloaded using the Uniprot API, where the SNP was swapped into the sequence. In total 21,094 wild type sequences and 74,393 SNP sequences were processed with PSIPRED on the HPC Wales cluster. For each SNP the probability change between each state of the position containing the SNP and the corresponding wildtype states were used as well as the overall predictions to produce 5 features for the final training data. The following bash script loads SNP co-ordinates from the humvar dataset into the uniprot API to retrieve protein sequences for each SNP:

```
1
2  #!/bin/bash
3
4  #download fasta seqs given file of uniprot ids
5
6  # SNP input file
7  file=$1
8  # Output file name minus extension
9  name=$2
10
11 # assign 4 columns in SNP input file to variables: protien ID, position, and alleles
12
13 ids=($(cat ${file} | awk '{print $1}'))
14 pos=($(cat ${file} | awk '{print $2}'))
15 wild=($(cat ${file} | awk '{print $3}'))
16 sub=($(cat ${file} | awk '{print $4}'))
17
18
19 # get ref fasta for each line in file, with custom header attached
20 # use cURL to retrieve from the uniport REST URL
21
22 for i in "${!ids[@]}" ; do
23       echo "#${ids[i]}_${pos[i]}_${wild[i]}_${sub[i]}";
24       curl -sS "http://www.uniprot.org/uniprot/"${ids[i]}".fasta";
25 done |
26 sed '/^>/ d' |
27 sed -r 's/[#]+/>/g' |
28 perl -npe 'chomp if ($.!=1 && !s/^>/\n>/)' > $name.snp.fasta
```

Figure 5.3: Bash script to retrieve fasta sequences for a file given in SNP format.

The awk script in figure 5.4 replaces the wild type amino acid at the SNP position with the mutated amino acid.

The sequences containing the wild type and the SNPs were then processed using

```
# field separator defined as _ i.e. fasta headers
BEGIN { FS="_" }
# get fasta headers and store components into array
/^>/ {
    id=$1;p=$2; wild=$3;subs=$4; c=$NF; next
}
{
    # start of sequence
    s=1
    e=length($0)  #end of sequence
    #substring up to mutation, substitution, substring after mutation
    print id"_"p"_"wild"_"subs">\n"substr($0,s,p-1) c substr($0,p+1,e)
}
```

Figure 5.4: An AWK script that replaces the wild type amino acid with the mutation

```
#!/bin/bash
#SBATCH --job-name psipred-array  # name of job as appears in queue
#SBATCH --time 01-21:00           # length of time for each job to run
#SBATCH -o psibatchout.$I         # standard output of job
#SBATCH -e psibatcherr.$J         # error log
#SBATCH --array=1061-1080         # job array i.e. parallel process multiple jobs
#SBATCH --mem-per-cpu=4000        # memory per cpu
#SBATCH --ntasks=128              # number of nodes
#SBATCH --mail-user=user@mail.com # notify job is complete via email
module load compiler/gnu/4.8.0    # compiler
module load R/3.2.3               # external dependencies

# psipred code
code=${HOME}/Phd/script_dev/rfpipeline.sh

# input sequence file
data_file="humvarids_${SLURM_ARRAY_TASK_ID}.fasta"
# declare the file about to be used
echo ${data_file}
# run psipred on input file
${code} ${data_file}
```

Figure 5.5: A SLURM job script run on the HPC Wales Portal that calls the psipred commandline facility. The script logs a job in a queue containing any other jobs users submit across HPC clusters, where parameters such as compiler, number of cores and how long the script should be allowed to run for. It takes 45 hours (time parameter 01-21:00) to process 300 sequences of varying length.

PSIPRED. Each sequence took on average 20 minutes to process on a standard desktop, therefore this task was completed by running PSIPRED on the HPC Wales cluster which reduced the time taken to 2-3 minutes. The SLURM script in 5.5 is the SLURM job schedule script that specifies a variety of parameters needed such as number of cores, run time and memory required to run psipred. The fair usage limit on the HPC Wales cluster allowed for a scheduled job to run up to 48 hours, which is the equivalent of processing 300 sequences, therefore the 21,094 wild type sequences and 74,393 SNP sequences were split into 318 jobs.

Each PSIPRED output file for a given SNP was compared to it's corresponding wild type protein using the bash script in figure 5.6. A sample comparison is given in Figure 5.7 which shows how the secondary structure prediction changes not only at

```
# Create index of all faster headers in humvar files
for i in *.fasta; do
        IFS=_ read -ra arr <"$i"
        mv $i `echo "${arr[0]}_${arr[1]}_$i" | sed -e 's/>//g'`
done


# get protein name, snp position and file id number from filename
IFS=$'\n' fa=( $(ls *.fasta | awk -F'[_.-]' '{print $1" "$2" "$5}' ) )

# use file id number to find .ss file (secondary structure file)

for i in "${fa[@]}" ; do
echo "$i" | xargs -n 3 bash -c 'cat *-$2.fasta.ss |sed "s/$/ $0 $1 $2/" | nl -v $2'
done > master.ss
# Extract SNP line where amino acid positions are equal in both files
awk '$2 == $9' master.ss | sed 's/ \{1,\}/,/g' | sed 's/^,//' > master.csv
```

Figure 5.6: A bash script to process output of psipred.

the SNP position, but also in neighbouring SNPs. This is due to the underlying PSIBLAST alignment used when processing the wild type sequence and the mutated sequence, where a proportion of candidate protein sequences used in the secondary structure prediction in both cases will differ, however the largest changes are generally at the SNP position or next to it.

M C D A K V M R K C Q V - Wild protein sequence
M C D A K J M R K C Q V - Sequence containing SNP

PSIPRED output wild

Position	Amino Acid	Prediction	C	H	S
1227	M	H	0.027	0.984	0.021
1228	C	H	0.054	0.915	0.021
1229	D	H	0.007	0.996	0.003
1230	A	H	0.007	0.999	0
1231	K	H	0.017	0.993	0.001
1232	**V**	**H**	**0.007**	**0.998**	**0.001**
1233	M	C	0.998	0.005	0.002
1234	R	C	0.982	0.027	0.002
1235	K	C	0.966	0.058	0.007
1236	C	C	0.999	0	0.001
1237	Q	C	0.915	0.054	0.021
1238	V	S	0.040	0.100	0.860

PSIPRED output SNP

Position	Amino Acid	Prediction	C	H	S
1227	M	H	0.017	0.984	0.031
1228	C	H	0.054	0.915	0.021
1229	D	H	0.007	0.996	0.003
1230	A	H	0.013	0.087	0
1231	K	C	0.064	0.195	0.015
1232	**J**	**C**	**0.805**	**0.190**	**0.005**
1233	M	C	0.967	0.023	0.002
1234	R	C	0.955	0.027	0.005
1235	K	C	0.966	0.058	0.007
1236	C	C	0.997	0.001	0.002
1237	Q	C	0.915	0.054	0.021
1238	V	S	0.040	0.100	0.860

Figure 5.7: Comparison of two sample PSIPRED output files, where the left shows predictions for the wild type protein and the right shows the same sequence with the SNP is inserted. Lines are colour coded by increasing difference in prediction probabilities between the wild type and SNP sequence, where red depicts the largest difference and yellow the smallest.

5.1.4 Results

After feature extraction the humvar dataset contained 14,266 pathogenic and 29,154 benign SNPs with all feature data. A 5-fold cross validation was performed in which for each fold 75% of the SNPs were randomly selected to form a training set, and the remaining 25% were used as an unseen test set on the trained classifiers. 6 classifiers were built using the training data: Random Forest, Decision Tree, Logistic Regression, Artificial Neural Network, Naive Bayes and Support Vector Machine (SVM). The R code in figure 5.8 shows the functions used to train and test each classifier.

Table 5.3 shows the results of each fold for each classifier and figure 5.9 shows a Receiver Operator Curve (ROC) plotting the sensitivity vs specificity of each classifier for it's mean scoring model in terms of accuracy across all scoring thresholds (normalized between 0 and 1) generated by the model. Figure 5.10 shows the importance of each feature when used in the Random Forest model to discriminate between pathogenic SNPs and neutral SNPs. The importance is measured by the difference in accuracy when excluding a feature and re-running the model compared to the model with all features used.

```r
# Train Random Forest, Logistic Regression, Support Vector Machine
# Artificial Neural Network, Decision Tree and Naive Bayes classifiers


# Random Forest
# load Random Forest package
require(randomForest)
# Train the model
rf.mdl <- randomForest(label ~.,data=train,importance=TRUE)
# Predict probability belong to each class
rf.prob<-predict(rf.mdl, test,type = "prob")
# Give prediction output: Disease or Polymophism
rf.pd<-predict(rf.mdl, test)


# Logistic Regression
# glm function is in base R
# Use glm function with family = "binomial" for logistic regression
lr.mdl = glm(label ~ ., data=train, family = binomial("logit"))
# Give prediction based on LG response
lr.prob<-predict(lr.mdl, test[,2:ncol(test)],type="response")


# C45 Decision Tree
# Load the rpart R package
library(rpart)
# Train decision tree
dt.mdl <- rpart(label ~ .,method="class", data=train)
# Give prediction
dt.prob<-predict(dt.mdl, test[,2:ncol(test)],type="prob")


# Support Vector Machine
# Load e1071 R package
library(e1071)
# Train SVM model
svm <- svm(label ~ ., data = train)
predict(svm, test[,2:ncol(test)], type = "class")


# Neural Network
# load the nnet R package
library(nnet)
# Train Neural Network
nn<-nnet(label ~ ., data = train, size = 3, rang = 0.1,
         decay = 5e-4, maxit = 200)
# Give prediction
nn.prob<-predict(nn, test[,2:ncol(test)])


# Naive Bayes
# load the e1071 R package
library(e1071)
# Train Naive Bayes
nb <- naiveBayes(label ~ ., data = train)
# Give prediction
predict(nb, test[,2:ncol(test)], type = "class")
```

Figure 5.8: R script used to train and test a Random Forest classifier

Table 5.3: 5 fold cross-validation of the 6 classifiers.

Algorithm	Fold	Disease	Neutral	TP	FP	FN	TN	TPR	FPR	Accuracy
Random Forest	1	4137	5426	3778	257	359	5169	93.63	93.51	93.56
	2	4045	5518	3687	249	358	5269	93.67	93.64	93.65
	3	4127	5436	3813	269	314	5167	93.41	94.27	93.9
	4	4125	5438	3791	265	334	5173	93.47	93.93	93.74
	5	4155	5408	3806	250	349	5158	93.84	93.66	93.74
Logistic Regression	1	4211	5352	3692	343	519	5009	91.5	90.61	90.99
	2	4109	5454	3593	343	516	5111	91.29	90.83	91.02
	3	4195	5368	3725	357	470	5011	91.25	91.42	91.35
	4	4154	5409	3680	376	474	5033	90.73	91.39	91.11
	5	4153	5410	3685	371	468	5039	90.85	91.5	91.23
Neural Network	1	4179	5384	3682	353	497	5031	91.25	91.01	91.11
	2	4077	5486	3587	349	490	5137	91.13	91.29	91.23
	3	4141	5422	3742	340	399	5082	91.67	92.72	92.27
	4	4133	5430	3681	375	452	5055	90.75	91.79	91.35
	5	4162	5401	3701	355	461	5046	91.25	91.63	91.47
Naïve Bayes	1	4744	4819	3717	318	1027	4501	92.12	81.42	85.94
	2	4652	4911	3614	322	1038	4589	91.82	81.55	85.78
	3	4746	4817	3755	327	991	4490	91.99	81.92	86.22
	4	4676	4887	3702	354	974	4533	91.27	82.31	86.11
	5	4769	4794	3773	283	996	4511	93.02	81.91	86.63
SVM	1	4241	5322	3775	260	466	5062	93.56	91.57	92.41
	2	4128	5435	3640	296	488	5139	92.48	91.33	91.8
	3	4220	5343	3768	314	452	5029	92.31	91.75	91.99
	4	4190	5373	3742	314	448	5059	92.26	91.86	92.03
	5	4239	5324	3769	287	470	5037	92.92	91.47	92.08
C45 Decision Tree	1	4333	5230	3685	350	648	4880	91.33	88.28	89.56
	2	4165	5398	3590	346	575	5052	91.21	89.78	90.37
	3	4370	5193	3715	367	655	4826	91.01	88.05	89.31
	4	4308	5255	3689	367	619	4888	90.95	88.76	89.69
	5	4353	5210	3706	350	647	4860	91.37	88.25	89.57

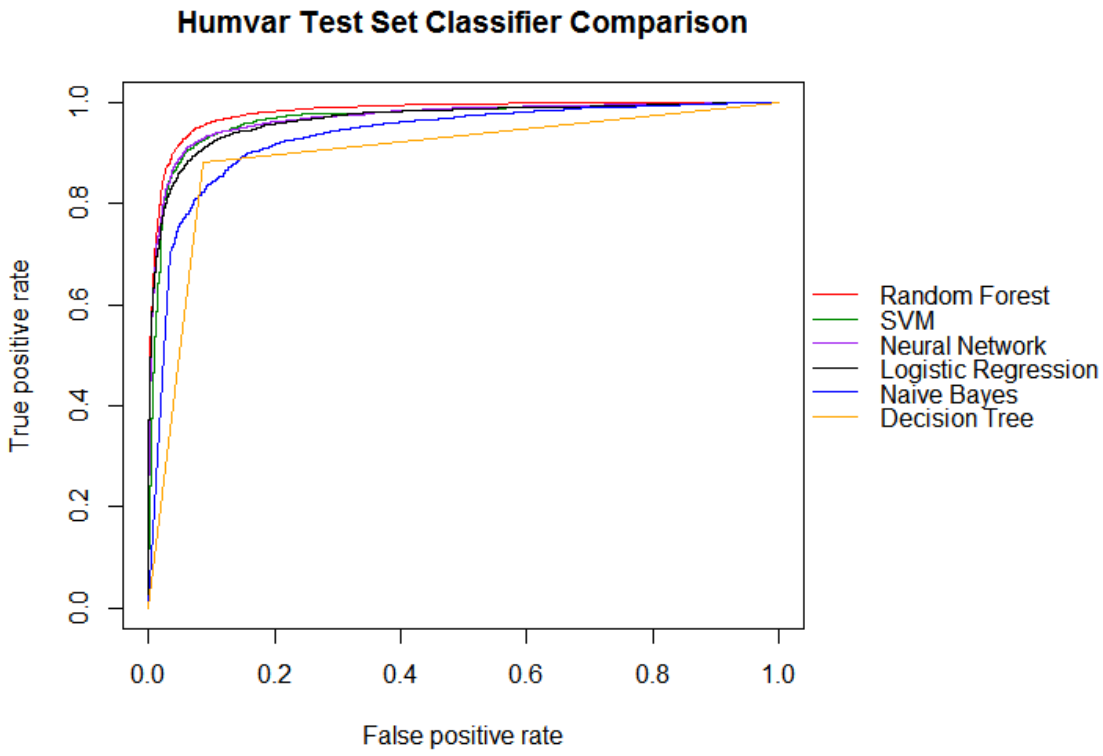

Figure 5.9: Feature importance ranked by the mean decrease in accuracy when each feature is excluded from the Random Forest model

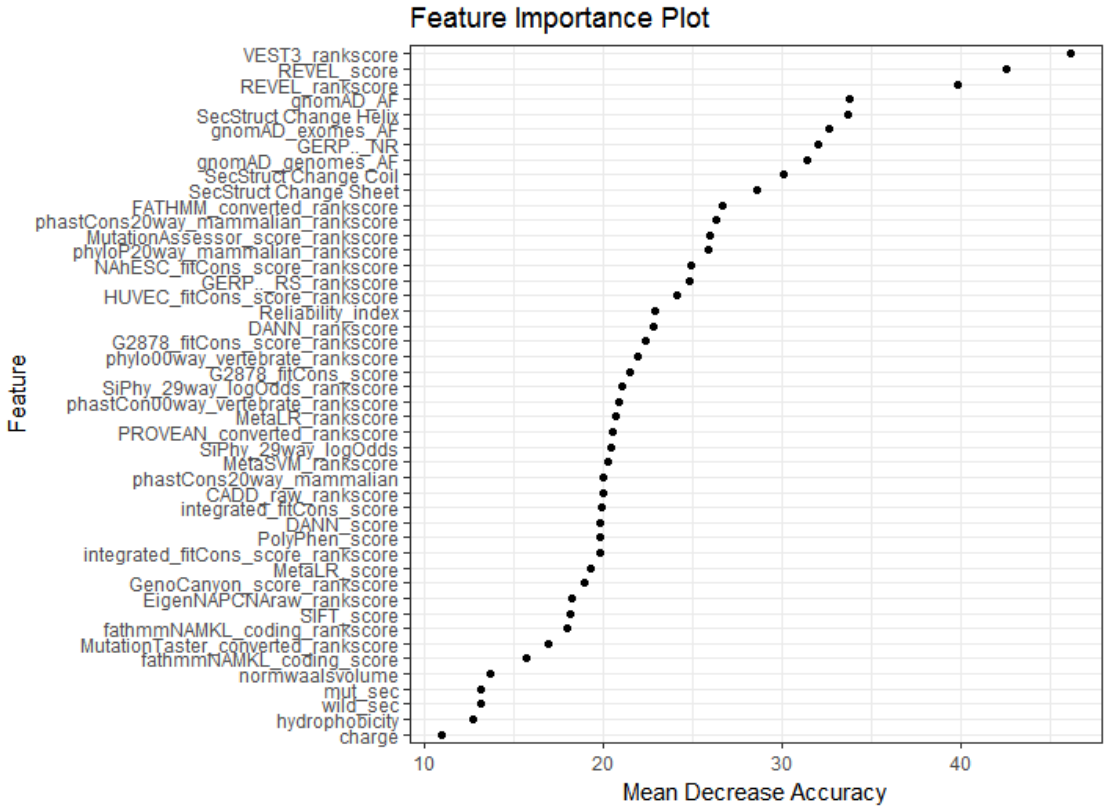

Figure 5.10: Feature importance ranked by the mean decrease in accuracy when each feature is excluded from the Random Forest model

5.1.5 Comparison of Random Forest to other classifiers

Random Forest was the most accurate classifier in every fold (93.56-93.9% accuracy) and was chosen to be compared against existing classifiers. The model used in the 3rd best fold was used (fold 3 or 4 both had 93.74 % accuracy) as this represented the mean accuracy in terms of all 5 Random Forest models. The results for the test set for each of these classifiers was collected as part of the feature extraction pipeline. The score for each classifier, for each SNP in the test set was compared to the class probability given by Random Forest and used to generate the ROC curve in Figure 5.11.

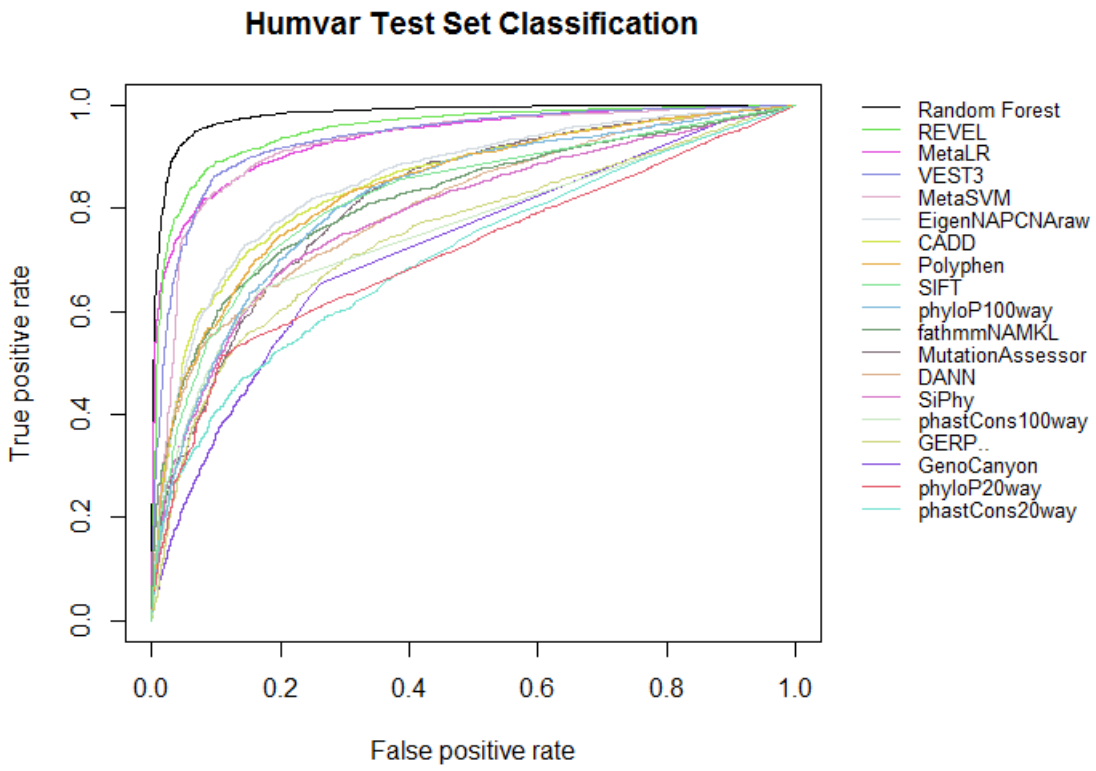

Figure 5.11: ROC curve comparing the classifier from this thesis (black) to scores from other classifiers when predicting disease/benign status on the humvar test set

Random Forest performed better than any of the other classifiers. Figure 5.12 shows the specificity of each algorithm when the sensitivity is set to 95% (for algorithms that could attain 95% sensitivity).

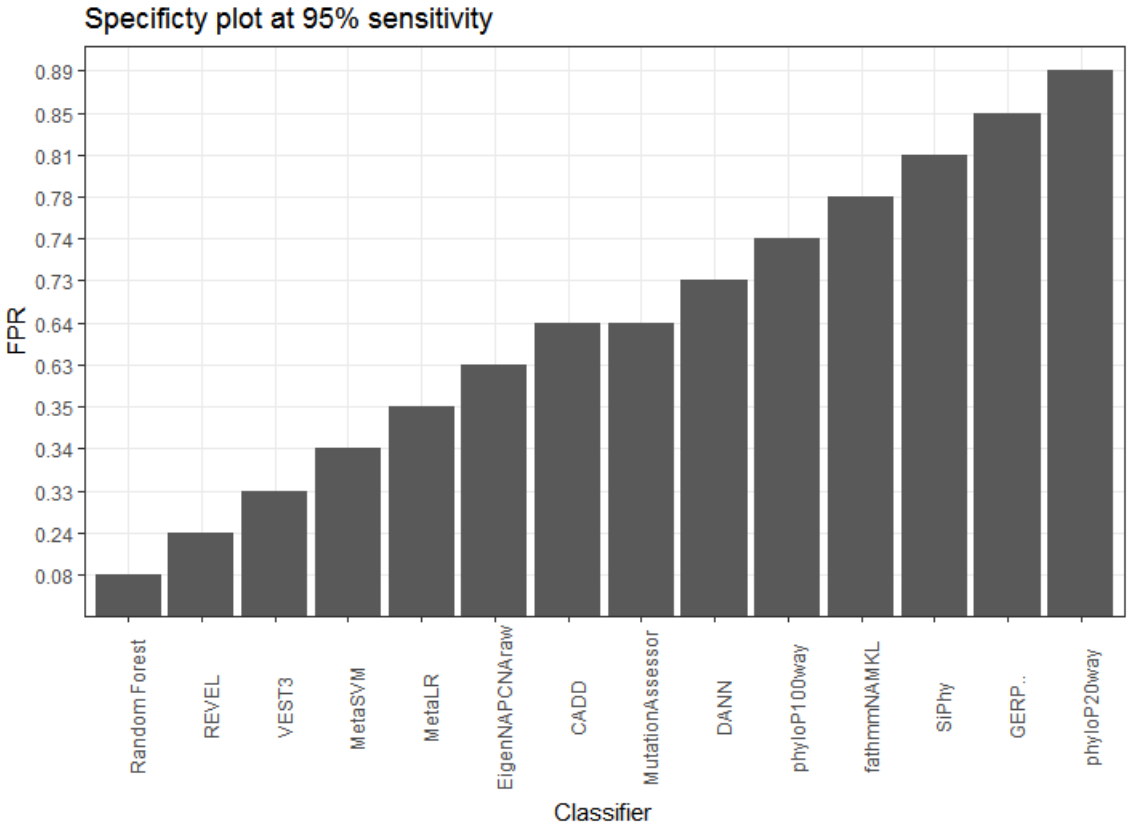

Figure 5.12: Specificity plot of each algorithm when sensitivity is set to 95%. Only algorithms that could achieve 95% sensitivity are presented

5.2 Functional Analysis of SNPs associated with Epilepsy

The Clinvar dataset was queried for all SNPs found in genes that contain mutations known to cause Epilepsy. Using the Clinvar clinical significance guidelines `https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/`, query terms "Benign" were used to identify non-pathogenic SNPs, and "Pathogenic" were used to identify pathogenic SNPs. In total 251 pathogenic and 50 benign SNPs were identified. Table 5.4 shows which genes were selected when querying Clinvar.

Table 5.4: Frequency table of genes associated with epilepsy sourced from Clinvar

Gene	Count	Gene	Count	Gene	Count
SCN1A	153	CPA6	3	STX1B	2
KCNQ3	14	EPM2A	3	CACNA1H	1
EFHC1	13	PLPBP	3	CLN8	1
LGI1	10	POLG	3	DDHD2	1
CHRNB2	9	PRICKLE1	3	GABRA1	1
SCN9A	9	SCN1B	3	GABRB3	1
ALDH7A1	8	SPATA5	3	GAL	1
RELN	7	CHRNA2	2	GLDC	1
SLC2A1	6	CNTNAP2	2	KCNC1	1
CHRNA4	5	GABRG2	2	MRI1	1
MEF2C	5	KCNMA1	2	NACC1	1
SCARB2	5	KCNQ2	2	PRDM8	1
ST3GAL5	5	KCNT1	2		
NHLRC1	4	SCN8A	2		

Each SNP was used as input to the pipeline to collect they required features for classification using the Random Forest classifier trained using the humvar dataset. Each SNP was scored using Random Forest and figure 5.13 shows that Random Forest achieved higher accuracy (92% accuracy, 93.2% sensitivity, and 86% specificity) when compared to other commonly used prediction scoring algorithms. Figure 5.14 shows the specificity of each classifier when set to 95% sensitivity (for classifiers that achieved at least 95% sensitivity), the confusion matrix in table 5.5 shows the overall predicted results and table 5.6 how many pathogenic SNPs were predicted for each gene.

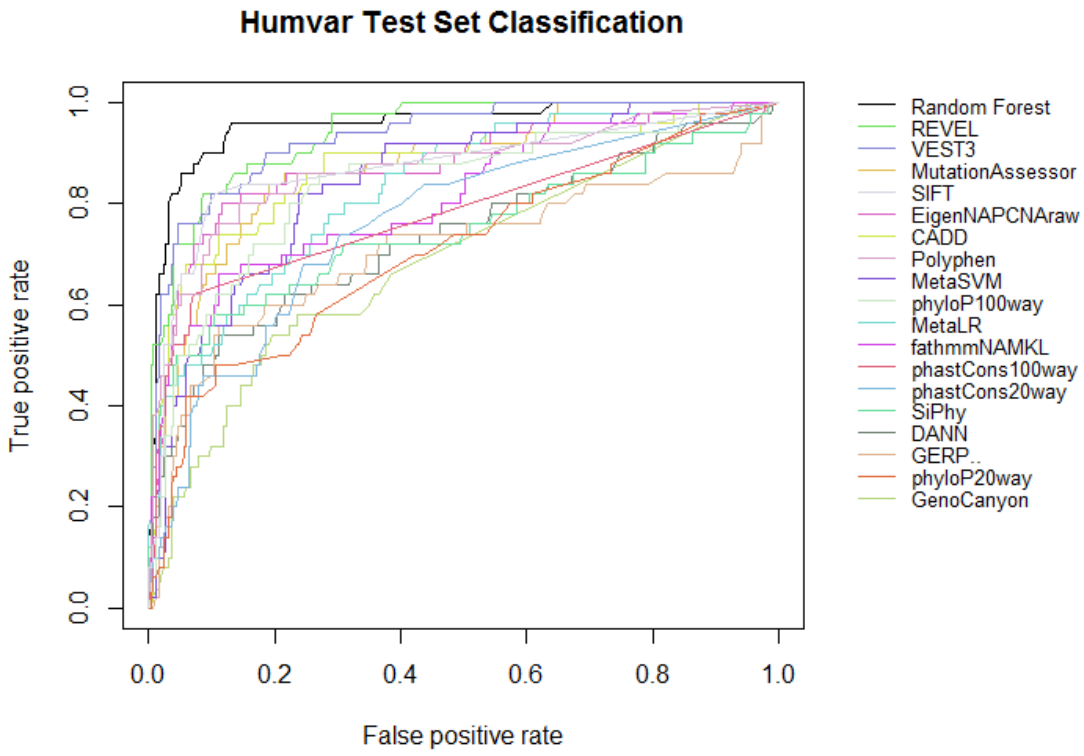

Figure 5.13: ROC curve comparing the classifier from this thesis (black) to rankscores from other classifiers when predicting disease/benign status for SNPs found in genes associated with epilepsy

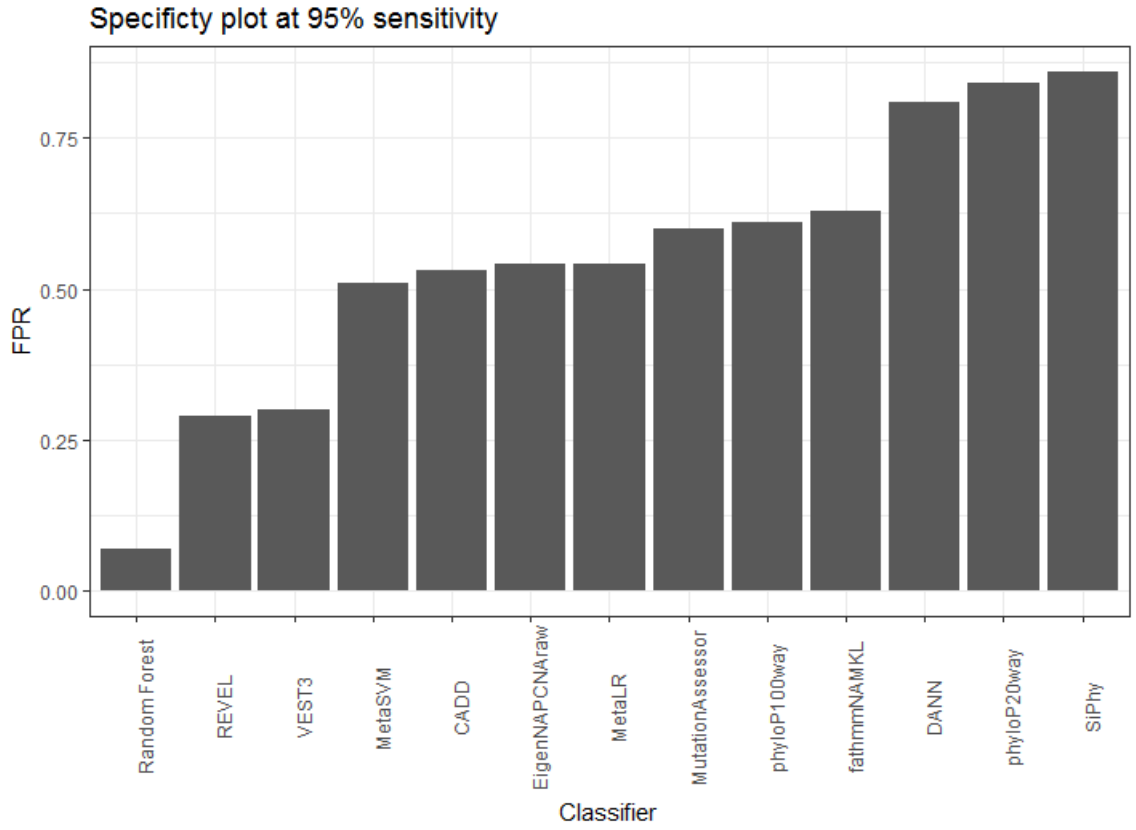

Figure 5.14: Specificity plot of each algorithm when sensitivity is set to 95%. Only algorithms that could achieve 95% sensitivity are presented

Table 5.5: Confusion matrix for Random Forest classifier showing the number of observed vs predicted classifications in 301 SNPs found in genes associated with epilepsy.

	Predicted	
Observed	Disease	Polymorphism
Disease	236	15
Polymorphism	7	43

5.3 Summary of Results

Predicting the effect of SNPs is an essential part of bioinformatics analysis that allows researchers to prioritize which SNPs should be analyzed using downstream processes in a laboratory setting. Many existing prediction scoring systems achieve high accuracy, but few offer both high sensitivity and specificity, as seen when comparing scoring systems on the humvar dataset. Scoring systems are mostly built from machine learning processes that use a variety of relevant protein features to train models, and as such some systems specialize in predicting the effect or certain SNPs, such as

Table 5.6: Comparison of true pathogenic SNPs and predicted pathogenic SNPs in each gene

Gene	True Pathogenic	Predicted Pathogenic	Gene	True Pathogenic	Predicted Pathogenic
SCN1A	150	149	SCARB2	2	2
LGI1	10	10	SCN1B	2	2
ALDH7A1	8	7	STX1B	2	2
CHRNB2	7	7	CLN8	1	2
RELN	7	6	DDHD2	1	1
SLC2A1	6	6	EFHC1	1	1
MEF2C	5	5	EPM2A	1	1
CHRNA4	4	4	GABRA1	1	1
NHLRC1	4	3	GABRB3	1	1
SCN9A	4	3	GAL	1	1
CPA6	3	3	GLDC	1	1
PLPBP	3	3	KCNC1	1	1
POLG	3	3	KCNQ2	1	1
PRICKLE1	3	2	MRI1	1	1
SPATA5	3	2	NACC1	1	1
ST3GAL5	3	2	PRDM8	1	1
GABRG2	2	2	SCN8A	1	0
KCNMA1	2	2	CACNA1H	0	0
KCNQ3	2	2	CHRNA2	0	0
KCNT1	2	2	CNTNAP2	0	0

ultra-rare SNPs or SNPs found in certain regions such as ion channels.

The approach taken in this chapter incorporates as much knowledge from existing scoring systems and commonly used features, as well as bespoke features derived with the use of various bioinformatics software. The aim is to train a classifier to achieve both high sensitivity and specificity in the capability to predict pathogenic SNPs. An automated feature extraction pipeline was built to allow ease of use when processing a large number of SNPs, as demonstrated during the classifier training process of this chapter. Multiple machine learning techniques were explored and compared to 14 other commonly- used prediction software, in which the Random Forest classifier trained in this study was able to achieve the highest accuracy amongst all prediction software.

A comparison of existing prediction software and the Random Forest classifier was also conducted for epilepsy specific SNPs, where an even larger increase in accuracy over the existing software was seen. Various other disease specific studies report reduced accuracy in SNP prediction using existing prediction software, which is hypothesized to be the use of non-disease specific training sets used to train classifiers. The results in this chapter show that epilepsy SNPs are also difficult to classify and show reduced accuracy when compared to the humvar dataset. However, the large increase in accuracy when compared to other software may be due to the inclusion of

many different features from prediction software that may specialize in certain areas. This could also possibly due to the inclusion of structural features in the training set, where various SNPs that cause epilepsy occur in transmembrane proteins. In conclusion, the classifier trained in this chapter performs better than commonly used SNPs prediction software in a large non-disease specific SNP dataset, and performs even better than prediction software in epilepsy specific SNPs.

Chapter 6

Discussion

6.1 SAIL studies

The case studies presented in the first results chapter demonstrate the ability to conduct powerful population level retrospective healthcare studies. Linkage of national datasets such as those held within the SAIL Databank can be used to facilitate novel studies in diseases such as epilepsy, where it is possible to measure the burden of epilepsy in terms of both health and social outcomes. The most important aspect of these results is the ability to define an accurate epilepsy cohort within the SAIL databank so that further downstream research can take place. The validation study to determine people with Epilepsy from GP records shows that data in SAIL can be compared to gold standard external datasets to achieve higher sensitivity, and arguably more importantly very high specificity to ensure only people with epilepsy are used for further studies, and a clear control group can also be defined.

While much research in epilepsy rightly focusses on seizure control, the SAIL databank can be used effectively to measure the social impact of living with epilepsy. It has long been thought that people living with epilepsy often suffer in terms of social deprivation because of factors such as employment and being able to hold a driving license. These results were able to measure the prevalence and incidence of epilepsy across all deciles of WIMD scores to which a strong trend of social deprivation and a diagnosis of epilepsy was found. Importantly, the results were able to address the question of social deprivation in terms of social drift or social causation by comparing social deprivation at time of diagnosis to a 10 year follow up period in which it could not be concluded that there is a strong trend in social drift following a diagnosis of Epilepsy. It is findings such as these that can be valuable to patients with a new diagnosis of epilepsy by giving some assurance that their diagnosis will not make

much difference in terms of social status, while highlighting that extra support is needed given the existing deprivation profile of people with epilepsy.

It is also possible to link electronic healthcare records held within SAIL to administrative data sets such as the Department for Children, Education, Lifelong Learning and Skills in Wales to measure not just the social impact of someone who has epilepsy, but the educational outcomes of children born to mothers with epilepsy. The results in this chapter show that mothers being prescribed sodium valproate during pregnancy are observed to have children that have poorer attainment in national school tests when compared to a control cohort. This result emphasizes why seizure control is such a large research focus in epilepsy, but why it must also take into account situations not limited to effective seizure control. Other UK studies and audits have found that while sodium valproate prescribing is declining in women of child bearing age, it still remains high given the already known risks of reduced IQ and cognition of offspring exposed to sodium valproate *inutero*. These results also add the social impact of epilepsy in school results of children born to mothers with epilepsy, where it is clear these should be an increased focus on limiting sodium valproate prescriptions for women thinking of having children.

There are however clear limitations in all of these studies, which mainly entail important information not available form electronic health care records or linked data, rather than small sample sizes. For example it is not possible to determine the severity of epilepsy in the SAIL databank as it is often not recorded formal in GP records, and while it is possible to ascertain what antiepileptic drug has been prescribed to a patient, the exact daily dose is not recorded. The lack of some of these variables means it is not possible explore the effects of poor seizure control during pregnancy on cognitive outcomes of children, or to meausre the variance in social deprivation of people with epilepsy in terms of their seizure control. Even the definition of epilepsy found in GP records does accurately describe the type of epilepsy, such as focal or generalised epilepsy, and so it is not possible to study outcomes within sub-groups of people with epilepsy. To strengthen the impact of linked healthcare data in epilepsy research, these data gaps must be addressed in order to answer more nuanced research questions.

6.1.1 Social deprivation and epilepsy

The aim of this study was to investigate the relationship between social deprivation of people diagnosed with epilepsy, in particular if high levels of deprivation of chronic diseases such as epilepsy are due to social drift of social causation.

8,100,232 person years of healthcare records were used to calculated the prevalence

and incidence of epilepsy in Wales and link them to the Welsh Index of Multiple Deprivation. The overall epilepsy prevalence using a combination of epilepsy diagnosis codes and AED prescriptions was 0.77% and incidence of 29.5/100 000 per year, which is comparable with other studies in developed countries [253], [51], [254], [49]. By comparing the prevalence of epilepsy in each WIMD decile, a strong association was shown between increased social deprivation as well as higher incidence of epilepsy. Both prevalence and incidence are doubled in the most deprived population decile compared to the least deprived decile (see table 3.5.

The increase in epilepsy incidence with increasingly deprived WIMD deciles initially suggested that the cause of higher epilepsy prevalence in more deprived deciles would be due to the movement from less deprived areas to more deprived areas following an epilepsy diagnosis. However, the follow up cohort study supports the hypothesis that the increased epilepsy prevalence in deprived areas is likely due to social causation rather than social drift (see table 3.7). While the higher prevalence of parents with epilepsy in deprived areas will inevitably produce more children with epilepsy, it is possible that acquired or symptomatic may play a larger role than epilepsy of a genetic origin in more deprived areas. It is difficult to obtain the cause of epilepsy in the SAIL Databank (either genetic or acquired), but given that more deprived areas have increased rates of risk factors for acquired epilepsy, such as perinatal hypoxic injury, head trauma, and cerebrovascular disease [21, 52, 53] it is possible that living in more deprived areas leads to an increased risk of developing epilepsy.

Similar results have been reported in both the UK and internationally. A retrospective study in Wales found more patients with epilepsy living in deprived wards of residence as measured by the Townsend index [50], and the incidence of epilepsy in a prospective study across 20 GP practices in London and the South-East of England identified a strong association between epilepsy incidence and deprivation when comparing the Carstairs score between deprivation fifths, although the association was weaker inside London [51]. Another prospective study using adults in Iceland found that poorer socio-economic status is a risk factor for epilepsy [49].

Area level deprivation measures such as WIMD, Townsend and Carstairs score have limitations in both geographical granularity, and modelling all possible factors associated with social deprivation. As seen in the weaker deprivation effects found in London and the the weaker correlation in the geographical representation of major cities in Wales, area level deprivation is not entirely suitable to measure social deprivation, and individual level deprivation scores would benefit these types of studies. For example, it is possible in this study for two people living in the same area to have the same deprivation score, but in reality they would have different

"individual deprivation," when considering more than geographic location, such as ease of access to services and employment. Similarly for people developing epilepsy in densely populated areas such as major cities, we might not expect to see much movement and therefore the effects of social drift may appear weaker.

These results add further evidence to support the argument that social causation, rather than social drift, could be responsible for an increase in higher social deprivation of people with epilepsy. This provides the opportunity to identify risk factors for epilepsy that could be targeted in areas of higher deprivation, as well as providing further evidence of the health impact of living in socially deprived areas.

Further work remains however. The WIMD score is based on an LSOA level, and it would be useful for future studies to study the effects of epilepsy on deprivation at an individual basis. For example, two people living in the same LSOA will clearly have different levels of social deprivation. There is potential however to link patient records to earnings, benefits and tax records using the Administrative Data Research Center `https://adrn.ac.uk`, which was set up to help researchers link healthcare records with administrative data such as those held in the Department of Work and Pensions. These data could potentially allow researchers to study how deprivation changes following a diagnosis of epilepsy in more detail.

While it was possible calculate a measure of sensitivity 90.5% for the algorithm used to determine epilepsy when comparing to patients in the Cardiff epilepsy register, it was not possible calculate the specificity due to lack of a control group. The prevalence of 0.77% provides an estimate of specificity, however a further work would validate this algorithm by using a control group of people that are known to definitely not have epilepsy. It would also be interesting to study the effects of epilepsy severity on social deprivation. Epilepsy severity is not well recorded in GP records, and so there needs to be a focus on enriching datasets routinely collected healthcare records with more detailed information on disease status.

6.1.2 Validation of epilepsy algorithm using a gold standard dataset

This study aimed to validate different algorithms for selecting people with epilepsy using anonymous GP patient records. The previous deprivation study used an algorithm that took into account epilepsy diagnosis codes and repeat AED prescriptions which cases identified as epilepsy by the algorithm were compared to the an epilepsy register with a gold standard diagnosis. This study used epilepsy patients and patients that definitely did not have a diagnosis of epilepsy sourced from

Morriston hospital to perform a sensitivity/specificity analysis of three algorithms: Epilepsy diagnosis codes only, epilepsy diagnosis codes with a repeat AED prescription and an AED prescription only.

The results showed that by using both epilepsy diagnosis codes with a repeat AED, anonymised GP records can be used to accurately identify patients with epilepsy. This algorithm achieved sensitivity and specificity 84%, 87%, 79% and 99%, 98%, 100% for all patients, adults and children respectively. These figures are comparable with sensitivities and specificities from other epilepsy validation studies in different healthcare systems e.g. Australian, Italian and American studies achieved sensitivities of 82-90% and specificities of 94-100% [236], [237], [238]. The results also showed that using a repeat AED prescription only may be the best approach when identifying children with epilepsy as it achieved 98% for both sensitivity and specificty. This algorithm would not be suitable for use on adults as it only achieved a specificity of 61%. This can be explained by the widespread use of AEDs for indications other than epilepsy in adults (e.g. migraine, mood disorders and neuropathic pain). AEDs are seldom prescribed for indications other than epilepsy in children in the UK [239].

There was little difference in performance between using just an epilepsy diagnosis and both an epilepsy diagnosis and a repeat AED other than adding AEDs results in slightly higher specificity / lower sensitivity. GP diagnosis codes for epilepsy therefore seem reliable in their own right. Although this is expected, given that epilepsy diagnosis should be made in secondary care in the UK and later transcribed into the primary care record by GPs `https://www.nice.org.uk/guidance/cg137`, this has not been described in the literature and is an important result for future research involving GP epilepsy diagnosis codes. It would however be desirable to use an algorithm that maximizes specificity when selecting patients from anonymized GP records as it is a priority to be as certain as possible that someone identified with epilepsy does indeed have epilepsy, especially so if a cohort of epilepsy patients were to be compared against a control group for further study. Therefore using a repeat AED in addition to an epilepsy diagnosis is the preferred algorithm for identifying epilepsy in SAIL. Various reasons might include certain disorders such as non-epileptic attack disorder that present almost identical symptoms to that of an epileptic seizure which require detailed EEG examination to tell the difference. For EEGs to be accurate it requires a seizure to occur during an examination, and so it is difficult to conduct a thorough analysis, especially if seizures are infrequent, leading to some patients taking years to have a confirmation of non-epileptic attacks while being treated with AEDs.

The strengths of this study was that a gold standard dataset of patients with know

epilepsy diagnosed by an epilepsy specialist was available as well as carefully selected patient records for patients that definitely did not have epilepsy. These data provided a robust validation of the three algorithms which could immediately be used to identify many epilepsy cases from the 2.8 million people with an anonymized GP patient record in SAIL. Although the reference population was of gold standard, it was a relatively small population due to the resources needed to manually check medical records and test results. Also, these results are specific to primary care records in Wales and are not applicable to other healthcare systems or methods of ascertaining epilepsy cases (for example hospital discharge summaries). Other parts of the UK do have similar healthcare systems and although the results may be generalizable to the remainder of the UK further work needs to be done to prove this. Currently there is no facility to include EEG and imaging data within SAIL and so we could not include these in our ascertainment algorithms. Additionally it is impossible to identify people with epilepsy who do not attend their GP or have not been seen by a hospital specialist.

The reference epilepsy cohort was obtained from a secondary care epilepsy database which may have provided a bias towards people with more severe epilepsy, and thus more likely for an epilepsy diagnosis to be recorded in GP records. Also the group of people without epilepsy were sourced from patients who had attended general neurology clinics as a control group. This group therefore does not represent the 'general' population without epilepsy. However, this group of patients may be considered as a 'better test' of ascertainment algorithms as patients with other neurological conditions may be more likely to be incorrectly coded as having epilepsy than the general population. Conversely it is also possible (although unlikely in our opinion) that neurologists would not record a diagnosis of epilepsy in a general neurology clinic appointment with a different focus (e.g. headache).

6.1.3 Educational attainment of children born to mother's with epilepsy

This study aimed to compare the educational attainment of children born to mothers with epilepsy to a control group, with a particular focus on which AEDs mothers took during pregnancy. This study used Key Stage 1 results for mathematics, science and English/Welsh, a national school assessment for 2,196 children (440 with epilepsy, 1,756 control) at 7 years of age between 2003 and 2008 academic years.

The results showed an association between poorer school attainment and children exposed to valproate or AEDs in combination inutero. Compared with a matched control group, fewer children with mothers being prescribed sodium valproate during

pregnancy achieved the national minimum standard in CSI (-12.7% less than the control group), mathematics (-12.1%), language (-10.4%) and in science (-12.2%). Even fewer children with mothers being prescribed multiple AEDs during pregnancy achieved a national minimum standard: CSI (by -20.7% less than the control group), mathematics (-21.9%), language (-19.3%) and science (-19.4%)

The results support previous studies that provide consistent evidence that *inutero* exposure to sodium valproate and AEDs in combination are linked to adverse neurodevelopmental outcomes. In contrast there was no difference seen in children exposed to carbamazepine, lamotrigine or mothers that did not take drugs during pregnancy, however it is impossible to accurately test for significance in the lamotrigine group as the sample size is small. While mothers not prescribed any AED during pregnancy do not appear to give birth to children that have decreased educational attainment, it is possible that this group of women have less frequent seizures, thus reducing the risks to the unborn child associated with exposure to seizures *inutero*.

Other studies have also studied the effect of intuero exposure to sodium valproate and the effect on children's IQ. The NEAD study found a 9-point decrease in IQ in children at 3 and 6 years old who were born to mothers taking sodium valproate during pregnancy [27, 243] as well as decreased motor, emotional and behavioural/adaptive functioning in children at 3 years old [255]. Studies based on the UK Epilepsy and Pregnancy Register have associated sodium valproate with a decrease in cognitive development and early cognitive delay that suggests children are at a disadvantage well before school age [32, 256]. While this study finds a statistically non-significant trend in language at KS1, other studies have shown decreased language and verbal skills at early infant stage [257, 258]. Some of the studies mentioned have found that increased AED dosage plays a part in cognitive impairment, however due to a lack of dosage information in the SAIL databank this could not be explored. While some of the mentioned studies associate exposure to carbamazepine with some forms of cognitive impairment, there are also studies that suggest carbamazepine has no effect on intelligence; these results supports the latter with no evidence of decreased educational attainment at school age [259].

The strength of this study is the ability to select a large cohort of 440 children with national test results without major recruitment bias and compare to a large control group. Using a standardized national assessment as a measure of performance ensures that each child has the opportunity to be assessed based on the same curriculum, and as such these results would closer reflect the learning experience of children at this age compared to an IQ test. The main limitation of these results are not being

able to use maternal intelligence quotient (IQ) as a covariate as in [27, 243], which are not recorded in the SAIL databank.

The SAIL Databank also lacks information on various other potential covariates such as epilepsy severity and seizure frequency during pregnancy which may effect cognitive function of unborn child, or if the mother was taking folic acid during pregnancy as this is available "over the counter". The AED data was based on prescriptions, and so it is impossible to comment on adherence, however there is no reason to suspect that adherence differs greatly between different AEDs. It is also possible that mothers with poorly controlled seizures may have an effect on their child's education in terms of parental support outside of school settings, but this information is difficult to ascertain and is not available to any comprehensive standard within the SAIL databank. Another limitation of this study is that we are unable to report on AED dosage, although other studies have reported significant cognitive impairment even at low dosages of sodium valproate.

While these results highlights the risk of cognitive effects in the children of mothers prescribed sodium valproate or multiple AEDs, it is important to acknowledge that some epilepsies are difficult to treat without these treatment regimes. Despite this, these results add to the growing evidence that *inutero* exposure to certain AEDs can cause developmental problems in children, to which sodium valproate has recently been banned for use in women of child bearing age unless a pregnancy prevention programme is in place `https://www.gov.uk/guidance/valproate-use-by-women-and-girls`.

6.2 Natural Language Processing of epilepsy clinic letters

This study aimed to validate an NLP algorithm developed to extract epilepsy specific information from unstructured clinic letters. A rules-based system was built using an open source NLP framework to extract details of epilepsy diagnosis, seizure type, seizure frequency, status of EEG, MRI and CT investigations. The main purpose of the NLP algorithm is to enrich routinely collected data sets such as those demonstrated in the previous chapter using the SAIL databank.

The algorithm was able to extract epilepsy information from a corpus of 200 clinic letters, written by 6 different clinicians, with an overall precision, recall and F1 score of 91%, 81% and 85% on a per item basis. As expected, the algorithm performed best in extracting clinic date and date of birth (F1 scores of 98% and

99%) given that these fields consist of fixed format dates which are relatively easily to extract. In terms of epilepsy-specific information the algorithm performed best for medication (F1=95%), confirming a diagnosis of epilepsy (93%), epilepsy type (84%) and presence of focal seizures (81%). These items are frequently mentioned and presented in a relatively standard format e.g. medication is usually stated as drug name-strength-unit-frequency, and diagnosis appears at the top of letters in structured lists with or in text with clear references to the patient such as "she has focal epilepsy" or "her current medication is" within the main text.

For example, a letter may confirm temporal lobe epilepsy three times but only one mention of temporal lobe epilepsy is required to correctly classify that person's epilepsy. In this context extracting only one mention of temporal lobe epilepsy is just as useful as extracting all three. In the "per letter" test we, therefore, aggregated multiple mentions within a category in each letter to a binary decision based on the algorithm's ability to extract at least one true positive mention. In the above example if the algorithm had only correctly identified one of the three mentions of temporal lobe epilepsy we would have scored it as having a recall of 100% on a per letter basis but only 33% (1/3) on a per item basis. For the medication annotation, in the per letter approach, only a full list of the drugs prescribed with the respective doses was considered to be a positive outcome.

The algorithm was less accurate in identifying CT (57%), MRI (75%) and EEG results (78%), seizure frequency (66%) and generalised seizure terms (66%). The two main reasons for not picking up such terms were due to mapping issues to UMLS, or the highly varied ways these terms are reported in clinic letters. For example, UMLS contains terms such as, "EEG with irregular generalized spike and wave complexes", however, it is often the case that when reported in text there are variety of words between the EEG and the finding e.g. "EEG was found to show generalized spike and wave complexes" or "There was no evidence of generalised spike-wave complexes when reviewing her EEG", and so this problem was approached by creating custom gazetteers that map to smaller terms such as "spike and wave" or "EEG", and writing JAPE rules to associate the finding with the EEG. Similarly the reporting of seizure frequency is highly varied e.g. "she had 5 seizures since March last year" or "1-2 focal seizures every evening". Seizure frequency is also often reported with terms such as "events" and "episodes" rather than defined seizure types, hence additional JAPE rules were built to accommodate such terms as part of seizure frequency in the presence of an epilepsy diagnosis.

Although every item in a letter was compared to that of a clinician, it is practical to provide a binary decision for some categories. If epilepsy is confirmed 3 times in a

letter, the important information is that epilepsy is confirmed. For this reason the "per letter" score was developed where if 1 true positive for a epilepsy confirmed, epilepsy type, seizure type or abnormal/normal investigation, then the "per letter" score was given a true positive finding by the algorithm. The basis for this decision boundary is based on the high precision of the "per item" score, where there is high confidence a positive identified by the algorithm is a true positive. The "per letter" score achieved higher scores for precision, recall, and F1-score (96%, 87%, 91%) on a per letter basis. The "per letter" approach for categories containing multiple mentions could be used with higher confidence than on an individual mention basis, as well as providing a practical way to summarise information from clinic letters. Additionally a "per person" measure (results summarised over several letters) could be used to determine epilepsy status as there will normally be several letters per person over a period of time.

Other studies demonstrate that NLP is being increasingly used for clinical information extraction purposes [260]. Performance of specific phenotype extraction algorithms developed as part of the i2b2 project using cTAKES (Apache clinical Text Analysis and Knowledge Extraction System) and HITex (Health Information Text Extraction) showed that for an NLP approach high PPV (precision) and sensitivity (recall) was achieved for extracting the following phenotypes; Crohn's disease (98%,64%), Ulcerative Colitis (97%,68%) , MS (94%,68%), and Rheumatoid arthritis (89%,56%) [139]. As we aimed to extract epilepsy specific information other than a confirmed diagnosis, a recent study on patients with known MS identified from electronic healthcare records used NLP techniques to extract attributes specific to MS with high PPV and sensitivity, namely EDSS (Expanded Disability Status Scale) (97%,89%), T25FW (Timed 25 Foot Walk) (93%,87%), MS subtype (92%,74%) and age of onset (77%,64%) [140]. This study took into account items attributable only to the patient, as opposed to family members, which is an important distinction and interesting area of study in terms of identify potential risk factors for disease development. A study used clinic letters available at `www.mtsamples.com` to determine whether sentences containing disease and procedure information were attributable to a family member using the BioMedICUS NLP system, which achieved and overall precision, recall and F1-score of 91%, 94% and 92% [141].

There are however only a few published studies of clinical epilepsy information extraction systems. Cui et al developed the rule based Epilepsy Data Extraction and Annotation (EpiDEA) system which extracts epilepsy information from epilepsy monitoring unit discharge summaries. EpiDEA achieved an overall prevision, recall and F1 score of 94%, 84% and 89% when extracting EEG pattern, past medications and current medication from 104 discharge summaries from Cleveland, USA [143].

Cui et al also developed the rule-based Phenotype Exaction in Epilepsy (PEEP) pipeline [144]. PEEP extracted epileptogenic zone, seizure semiology, lateralising sign, interictal and ictal EEG pattern with an overall precision, recall and F1 score of 93%, 93% and 92% in a validation set of 262 epilepsy monitoring unit discharge summaries from Cleveland, USA. Sullivan et al used a machine based learning NLP pipeline to identify a rare epilepsy syndrome from discharge summaries and EEG reports in Phoenix USA and achieved a precision, recall and F1 score of 77%, 67% and 71% respectively.

The main strength of this study was the use of a gold standard dataset of de-identified clinic letters to accurately extract novel data types from free texts that are not well populated in electronic healthcare records. The algorithm was built using open source technology so that the algorithm is easily shareable and can be run on potentially millions of letters as NLP tasks can be parallelism. The algorithm was able to make use of two open source plugins that have been used for information extraction tasks previously, as well as widely used medical ontologies that produce easily interpretable annotations that can be adopted for healthcare research [261] [244]. The algorithm was developed to extract epilepsy specific information, however this aim was met by filtering out non-disease specific information. It is possible that the algorithm could be adapted for other diseases with relative ease and many rules were built to capture language rules in general, not just medical item tagging. These rules could be adapted for more nuanced tasks such as finding frequency of events such as depressive episodes or migraines. Another advantage was that all rules are programmed in a relatively simple scripting language (JAPE) where other NLP systems rely heavily on the ability to program in more complicated languages such as Java. For this reason it is possible that clinicians themselves are more likely to participate in writing their own rules and embed medical expertise more readily into the algorithm or algorithms developed in future, for example adding in custom dictionaries for colloquial terms or coding a particular phrase that is meaningful when reported in clinic letters.

The main limitation of this study was the sample size of letters used to develop and test the algorithm. This is a limitation across many NLP tasks that focus on information extraction as it is labour intensive to manually annotate letters in the detail required to develop information extraction systems. Most information extraction systems typically use hundreds of letters rather than thousands. However even though this study only used 200 letters to test the accuracy of the algorithm, 1925 individual items were compared to those of an epilepsy specialist. Another limitation was the use of one specialist to annotate the letters used for comparison, where it is possible that annotations are affected by reviewer fatigue, and it was not possible to produce an inter-annotation agreement score if multiple clinicians annotated the letters. The

address this a review of all letters that contained a disagreement between the human annotations and the algorithm annotations (174 letters out of the 200 letters) was conducted. The review showed that indeed the human annotator errors tended to be missing true positive items, which at first lead to a higher number of false positives produced by the algorithm where the human annotator corrected their annotations. In future multiple specialists are recommended for information extraction tasks.

While existing medical ontologies were used to tag items in the clinic letters, and rules were built to arrange which particular order of tagged terms describes a concept, medical ontologies such as UMLS, SNOMED-CT, READ and ICD were not built to aid information extraction from clinical free texts. Many descriptions of codes in these ontologies are too structured to reflect the language used in clinic letters and thus reduces recall in information extraction tasks. It is however important to produce annotations that adhere to existing medical ontologies, and so future work to address this would include developing methods to add some flexibility when matching terms in clinic letters to ontology descriptions. Much work has been done in so called "fuzzy matching" where sequences of words called "n-grams" are mapped to each other using word-vectorization and similarity measures [125] [262] [263]. Also it would be interesting to incorporate machine learning classification tasks where detailed information extraction is not needed. For example it might be possible to use machine learning to classify EEG reports as either normal or abnormal, without worrying about extracting every item that may indicate abnormalities.

6.3 Predicting pathogenicity of SNPs for large datasets

This study aimed to understand the different approaches for predicting pathogenicity of SNPs and to build a pipeline that sources a variety of SNP annotation and scoring data that is used to build an accurate classifier for SNPs. Many existing software have significantly better sensitivity that specificity, or vice-a-versa, and some specialize in capturing rare pathogenic variants or variants found within certain protein domains such as transmembrane regions. Most classifiers are non-disease specific, but some studies have shown that the accuracy reported on non-disease specific test sets are not matched in some disease only test sets. As well as building a non-disease specific SNP classifier, the classifier was also built to accurately classify epilepsy SNPs.

The results in this study showed that by using a mixture of common SNP annotations such as conservation scores, existing software scoring systems and bespoke features such as secondary structure prediction, the Random Forest classifier was able to

score higher than all of the highest performing software currently available on both a non-disease specific dataset as well as showing a further increase in accuracy when using an epilepsy dataset. The classifier also achieved both high sensitivity and specificity, which some classifiers achieved high sensitivity or specificity. Constructing a ROC curve and using the class probability of predictions allowed the Random Forest classifier to be compared at different sensitivity thresholds to compare the corresponding specificity across all other classifiers, in which at a sensitivity threshold of 95%, the Random Forest classifier achieved 97% specificity which was higher than the closest classifier, REVEL (92% specificity). Six different classifiers were compared during the development phase using a 5-fold cross validation coupled with feature selection techniques to try and improve the accuracy of each algorithm. Internal parameters of the classifiers were also tuned to achieve higher accuracy, although this did not have much of an effect. It was interesting that aside from using the Naive Bayes classifier, many of the algorithms showed higher accuracy than any of the other existing classifiers used for comparison. This suggests that the samples in the training set and the features used in the training process many have play more of a role in classification accuracy than a difference in choosing different classifiers.

Prediction accuracy of new software as presented in their original publications state high accuracy in non-specific disease datasets such as Humvar, but often vary in accuracy when used for a particular disease. Given that researching a specific disease is a common use case of research groups, it is important to know that the accuracy stated for an algorithm is not expected to be reproduced in other datasets. A study that compared the prediction accuracy of 17 different classifiers on SNPs in limb-girdle muscular distrophy (LGMS) showed that the Polyhen Humvar classifier achieved just 70.2% accuracy where in the Polyphen paper an accuracy of 86% is stated, with a similar trend for the majority of the other classifiers used in the comparison [264]. A study that selected 23 genes associated with immunity compared PolyPhen2, SIFT, MutationAssessor, Panther, CADD, and Condel classifiers in which only 20% of pathogenic SNPs were predicted as such, and over 45% of neutral SNPs were classified as pathogenic [265]. A larger non-disease specific study using 40,000 variants from the Phencode and dbSNP databases found that accuracy ranged from 15-65% across MutPred, nsSNPA-nalyzer, Panther, PhD-SNP, PolyPhen, PolyPhen2, SIFT,SNAP, and SNPsGO [266].

There have been various studies which have explored creating disease specific classifiers that have reported higher accuracy than non-disease specific classifiers. A cancer specific machine learning based classifier was developed using common SNP annotation for 6326 missense SNPs that are known to be drivers for various subtypes of cancer, achieving 93% accuracy [267]. Machine learning classifiers that dominate SNP

prediction require large datasets with many features, in which there are not enough pathogenic SNPs in the majority of individual disease areas to train a classifier by using only damaging SNPs in the disease of interest. This is certainly true for epilepsy, and as such the approach taken was to include many features other than traditional conservation scores. Some of the prediction scores used in the classifier include features that specify which protein domain a SNP is found in such as transmembrane or ion channel regions as well as incorporating the predicted secondary structure of the SNP location as well as modelling the difference in predicted secondary structures between sequences containing the wild type and the SNP. Many pathogenic SNPs in genes associated with epilepsy such as SCN1A are located in ion channels, where any structural change in these channels can effect cellular excitability and induce seizures [268]. It is possible that some features used in the RF classifier are able to contribute towards improving classification accuracy in epilepsy related SNPs.

6.3.1 Use of existing predictors as features

Including the prediction and scores of existing software as a feature for machine learning purposes is not a novel idea. Earlier attempts to improve classification accuracy when relatively few prediction software existed involved combining these scores and weighting them into a single score using a statistical approach. The Combined Annotation scoRing toOL (CAROL) normalizes polyphen and SIFT scores by their standard normal deviations to achieve a 1% increase on that of their respective scores using a test set of 1,959 pathogenic and 9,691 neutral SNPs [269].

6.3.2 Use of physiochemical properties and predicted secondary structure

Most SNP prediction programs that use secondary structure as features obtain this information from known protein databases and uses the secondary structure status where the SNP is located as reported in the wild type protein. Querying secondary structure in this manner produces many missing data as only a proportion of proteins have been studied and had their secondary structure profile reported. This study used secondary structure prediction to obtain a predication of the secondary structure for every SNP, and modelled the difference in prediction between wild type and SNP sequences. These differences were reported as the percentage change across three secondary structure domains, in which they were ranked as the 5th, 9th and 10th most important features (out of 47 features) used in the Random Forest model.

It is difficult to define "difference in secondary structure" between a wild type sequence, and a sequence containing a SNP. The secondary structure of both sequences are

predicted, not measured in a laboratory. The only thing that can be measured using PSIPRED is the likelihood of a single amino acid falling into the three categories of a beta sheet, alpha coil or helix. While this may be useful, particularly in some strict complexes of a protein, it is actually not important to this pipeline to determine what the effect of a SNP may have on a predicted secondary structure category. Firstly this is because it is not feasible to determine such a change for a SNP via prediction only, but mainly that the underlying methodology to secondary structure prediction is similar to sequence homology and evolutionary analysis of proteins - methods that already play a prominent role in functional prediction programs. Secondary structural prediction programs are extensions of multiple sequence alignments, where known secondary structures are incorporated later as a ground truth to assess structure based on these alignments. It is this inference between multiple sequence alignments and the disparity in structural prediction between a wild type sequence and a SNP sequence that is important.

6.4 Future Work

The SAIL case studies in this thesis showed that epilepsy patients can accurately be determined from GP records. The algorithm developed could be used to create an epilepsy register within the SAIL databank to facilitate further epilepsy research and be used in multiple studies. These studies have shown that socio-economic and national education datasets can be linked to epilepsy patients. One future project would be to link the educational outcomes of children who have epilepsy compared to a control group or other neurological conditions. It could be possible to study sub groups of children by taking into account what prescriptions they were taking during the school year.

There are limitations, however to the SAIL databank. Rich information found in clinic letters were largely absent from rountinely collected healthcare records. The next step would be to use the validated NLP application developed in chapter 4 to link rich epilepsy data from clinic records to the SAIL databank. This would require producing a version of the NLP application that could run on distributed systems. The lack of clinic letters available for NLP studies due to patient identifiers present is a major limitation that is not reflected with the SAIL databank. Forming a governance model to analyse identifiable patient data would be an important future project to help facilitate NLP research. Another barrier is not only the lack of clinic letters, but the lack of letter annotation by a clinician. Every NLP algorithm needs to be scored against a human annotator, and generating annotated training sets requires a lot of time. With crowd sourcing becoming popular it may be possible

to open up annotation tasks to a wider participation group through crowd sourcing platforms.

Some categories of information from the NLP chapter were difficult to obtain accurately. Seizure frequency is an important piece of information that provides a level of severity of epilepsy, and would be an important covariate in an epilepsy study. The language describing seizures frequency and investigation reports remain to varied to capture all of the nuances with a rule based approach. Machine learning is one way to add flexibility into capturing these categories, and one immediate improvement would be to create methods of matching short phrases and slang terms to UMLS concept descriptions so that NLP applications aren't relying on exact phrase matching. Much work has been done on word embeddings to infer semantic similarity of words, however some future work in this area could expand this to phrases.

SNP prediction tools are numerous and show effectiveness in different disease areas. The pipeline developed in chapter 6 successfully incorporated existing SNP prediction tools as features alongside bespoke protein features to produce higher accuracy than all common SNP scoring tools in a disease non-specific dataset as well as epilepsy specific SNPs. With advances in technology powering big data, accurate tools will become an increasingly important tool in prioritising which SNPs warrant further research/ Future work would be to set up the pipeline developed as a web service to be used freely for research. This could help laboratories focus on building assays for a smaller group of SNPs. The pipeline could also be used to prioritise SNPs that may be linked to the SAIL databank. The ability to identify patients with certain conditions and co-morbidities and link them to SNPs may allow researchers to discover associations between SNPs and disease. Future work will include linking exomes and genomes from Welsh patients collected as part of the Wales Epilepsy Research Network to the SAIL databank. These patients have well defined epilepsy genotypes and it would be interesting to explore any potential trends in co-morbidities within subgroups of epilepsy syndromes such as exploring if different SNPs for the same epilepsy syndrome have a a higher propensity form experiencing co-morbidities such as migraines.

6.5 Conclusions

Epilepsy is a common disease that can have an impact on health and social well being. Epilepsy can also affect family members such as children born to parents with epilepsy. Advances in big data have provided the opportunity to explore the impact of epilepsy on at population levels to uncover trends in healthcare data for people with epilepsy where insufficient data previously existed.

The SAIL Databank not only provides national healthcare datasets, but it also socio-economic and administrative datasets that can all be linked anonymously. Data linkage is a powerful tool that allows datasets to be aligned to produce novel research, and the ability to do so anonymously speeds up research in terms of using a governance model that is exempt from ethical approval. A study using 8.1 million person-years of data was used to identify a strong trend between increased social deprivation and the incidence and prevalence of epilepsy. The study found that increased deprivation was likely to due to social causation rather than social drift. This finding was contrary to other studies that suggest chronic conditions such as epilepsy cause social drift after diagnosis.

A study of GP coding habits showed that it is possible to use GP records as a data source for identifying patients with epilepsy to be used for epidemiology studies. An algorithm was developed that found a combination of repeat AED prescriptions and a diagnosis of epilepsy was the most effective way to identify epilepsy from GP records, which gave 88% sensitivity and 98% specificity. This algorithm was used to select mother's with epilepsy who had children with Key Stage 1 education data to compare attainment between a large control group. This study found that for children exposed to sodium valproate or a combination of AEDs *inutero* there was a 12.7% and 19.8% decrease in attainment when compared to the control group.

Each SAIL study showed that there were various limitations in the data i.e. not having epilepsy type or AED daily dose. The Natural Language Processing study showed that there is potential to source rich patient information that is typically missing from routinely collected data. A validation of 200 epilepsy clinic letters showed that a rule-based NLP application can accurately identify patients with epilepsy (88.5%), epilepsy type (84.5%) and prescriptions (95%). Other categories such as seizure frequency and investigation outcomes were more difficult to capture, however high specificity is reported across all categories. Aggregating all mentioned per category per letter achieved even higher accuracy and would be a practical approach to analysing large volumes of letters.

A SNP prediction pipeline was developed using the Random Forest machine learning classifier to determine the pathogenicity of SNPs. Validation on a large disease non-specific SNP dataset showed that the Random Forest classifier produced more accurate results than all other commonly used SNP prediction software, and was able to achieve 95% sensitivity with a specificity of 92%. The classifier also achieved the highest accuracy on a dataset of 301 SNPs reported in epilepsy genes, and confirms findings in other studies than disease specific SNP datasets can pose a more difficult challenge in terms of predicting SNP pathogenicity.

Appendices

Appendix item 1: Read codes used for epilepsy definition

Table 6.1: Read codes used to signify a diagnosis of epilepsy.

Code	Description	Code	Description
F25B.	Alcohol-induced epilepsy	F25y2	Localisation related epilepsy
F25y4	Benign Rolandic epilepsy	F25D.	Menstrual epilepsy
F2545	Complex partial epileptic seizure	F2511	Neonatal myoclonic epilepsy
F25y3	Complex partial status epilepticus	667B.	Nocturnal epilepsy
F25y0	Cursive (running) epilepsy	F25y.	Other forms of epilepsy
F25C.	Drug-induced epilepsy	F25yz	Other forms of epilepsy NOS
F259.	Early infant epileptic encephalopathy wth suppression bursts	F251y	Other specified generalised convulsive epilepsy
F25..	Epilepsy	F250y	Other specified generalised nonconvulsive epilepsy
1O30.	Epilepsy confirmed	F25y5	Panayiotopoulos syndrome
F25z.	Epilepsy NOS	F254.	Partial epilepsy with impairment of consciousness
F2544	Epileptic automatism	F254z	Partial epilepsy with impairment of consciousness NOS
F2503	Epileptic seizures - akinetic	F255.	Partial epilepsy without impairment of consciousness
F2502	Epileptic seizures - atonic	F255z	Partial epilepsy without impairment of consciousness NOS
F2512	Epileptic seizures - clonic	F255y	Partial epilepsy without impairment of consciousness OS
F2513	Epileptic seizures - myoclonic	F2500	Petit mal (minor) epilepsy
F2514	Epileptic seizures - tonic	F252.	Petit mal status
F25y1	Gelastic epilepsy	F25F.	Photosensitive epilepsy
F251.	Generalised convulsive epilepsy	F258.	Post-ictal state
F251z	Generalised convulsive epilepsy NOS	F2541	Psychomotor epilepsy
F250.	Generalised nonconvulsive epilepsy	F2542	Psychosensory epilepsy
F250z	Generalised nonconvulsive epilepsy NOS	F2501	Pykno-epilepsy
F2510	Grand mal (major) epilepsy	F2561	Salaam attacks
F2516	Grand mal seizure	F2551	Sensory induced epilepsy
F253.	Grand mal status	F2556	Simple partial epileptic seizure
F2560	Hypsarrhythmia	F2552	Somatosensory epilepsy
F256z	Infantile spasms NOS	F25E.	Stress-induced epilepsy
F2504	Juvenile absence epilepsy	F2515	Tonic-clonic epilepsy
F25A.	Juvenile myoclonic epilepsy	SC200	Traumatic epilepsy
F257.	Kojevnikov's epilepsy	F2555	Unilateral epilepsy
F2505	Lennox-Gastaut syndrome	F2553	Visceral reflex epilepsy
F2543	Limbic system epilepsy	F2554	Visual reflex epilepsy

Code	Description	Code	Description
dn3e.*	ARBIL MR 200mg m/r tablets	dnp7.	LYRICA 300mg capsules
dn3f.	*ARBIL MR 400mg m/r tablets	dnp2.	LYRICA 50mg capsules
dn2..	*BECLAMIDE	dnp3.	LYRICA 75mg capsules
dn2z.	*BECLAMIDE 500mg tablets	dn6..	METHYLPHENOBARBITAL
dnc1.	*CLOBAZAM SLS 10mg capsules	dn6z.	METHYLPHENOBARBITONE 200mg tablets

Table 6.2 AED Read codes — *Continued . . .*

Code	Description	Code	Description
do1z.	*DIAZEPAM 20mg/4mL injection	dn6x.	METHYLPHENOBARBITONE 30mg tablets
do1B.	*DIAZEPAM 20mg/5mL RecTubes	dn6y.	METHYLPHENOBARBITONE 60mg tablets
dn53.	*EMESIDE 250mg capsules	dna1.	MYSOLINE 250mg tablets
dn3A.	*EPIMAZ 100mg tablets	dna2.	MYSOLINE 250mg/5mL oral suspension
dn3B.	*EPIMAZ 200mg tablets	dna3.	MYSOLINE 50mg tablets
dn3C.	*EPIMAZ 400mg tablets	dnj4.	NEURONTIN 100mg capsules
dn51.	*ETHOSUXIMIDE 250mg capsules	dnj5.	NEURONTIN 300mg capsules
dn52.	*ETHOSUXIMIDE 250mg/5mL elixir	dnj9.	NEURONTIN 300mg capsules/600mg tablets titration pack
dn5y.	*ETHOSUXIMIDE 250mg/5mL elixir	dnj6.	NEURONTIN 400mg capsules
dn79.	*GARDENAL 200mg/1mL injection	dnj7.	NEURONTIN 600mg tablets
dn7a.	*LUMINAL 15mg tablets	dnj8.	NEURONTIN 800mg tablets
dn7b.	*LUMINAL 30mg tablets	dng2.	NOOTROPIL 1.2g tablets
dn7c.	*LUMINAL 60mg tablets	dng3.	NOOTROPIL 33
dn21.	*NYDRANE 500mg tablets	dng1.	NOOTROPIL 800mg tablets
dnba.	*ORLEPT 200mg e/c tablets	dnbA.	ORLEPT 200mg/5mL sugar free liquid
dnbb.	*ORLEPT 500mg e/c tablets	dnb9.	ORLEPT STARTER PACK 200mg e/c tablets x10
do52.	*PARALDEHYDE injection 10mL	dnm..	OXCARBAZEPINE
do51.	*PARALDEHYDE injection 5mL	dnmx.	OXCARBAZEPINE 150mg tablets
dn98.	*PENTRAN 100mg tablets	dnmy.	OXCARBAZEPINE 300mg tablets
dn97.	*PENTRAN 50mg tablets	dnmz.	OXCARBAZEPINE 600mg tablets
dn63.	*PROMINAL 200mg tablets	dnmw.	OXCARBAZEPINE 60mg/mL sugar free oral suspension
dn61.	*PROMINAL 30mg tablets	do5..	PARALDEHYDE
dn62.	*PROMINAL 60mg tablets	dn7..	PHENOBARBITAL
do13.	*STESOLID 20mg/4mL injection	dn74.	PHENOBARBITAL 100mg tablets
dn3H.	*TERIL CR 200mg m/r tablets	dn71.	PHENOBARBITAL 15mg tablets
dn3I.	*TERIL CR 400mg m/r tablets	dn7d.	PHENOBARBITAL 15mg/5mL elixir
dn55.	*ZARONTIN 250mg capsules	dn78.	PHENOBARBITAL 200mg/1mL injection
dn1y.	ACETAZOLAMIDE [EP] 250mg tablets	dn72.	PHENOBARBITAL 30mg tablets
dn1z.	ACETAZOLAMIDE [EP] 500mg injection	dn73.	PHENOBARBITAL 60mg tablets
dn1x.	ACETAZOLAMIDE [EP] 500mg m/r capsules	dn77.	PHENOBARBITONE 15mg/10mL elixir
dn1..	ACETAZOLAMIDE [EPILEPSY]	dn75.	PHENOBARBITONE SODIUM 30mg tablets
do41.	ATIVAN [EP] 4mg/mL injection	dn76.	PHENOBARBITONE SODIUM 60mg tablets
dn3J.	CARBAGEN SR 200mg m/r tablets	dn8..	PHENYTOIN

Table 6.2 AED Read codes — *Continued ...*

Code	Description	Code	Description
dn3K.	CARBAGEN SR 400mg m/r tablets	dn92.	PHENYTOIN 100mg tablets
dn3..	CARBAMAZEPINE	dn8y.	PHENYTOIN 30mg/5mL suspension
dn3y.	CARBAMAZEPINE 100mg chewable tablets	dn83.	PHENYTOIN 50mg chewable tablets
dn31.	CARBAMAZEPINE 100mg tablets	dn91.	PHENYTOIN 50mg tablets
dn3z.	CARBAMAZEPINE 100mg/5mL sugar free liquid	dn8z.	PHENYTOIN 90mg/5mL sugar free suspension
dn3v.	CARBAMAZEPINE 125mg suppositories	dn9..	PHENYTOIN SODIUM
dn3x.	CARBAMAZEPINE 200mg chewable tablets	do6..	PHENYTOIN SODIUM [STATUS EPILEPSY]
dn3a.	CARBAMAZEPINE 200mg m/r tabs	dn9z.	PHENYTOIN SODIUM 100mg capsules
dn32.	CARBAMAZEPINE 200mg tablets	do6z.	PHENYTOIN SODIUM 250mg/5mL injection
dn3w.	CARBAMAZEPINE 250mg suppositories	dn9x.	PHENYTOIN SODIUM 25mg caps
dn3b.	CARBAMAZEPINE 400mg m/r tabs	dn9w.	PHENYTOIN SODIUM 300mg capsules
dn33.	CARBAMAZEPINE 400mg tablets	dn9y.	PHENYTOIN SODIUM 50mg capsules
dnc..	CLOBAZAM [EPILEPSY ONLY]	dng..	PIRACETAM
do3..	CLOMETHIAZOLE EDISYLATE [CENTRAL NERVOUS SYSTEM USE]	dng5.	PIRACETAM 1.2g tablets
do3z.	CLOMETHIAZOLE EDISYLATE 8mg/mL intravenous infusion	dng6.	PIRACETAM 333.3mg/mL oral solution
dn4..	CLONAZEPAM [EPILEPSY CONTROL]	dng4.	PIRACETAM 800mg tablets
do2..	CLONAZEPAM [STATUS EPILEPSY]	dnp..	PREGABALIN
dn4w.	CLONAZEPAM 0.5mg/5mL sugar free oral solution	dnpv.	PREGABALIN 100mg capsules
do2z.	CLONAZEPAM 1mg/1mL injection	dnpw.	PREGABALIN 150mg capsules
dn4z.	CLONAZEPAM 2mg tablets	dnpu.	PREGABALIN 200mg capsules
dn4x.	CLONAZEPAM 2mg/5mL sugar free oral solution	dnps.	PREGABALIN 225mg capsules
dn4y.	CLONAZEPAM 500microgram tablets	dnpz.	PREGABALIN 25mg capsules
dn...	CONTROL OF EPILEPSY	dnpt.	PREGABALIN 300mg capsules
dnh1.	CONVULEX 150mg e/c capsules	dnpy.	PREGABALIN 50mg capsules
dnh2.	CONVULEX 300mg e/c capsules	dnpx.	PREGABALIN 75mg capsules
dnh3.	CONVULEX 500mg e/c capsules	dna..	PRIMIDONE
dnh7.	DEPAKOTE 250mg e/c tablets	dnay.	PRIMIDONE 250mg tablets
dnh8.	DEPAKOTE 500mg e/c tablets	dnaz.	PRIMIDONE 250mg/5mL oral suspension
dns1.	DIACOMIT 250mg capsules	dnax.	PRIMIDONE 50mg tablets

Table 6.2 AED Read codes — *Continued . . .*

Code	Description	Code	Description
dns3.	DIACOMIT 250mg/sachet powder for oral suspension	dni2.	PRO-EPANUTIN 750mg/10mL injection concentrate
dns2.	DIACOMIT 500mg capsules	dnv..	RETIGABINE
dns4.	DIACOMIT 500mg/sachet powder for oral suspension	dnv8.	RETIGABINE 100mg tablets
dn12.	DIAMOX [EP] 250mg tablets	dnv9.	RETIGABINE 200mg tablets
dn13.	DIAMOX [EP] 500mg injection	dnvA.	RETIGABINE 300mg tablets
dn11.	DIAMOX [EP] 500mg m/r capsules	dnvB.	RETIGABINE 400mg tablets
do11.	DIAZEMULS [EP] 10mg/2mL injection	dnv7.	RETIGABINE 50mg tablets
do1..	DIAZEPAM [EPILEPSY USE]	dnvC.	RETIGABINE 50mg+100mg tablets initiation pack
do1y.	DIAZEPAM 10mg/2.5mL rectal solution	do21.	RIVOTRIL 1mg/1mL injection
do19.	DIAZEPAM 10mg/2.5mL RecTubes	dn42.	RIVOTRIL 2mg tablets
do1v.	DIAZEPAM 10mg/2mL emulsion injection	dn41.	RIVOTRIL 500micrograms tablets
do1w.	DIAZEPAM 10mg/2mL injection	dnr..	RUFINAMIDE
do1t.	DIAZEPAM 2.5mg/1.25mL rectal solution	dnrz.	RUFINAMIDE 100mg tablets
do1A.	DIAZEPAM 2.5mg/1.25mL RecTubes	dnry.	RUFINAMIDE 200mg tablets
do1u.	DIAZEPAM 20mg/5mL rectal solution	dnrx.	RUFINAMIDE 400mg tablets
do1x.	DIAZEPAM 5mg/2.5mL rectal solution	dne4.	SABRIL 500mg powder sachets
do18.	DIAZEPAM 5mg/2.5mL RecTubes	dne2.	SABRIL 500mg tablets
dn54.	EMESIDE 250mg/5mL syrup	dnb..	SODIUM VALPROATE
do61.	EPANUTIN [EP] 250mg/5mL injection	dnbv.	SODIUM VALPROATE 100mg crushable tablets
dn95.	EPANUTIN 100mg capsules	dnbo.	SODIUM VALPROATE 100mg/sachet m/r granules
dn93.	EPANUTIN 25mg capsules	dnbJ.	SODIUM VALPROATE 150mg m/r capsules
dn96.	EPANUTIN 300mg capsules	dnbN.	SODIUM VALPROATE 1g/10mL solution for injection
dn81.	EPANUTIN 30mg/5mL suspension	dnbM.	SODIUM VALPROATE 1g/sachet m/r granules
dn94.	EPANUTIN 50mg capsules	dnbw.	SODIUM VALPROATE 200mg crushable tablets
dn82.	EPANUTIN 50mg Infatabs	dnb7.	SODIUM VALPROATE 200mg e/c tablets
dnb1.	EPILIM 100mg crushable tablets	dnbr.	SODIUM VALPROATE 200mg m/r tablets
dnb2.	EPILIM 200mg e/c tablets	dnby.	SODIUM VALPROATE 200mg/5mL sugar free liquid
dnb4.	EPILIM 200mg/5mL sugar free liquid	dnbz.	SODIUM VALPROATE 200mg/5mL syrup

Table 6.2 AED Read codes — *Continued ...*

Code	Description	Code	Description
dnb5.	EPILIM 200mg/5mL syrup	dnbp.	SODIUM VALPROATE 250mg/sachet m/r granules
dnb3.	EPILIM 500mg e/c tablets	dnbK.	SODIUM VALPROATE 300mg m/r capsules
dnbc.	EPILIM CHRONO 200 m/r tablets	dnbs.	SODIUM VALPROATE 300mg m/r tablets
dnbd.	EPILIM CHRONO 300 m/r tablets	dnbE.	SODIUM VALPROATE 300mg/3mL solution for injection
dnbe.	EPILIM CHRONO 500 m/r tablets	dnbu.	SODIUM VALPROATE 400mg/4mL injection
dnbQ.	EPILIM CHRONOSPHERE 100mg/sachet m/r granules	dnb8.	SODIUM VALPROATE 500mg e/c tablets
dnbU.	EPILIM CHRONOSPHERE 1g/sachet m/r granules	dnbt.	SODIUM VALPROATE 500mg m/r tablets
dnbR.	EPILIM CHRONOSPHERE 250mg/sachet m/r granules	dnbx.	SODIUM VALPROATE 500mg tablets
dnbS.	EPILIM CHRONOSPHERE 500mg/sachet m/r granules	dnbL.	SODIUM VALPROATE 500mg/sachet m/r granules
dnbP.	EPILIM CHRONOSPHERE 50mg/sachet m/r granules	dnbn.	SODIUM VALPROATE 50mg/sachet m/r granules
dnbT.	EPILIM CHRONOSPHERE 750mg/sachet m/r granules	dnbq.	SODIUM VALPROATE 750mg/sachet m/r granules
dnb6.	EPILIM IV 400mg/4mL injection	do...	STATUS EPILEPTICUS DRUGS
dnbF.	EPISENTA 150mg m/r capsules	do12.	STESOLID [EP] 10mg/2mL injection
dnbO.	EPISENTA 1g/10mL solution for injection	do15.	STESOLID 10mg/2.5mL rectal solution
dnbI.	EPISENTA 1g/sachet m/r granules	do14.	STESOLID 5mg/2.5mL rectal solution
dnbG.	EPISENTA 300mg m/r capsules	dnsw.	STIRIPENDOL 500mg/sachet powder for oral suspension
dnbD.	EPISENTA 300mg/3mL solution for injection	dns..	STIRIPENTOL
dnbH.	EPISENTA 500mg/sachet m/r granules	dnsz.	STIRIPENTOL 250mg capsules
dnbB.	EPIVAL CR 300mg m/r tablets	dnsx.	STIRIPENTOL 250mg/sachet powder for oral suspension
dnbC.	EPIVAL CR 500mg m/r tablets	dnsy.	STIRIPENTOL 500mg capsules
dnu..	ESLICARBAZEPINE	dn3c.	TEGRETOL 100mg chewable tablets
dnu2.	ESLICARBAZEPINE ACETATE 800mg tablets	dn34.	TEGRETOL 100mg tablets
dn5..	ETHOSUXIMIDE	dn37.	TEGRETOL 100mg/5mL sugar free liquid
dn5x.	ETHOSUXIMIDE 250mg capsules	dn3D.	TEGRETOL 125mg suppositories
dn5z.	ETHOSUXIMIDE 250mg/5mL syrup	dn3d.	TEGRETOL 200mg chewable tablets
dni..	FOSPHENYTOIN SODIUM	dn35.	TEGRETOL 200mg tablets

Table 6.2 AED Read codes — *Continued . . .*

Code	Description	Code	Description
dni1.	FOSPHENYTOIN SODIUM 750mg/10mL injection concentrate	dn3E.	TEGRETOL 250mg suppositories
dnj..	GABAPENTIN	dn36.	TEGRETOL 400mg tablets
dnj1.	GABAPENTIN 100mg capsules	dn38.	TEGRETOL RETARD 200mg m/r tabs
dnj2.	GABAPENTIN 300mg capsules	dn39.	TEGRETOL RETARD 400mg m/r tabs
dnjx.	GABAPENTIN 300mg capsules/600mg tablets titration pack	dnl..	TIAGABINE
dnj3.	GABAPENTIN 400mg capsules	dnl2.	TIAGABINE 10mg tablets
dnjy.	GABAPENTIN 600mg tablets	dnl3.	TIAGABINE 15mg tablets
dnjz.	GABAPENTIN 800mg tablets	dnl1.	TIAGABINE 5mg tablets
dnl5.	GABITRIL 10mg tablets	dn3F.	TIMONIL RETARD 200mg m/r tablets
dnl6.	GABITRIL 15mg tablets	dn3G.	TIMONIL RETARD 400mg m/r tablets
dnl4.	GABITRIL 5mg tablets	dnk5.	TOPAMAX 100mg tablets
do31.	HEMINEVRIN [CNS] 8mg/mL intravenous infusion	dnk6.	TOPAMAX 200mg tablets
dnr1.	INOVELON 100mg tablets	dnk8.	TOPAMAX 25mg tablets
dnr2.	INOVELON 200mg tablets	dnk4.	TOPAMAX 50mg tablets
dnr3.	INOVELON 400mg tablets	dnkB.	TOPAMAX SPRINKLE 15mg capsules
dno5.	KEPPRA 100mg/mL s/f oral solution	dnkC.	TOPAMAX SPRINKLE 25mg capsules
dno3.	KEPPRA 1g tablets	dnkE.	TOPAMAX SPRINKLE 50mg capsules
dno1.	KEPPRA 250mg tablets	dnk..	TOPIRAMATE
dno2.	KEPPRA 500mg tablets	dnk2.	TOPIRAMATE 100mg tablets
dno6.	KEPPRA 500mg/5mL solution for injection	dnk9.	TOPIRAMATE 15mg beads in capsules
dno4.	KEPPRA 750mg tablets	dnk3.	TOPIRAMATE 200mg tablets
dnt..	LACOSAMIDE	dnkA.	TOPIRAMATE 25mg beads in capsules
dntA.	LACOSAMIDE 100mg tablets	dnk7.	TOPIRAMATE 25mg tablets
dntB.	LACOSAMIDE 150mg tablets	dnkD.	TOPIRAMATE 50mg beads in capsules
dnt8.	LACOSAMIDE 15mg/1mL sugar free liquid	dnk1.	TOPIRAMATE 50mg tablets
dntC.	LACOSAMIDE 200mg tablets	dnm1.	TRILEPTAL 150 tablets
dnt7.	LACOSAMIDE 200mg/20mL solution for injection	dnm2.	TRILEPTAL 300 tablets
dnt9.	LACOSAMIDE 50mg tablets	dnm3.	TRILEPTAL 600 tablets
dnf9.	LAMICTAL 100mg dispersible tablets	dnm4.	TRILEPTAL 60mg/mL sugar free oral suspension
dnf4.	LAMICTAL 100mg tablets	dnv2.	TROBALT 100mg tablets
dnfD.	LAMICTAL 200mg tablets	dnv3.	TROBALT 200mg tablets
dnf8.	LAMICTAL 25mg dispersible tablets	dnv4.	TROBALT 300mg tablets
dnf6.	LAMICTAL 25mg tablets	dnv5.	TROBALT 400mg tablets
dnfJ.	LAMICTAL 2mg dispersible tablets	dnv1.	TROBALT 50mg tablets
dnf3.	LAMICTAL 50mg tablets	dnv6.	TROBALT tablets initiation pack
dnf7.	LAMICTAL 5mg dispersible tablets	do16.	VALIUM [EP] 10mg/2mL injection

Table 6.2 AED Read codes — *Continued ...*

Code	Description	Code	Description
dnfF.	LAMICTAL MONOTHERAPY 25mg starter pack	do17.	VALIUM [EP] 20mg/4mL injection
dnfH.	LAMICTAL NON-VALPROATE ADD-ON 50mg starter pack	dnh..	VALPROIC ACID
dnfG.	LAMICTAL VALPROATE ADD-ON 25mg starter pack	dnh4.	VALPROIC ACID 150mg e/c capsules
dnf..	LAMOTRIGINE	dnhz.	VALPROIC ACID 250mg e/c tablets
dnfC.	LAMOTRIGINE 100mg dispersible tablets	dnh5.	VALPROIC ACID 300mg e/c capsules
dnf2.	LAMOTRIGINE 100mg tablets	dnh6.	VALPROIC ACID 500mg e/c capsules
dnfE.	LAMOTRIGINE 200mg tablets	dnhy.	VALPROIC ACID 500mg e/c tablets
dnfB.	LAMOTRIGINE 25mg dispersible tablets	dne..	VIGABATRIN
dnf5.	LAMOTRIGINE 25mg tablets	dne3.	VIGABATRIN 500mg powder sachets
dnfz.	LAMOTRIGINE 2mg dispersible tablets	dne1.	VIGABATRIN 500mg tablets
dnf1.	LAMOTRIGINE 50mg tablets	dnt4.	VIMPAT 100mg tablets
dnfA.	LAMOTRIGINE 5mg dispersible tablets	dnt5.	VIMPAT 150mg tablets
dno..	LEVETIRACETAM	dnt2.	VIMPAT 15mg/1mL sugar free liquid
dnov.	LEVETIRACETAM 100mg/mL s/f oral solution	dnt6.	VIMPAT 200mg tablets
dnox.	LEVETIRACETAM 1g tablets	dnt1.	VIMPAT 200mg/20mL solution for injection
dnoz.	LEVETIRACETAM 250mg tablets	dnt3.	VIMPAT 50mg tablets
dnoy.	LEVETIRACETAM 500mg tablets	dn56.	ZARONTIN 250mg/5mL syrup
dnou.	LEVETIRACETAM 500mg/5mL solution for injection	dnu1.	ZEBINIX 800mg tablets
dnow.	LEVETIRACETAM 750mg tablets	dnq6.	ZONEGRAN 100mg capsules
do4..	LORAZEPAM [EPILEPSY]	dnq4.	ZONEGRAN 25mg capsules
dnp4.	LYRICA 100mg capsules	dnq5.	ZONEGRAN 50mg capsules
dnp5.	LYRICA 150mg capsules	dnq..	ZONISAMIDE
dnp6.	LYRICA 200mg capsules	dnq3.	ZONISAMIDE 100mg capsules
dnp8.	LYRICA 225mg capsules	dnq1.	ZONISAMIDE 25mg capsules
dnp1.	LYRICA 25mg capsules	dnq2.	ZONISAMIDE 50mg capsule

Bibliography

[1] W Owen Pickrell, Arron S Lacey, Rhys H Thomas, Ronan A Lyons, Phil EM Smith, and Mark I Rees. Trends in the first antiepileptic drug prescribed for epilepsy between 2000 and 2010. *Seizure-European Journal of Epilepsy*, 23(1): 77–80, 2014.

[2] Erik F Tjong Kim Sang and Sabine Buchholz. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, pages 127–132. Association for Computational Linguistics, 2000.

[3] Arron S Lacey, William Owen Pickrell, Rhys H Thomas, Mike P Kerr, Cathy P White, and Mark I Rees. Educational attainment of children born to mothers with epilepsy. *J Neurol Neurosurg Psychiatry*, pages jnnp–2017, 2018.

[4] Owen Pickrell, Arron Lacey, Owen Bodger, Joanne Demmler, Rhys Thomas, Ronan Lyons, Phil Smith, Mark Rees, and Mike Kerr. Epilepsy prevalence, incidence and socioeconomic deprivation. *Accepted into Epilepsia*, 2014.

[5] Beata Fonferko-Shadrach, Arron S Lacey, Catharine P White, HW Rob Powell, Inder MS Sawhney, Ronan A Lyons, Phil EM Smith, Mike P Kerr, Mark I Rees, and W Owen Pickrell. Validating epilepsy diagnoses in routinely collected data. *Seizure-European Journal of Epilepsy*, 52:195–198, 2017.

[6] Anthony K Ngugi, Christian Bottomley, Immo Kleinschmidt, Josemir W Sander, and Charles R Newton. Estimation of the burden of active and life-time epilepsy: a meta-analytic approach. *Epilepsia*, 51(5):883–890, 2010.

[7] JW Sander and SD Shorvon. Epidemiology of the epilepsies. *Journal of neurology, neurosurgery, and psychiatry*, 61(5):433, 1996.

[8] Matilde Leonardi and T Bedirhan Ustun. The global burden of epilepsy. *Epilepsia*, 43(s6):21–25, 2002.

[9] Robert S Fisher, Carlos Acevedo, Alexis Arzimanoglou, Alicia Bogacz, J Helen Cross, Christian E Elger, Jerome Engel, Lars Forsgren, Jacqueline A French, Mike Glynn, et al. Ilae official report: a practical clinical definition of epilepsy. *Epilepsia*, 55(4):475–482, 2014.

[10] Anne T Berg, Samuel F Berkovic, Martin J Brodie, Jeffrey Buchhalter, J Helen Cross, Walter van Emde Boas, Jerome Engel, Jacqueline French, Tracy A Glauser, Gary W Mathern, et al. Revised terminology and concepts for organization of seizures and epilepsies: report of the ilae commission on classification and terminology, 2005–2009. *Epilepsia*, 51(4):676–685, 2010.

[11] Jerome Engel Jr. Ilae classification of epilepsy syndromes. *Epilepsy research*, 70:5–10, 2006.

[12] Anna L Peljto, Christie Barker-Cummings, Vincent M Vasoli, Cynthia L Leibson, W Allen Hauser, Jeffrey R Buchhalter, and Ruth Ottman. Familial risk of epilepsy: a population-based study. *Brain*, 137(3):795–805, 2014.

[13] D Janz, G Beck-Mannagetta, and T Sander. Do idiopathic generalized epilepsies share a common susceptibility gene? *Neurology*, 42(4 Suppl 5):48–55, 1992.

[14] Samuel F Berkovic, R Anne Howell, David A Hay, and John L Hopper. Epilepsies in twins: genetics of the major epilepsy syndromes. *Annals of neurology*, 43(4): 435–445, 1998.

[15] G Beck-Mannagetta and D Janz. Syndrome-related genetics in generalized epilepsy. *Epilepsy research. Supplement*, 4:105, 1991.

[16] Patrick Cossette, Lidong Liu, Katéri Brisebois, Haiheng Dong, Anne Lortie, Michel Vanasse, Jean-Marc Saint-Hilaire, Lionel Carmant, Andrei Verner, Wei-Yang Lu, et al. Mutation of gabra1 in an autosomal dominant form of juvenile myoclonic epilepsy. *Nature genetics*, 31(2):184–189, 2002.

[17] Deb K Pal and David A Greenberg. Major susceptibility genes for common idiopathic epilepsies: Elp4 in rolandic epilepsy and brd2 in juvenile myoclonic epilepsy. 2012.

[18] John C Mulley, Ingrid E Scheffer, Steven Petrou, Leanne M Dibbens, Samuel F Berkovic, and Louise A Harkin. Scn1a mutations and epilepsy. *Human mutation*, 25(6):535–542, 2005.

[19] Louise A Harkin, Jacinta M McMahon, Xenia Iona, Leanne Dibbens, James T Pelekanos, Sameer M Zuberi, Lynette G Sadleir, Eva Andermann, Deepak

Gill, Kevin Farrell, et al. The spectrum of scn1a-related infantile epileptic encephalopathies. *Brain*, 130(3):843–852, 2007.

[20] Samuel Steer, William O Pickrell, Michael P Kerr, and Rhys H Thomas. Epilepsy prevalence and socioeconomic deprivation in england. *Epilepsia*, 55 (10):1634–1641, 2014.

[21] Charles R Newton and Hector H Garcia. Epilepsy in poor regions of the world. *The Lancet*, 380(9848):1193–1201, 2012.

[22] Elias Olafsson and W Allen Hauser. Prevalence of epilepsy in rural iceland: A population-based study. *Epilepsia*, 40(11):1529–1534, 1999.

[23] MICHAEL J McLEAN and ROBERT L Macdonald. Sodium valproate, but not ethosuximide, produces use-and voltage-dependent limitation of high frequency repetitive firing of action potentials of mouse central neurons in cell culture. *Journal of Pharmacology and Experimental therapeutics*, 237(3):1001–1011, 1986.

[24] Cynthia L Harden. New antiepileptic drugs. *Neurology*, 44(5):787–787, 1994.

[25] Marc A Dichter and Martin J Brodie. New antiepileptic drugs. *New England Journal of Medicine*, 334(24):1583–1590, 1996.

[26] Anthony G Marson, Asya M Al-Kharusi, Muna Alwaidh, Richard Appleton, Gus A Baker, David W Chadwick, Celia Cramp, Oliver C Cockerell, Paul N Cooper, Julie Doughty, et al. The sanad study of effectiveness of valproate, lamotrigine, or topiramate for generalised and unclassifiable epilepsy: an unblinded randomised controlled trial. *The Lancet*, 369(9566):1016–1026, 2007.

[27] Kimford J Meador, Gus A Baker, Nancy Browning, Jill Clayton-Smith, Deborah T Combs-Cantrell, Morris Cohen, Laura A Kalayjian, Andres Kanner, Joyce D Liporace, Page B Pennell, et al. Cognitive function at 3 years of age after fetal exposure to antiepileptic drugs. *New England Journal of Medicine*, 360(16):1597–1605, 2009.

[28] Carol S Camfield, Sheila Chaplin, Anna-Beth Doyle, Stanley H Shapiro, Carl Cummings, and Peter R Camfield. Side effects of phenobarbital in toddlers; behavioral and cognitive aspects. *The Journal of pediatrics*, 95(3):361–365, 1979.

[29] TA Ketter, RM Post, and WH Theodore. Positive and negative psychiatric effects of antiepileptic drugs in patients with seizure disorders. *Neurology*, 53(5 Suppl 2):S53–67, 1999.

[30] Hanne Dinesen, Lennart Gram, Teis Andersen, and Mogens Dam. Weight gain during treatment with valproate. *Acta neurologica scandinavica*, 70(2):65–69, 1984.

[31] William Owen Pickrell, Arron S Lacey, Rhys H Thomas, Philip EM Smith, and Mark I Rees. Weight change associated with antiepileptic drugs. *J Neurol Neurosurg Psychiatry*, 84(7):796–799, 2013.

[32] Rebekah Shallcross, Rebecca L Bromley, B Irwin, LJ Bonnett, J Morrow, and GA Baker. Child development following in utero exposure levetiracetam vs sodium valproate. *Neurology*, 76(4):383–389, 2011.

[33] Lars Skou Elkjær, Bodil Hammer Bech, Yuelian Sun, Thomas Munk Laursen, and Jakob Christensen. Association between prenatal valproate exposure and performance on standardized language and mathematics tests in school-aged children. *JAMA neurology*, 2018.

[34] Christopher JL Murray, Theo Vos, Rafael Lozano, Mohsen Naghavi, Abraham D Flaxman, Catherine Michaud, Majid Ezzati, Kenji Shibuya, Joshua A Salomon, Safa Abdalla, et al. Disability-adjusted life years (dalys) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010. *The lancet*, 380(9859):2197–2223, 2012.

[35] Hanneke M De Boer. "out of the shadows": a global campaign against epilepsy. *Epilepsia*, 43(s6):7–8, 2002.

[36] SD Lhatoo and JWAS Sander. The epidemiology of epilepsy and learning disability. *Epilepsia*, 42(s1):6–9, 2001.

[37] Gregory Stores. School-children with epilepsy at risk for learning and behaviour problems. *Developmental Medicine & Child Neurology*, 20(4):502–508, 1978.

[38] Joan K Austin, David W Dunn, Cynthia S Johnson, and Susan M Perkins. Behavioral issues involving children and adolescents with epilepsy and the impact of their families: recent research data. *Epilepsy & Behavior*, 5:33–41, 2004.

[39] Lorie D Hamiwka, G Yu Cara, Lorraine A Hamiwka, Elisabeth MS Sherman, Blaire Anderson, and Elaine Wirrell. Are children with epilepsy at greater risk for bullying than their peers? *Epilepsy & Behavior*, 15(4):500–505, 2009.

[40] Irene M Elliott, Lucyna Lach, and Mary Lou Smith. I just want to be normal: a qualitative study exploring how children and adolescents view the impact of

intractable epilepsy on their quality of life. *Epilepsy & behavior*, 7(4):664–678, 2005.

[41] Michael Seidenberg, Niels Beck, Michael Geisser, Bruno Giordani, J Chris Sackellares, Stanley Berent, FE Dreifuss, and Thomas J Boll. Academic achievement of children with epilepsy. *Epilepsia*, 27(6):753–759, 1986.

[42] R Schulz, Hans Otto Lüders, S Noachtar, T May, A Sakamoto, H Holthausen, and P Wolf. Amnesia of the epileptic aura. *Neurology*, 45(2):231–235, 1995.

[43] H Meinardi, RA Scott, R Reis, JWAS Sander, ILAE Commission on the Developing World, et al. The treatment gap in epilepsy: the current situation and ways forward. *Epilepsia*, 42(1):136–149, 2001.

[44] Rajendra Kale. Bringing epilepsy out of the shadows. *BMJ: British Medical Journal*, 315(7099):2, 1997.

[45] World Health Organization et al. Atlas: epilepsy care in the world. 2005.

[46] Joan Petersilia. Invisible victims-violence against persons with developmental disabilities. *Hum. Rts.*, 27:9, 2000.

[47] Kate Jacoby and Ann Jacoby. Epilepsy and insurance in the uk: an exploratory survey of the experiences of people with epilepsy. *Epilepsy & Behavior*, 5(6): 884–893, 2004.

[48] Ann Jacoby, Joanne Gorry, and Gus A Baker. Employers' attitudes to employment of people with epilepsy: still the same old story? *Epilepsia*, 46(12):1978–1987, 2005.

[49] Dale C Hesdorffer, Hong Tian, Kishlay Anand, W Allen Hauser, Petur Ludvigsson, Elias Olafsson, and Olafur Kjartansson. Socioeconomic status is a risk factor for epilepsy in icelandic adults but not in children. *Epilepsia*, 46 (8):1297–1303, 2005.

[50] Christopher LI Morgan, Zahir Ahmed, and Michael P Kerr. Social deprivation and prevalence of epilepsy and associated health usage. *Journal of Neurology, Neurosurgery & Psychiatry*, 69(1):13–17, 2000.

[51] Dominic C Heaney, Bridget K MacDonald, Alex Everitt, Simon Stevenson, Giovanni S Leonardi, Paul Wilkinson, and Josemir W Sander. Socioeconomic variation in incidence of epilepsy: prospective community based study in south east england. *Bmj*, 325(7371):1013–1016, 2002.

[52] Debbie A Lawlor, George Davey Smith, Rita Patel, and Shah Ebrahim.

Life-course socioeconomic position, area deprivation, and coronary heart disease: findings from the british women's heart and health study. *American journal of public health*, 95(1):91–97, 2005.

[53] Chien-Chang Liao, Huai-Chia Chang, Chun-Chieh Yeh, Yi-Chun Chou, Wen-Ta Chiu, and Ta-Liang Chen. Socioeconomic deprivation and associated risk factors of traumatic brain injury in children. *Journal of trauma and acute care surgery*, 73(5):1327–1331, 2012.

[54] L Anne Jeffreys, Andrew L Clark, and Marek Koperski. Practice nurses' workload and consultation patterns. *Br J Gen Pract*, 45(397):415–418, 1995.

[55] New clinical classification system will streamline computerised medical records. 1990.

[56] World Health Organization et al. History of the development of the icd. *World Health Organization*, 2006.

[57] World Health Organization. International statistical classification of diseases and related health problems. 1, 2004.

[58] Stephen J Gillam, A Niroshan Siriwardena, and Nicholas Steel. Pay-for-performance in the united kingdom: impact of the quality and outcomes framework—a systematic review. *The Annals of Family Medicine*, 10(5):461–468, 2012.

[59] Melanie Calvert, Aparna Shankar, Richard J McManus, Helen Lester, and Nick Freemantle. Effect of the quality and outcomes framework on diabetes care in the united kingdom: retrospective cohort study. *Bmj*, 338:b1870, 2009.

[60] Congenital Anomaly Register. Information service for wales (caris). *CARIS review 2012*, 1998.

[61] Ronan A Lyons, Kerina H Jones, Gareth John, Caroline J Brooks, Jean-Philippe Verplancke, David V Ford, Ginevra Brown, and Ken Leake. The sail databank: linking multiple health and social care datasets. *BMC Medical Informatics and Decision Making*, 9(1):3, 2009.

[62] David V Ford, Kerina H Jones, Jean-Philippe Verplancke, Ronan A Lyons, Gareth John, Ginevra Brown, Caroline J Brooks, Simon Thompson, Owen Bodger, Tony Couch, et al. The sail databank: building a national architecture for e-health research and evaluation. *BMC health services research*, 9(1):157, 2009.

[63] M Hyatt, SE Rodgers, S Paranjothy, D Fone, and RA Lyons. The wales electronic cohort for children (wecc) study. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 96(Suppl 1):Fa18–Fa18, 2011.

[64] Shantini Paranjothy, Frank Dunstan, William J Watkins, Melanie Hyatt, Joanne C Demmler, Ronan A Lyons, and David Fone. Gestational age, birth weight, and risk of respiratory hospital admission in childhood. *Pediatrics*, 132 (6):e1562–e1569, 2013.

[65] Belinda J Gabbe, Caroline Brooks, Joanne C Demmler, Steven Macey, Melanie A Hyatt, and Ronan A Lyons. The association between hospitalisation for childhood head injury and academic performance: evidence from a population e-cohort study. *J Epidemiol Community Health*, pages jech–2013, 2014.

[66] Shantini Paranjothy, Annette Evans, Amrita Bandyopadhyay, David Fone, Behnaz Schofield, Ann John, Mark A Bellis, Ronan A Lyons, Daniel Farewell, and Sara Jayne Long. Risk of emergency hospital admission in children associated with mental disorders and alcohol misuse in the household: an electronic birth cohort study. *The Lancet Public Health*, 2018.

[67] Narayan P Iyer, David F Tucker, Selwyn H Roberts, Marsham Moselhi, Margery Morgan, and Jean W Matthes. Outcome of fetuses with turner syndrome: a 10-year congenital anomaly register based study. *The Journal of Maternal-Fetal & Neonatal Medicine*, 25(1):68–73, 2012.

[68] Hayley A Hutchings, Annette Evans, Peter Barnes, Joanne Demmler, Martin Heaven, Melanie A Hyatt, Michelle James-Ellison, Ronan A Lyons, Alison Maddocks, Shantini Paranjothy, et al. Do children who move home and school frequently have poorer educational outcomes in their early years at school? an anonymised cohort study. *PloS one*, 8(8):e70601, 2013.

[69] Ann John, Joanne McGregor, David Fone, Frank Dunstan, Rosie Cornish, Ronan A Lyons, and Keith R Lloyd. Case-finding for common mental disorders of anxiety and depression in primary care: an external validation of routinely collected data. *BMC medical informatics and decision making*, 16(1):35, 2016.

[70] A John, AL Marchant, DL Fone, JI McGregor, MS Dennis, JOA Tan, and K Lloyd. Recent trends in primary-care antidepressant prescribing to children and young people: an e-cohort study. *Psychological medicine*, 46(16):3315–3327, 2016.

[71] Ann John, Michael Dennis, L Kosnes, David Gunnell, J Scourfield, DV Ford, and K Lloyd. Suicide information database-cymru: a protocol for a population-based,

routinely collected data linkage study to explore risks and patterns of healthcare contact prior to suicide to identify opportunities for intervention. *BMJ open*, 4 (11):e006780, 2014.

[72] Bethan Bowden, Ann John, Laszlo Trefan, Jennifer Morgan, Daniel Farewell, and David Fone. Risk of suicide following an alcohol-related emergency hospital admission: An electronic cohort study of 2.8 million people. *PloS one*, 13(4): e0194772, 2018.

[73] David V Ford, Kerina H Jones, Rod M Middleton, Hazel Lockhart-Jones, Inocencio DC Maramba, Gareth J Noble, Lisa A Osborne, and Ronan A Lyons. The feasibility of collecting information from people with multiple sclerosis for the uk ms register via a web portal: characterising a cohort of people with ms. *BMC medical informatics and decision making*, 12(1):73, 2012.

[74] Kerina H Jones, David V Ford, Philip A Jones, Ann John, Rodden M Middleton, Hazel Lockhart-Jones, Lisa A Osborne, and J Gareth Noble. A large-scale study of anxiety and depression in people with multiple sclerosis: a survey via the web portal of the uk ms register. *PLoS One*, 7(7):e41910, 2012.

[75] Kerina H Jones, Philip A Jones, Rodden M Middleton, David V Ford, Katie Tuite-Dalton, Hazel Lockhart-Jones, Jeffrey Peng, Ronan A Lyons, Ann John, and J Gareth Noble. Physical disability, anxiety and depression in people with ms: an internet-based survey via the uk ms register. *PLoS One*, 9(8):e104604, 2014.

[76] Kerina H Jones, David V Ford, Philip A Jones, Ann John, Rodden M Middleton, Hazel Lockhart-Jones, Jeffrey Peng, Lisa A Osborne, and J Gareth Noble. How people with multiple sclerosis rate their quality of life: an eq-5d survey via the uk ms register. *PLoS One*, 8(6):e65640, 2013.

[77] Kelly Morgan, Mohammed Rahman, Mark Atkinson, Shang-Ming Zhou, Rebecca Hill, Ashrafunnesa Khanom, Shantini Paranjothy, and Sinead Brophy. Association of diabetes in pregnancy with child weight at birth, age 12 months and 5 years–a population-based electronic cohort study. *PloS one*, 8(11):e79803, 2013.

[78] Adrian Sayers, Daniel Thayer, John N Harvey, Stephen Luzio, Mark D Atkinson, Robert French, Justin T Warner, Colin M Dayan, Susan F Wong, and John W Gregory. Evidence for a persistent, major excess in all cause admissions to hospital in children with type-1 diabetes: results from a large welsh national matched community cohort study. *BMJ open*, 5(4):e005644, 2015.

[79] R Robson, AS Lacey, SD Luzio, Hugo Van Woerden, ML Heaven, M Wani, JPJ Halcox, L Castilla-Guerra, J Dawson, and Jonathan Hewitt. Hba1c measurement and relationship to incident stroke. *Diabetic Medicine*, 33(4):459–462, 2016.

[80] Kymberley Thorne, John G Williams, Ashley Akbari, and Stephen E Roberts. The impact of social deprivation on mortality following acute myocardial infarction, stroke or subarachnoid haemorrhage: A record linkage study. *BMC cardiovascular disorders*, 15(1):71, 2015.

[81] Majd B Protty, Arron Lacey, Jamie Hayes, and Phillip Freeman. Statins for secondary prevention: clinical use in patients with acute coronary syndrome in wales. *Future cardiology*, 13(2):137–141, 2017.

[82] Majd Protty, Phillip Freeman, Omar Aldalati, Arron Lacey, William King, Richard Anderson, and Dave Smith. Severe symptomatic aortic stenosis: Medical treatment versus transcatheter aortic valve implantation: A real world analysis of admission profiles, cost, and mortality using the secure anonymised information linkage (sail) databank. *Journal of the American College of Cardiology*, 65(10 Supplement):A1867, 2015.

[83] William King, Arron Lacey, James White, Daniel Farewell, Frank Dunstan, and David Fone. Equity in healthcare for coronary heart disease, wales (uk) 2004–2010: A population-based electronic cohort study. *PloS one*, 12(3): e0172618, 2017.

[84] William Owen Pickrell, Arron S Lacey, Rhys H Thomas, Philip EM Smith, and Mark I Rees. Weight change associated with antiepileptic drugs. *Journal of Neurology, Neurosurgery & Psychiatry*, pages jnnp–2012, 2012.

[85] Hayley C Gorton, Roger T Webb, Matthew J Carr, Marcos DelPozo-Banos, Ann John, and Darren M Ashcroft. Risk of unnatural mortality in people with epilepsy. *JAMA neurology*, 2018.

[86] Procedures as a representation of data in a computer program for understanding natural language.

[87] Wendy G Lehnert. The process of question answering. Technical report, DTIC Document, 1977.

[88] Richard Cullingford. Sam. In *Readings in natural language processing*, pages 627–649. Morgan Kaufmann Publishers Inc., 1986.

[89] Lalit R Bahl, Peter F Brown, Peter V de Souza, and Robert L Mercer. A

tree-based statistical language model for natural language speech recognition. In *Readings in Speech Recognition*, pages 507–514. Elsevier, 1990.

[90] Henry Kučera and Winthrop Nelson Francis. *Computational analysis of present-day American English*. Dartmouth Publishing Group, 1967.

[91] John M Sinclair. *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary*. Collins Elt, 1987.

[92] Barbara B Greene and Gerald M Rubin. Automated grammatical tagging of english. 1971.

[93] Stig Johansson. The lob corpus of british english texts: presentation and comments. *ALLC journal*, 1(1):25–36, 1980.

[94] Kenneth Ward Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the second conference on Applied natural language processing*, pages 136–143. Association for Computational Linguistics, 1988.

[95] Steven J DeRose. Grammatical category disambiguation by statistical optimization. *Computational linguistics*, 14(1):31–39, 1988.

[96] Julian Kupiec. Robust part-of-speech tagging using a hidden markov model. *Computer Speech & Language*, 6(3):225–242, 1992.

[97] Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing*, pages 152–155. Association for Computational Linguistics, 1992.

[98] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

[99] Stefano Federici, Simonetta Montemagni, and Vito Pirrelli. Shallow parsing and text chunking: a view on underspecification in syntax. *Cognitive science research paper-university of Sussex CSRP*, pages 35–44, 1996.

[100] Lance A Ramshaw and Mitchell P Marcus. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer, 1999.

[101] Walter Daelemans, Sabine Buchholz, and Jorn Veenstra. Memory-based shallow parsing. *arXiv preprint cs/9906005*, 1999.

[102] Tong Zhang, Fred Damerau, and David Johnson. Text chunking using regularized winnow. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 539–546. Association for Computational Linguistics, 2001.

[103] Dayne Freitag and Andrew McCallum. Information extraction with hmm structures learned by stochastic optimization. *AAAI/IAAI*, 2000:584–589, 2000.

[104] Daniel M Bikel, Richard Schwartz, and Ralph M Weischedel. An algorithm that learns what's in a name. *Machine learning*, 34(1-3):211–231, 1999.

[105] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[106] Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[107] Qinghua Zou, Wesley W Chu, Craig Morioka, Gregory H Leazer, and Hooshang Kangarloo. Indexfinder: a method of extracting key concepts from clinical texts for indexing. In *AMIA Annual Symposium Proceedings*, volume 2003, page 763. American Medical Informatics Association, 2003.

[108] Ralph Grishman and Beth Sundheim. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, volume 1, 1996.

[109] Mark A Przybocki, Jonathan G Fiscus, John S Garofolo, and David S Pallett. 1998 hub-4 information extraction evaluation. In *Proc. DARPA Broadcast News Workshop,(Herndon, Va, USA)*, pages 13–18, 1999.

[110] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In *LREC*, 2002.

[111] George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, volume 2, page 1, 2004.

[112] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[113] Steven Bird and Edward Loper. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics, 2004.

[114] David Ferrucci and Adam Lally. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348, 2004.

[115] Apache OpenNLP. Apache software foundation. *URL http://opennlp. apache. org*, 2011.

[116] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[117] Hamish Cunningham. Gate, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254, 2002.

[118] Wei-Dong Jackie Zhu, Bob Foyle, Daniel Gagné, Vijay Gupta, Josemina Magdalen, Amarjeet S Mundi, Tetsuya Nasukawa, Mark Paulis, Jane Singer, Martin Triska, et al. *IBM Watson Content Analytics: Discovering Actionable Insight from Your Content*. IBM Redbooks, 2014.

[119] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL `https://aclweb.org/anthology/D/D15/D15-1162`.

[120] I Feinerer and K Hornik. opennlp: opennlp interface. *R package version 0.0-8*, 2010.

[121] David Meyer, Kurt Hornik, and Ingo Feinerer. Text mining infrastructure in r. *Journal of statistical software*, 25(5):1–54, 2008.

[122] Peter Kluegl, Martin Toepfer, Philip-Daniel Beck, Georg Fette, and Frank Puppe. Uima ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering*, 22(1):1–40, 2016.

[123] R Rehurek and P Sojka. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2), 2011.

[124] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[125] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[126] James W Perry, Allen Kent, and Madeline M Berry. Machine literature searching x. machine language; factors underlying its design and development. *Journal of the Association for Information Science and Technology*, 6(4):242–254, 1955.

[127] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[128] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics, 2012.

[129] Brett R South, Shuying Shen, Jianwei Leng, Tyler B Forbush, Scott L DuVall, and Wendy W Chapman. A prototype tool set to support machine-assisted annotation. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 130–139. Association for Computational Linguistics, 2012.

[130] Naomi Sager, Carol Friedman, and Margaret S Lyman. *Medical language processing: computer management of narrative data*. Addison-Wesley Longman Publishing Co., Inc., 1987.

[131] Naomi Sager, Margaret Lyman, Ngo Thanh Nhan, and Leo J Tick. Medical language processing: applications to patient data representation and automatic encoding. *Methods of information in medicine*, 34(1-2):140–146, 1995.

[132] Lynette Hirschman, Naomi Sager, and Margaret Lyman. Automatic application of health care criteria to narrative patient records. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 105. American Medical Informatics Association, 1979.

[133] Carol Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. A general natural-language text processor for clinical

radiology. *Journal of the American Medical Informatics Association*, 1(2): 161–174, 1994.

[134] Carol Friedman, George Hripcsak, William DuMouchel, Stephen B Johnson, and Paul D Clayton. Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1(1):83–108, 1995.

[135] Nilesh L Jain and Carol Friedman. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. In *Proceedings of the AMIA Annual Fall Symposium*, page 829. American Medical Informatics Association, 1997.

[136] Carol Friedman, Pauline Kra, Hong Yu, Michael Krauthammer, and Andrey Rzhetsky. Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. In *ISMB (supplement of bioinformatics)*, pages 74–82, 2001.

[137] Marcelo Fiszman, Wendy W Chapman, Dominik Aronsky, R Scott Evans, and Peter J Haug. Automatic detection of acute bacterial pneumonia from chest x-ray reports. *Journal of the American Medical Informatics Association*, 7(6): 593–604, 2000.

[138] Peter Haug, Spence Koehler, Lee Min Lau, Ping Wang, Roberto Rocha, and Stan Huff. A natural language understanding system combining syntactic and semantic techniques. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 247. American Medical Informatics Association, 1994.

[139] Katherine P Liao, Tianxi Cai, Guergana K Savova, Shawn N Murphy, Elizabeth W Karlson, Ashwin N Ananthakrishnan, Vivian S Gainer, Stanley Y Shaw, Zongqi Xia, Peter Szolovits, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *bmj*, 350:h1885, 2015.

[140] Vincent Damotte and Pierre-Antoine Gourraud. Electronic medical records in multiple sclerosis research. *Clinical and Experimental Neuroimmunology*, 9(1): 13–18, 2018.

[141] Robert Bill, Serguei Pakhomov, Elizabeth S Chen, Tamara J Winden, Elizabeth W Carter, and Genevieve B Melton. Automated extraction of family history information from clinical notes. In *AMIA Annual Symposium Proceedings*, volume 2014, page 1709. American Medical Informatics Association, 2014.

[142] Lifeng Chen and Carol Friedman. Extracting phenotypic information from the literature via natural language processing. In *Medinfo*, pages 758–762, 2004.

[143] Licong Cui, Alireza Bozorgi, Samden D Lhatoo, Guo-Qiang Zhang, and Satya S Sahoo. Epidea: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1191. American Medical Informatics Association, 2012.

[144] Licong Cui, Satya S Sahoo, Samden D Lhatoo, Gaurav Garg, Prashant Rai, Alireza Bozorgi, and Guo-Qiang Zhang. Complex epilepsy phenotype extraction from narrative clinical discharge summaries. *Journal of biomedical informatics*, 51:272–279, 2014.

[145] Ryan Sullivan, Robert Yao, Randa Jarrar, Jeffrey Buchhalter, and Graciela Gonzalez. Text classification towards detecting misdiagnosis of an epilepsy syndrome in a pediatric population. In *AMIA Annual Symposium Proceedings*, volume 2014, page 1082. American Medical Informatics Association, 2014.

[146] George Karystianis, Therese Sheppard, William G Dixon, and Goran Nenadic. Modelling and extraction of variability in free-text medication prescriptions from an anonymised primary care electronic medical record research database. *BMC medical informatics and decision making*, 16(1):18, 2015.

[147] Stuart McTaggart, Clifford Nangle, Jacqueline Caldwell, Samantha Alvarez-Madrazo, Helen Colhoun, and Marion Bennie. Use of text-mining methods to improve efficiency in the calculation of drug exposure to support pharmacoepidemiology studies. *International journal of epidemiology*, 47(2): 617–624, 2018.

[148] Peter L Elkin, James J Cimino, Henry J Lowe, David B Aronow, Tom H Payne, Pierre S Pincetl, and G Octo Barnett. Mapping to mesh: The art of trapping mesh equivalence from within narrative text. In *Proceedings of the annual symposium on computer application in medical care*, page 185. American Medical Informatics Association, 1988.

[149] Adam Wilcox, George Hripcsak, and Carol Friedman. Using knowledge sources to improve classification of medical text reports. In *Proceedings of Workshop on Text Mining KDD-2000*, 2000.

[150] Randolph A Miller, Filip M Gieszczykiewicz, John K Vries, and Gregory F Cooper. Chartline: providing bibliographic references relevant to patient charts using the umls metathesaurus knowledge sources. In *Proceedings of the Annual*

Symposium on Computer Application in Medical Care, page 86. American Medical Informatics Association, 1992.

[151] William R Hersh, David H Hickam, R Brian Haynes, and K Ann McKibbon. A performance and failure analysis of saphire with a medline test collection. *Journal of the American Medical Informatics Association*, 1(1):51–60, 1994.

[152] Prakash Nadkarni, Roland Chen, and Cynthia Brandt. Umls concept indexing for production databases: a feasibility study. *Journal of the American Medical Informatics Association*, 8(1):80–91, 2001.

[153] Michele Cargill, David Altshuler, James Ireland, Pamela Sklar, Kristin Ardlie, Nila Patil, Charles R Lane, Esther P Lim, Nilesh Kalyanaraman, James Nemesh, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature genetics*, 22(3):231, 1999.

[154] Francis S Collins, Lisa D Brooks, and Aravinda Chakravarti. A dna polymorphism discovery resource for research on human genetic variation. *Genome research*, 8(12):1229–1231, 1998.

[155] Wen-Hsiung Li and Lori A Sadler. Low nucleotide diversity in man. *Genetics*, 129(2):513–523, 1991.

[156] 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[157] Kelly A Frazer, Sarah S Murray, Nicholas J Schork, and Eric J Topol. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241–251, 2009.

[158] Gabor Marth, Raymond Yeh, Matthew Minton, Rachel Donaldson, Qun Li, Shenghui Duan, Ruth Davenport, Raymond D Miller, and Pui-Yan Kwok. Single-nucleotide polymorphisms in the public domain: how useful are they? *Nature genetics*, 27(4):371, 2001.

[159] Elizabeth Pennisi. Encode project writes eulogy for junk dna, 2012.

[160] Michael N Weedon, Inês Cebola, Ann-Marie Patch, Sarah E Flanagan, Elisa De Franco, Richard Caswell, Santiago A Rodríguez-Seguí, Charles Shaw-Smith, Candy HH Cho, Hana Lango Allen, et al. Recessive mutations in a distal ptf1a enhancer cause isolated pancreatic agenesis. *Nature genetics*, 46(1):61, 2014.

[161] Kaiyu Jiang, Lisha Zhu, Michael J Buck, Yanmin Chen, Bradley Carrier, Tao Liu, and James N Jarvis. Disease-associated single-nucleotide polymorphisms

from noncoding regions in juvenile idiopathic arthritis are located within or adjacent to functional genomic elements of human neutrophils and cd4+ t cells. *Arthritis & Rheumatology*, 67(7):1966–1977, 2015.

[162] Kyle Kai-How Farh, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J Housley, Samantha Beik, Noam Shoresh, Holly Whitton, Russell JH Ryan, Alexander A Shishkin, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337, 2015.

[163] Michael Krawczak, Edward V Ball, Iain Fenton, Peter D Stenson, Shaun Abeysinghe, Nick Thomas, and David N Cooper. Human gene mutation database–a biomedical information and research resource. *Human mutation*, 15 (1):45, 2000.

[164] C Sue Richards, Sherri Bale, Daniel B Bellissimo, Soma Das, Wayne W Grody, Madhuri R Hegde, Elaine Lyon, and Brian E Ward. Acmg recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genetics in Medicine*, 10(4):294, 2008.

[165] Callum J Bell, Darrell L Dinwiddie, Neil A Miller, Shannon L Hateley, Elena E Ganusova, Joann Mudge, Ray J Langley, Lu Zhang, Clarence C Lee, Faye D Schilkey, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Science translational medicine*, 3(65):65ra4–65ra4, 2011.

[166] DG MacArthur, TA Manolio, DP Dimmock, HL Rehm, J Shendure, GR Abecasis, DR Adams, RB Altman, SE Antonarakis, EA Ashley, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497):469, 2014.

[167] Bat-sheva Kerem, Johanna M Rommens, Janet A Buchanan, Danuta Markiewicz, Tara K Cox, Aravinda Chakravarti, Manuel Buchwald, and Lap-Chee Tsui. Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245(4922):1073–1080, 1989.

[168] James F Gusella, Nancy S Wexler, P Michael Conneally, Susan L Naylor, Mary Anne Anderson, Rudolph E Tanzi, Paul C Watkins, Kathleen Ottina, Margaret R Wallace, Alan Y Sakaguchi, et al. A polymorphic dna marker genetically linked to huntington's disease. *Nature*, 306(5940):234, 1983.

[169] Anna Sillén, Jorge Andrade, Lena Lilius, Charlotte Forsell, Karin Axelman, Jacob Odeberg, Bengt Winblad, and Caroline Graff. Expanded high-resolution

genetic study of 109 swedish families with alzheimer's disease. *European Journal of Human Genetics*, 16(2):202, 2008.

[170] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in medicine*, 17 (5):405, 2015.

[171] Elizabeth M Smigielski, Karl Sirotkin, Minghong Ward, and Stephen T Sherry. dbsnp: a database of single nucleotide polymorphisms. *Nucleic acids research*, 28(1):352–355, 2000.

[172] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467, 1977.

[173] Robert D Fleischmann, Mark D Adams, Owen White, Rebecca A Clayton, Ewen F Kirkness, Anthony R Kerlavage, Carol J Bult, Jean-Francois Tomb, Brian A Dougherty, Joseph M Merrick, et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512, 1995.

[174] André Goffeau, Bart G Barrell, Howard Bussey, RW Davis, Bernard Dujon, Heinz Feldmann, Francis Galibert, JD Hoheisel, Cr Jacq, Michael Johnston, et al. Life with 6000 genes. *Science*, 274(5287):546–567, 1996.

[175] The C. elegans Sequencing Consortium. Genome sequence of the nematode c. elegans: a platform for investigating biology. *Science*, pages 2012–2018, 1998.

[176] Mark D Adams, Susan E Celniker, Robert A Holt, Cheryl A Evans, Jeannine D Gocayne, Peter G Amanatides, Steven E Scherer, Peter W Li, Roger A Hoskins, Richard F Galle, et al. The genome sequence of drosophila melanogaster. *Science*, 287(5461):2185–2195, 2000.

[177] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507): 1304–1351, 2001.

[178] Nayanah Siva. 1000 genomes project, 2008.

[179] Genomics England. The 100,000 genomes project. *The*, 100:0–2, 2016.

[180] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. 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, 2015.

[181] Konrad J Karczewski, Ben Weisburd, Brett Thomas, Matthew Solomonson, Douglas M Ruderfer, David Kavanagh, Tymor Hamamsy, Monkol Lek, Kaitlin E Samocha, Beryl B Cummings, et al. The exac browser: displaying reference data information from over 60 000 exomes. *Nucleic acids research*, 45(D1): D840–D845, 2016.

[182] Elaine R Mardis. The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3):133–141, 2008.

[183] David A Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G Thomas Roth, et al. The complete genome of an individual by massively parallel dna sequencing. *nature*, 452(7189):872, 2008.

[184] Shamil Sunyaev, Vasily Ramensky, and Peer Bork. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends in Genetics*, 16(5):198–200, 2000.

[185] Carles Ferrer-Costa, Modesto Orozco, and Xavier de la Cruz. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties1. *Journal of molecular biology*, 315(4):771–786, 2002.

[186] Pauline C Ng and Steven Henikoff. Predicting deleterious amino acid substitutions. *Genome research*, 11(5):863–874, 2001.

[187] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003.

[188] James Ulrich Bowie and Robert T Sauer. Identifying determinants of folding and activity for a protein of unknown structure. *Proceedings of the National Academy of Sciences*, 86(7):2152–2156, 1989.

[189] Alyssa Bentley, Bridget MacLennan, Jonathan Calvo, and Charles R Dearolf. Targeted recovery of mutations in drosophila. *Genetics*, 156(3):1169–1173, 2000.

[190] Yijing Chen, Della Yee, Katherine Dains, Aurobindo Chatterjee, James Cavalcoli, Elizabeth Schneider, Jinsop Om, Richard P Woychik, and Terry Magnuson. Genotype-based screen for enu-induced mutations in mouse embryonic stem cells. *Nature genetics*, 24(3):314, 2000.

[191] Peter Markiewicz, Lynn G Kleina, Christina Cruz, Susannah Ehret, and Jeffrey H Miller. Genetic studies of the lac repressor. xiv. analysis of 4000 altered escherichia coli lac repressors reveals essential and non-essential residues, as well as" spacers" which do not require a specific sequence. *Journal of molecular biology*, 240(5):421–433, 1994.

[192] Jörg Suckow, Peter Markiewicz, Lynn G Kleina, Jeffrey Miller, Brigitte Kisters-Woike, and Benno Müller-Hill. Genetic studies of the lac repressor xv: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *Journal of molecular biology*, 261(4):509–523, 1996.

[193] Daniel D Loeb, Ronald Swanstrom, Lorraine Everitt, Marianne Manchester, Susan E Stamper, and Clyde A Hutchison III. Complete mutagenesis of the hiv-1 protease. *Nature*, 340(6232):397, 1989.

[194] Dale Rennell, Suzanne E Bouvier, Larry W Hardy, and Anthony R Poteete. Systematic mutation of bacteriophage t4 lysozyme. *Journal of molecular biology*, 222(1):67–88, 1991.

[195] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[196] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4): 248, 2010.

[197] Shamil R Sunyaev, Frank Eisenhaber, Igor V Rodchenkov, Birgit Eisenhaber, Vladimir G Tumanyan, and Eugene N Kuznetsov. Psic: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein engineering*, 12(5):387–394, 1999.

[198] Anders Krogh, BjoÈrn Larsson, Gunnar Von Heijne, and Erik LL Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes1. *Journal of molecular biology*, 305(3):567–580, 2001.

[199] Andrei Lupas, Marc Van Dyke, and Jeff Stock. Predicting coiled coils from protein sequences. *Science*, pages 1162–1164, 1991.

[200] Henrik Nielsen, Jacob Engelbrecht, Søren Brunak, and Gunnar von Heijne. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein engineering*, 10(1):1–6, 1997.

[201] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[202] Helen M Berman, Tammy Battistuz, Talapady N Bhat, Wolfgang F Bluhm, Philip E Bourne, Kyle Burkhardt, Zukang Feng, Gary L Gilliland, Lisa Iype, Shri Jain, et al. The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, 58(6):899–907, 2002.

[203] Vasily Ramensky, Peer Bork, and Shamil Sunyaev. Human non-synonymous snps: server and survey. *Nucleic acids research*, 30(17):3894–3900, 2002.

[204] Daniel Quang, Yifei Chen, and Xiaohui Xie. Dann: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, 31(5): 761–763, 2014.

[205] Chengliang Dong, Peng Wei, Xueqiu Jian, Richard Gibbs, Eric Boerwinkle, Kai Wang, and Xiaoming Liu. Comparison and integration of deleteriousness prediction methods for nonsynonymous snvs in whole exome sequencing studies. *Human molecular genetics*, 24(8):2125–2137, 2014.

[206] Boris Reva, Yevgeniy Antipin, and Chris Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, 39(17):e118–e118, 2011.

[207] Nilah M Ioannidis, Joseph H Rothstein, Vikas Pejaver, Sumit Middha, Shannon K McDonnell, Saurabh Baheti, Anthony Musolf, Qing Li, Emily Holzinger, Danielle Karyadi, et al. Revel: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics*, 99(4):877–885, 2016.

[208] Hashem A Shihab, Julian Gough, David N Cooper, Peter D Stenson, Gary LA Barker, Keith J Edwards, Ian NM Day, and Tom R Gaunt. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models. *Human mutation*, 34(1):57–65, 2013.

[209] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310, 2014.

[210] Hannah Carter, Christopher Douville, Peter D Stenson, David N Cooper, and Rachel Karchin. Identifying mendelian disease genes with the variant effect scoring tool. *BMC genomics*, 14(3):S3, 2013.

[211] Emidio Capriotti and Piero Fariselli. Phd-snpg: a webserver and lightweight tool for scoring single nucleotide variants. *Nucleic acids research*, 45(W1): W247–W252, 2017.

[212] Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, 42(D1):D980–D985, 2013.

[213] D Fredman, Marianne Siegfried, Yan P. Yuan, Peer Bork, Heikki Lehväslaiho, and Anthony J Brookes. Hgvbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic acids research*, 30(1):387–391, 2002.

[214] Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl 1): D514–D517, 2005.

[215] Kevin G Becker, Kathleen C Barnes, Tiffani J Bright, and S Alex Wang. The genetic association database. *Nature genetics*, 36(5):431–432, 2004.

[216] David N Cooper, Edward V Ball, and Michael Krawczak. The human gene mutation database. *Nucleic acids research*, 26(1):285–287, 1998.

[217] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215 (3):403–410, 1990.

[218] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[219] Julie D Thompson, Desmond G Higgins, and Toby J Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22):4673–4680, 1994.

[220] Robert C Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 2010.

[221] Robert D Finn, Jody Clements, and Sean R Eddy. Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, 39(suppl_2): W29–W37, 2011.

[222] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1):110–121, 2010.

[223] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050, 2005.

[224] Manuel Garber, Mitchell Guttman, Michele Clamp, Michael C Zody, Nir Friedman, and Xiaohui Xie. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, 25(12):i54–i62, 2009.

[225] Liam J McGuffin, Kevin Bryson, and David T Jones. The psipred protein structure prediction server. *Bioinformatics*, 16(4):404–405, 2000.

[226] Shide Liang, Chi Zhang, Song Liu, and Yaoqi Zhou. Protein binding site prediction using an empirical scoring function. *Nucleic acids research*, 34(13): 3698–3707, 2006.

[227] Alexei Vazquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Global protein function prediction from protein-protein interaction networks. *Nature biotechnology*, 21(6):697, 2003.

[228] Xiaoming Liu, Xueqiu Jian, and Eric Boerwinkle. dbnsfp: a lightweight database of human nonsynonymous snps and their functional predictions. *Human mutation*, 32(8):894–899, 2011.

[229] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome biology*, 17(1):122, 2016.

[230] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, 2012.

[231] Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation

of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

[232] Henk Harkema, John N Dowling, Tyler Thornblade, and Wendy W Chapman. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics*, 42(5):839–851, 2009.

[233] John Moult, Jan T Pedersen, Richard Judson, and Krzysztof Fidelis. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 23(3), 1995.

[234] David T Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, 292(2):195–202, 1999.

[235] William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

[236] Michael Tan, Ian Wilson, Vanessa Braganza, Sophia Ignatiadis, Ray Boston, Vijaya Sundararajan, Mark J Cook, and Wendyl J D'Souza. Development and validation of an epidemiologic case definition of epilepsy for use with routinely collected australian health data. *Epilepsy & Behavior*, 51:65–72, 2015.

[237] C Franchi, G Giussani, P Messina, M Montesano, S Romi, A Nobili, I Fortino, A Bortolotti, L Merlino, E Beghi, et al. Validation of healthcare administrative data for the diagnosis of epilepsy. *J Epidemiol Community Health*, pages jech–2013, 2013.

[238] E Wayne Holden, Elizabeth Grossman, Hoang Thanh Nguyen, Margaret J Gunter, Becky Grebosky, Ann Von Worley, Leila Nelson, Scott Robinson, and David J Thurman. Developing a computer algorithm to identify epilepsy cases in managed care organizations. *Disease Management*, 8(1):1–14, 2005.

[239] Wilhelmine Hadler Meeraus, Irene Petersen, Richard Frank Chin, Felicity Knott, and Ruth Gilbert. Childhood epilepsy recorded in primary care in the uk. *Archives of disease in childhood*, 98(3):195–202, 2013.

[240] Vanessa Delgado Nunes, Laura Sawyer, Julie Neilson, Grammati Sarri, and J Helen Cross. Diagnosis and management of the epilepsies in adults and children: summary of updated nice guidance. *Bmj*, 344:e281, 2012.

[241] Rebecca Louise Bromley, George E Mawer, Maria Briggs, Christopher Cheyne, Jill Clayton-Smith, Marta García-Fiñana, Rachel Kneen, Sam B Lucas, Rebekah Shallcross, Gus A Baker, et al. The prevalence of neurodevelopmental disorders

in children prenatally exposed to antiepileptic drugs. *J Neurol Neurosurg Psychiatry*, 84(6):637–643, 2013.

[242] Karl Titze, Sabine Koch, Hans Helge, Ulrike Lehmkuhl, Hellgard Rauh, and Hans-Christoph Steinhausen. Prenatal and family risks of children born to mothers with epilepsy: effects on cognitive development. *Developmental Medicine & Child Neurology*, 50(2):117–122, 2008.

[243] Kimford J Meador, Gus A Baker, Nancy Browning, Morris J Cohen, Rebecca L Bromley, Jill Clayton-Smith, Laura A Kalayjian, Andres Kanner, Joyce D Liporace, Page B Pennell, et al. Fetal antiepileptic drug exposure and cognitive outcomes at age 6 years (nead study): a prospective observational study. *The Lancet Neurology*, 12(3):244–252, 2013.

[244] Giouliana Kadra, Robert Stewart, Hitesh Shetty, Richard G Jackson, Mark A Greenwood, Angus Roberts, Chin-Kuo Chang, James H MacCabe, and Richard D Hayes. Extracting antipsychotic polypharmacy data from electronic health records: developing and evaluating a novel process. *BMC psychiatry*, 15 (1):166, 2015.

[245] Brad Gulko, Ilan Gronau, Melissa J Hubisz, and Adam Siepel. Probabilities of fitness consequences for point mutations across the human genome. *bioRxiv*, page 006825, 2014.

[246] Qiongshi Lu, Yiming Hu, Jiehuan Sun, Yuwei Cheng, Kei-Hoi Cheung, and Hongyu Zhao. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Scientific reports*, 5:10576, 2015.

[247] Iuliana Ionita-Laza, Kenneth McCallum, Bin Xu, and Joseph D Buxbaum. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature genetics*, 48(2):214, 2016.

[248] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56, 2012.

[249] Yongwook Choi and Agnes P Chan. Provean web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, 31(16): 2745–2747, 2015.

[250] Inna Dubchak, Ilya Muchnik, Christopher Mayor, Igor Dralyuk, and Sung-Hou Kim. Recognition of a protein fold in the context of the scop classification. *Proteins: Structure, Function, and Bioinformatics*, 35(4):401–407, 1999.

[251] Jenny Eachus, Mark Williams, Philip Chan, George Davey Smith, Matthew Grainge, Jenny Donovan, and Stephen Frankel. Deprivation and cause specific morbidity: evidence from the somerset and avon survey of health. *Bmj*, 312 (7026):287–292, 1996.

[252] Walter F Stewart, Richard B Lipton, David D Celentano, and Michael L Reed. Prevalence of migraine headache in the united states: relation to age, income, race, and other sociodemographic factors. *Jama*, 267(1):64–69, 1992.

[253] Anthony K Ngugi, SM Kariuki, C Bottomley, I Kleinschmidt, JW Sander, and CR Newton. Incidence of epilepsy a systematic review and meta-analysis. *Neurology*, 77(10):1005–1012, 2011.

[254] Mohammad Reza Mohammadi, Ahmad Ghanizadeh, Haratoun Davidian, Mohammad Mohammadi, and Maryam Norouzian. Prevalence of epilepsy and comorbidity of psychiatric disorders in iran. *Seizure-European Journal of Epilepsy*, 15(7):476–482, 2006.

[255] Morris J Cohen, Kimford J Meador, Nancy Browning, Gus A Baker, Jill Clayton-Smith, Laura A Kalayjian, Andres Kanner, Joyce D Liporace, Page B Pennell, Michael Privitera, et al. Fetal antiepileptic drug exposure: motor, adaptive, and emotional/behavioral functioning at age 3 years. *Epilepsy & Behavior*, 22(2):240–246, 2011.

[256] Rebecca L Bromley, George Mawer, Jenna Love, James Kelly, Laura Purdy, Lauren McEwan, Maria Briggs, Jill Clayton Smith, Xin Sin, and Gus A Baker. Early cognitive development in children born to women with epilepsy: a prospective report. *Epilepsia*, 51(10):2058–2065, 2010.

[257] Kimford J Meador, Gus A Baker, Nancy Browning, Morris J Cohen, Jill Clayton-Smith, Laura A Kalayjian, Andres Kanner, Joyce D Liporace, Page B Pennell, Michael Privitera, et al. Foetal antiepileptic drug exposure and verbal versus non-verbal abilities at three years of age. *Brain*, 134(2):396–404, 2011.

[258] Rebecca L Bromley, Rebecca Calderbank, Christopher P Cheyne, Claire Rooney, Penny Trayner, Jill Clayton-Smith, Marta García-Fiñana, Beth Irwin, James Irvine Morrow, Rebekah Shallcross, et al. Cognition in school-age children exposed to levetiracetam, topiramate, or sodium valproate. *Neurology*, 87(18): 1943–1953, 2016.

[259] Eija Gaily, Elisa Kantola-Sorsa, Vilho Hiilesmaa, M Isoaho, Riitta Matila, Mervi Kotila, T Nylund, A Bardy, E Kaaja, and M-L Granström. Normal

intelligence in children with prenatal exposure to carbamazepine. *Neurology*, 62 (1):28–32, 2004.

[260] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. Clinical information extraction applications: A literature review. *Journal of biomedical informatics*, 2017.

[261] Honghan Wu, Giulia Toti, Katherine I Morley, Zina Ibrahim, Amos Folarin, Ismail Kartoglu, Richard Jackson, Asha Agrawal, Clive Stringer, Darren Gale, et al. Semehr: surfacing semantic data from clinical notes in electronic health records for tailored care, trial recruitment, and clinical research. *The Lancet*, 390:S97, 2017.

[262] Ray R Larson. Introduction to information retrieval. *Journal of the American Society for Information Science and Technology*, 61(4):852–853, 2010.

[263] Deepak Ravichandran, Patrick Pantel, and Eduard Hovy. Randomized algorithms and nlp: using locality sensitive hash function for high speed noun clustering. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 622–629. Association for Computational Linguistics, 2005.

[264] Lauren C Walters-Sen, Sayaka Hashimoto, Devon Lamb Thrush, Shalini Reshmi, Julie M Gastier-Foster, Caroline Astbury, and Robert E Pyatt. Variability in pathogenicity prediction programs: impact on clinical diagnostics. *Molecular genetics & genomic medicine*, 3(2):99–110, 2015.

[265] Lisa A Miosge, Matthew A Field, Yovina Sontani, Vicky Cho, Simon Johnson, Anna Palkova, Bhavani Balakishnan, Rong Liang, Yafei Zhang, Stephen Lyon, et al. Comparison of predicted and actual consequences of missense mutations. *Proceedings of the National Academy of Sciences*, 112(37):E5189–E5198, 2015.

[266] Janita Thusberg, Ayodeji Olatubosun, and Mauno Vihinen. Performance of mutation pathogenicity prediction methods on missense variants. *Human mutation*, 32(4):358–368, 2011.

[267] Emidio Capriotti and Russ B Altman. Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC bioinformatics*, 12(Suppl 4):S3, 2011.

[268] Christoph Lossin. A catalog of scn1a variants. *Brain and Development*, 31(2): 114–130, 2009.

[269] Margarida C Lopes, Chris Joyce, Graham RS Ritchie, Sally L John, Fiona Cunningham, Jennifer Asimit, and Eleftheria Zeggini. A combined functional annotation score for non-synonymous variants. *Human heredity*, 73(1):47–51, 2012.