THE UNIVERSITY OF SHEFFIELD

DOCTORAL THESIS

# Improving diagnostic procedures for epilepsy through automated recording and analysis of patients' history

*Author:*
Nathan Pevy

*Supervisor:*
Professor Markus REUBER
Professor Heidi CHRISTENSEN
Dr Traci WALKER

*A thesis submitted in fulfillment of the requirements*
*for the degree of Doctor of Philosophy*

*in the*

Department of Neuroscience

April 14, 2023

# *Abstract*

Doctor of Philosophy

**Improving diagnostic procedures for epilepsy through automated recording and analysis of patients' history**

by Nathan Pevy

Transient loss of consciousness (TLOC) is a time-limited state of profound cognitive impairment characterised by amnesia, abnormal motor control, loss of responsiveness, a short duration and complete recovery. Most instances of TLOC are caused by one of three health conditions: epilepsy, functional (dissociative) seizures (FDS), or syncope. There is often a delay before the correct diagnosis is made and 10-20% of individuals initially receive an incorrect diagnosis. Clinical decision tools based on the endorsement of TLOC symptom lists have been limited to distinguishing between two causes of TLOC. The Initial Paroxysmal Event Profile (iPEP) has shown promise but was demonstrated to have greater accuracy in distinguishing between syncope and epilepsy or FDS than between epilepsy and FDS. The objective of this thesis was to investigate whether interactional, linguistic, and communicative differences in how people with epilepsy and people with FDS describe their experiences of TLOC can improve the predictive performance of the iPEP. An online web application was designed that collected information about TLOC symptoms and medical history from patients and witnesses using a binary questionnaire and verbal interaction with a virtual agent. We explored potential methods of automatically detecting these communicative differences, whether the differences were present during an interaction with a VA, to what extent these automatically detectable communicative differences improve the performance of the iPEP, and the acceptability of the application from the perspective of patients and witnesses. The two feature sets that were applied to previous doctor-patient interactions, features designed to measure formulation effort or detect semantic differences between the two groups, were able to predict the diagnosis with an accuracy of 71% and 81%, respectively. Individuals with epilepsy or FDS provided descriptions of TLOC to the VA that were qualitatively like those observed in previous research. Both feature sets were effective predictors of the diagnosis when applied to the web application recordings (85.7% and 85.7%). Overall, the accuracy of machine learning models trained for the three-way classification between epilepsy, FDS, and syncope using the iPEP responses from patients that were collected through the web application was worse than the performance observed in previous research (65.8% vs 78.3%), but the performance was increased by the inclusion of features extracted from the spoken descriptions on TLOC (85.5%). Finally, most participants who provided feedback reported that the online application was acceptable. These findings suggest that it is feasible to differentiate between people with epilepsy and people with FDS using an automated analysis of spoken seizure descriptions. Furthermore, incorporating these features into a clinical decision tool for TLOC can improve the predictive performance by improving the differential diagnosis between these two health conditions. Future research should use the feedback to improve the design of the application and increase perceived acceptability of the approach.

# *Acknowledgements*

I would like to thank my supervisors Professor Markus Reuber, Professor Heidi Christensen, and Dr Traci Walker for their continued support and guidance throughout my PhD. I am incredibly grateful to Epilepsy Research UK for funding my PhD project and providing me with this opportunity.

I appreciate the Speech and Hearing Group and Department of Computer Science for welcoming me into their team and for their technical guidance. In particular, I would like to thank Dr Bahman Mirheidari for his valued guidance. Furthermore, I appreciate the members of the Conversation Analysis Data Sessions in the Division of Human Communication Sciences for their insight and the provoking discussions about human communication.

I am grateful to my academic advisors, Dr Haris Stavroulakis and Dr Dan Blackburn, and my personal tutor, Dr Lynne Prince, for their academic guidance and support.

I would like to thank my family for their continued support and encouragement.

Last but not least, I would like to thank Epilepsy Research UK for funding my PhD.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| **TLOC** | Transient Loss of Consciousness |
| **EEG** | Electroencephalography |
| **ECG** | Electrocardiogram |
| **MRI** | Magnetic Resonance Imaging |
| **CA** | Conversation Analysis |
| **PEP** | Paroxysmal Event Profile |
| **PEO** | Paroxysmal Event Observer |
| **iPEP** | initial Paroxysmal Event Profile |
| **LPCC** | Linear Prediction Cepstrum Coefficients |
| **MFCC** | Mel-Frequency Cepstrum Coefficients |
| **GFCC** | Gammatone Frequency Cepstrum Coefficients |
| **LOOCV** | Leave one out cross validation |
| **SVM** | Support Vector Machine |
| **KNN** | K-Nearest Neigbour |
| **TP** | True Positive |
| **TN** | True Negative |
| **FP** | False Positive |
| **FN** | False Negative |
| **XML** | Extensible Markup Language |
| **SD** | Standard Deviation |
| **LIWC** | Linguistic Inquiry and Word Count |
| **TRP** | Transition relevant place |
| **TAM** | Technology Acceptance Model |

# Chapter 1

# Introduction

## 1.1 Transient Loss of Consciousness

Transient loss of consciousness (TLOC) is defined as a loss of awareness characterised by amnesia, abnormal motor control, loss of responsiveness and a short duration with a full recovery and no obvious cause (Brignole et al., 2018). Many individuals who experience TLOC are initially assessed in Primary and Emergency Care services, but the symptoms of TLOC have typically subsided before they are seen by a medical professional. Over 90% of presentations with TLOC are due to one of three health conditions: epilepsy, functional (dissociative) seizures (FDS) and syncope (Kotsopoulos et al., 2003). Although there are interictal tests that can indicate a potential cause of TLOC, these tests have low sensitivity and are specific to a single cause of TLOC. In most cases, these tests fail to provide a clear indication of the cause of TLOC (Malmgren, Reuber, and Appleton, 2012). Unfortunately, many people who visit Primary and Emergency Care Services do not receive a diagnosis (Briggs et al., 2017), and approximately 20% of individuals receive the wrong diagnosis (Xu et al., 2016). Individuals who receive the wrong initial diagnosis may be referred for tests that are irrelevant and delay the patient's access to appropriate medical treatment. Such delays and referrals to the incorrect specialist service is not only unpleasant for the patient and expensive for the NHS, but they can also be dangerous in cases where – if untreated - the cause of TLOC can endanger life.

### 1.1.1 Epilepsy

Epilepsy is a neurological health condition that causes individuals to experience unprovoked seizures. It is estimated that around 7.6 per 1,000 people will experience an epileptic seizure within their lifetime (Fiest et al., 2017). An epileptic seizure happens when pathological electrical activity in the brain increases beyond a "seizure threshold" because there is an imbalance between the level of excitation and inhibition in neurons causing synchronised neuronal network oscillations across different regions of the brain (Staley, 2015). There are a broad range of aetiological factors that can cause epileptic seizures, such as structural abnormalities, genetic abnormalities, infectious diseases, and metabolic or immune disorders (Scheffer et al., 2017). This means that the diagnostic process needs to extend beyond the differentiation of epilepsy from other attack disorders and encompass the identification of possible contributing causes which may require treatment in their own right. What is more, medical professions must consider the particular neuronal mechanisms in the brain that are giving rise to epileptic activity because this can influence their choice of optimal treatment (Shorvon, 2011).

These causes and neuronal mechanisms can determine the physical and psychological manifestations of the seizure because they influence the specific neural networks involved in the seizure and how widespread the seizure activity is across the brain. The International League Against Epilepsy published guidance on how to define an epileptic seizure based upon certain characteristics (Fisher, 2017). The classification of epileptic seizures is dependent on three important things. Firstly, the description of the seizure is influenced by whether the electrical activity originated in a single brain hemisphere (a focal seizure) or across both hemispheres (a generalised seizure). Secondly, focal seizures can be further classified based upon whether the patient retained awareness during the onset of the seizure or experienced an immediate impairment in awareness. Finally, epileptic seizures can be characterised based on the presence and description of motor or non-motor activity that occurred during the onset of the seizure, which patients may or may not be aware of. Examples of non-motor activity include unusual thoughts, feelings, and sensations that are caused by the spread of electrical activity during the seizure. These experiences are commonly known as auras and can help to identify where the epileptic seizure originated within the brain (Chowdhury et al., 2021; Bien et al., 2000). Identification of the type of seizure is made using information collected from the patient's history, any witness to the seizure, concurrent video and EEG recordings during seizures (video-EEG), and MRI scans (NICE, 2022). In addition to providing a consistent language that medical professionals can use to describe epileptic seizures, these important distinctions highlight the variability in the presentations of epilepsy. The diversity in the type of seizures people can experience overlaps with the clinical presentation of other health conditions and demonstrates the diagnostic challenges associated with TLOC.

### 1.1.2 Functional (Dissociative) Seizures

Functional (Dissociative) Seizures (FDS) are seizures that involve alterations in movement, thoughts, sensation, and consciousness that superficially resemble the manifestations of epileptic seizures or syncope but that are not associated with the abnormal electrical activity observed in epileptic seizures or other physiological changes sufficient to cause TLOC (such as changes in blood pressure or heart rate) (LaFrance Jr et al., 2013). The estimated prevalence of FDS is approximately 33 people per 100,000 (Benbadis and Hauser, 2000). The semiological presentations of FDS can include motor activity (for example tonic-clonic, tremors, and body rigidity), non-motor activity (e.g., unresponsiveness and lack of awareness), different emotional and cognitive experiences, and a mixture of different presentations (Asadi-Pooya, 2019).

Although research suggests that FDS are usually a (mal-) adaptive psychological response to distressing internal or external triggers, no concrete theory of the origin has been accepted across the field. Brown and Reuber (2016a) conducted a review of different explanatory models of FDS and the supporting evidence. Their review identified multiple models that have been proposed as causal explanations of FDS: FDS as a consequence of dissociation induced by recalling past memories, FDS as an automated response to threat that is hardwired into the brain, and FDS as a behaviour that has been learned from witnessing other people's seizures in the past. Although they identified research that could support each theory, the findings in support of the different theories were not consistent across all studies and the authors argued that the overall study quality of the research was low, leading the authors to conclude that there was insufficient evidence to make a concrete decision regarding the validity of the theories.

The researchers put forward their own theory, building on the Integrative Cognitive Model of medically unexplainable symptoms (Brown and Reuber, 2016b). This theory posits that individuals with FDS have a mental representation of a seizure that can be formed by a range of experiences, for example through experience of witnessing seizures, learning about seizures, and misinterpreting their own bodily sensations and mistaking them as a threat or the onset of a seizure. The mental representation can be activated by internal or external stressors, which causes the FDS process to start. The theory states that the cognitive system places undue weight on the prediction that a seizure is about to happen which activates the mental representation of a seizure and causes the seizure sequence to happen in a similar way as an automatic, well-rehearsed behavioural response. The mental representation, also known as a seizure scaffold, is suggested to be a fluid construct formed through a network of neural connections that can be updated over time based on new experience. Updates to the network can include changes to the cognitive, emotional, and behavioural representations of the seizure scaffold and the potential triggers.

### 1.1.3 Syncope

Reflex syncope occurs when there is a reduction in blood pressure because of marked changes in heart rate or dilatation of blood vessels. It is caused by changes in the activity of the sympathetic and parasympathetic nerves that regulate heart rate (Adkisson and Benditt, 2017). Reflexive syncope can be further subdivided into three different types: vasovagal syncope, carotid sinus syndrome, and situational syncope (Brignole et al., 2018).

While maintaining an upright posture, individuals experience a reduction in the amount of blood returned to the heart because of increased pooling of blood in the lower parts of the body. Vasovagal syncope can occur when - while upright - an individual is exposed to a stimulus causing a sudden increase in heart rate. The concurrent reduction of blood available to the heart and the sudden increase in heart rate causes a response characterised by a sudden withdrawal of sympathetic activation and an increase in parasympathetic activation. This triggers a reduction in heart rate which can result in loss of consciousness (Adkisson and Benditt, 2017).

Carotid sinus syndrome is caused by the physical manipulation of the carotid sinus region located within the neck. The manipulation causes a malfunction of the baroreceptor reflex that reduces blood pressure and results in a loss of consciousness (Adkisson and Benditt, 2017).

Situational syncope is caused by a broad range of situational factors that can trigger syncope. These factors include stimuli associated with the respiratory tract (e.g., coughing, sneezing, and laughing), the gastrointestinal tract (e.g., swallowing and defecation), and the genitourinary tract (e.g., micturition) (Adkisson and Benditt, 2017).

Orthostatic hypotension can occur when an individual rises to the standing position and is unable to prevent excessive pooling of blood in the lower parts of the body that is usually prevented by physiological, neurological, cardiac, vascular, and muscular responses (Bradley and Davis, 2003). Orthostatic hypotension can occur when there is an interference with these responses, for instance induced by drugs, reductions in blood volume caused by excessive vomiting and diarrhoea, or changes to the neurological systems responsible for this reflex (Brignole et al., 2018).

Cardiac syncope is caused by a problem in the heart that impairs blood flow and reduces the amount of oxygen and nutrients that are supplied to the brain. The problems can be structural or caused by heart rhythm changes (Koene, Adkisson, and Benditt, 2017; Mizrachi and Sitammagari, 2018). Cardiac syncope is the most dangerous form of syncope. It is associated with an estimated annual mortality of 30% (Waytz, Cifu, and Stern, 2018) in part related to the cardiac problems that give rise to syncope (Koene, Adkisson, and Benditt, 2017).

Syncope can be associated with a range of manifestations before loss of consciousness occurs, during the period of impairment of awareness, and after the patient has regained consciousness. Wieling et al. (2009) conducted a review of the manifestations of syncope that have been identified using experimental methods of inducing syncope in volunteers. One example of these methods is the 'fainting lark': a procedure where participants are asked to squat with their knees fully bent while taking approximately 20 rapid, deep breaths and then asked to return to the standing position (Lempert, Bauer, and Schmidt, 1994). The review found that the prodromal phase of syncope can involve light-headedness, a darkened or loss of vision, pallor, blank staring, an inability to move, palpitations, hyperventilation, pupillary dilation, feeling physically uncomfortable, and automatic behaviours, such as appearing drunk. During syncope, people may present as flaccid or rigid (if their EEG recording becomes flat), stare blankly, bite the tip of their tongue, or exhibit a reduction or stop in heart rate, automatic movements, myoclonic jerks, movement of eyes or head, and urinary incontinence. One of the characteristic features of syncope is that it is typically of under 30 seconds duration and that individuals make a rapid recovery (Brignole et al., 2018). However, the post-ictal phase may involve fatigue, pallor, nausea, weakness, visual or auditory hallucinations, confusion, and emotional responses. This review highlights the breadth of manifestations of syncope, many of which may also occur in TLOC related to other causes, demonstrating the challenges associated with differential diagnosis.

## 1.2 Differential Diagnosis

The differential diagnostic process for TLOC typically begins when a patient presents to Primary or Emergency Care Services. General Practitioners and Emergency Physicians often have limited expertise in the differential diagnosis of TLOC and they have limited access to the physical tests. Their primary objective is typically to determine whether the patient is currently at risk of death (Brignole et al., 2018) and to refer patients to specialist neurologists or cardiologists for more thorough investigations within two weeks of a first seizure or remission (NICE, 2022). Nevertheless, patients will require an initial working diagnosis to determine the route of referral. Approximately 20% of patients receive an incorrect diagnosis at this stage, resulting in a patient being sent to the wrong specialist department (Xu et al., 2016). A neurologist or cardiologist will usually conduct an expert assessment of the patient's medical history, the manifestations of their episode(s) of TLOC, and any witness accounts or video recordings of the seizure, conduct a physical examination, and use ECG to identify any cardiac related conditions (Plug and Reuber, 2009; NICE, 2022). This assessment may be followed by a referral for further tests (Toerien, Jackson, and Reuber, 2020) whose primary purpose is not the differentiation between the causes of TLOC but the search for underlying causes of the further sub-differentiation of the different types of epilepsy or syncope (NICE, 2022).

### 1.2.1 Evaluating the medical history

A thorough analysis of the patient's history and the clinical characteristics of their TLOC experiences is one the most important methods for determining the cause (Plug and Reuber, 2009; NICE, 2022). Professionals often segment the experience of TLOC into multiple stages to determine relevant clinical characteristics that can assist the differential diagnosis: what was happening before the attack, what happened during the attack, and what happened after the attack (Malmgren, Reuber, and Appleton, 2012).

The situational characteristics of TLOC refers to potential triggers or particular circumstances that increase the likelihood of an attack. Given the nature of syncope, there are a range of triggers or circumstances that may indicate this particular diagnosis, for example prolonged standing, extreme coughing, urination, defecation, physical exertion, drug use, blood loss, venepuncture or other invasive medical procedures, lack of sleep and fatigue, emotional circumstances, pain, and illness with fever (Lempert, Bauer, and Schmidt, 1994; Colman et al., 2004). For many individuals, epileptic seizures are considered to have no triggers; however, there are some people whose epileptic seizures are considered "reflexive" because they are caused by specific stimuli such as flashing light, decision making, reading, writing, being startled, somatosensory stimulation, proprioception, auditory stimuli, eating, and vestibular stimulation (Xue and Ritaccio, 2006). Furthermore, many individuals report triggers for their seizures, of which stress, sleep deprivation, sleep, fever, and fatigue are most frequently identified (Frucht et al., 2000). There is a similar level of diversity regarding triggers for FDS, with many individuals with such seizures reporting no trigger or emotional stress as a trigger for some or all of their attacks (Reuber et al., 2011). Furthermore, FDS can be triggered through procedures such as hyperventilation, photic stimulation, verbal suggestion, a template massage, and placebo injections (Hingray et al., 2016), demonstrating that there are a broad range of potential triggers for FDS.

Although there is a diverse range of potential manifestations of TLOC and large amounts of overlap between TLOC due to different causes, research has identified some manifestations which are more likely to be associated with one or other cause. For instance, some individuals with epilepsy and FDS report subjective symptoms that precede losing consciousness, for example déjà vu, auditory or gustatory hallucinations, sensations in their abdomen, whereas individuals with syncope often report a cluster of symptoms involving feeling hot, sweaty, lightheaded, and visual and auditory changes (Malmgren, Reuber, and Appleton, 2012). Individuals with epilepsy are more likely to experience oral lacerations as a result of having a seizure than those who have experienced syncope (Benbadis et al., 1995) and FDS (Oliva et al., 2008). Moreover, there are differences in the duration of the attack between different types of TLOC, with syncope typically being of the shortest duration (Brignole et al., 2018) and FDS the longest (Cragar et al., 2002). With regards to the motor manifestations which may be observed in TLOC, individuals with epilepsy, particularly those with bilateral tonic-clonic seizures, exhibit synchronised motor activity that gradually reduces in frequency towards the end of the seizure, whereas the motor activity observed in individuals with FDS often has a sudden onset and offset with little variations in frequency and individuals with syncope are more likely to exhibit brief myoclonic jerks (Malmgren, Reuber, and Appleton, 2012). Finally, individuals with FDS may be more likely to report panic symptoms compared to individuals with epilepsy or syncope (Rawlings et al., 2017a).

The symptoms that individuals experience after the attack may also give insight into the likely cause. Individuals with epilepsy and FDS often exhibit postictal disorientation and retrograde amnesia, which can be markedly longer in individuals with epilepsy (Malmgren, Reuber, and Appleton, 2012). In contrast, individuals with syncope often recover rapidly from postictal confusion and disorientation within a matter of seconds or a small number of minutes (Lempert, 1996).

The clinical characteristics of the attacks outlined in this section do not provide an exhaustive list of the possible manifestations of TLOC with differential diagnostic value. However, it is important to note that, while these different characteristics of TLOC can aid the differential diagnosis, there is no single characteristic that can reliably determine the cause of TLOC. It is the particular combination of features which allows experts to determine the most likely cause. There are also differences in how people describe the seizure that can guide the diagnostic decisions of experts (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008; Plug, Sharrack, and Reuber, 2009b; Plug, Sharrack, and Reuber, 2010; Robson et al., 2012; Robson, Drew, and Reuber, 2016), and these differences will be described in greater detail in a subsequent section.

### 1.2.2 Interictal EEG

The most recent NICE guidelines state that interictal EEG does not have a reliable role in determining whether a person has epilepsy (NICE, 2022). Therefore, it should not be used to exclude a diagnosis of epilepsy because the interictal EEG is not abnormal for many people with epilepsy. Its primary purpose is to guide the process of determining the type of epilepsy. There is a clinical profile for interictal EEG recordings that can be present in people with epilepsy, for example whether the activity is paroxysmal, involves an abrupt change in polarity with a short duration (<200ms), is morphological, negative in polarity, and followed by slow-wave cortical activity (Pillai and Sperling, 2006). Although interictal EEG can lead to a change in the initial diagnosis if very clear epileptiform discharges are found in a patient who was initially not thought to have epilepsy based on an analysis of the medical history, this only caused a change in the diagnosis for 1/158 patients in a prospective analysis (Angus-Leppan, 2008).

### 1.2.3 Structural brain abnormalities

Neuroimaging techniques such as Magnetic Resonance Imaging (MRI) can be used to detect structural abnormalities in the brain that can be associated with epilepsy. A broad range of structural abnormalities have been associated with epilepsy, for example atrophy of the hippocampus, malformations of the cerebral cortex and vascular system, brain tumours, and destructive brain lesions caused by a neurological insult (Fitsiori et al., 2019). Although structural brain abnormalities may provide insight into potential causes and treatments of an epileptic seizure, it is not possible to make differential diagnostic decisions based on neuroimaging alone because a structural brain abnormality, although present, may not be causing seizures and asymptomatic or incidental brain abnormalities may be found in patients with syncope and dissociative seizures (Malmgren, Reuber, and Appleton, 2012).

### 1.2.4 Electrocardiogram

An electrocardiogram (ECG) is used to measure the electrical activity of the heart. Electrodes are placed on the chest and surrounding body that record the electrical activity. It is recommended that a routine (12 lead) ECG is used for everyone who presents with a seizure (NICE, 2022) because it can detect abnormalities that predispose people to arrhythmias and TLOC (Dovgalyuk et al., 2007). Furthermore, prolonged ECG can be used to capture brief, asymptomatic cardiac arrhythmias or episodes of TLOC in patients with unexplained recurrent syncope - especially if collected using implantable loop recorders (Bisignani et al., 2019).

### 1.2.5 Video Telemetry

The diagnosis of epilepsy or FDS is documented most definitively using a method called video telemetry (video-EEG) that involves monitoring a patient in a hospital or home setting until they have a seizure and simultaneously recording video and EEG during an event to capture typical seizure-related observable manifestations and to explore whether behavioural changes are associated with ictal electrical brain activity (Noachtar and Rémi, 2009; Kinney, Kovac, and Diehl, 2019). Video-EEG can be used to document a diagnosis of FDS by showing behavioural patterns typically associated with this cause of TLOC in the absence of the changes of electrical brain activity which characterise epilepsy (Brown and Reuber, 2016a; Kinney, Kovac, and Diehl, 2019). Although video telemetry is considered the most reliable test, it cannot be used to diagnose all patients because most patients first presenting with TLOC would not have a sufficiently high number of events to make video-EEG recordings feasible (Mohan, Markand, and Salanov, 1996; Cascino, 2002).

### 1.2.6 Tilt-Table Testing

The Tilt-Table test is a method of detecting vasovagal or orthostatic syncope by recording the patient's blood pressure and heart rate during a superficial transition from lying to standing (Kohno, Adkisson, and Benditt, 2018). At the start of the procedure, the patient is strapped to a bed in the supine position. The bed is subsequently raised, typically to an angle of approximately 60-80 degrees, to investigate whether the transition from laying to standing reduces the patient's blood pressure and causes them to faint. The method has been demonstrated as a useful method for detecting susceptibility to vasovagal syncope, but the interpretation of the Tilt-Table test requires expertise (Kohno, Adkisson, and Benditt, 2018). Furthermore, the tilt-table test can help to differentiate between syncope and FDS resembling syncope by demonstrating that, in patients with FDS, typical TLOC manifestations occur in the absence of significant changes of heart rate or blood pressure (Tannemaat et al., 2013).

### 1.2.7 Serological Biomarkers

Biomarkers are objective measurements that provide insight into the status of biological processes (Sueri et al., 2018). Many health conditions are caused by changes in biological processes that can be detected by corresponding changes in known biomarkers. Researchers have identified several serological biomarkers that may indicate that a person has just recovered from an epileptic seizure, for example an increase in the serum levels of prolactin, Interleukin-6, and creatine kinase, measured minutes or hours after a seizure (Sueri et al., 2018). Furthermore, recent research has suggested that elevated levels of Neurogranin may be indicative of a postictal state after an epileptic seizure (Kalkan et al., 2022). Although prolactin and creatine kinase have a high specificity, the utility for differential diagnosis can be restricted to tonic clonic seizures (Brigo et al., 2015; Wang et al., 2021) and a high specificity does not assist the differential diagnosis of other causes of TLOC, for example FDS and syncope.

### 1.2.8 Summary

This section has highlighted the challenges associated with the differential diagnosis of TLOC. TLOC is associated with many different symptoms that may be more or less indicative of a particular cause but cannot be used to make a differential diagnosis alone. Although there are tests that can detect the cause of TLOC, many of these tests are only able to detect one of the three causes of TLOC and are of limited diagnostic utility during interictal periods. This is problematic for individuals who receive an incorrect diagnosis from Primary and Emergency Care Services because the tests they are referred for may be unable to detect the true cause of TLOC, which delays the time that it takes to receive a final diagnosis. Therefore, identifying methods that can consider a range of the symptoms that patients experience and predict the cause of TLOC in Primary and Emergency Care Services could help improve the treatment pathway.

## 1.3 Current research methods

### 1.3.1 Clinical Decision Tools

Individuals who experience TLOC will typically first present in non-expert settings, i.e. Primary Care or Emergency Departments. Given that the testing and treatment pathways are targeted to one specific diagnosis and that the differential diagnosis requires a high level of expertise in the interpretation of TLOC descriptions and medical history, an important first step is to determine the most likely cause of the TLOC. This will ensure individuals are referred to the most appropriate service. Unfortunately, many individuals initially receive the wrong diagnosis. In a retrospective analysis of 1506 adult patients referred to a neurologist from Primary Care with a diagnosis of epilepsy across a 16-year period, 194 (12%) were subsequently diagnosed with syncope (Josephson, Rahey, and Sadler, 2007). Furthermore, a review of 27 studies investigating the rate of false positives among diagnoses of epilepsy found that between 2-71% of individuals with a diagnosis of syncope or FDS were misdiagnosed in Primary and Emergency Care Services, with a median value of 20%. Many of these people were exposed to negative consequences of an epilepsy diagnosis, for example the unpleasant side effects of anti-epileptic medication and prolonged driving restrictions (Xu et al., 2016). Many individuals with a diagnosis of unexplained syncope in cardiology departments may also have a functional neurological disorder diagnosis (Iglesias et al., 2009) and may need referral to other services (Raj et al., 2014). Referrals to the wrong service can delay the time to receiving the correct medical diagnosis and treatment. Therefore, the identification of methods that can reduce the rate of misdiagnosis and guide referral pathways may help to improve patient care and reduce the unnecessary costs associated with delivering inappropriate tests and treatments.

A thorough analysis of the patient's history by a specialised neurologist often leads to the correct diagnosis before any physical tests are undertaken (Angus-Leppan, 2008). Given that experts can make accurate diagnoses based on the patient's history, researchers have investigated whether it is possible to create a clinical decision tool to accurately stratify TLOC patients in Primary Care Services (Wardrope, Newberry, and Reuber, 2018). A clinical decision tool can combine multiple variables about the medical history into a single tool that can aid clinicians to make diagnostic decisions (Stiell and Bennett, 2007). Given that there are many different clinical variables that provide insight into the cause of TLOC, a clinical decision tool may make it easier to process a large quantity of clinical data when considering the most likely cause and the most appropriate referral pathway.

Another approach has been to create rapid access triage clinics where trained clinicians stratify patients based upon risk scores using a web-based questionnaire and electrocardiographic monitoring (Petkar et al., 2011). This approach effectively identified high risk patients with cardiac abnormalities and provided a treatment plan for these patients. Other approaches have involved devices with objective scoring tools that can be completed by clinicians using information collected from the patient's medical record, medical history interviews, observations, and symptom questionnaires (Baroni et al., 2021; Kerr et al., 2020; Wardrope, Newberry, and Reuber, 2018).

Although many of these methods have shown effectiveness at predicting the underlying diagnosis, many of the tools can only be applied once patients have been referred to specialist neurology or cardiology departments and most are limited to differentiating between two of the three most prominent causes of TLOC (epilepsy and FDS). Wardrope, Newberry, and Reuber (2018) conducted a review of the research investigating potential clinical decision tools for patients presenting with TLOC and found that only two studies included patients with syncope in their sample. In order for a clinical decision tool for TLOC to be effective in a primary care setting, it should be able to distinguish between all three common diagnoses. Of the two studies that included syncope in their sample, one study aimed to stratify patients based on the symptoms that they reported experiencing during the TLOC (Reuber et al., 2016) and the second focused on the reported behavioural characteristics of the TLOC that was observed by a witness (Chen et al., 2019).

Reuber et al. (2016) aimed to investigate whether patient reportable TLOC features could help to differentiate between epilepsy, FDS and syncope using an 86 item questionnaire that was administered to 300 patients who had previously received a diagnosis from a specialist neurology clinic (100 syncope, 100 FDS and 100 epilepsy). A paper copy of the questionnaires was sent to the participants by post alongside an envelope to return the completed questionnaires. The items measured the extent that the patient had experienced a symptom using a 5-point Likert scale (1 = "always" - 5 = "never") and 7 additional demographic and clinical features. The responses significantly differed across all three groups for 57 of the items. Overall, the data was able to successfully identify the correct diagnosis for 91% of patients with syncope, 66% of patients with epilepsy and 78% of FDS patients. These findings suggest that a symptom checklist can have some diagnostic utility for detecting the cause of TLOC, particularly for cases of syncope, but that some symptoms may exhibit more distinct group differences than others.

The extensive group differences across the symptoms could be explained by different latent variables. Reuber et al. (2016) conducted an exploratory factor analysis on the 86 patient reported TLOC features to find out more about the nature of the differences between TLOC experiences depending on the cause of the events. They identified differences in factors they named "feeling overpowered", "mind/body/world disconnection" and "catastrophic experience" between the three health conditions. Patients with FDS endorsed TLOC features contributing to these factors more strongly than those with epilepsy or syncope. Theories about FDS suggest that seizures can be related to threat processing or the activation of memories of previous traumatic experiences (Brown and Reuber, 2016a). This may explain why patients with FDS report higher levels of “mind/body/world disconnection” and “catastrophic experience”. Responses from patients with syncope were significantly less related to the factor “amnesia” compared to patients with epilepsy; and patients with epilepsy were significantly less likely to score highly on the factor “sensory experience” than those with syncope and FDS, suggesting that the sensory experience symptoms within the checklist are more applicable to the experiences of people with FDS and syncope. These findings provide more insight into how the symptom checklist can be used to predict each diagnosis, but further research is required to improve the predictive performance.

Many patients experience TLOC in the presence of others and research suggests that there are diagnostically relevant behavioural characteristics which can help with the differential diagnosis of TLOC (Kinney, Kovac, and Diehl, 2019). Chen et al. (2019) investigated whether witness reports could improve the diagnostic performance achievable by using patient-provided data alone. They collected responses of 249 witnesses from a separate 31 item witness questionnaire and found that 24 items significantly differed between epilepsy, syncope and FDS. Combining the patient questionnaire (Reuber et al., 2016) and witness questionnaire responses, the two questionnaires were able to accurately identify the correct diagnosis using logistic regression in 80% of patients with epilepsy, 79% with FDS and 92% with syncope. Therefore, the inclusion of witness reports in a clinical decision tool may be important for increasing accuracy of the diagnosis, especially for patients with epilepsy who showed the largest increase in accuracy (of 14%).

Non-linear machine learning methods may improve the classification accuracy of the questionnaires based on patient and witness provided data compared to linear methods. Wardrope et al. (2020a) used dichotomised responses ("ever” or “never”) from the PEP and PEO and a machine learning method “Random Forest" to classify patients. The Random Forest algorithm has an inherent feature selection mechanism that reduces the number of variables and improves the diagnostic prediction, which was used on each individual dataset. They identified 34 items from the patient only responses that were able to accurately classify 78.3% of patients (83.8% syncope, 81.5% epilepsy and 67.9% FDS) and 36 items from the patient and witness responses that were able to accurately classify 86% of patients (100% syncope, 85.7% epilepsy and 75% FDS). This new model was called the iPEP procedure. The classification accuracy for the three-way differentiation using patient and witness data was higher when the data were analysed using this method rather than using a regression approach.

However, the percentage of FDS patients that were correctly classified decreases when the data were analysed using the Random Forest method compared to the original logistic regression, from 79% (Chen et al., 2019) to 75% (Wardrope et al., 2020a). One potential explanation is that the variables were dichotomised. In the original patient responses, FDS patients reported a larger number of different ictal symptoms overall but they reported that their seizure experiences were less stereotyped, indicated by a lower amount of extreme values (“always” or “never”) and a greater number of intermediate responses (“rarely”, “sometimes”, “often”) (Reuber et al., 2016). Dichotomising the data may have reduced this pattern in the dataset, subsequently reducing the classification accuracy. Future research should investigate how the FDS classification accuracy changes when patients are presenting with binary symptom questionnaires.

A second potential critique of the validation work carried out on the iPEP is that the results may not be generalizable to all patients, especially newly presenting patients who have only experienced one episode or a small number of episodes of TLOC. The patients on which the modelling of the iPEP performance reported by Wardrope et al. (2020a) was based were aware of their diagnosis of epilepsy, syncope or FDS when they provided the responses to the iPEP. Patients who already have a diagnosis may be more familiar with the signs and symptoms of their condition and in a position to describe their TLOC experiences in greater detail. This may have influenced their responses. If the iPEP were to be implemented in an emergency or Primary Care setting (or at the interface between these and specialist care settings), patients will not have a secure diagnosis and would only be able to draw on a smaller number of TLOC experiences. Consequently, to evaluate its usefulness as a screening, stratification or diagnostic tool in an emergency room or Primary Care setting, the iPEP will need to be validated in a new patient population of patients at the point of initial referral for a specialist evaluation of TLOC.

In this section, we have introduced and discussed a symptom questionnaire that can reliably differentiate between epilepsy, FDS, and syncope. The iPEP could be used to guide referral pathways because it can be used in Primary and Emergency Care Services, but future research should investigate the predictive performance of the tool when used by patients first presenting to health services and administered in a binary format before any firm conclusions can be made. Although the early modelling research suggests the iPEP may identify the correct diagnosis more than pre-existing methods used in Primary and Emergency Care Services, which have a misdiagnosis rate of  20% (Xu et al., 2016), there is still scope for improvement, particularly for the challenging differentiation between epilepsy and FDS. Therefore, future research should explore additional features that can be incorporated into the clinical decision tool that can reliably detect these two health conditions.

### 1.3.2 Conversation Analysis and TLOC

Speech in medical interactions can achieve more than sharing information. What a patient says and where they say it in the conversation has interactional significance (Stivers, 2002). Research aiming to understand communication in healthcare interactions often uses conversation analysis (CA), which is a method of understanding the dynamics of human conversation through the micro-analysis of recorded conversations (Jefferson et al., 1983). CA aims to understand the norms that govern conversation through identifying patterns that occur across multiple conversations and contexts (Sacks, 1974). The method relies on the assumption that what people say is largely influenced by what was said prior to their turn at talk (Sacks, 1974) and the social norms of conversation (Sacks, 1992).

Previous conversation analytic research shows that patients with epilepsy and FDS describe their symptoms differently in medical interactions (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). The original research followed interview guidelines that were established during the pilot phase of the first study using CA in this setting (Schwabe et al., 2008). The guidelines included a list of questions that the neurologist was instructed to ask the patient and encouraged the neurologist to allow the patient to speak freely and refrain from interrupting. Patients were asked what they hoped to get from the consultation and to describe their first, worst, and last seizure. Transcripts of the interaction were analysed to detect linguistic and interactional differences.

One of the differences found related to the amount of information that people provided about the seizure experience. People with epilepsy provided more information about the subjective symptoms that they experienced and attempted to provide a coherent narrative of the events that surrounded the unconscious period by detailing their memories before and after the "gap" and attempting to reconstruct what happened while they were unconscious (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). In contrast, people with FDS often equated the unconscious period and the seizure and made generalised statements highlighting that they do not know anything about what happened (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008).

The two groups also differed in the extent to which they displayed effort to describe and redescribe what they experience, also known as formulation effort (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). People with epilepsy displayed more formulation effort during descriptions of their subjective symptoms and the unconscious period compared to people with FDS. These differences included more hesitations, repetitions, restarts, and repairs.

There were also differences in how people conceptualised the seizure experience during the interaction. People with FDS were more likely to use the diagnostic label "blackout" compared to people with epilepsy (Plug, Sharrack, and Reuber, 2010). The two groups were different in the consistency of the metaphorical conceptualisation they used to describe the seizure experience. People with epilepsy were more likely to conceptualise the seizure experience as an external agent that they fought or struggle against ("they just creep up on you and they get you"), whereas people with FDS were less consistent in their use of metaphor and were more likely to describe the seizure experience as a place or space that they went to ("that's it, I'm gone") (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008; Plug, Sharrack, and Reuber, 2009b). Together, these features were found to amount to distinctively different linguistic profiles for people with epilepsy and people with FDS.

The original research findings were extended by studies investigating the role of accompanying others during routine clinical encounters at a seizure clinic (Robson et al., 2012; Robson, Drew, and Reuber, 2016). This research was different to the original studies because the neurologists did not follow the interview guidelines from Schwabe, Howell, and Reuber (2007) and Schwabe et al. (2008). People with FDS were found to make more catastrophising third party references, in contrast to the normalising third party references more typically observed in the talk of people with epilepsy (Robson et al., 2012). The extensive interactional resistance demonstrated by people with FDS was also shown to increase the involvement of accompanying others during routine clinical consultations (Robson, Drew, and Reuber, 2016). These findings demonstrate that linguistic and interactional differences between the talk of people with FDS and people with epilepsy are still present when the interviews are conducted differently, and their patterns of talking influence the contributions of other people partaking in the interaction.

These linguistic profiles can aid the differential diagnosis of seizures. Reuber et al. (2009) used the qualitative research findings to create a Diagnostic Scoring Aid (DSA) that could be used by linguists to detect linguistic differences between people with epilepsy and people with FDS and to predict the diagnosis. The DSA consists of 17 items that are rated with a score of 1, 0, or -1 (Table 1.1). Two linguists who did not know the diagnosis of the twenty participants used in the study were trained to use the DSA. Two cut-off values were selected, one for each linguist, that produced the optimum diagnostic decisions for the whole sample (4.5 and 7.5) and were able to predict the diagnosis with an accuracy between 75-80%. No single item was indicative of a particular diagnosis for all participants, but people with epilepsy typically scored higher across the whole DSA. These findings indicate that there is no single linguistic difference that can differentiate between people with epilepsy and people with FDS, but the presence of multiple linguistic differences can help to make the diagnosis.

TABLE 1.1: The Diagnostic Scoring Aid taken from Reuber et al. (2009). P = Patient. I = Interviewer

| Item | Description | Observation | Score |
|---|---|---|---|
| 1 | General focus on seizure experience (rather than seizure situations or consequences) | Introduced by the P | 1 |
| | | Introduced by the I, followed by P | 0 |
| | | Introduced by I, lost by P | -1 |
| 2 | Description of subjective seizure symptoms | Volunteered | 1 |
| | | Offered only when prompted | 0 |
| | | Prompting unanswered | -1 |
| 3 | Description of seizure suppression attempts | Volunteered | 1 |
| | | Not described/ only on prompting | 0 |
| | | Prompting unanswered | -1 |
| 4 | Description of 'gaps' (phases of reduced self-control or recollection) | Volunteered | 1 |
| | | Offered when prompted | 0 |
| | | Prompting unanswered/ 'holistic' statements only | -1 |
| 5 | Response to challenge of statements about 'gaps' | Elaboration or reformulation of previous description | 1 |
| | | Not described/ only on prompting | 0 |
| | | Prompting unanswered | -1 |
| 6 | Description of individual seizure episodes (possible 'focusing resistance': interactional resistance to focus on particular seizures) | Volunteered | 1 |
| | | Not offered / episodes explicitly not distinguishable | 0 |
| | | Not offered, no explicit denial of ability to distinguish episodes distinguish episodes | -1 |
| 7 | Subjective seizure symptoms | Described in great detail | 1 |
| | | Little or some detail | 0 |
| | | (Listed but) not described in detail | -1 |
| 8 | Relative importance of subjective seizure symptoms | Treated as central to description | 1 |
| | | More or equal attention to circumstantial details | 0 |
| | | Not described beyond brief statements | -1 |
| 9 | Relative importance of 'gaps' (phases of reduced self-control or recollection) | One of several elements of seizures | 1 |
| | | Prominent element of seizure episodes | 0 |

Continued on next page

**Table 1.1 – continued from previous page**

| Item | Description | Observation | Score |
|---|---|---|---|
| | | Defining element of seizures | -1 |
| 10 | Contouring of 'gaps' in seizure trajectory (eg. detailing of last memory before / first after seizure) | Clear attempt to contour 'gaps' | 1 |
| | | Some attempt to contour 'gaps' | 0 |
| | | No contouring of gaps / no clear seizure trajectory | -1 |
| 11 | Reconstruction of 'gaps' (eg. filling own memory gaps with own recollections / witness accounts) | Clear attempts to fill 'gaps' with own recollections | 1 |
| | | Some attempts to reconstruct 'gaps' with own recollections | 0 |
| | | No attempts to reconstruct gaps using own recollections | -1 |
| 12 | 'Formulation effort' associated with description of subjective seizure symptoms ('formulation effort' includes restarts, reformulations, neologisms) | With marked formulation effort | 1 |
| | | With some or little formulation effort | 0 |
| | | No description beyond brief | -1 |
| 13 | Negations in descriptions of seizure experience (absolute: 'I don't remember anything, contextualised: I remember X but not Y') | Contextualised negations only | 1 |
| | | With some absolute negations | 0 |
| | | With pervasive absolute negations | -1 |
| 14 | 'Formulation effort' associated with description of 'gaps' | With marked formulation effort | 1 |
| | | With some/little formulation effort | 0 |
| | | No description beyond 'holistic' statements | -1 |
| 15 | Metaphoric seizure conceptualisation | Consistent across seizures | 1 |
| | | With variations across seizures | 0 |
| | | No coherent conceptualisation | -1 |
| 16 | External or internal conceptualisation of seizures | Consistent seizure conceptualisation as external | 1 |
| | | Seizures sometimes conceptualised as external | 0 |
| | | Seizures not conceptualised as external | -1 |
| 17 | Conceptualisation of seizures as a fight or or struggle | Seizures repeatedly conceptualised as a fight struggle | 1 |
| | | Seizures sometimes conceptualised as a fight or struggle | 0 |
| | | Seizures not conceptualised as a fight or struggle | -1 |

The accuracy of the DSA has been supported by further research studies applying the DSA to different samples. Using a sample of 10 Italian participants, 5 people with epilepsy and 5 people with FDS, a trained linguist identified the video-EEG diagnosis of 90% percent of participants (Cornaggia et al., 2012). Two psychologists applied the DSA to another Italian sample containing 49 patients with epilepsy and 12 patients with FDS (Papagno et al., 2017). They correctly identified 80.3% of patients with a sensitivity of 0.795 and a specificity of 0.83. Most, but not all, diagnoses were made using video-EEG. A neurologist and linguist applied the DSA to a sample of 12 Chinese patients and correctly identified 83% (Yao et al., 2017). In a French sample containing 13 patients with FDS and 19 patients with epilepsy, two neurologists correctly identified 84% and 88% of patients, respectively (Biberon et al., 2020). These findings demonstrate that the DSA is able to retain an accuracy of approximately 80-90% across different samples, suggesting that the scoring aid can be applied outside of the original research context.

Although the DSA is the most widely applied scoring aid for detecting the linguistic differences between people with epilepsy and people with FDS, researchers have created more condensed and simplified scoring tables. Biberon et al. (2020) aimed to detect the items from the DSA that were the most effective at predicting the diagnosis in order to create a simplified scoring table. They selected the following items from the DSA (Table 1.1): items 4, 5, 7, 12, and 15. These items focussed on how people described the unconscious period and subjective symptoms, the amount of formulation effort present during descriptions of subjective symptoms, and the consistency of metaphoric conceptulations. Two raters were able to identify 88-91% of cases using these items alone. In addition, Beghi et al. (2020) investigated a novel 5 item scoring table using a sample of 35 participants, which was able to identify the diagnosis of 82.9% when the optimum cut-off was calculated. Furthermore, neurologists who had undertaken a training course to detect the linguistic profiles during clinical interactions were able to score patients on a condensed scoring table immediately after consultations and correctly predicted the diagnosis made on clinical grounds in 81.8% of the 33 participants in the study (Jenkins et al., 2016). Therefore, it is possible to make accurate predictions about whether someone has epilepsy or FDS using fewer linguistic scoring items.

In this section, we have demonstrated that the talk of people with epilepsy or FDS interacting with a neurologist can be described with distinct linguistic profiles. The differences captured in these profiles can be detected by trained professionals and can be used to predict the diagnosis with a high accuracy. It is of particular importance for the objectives of this thesis that research has demonstrated a similarly high level of accuracy when fewer elements of the linguistic profile were measured. This demonstrates that it may not be necessary to measure all facets of the linguistic profile. A full conversation analytic workup of every clinic interaction is unfeasible; therefore, automating the collection and analysis of spoken descriptions of TLOC may allow a reliable and efficient method of extracting features of these linguistic profiles, especially considering that previous research has shown that accurate predictions can be made using fewer features (Biberon et al., 2020). Moreover, these features could be combined with the iPEP to potentially improve the challenging differentiation between people with epilepsy and people with FDS.

### 1.3.3 Thesis Overview

The introductory chapter has shown that the different diagnostic pathways for TLOC are highly specific for a single health condition. A clinical decision tool capable of predicting the most likely cause of TLOC would speed up the time taken to receive the correct diagnosis because people could be referred for the most appropriate specific tests more quickly. Furthermore, the tool could guide the interpretation of future test results by providing a pre-test probability of a diagnosis which could inform the post-test probability. Previous research suggests that there is scope to improve currently proposed clinical decision tools, but more work is required to determine what features differentiate more clearly and are capable of improving the performance of current models. The initial research using CA in patients with seizures suggests that identifying and incorporating linguistic differences between how people with different diagnoses describe their experience of TLOC could help the differential diagnostic process.

Therefore, the objective of this thesis is to explore the feasibility of improving a currently proposed clinical decision tool by incorporating an automated analysis of spoken descriptions of TLOC inspired by the previous CA research. We have created an online application that people who have experienced TLOC (and witnesses if available) can use to share information about what they experienced. The application consists of a short binary questionnaire and an interaction with a virtual agent that asks them questions about what they experienced during their most recent attack. We evaluated feasibility based upon the capacity of the application to predict the cause of TLOC and the acceptability of the approach from the perspective of the patients and witnesses who have used it.

#### 1.3.3.1 Research Questions

The thesis aimed to answer the following research questions:

**Research Question 1**

Given that there was only one clinical decision tool that reliably differentiated between the three most common causes of TLOC when this PhD was designed (Wardrope, Newberry, and Reuber, 2018), the PhD aims to further validate the accuracy of the iPEP version of this tool (Wardrope et al., 2020a). The questionnaire was originally implemented as a five-point Likert scale that was delivered to patients who already had a gold-standard diagnosis. The responses were then dichotomised by the research team to investigate whether this increased the predictive performance of the questionnaire and might allow its use in patients presenting after only one event. We would like further to explore the validity of this questionnaire by answering the following question - **does the iPEP demonstrate a similar level of predictive performance when the questionnaire is administered in a binary format through an online application and when the sample includes people who have newly presented with TLOC?**

**Research Question 2**

The linguistic and interactional differences between people with epilepsy and people with FDS have always been measured by trained linguists. Humans can understand the context of a conversation and infer the meaning that transcends the raw semantic content. In contrast, automatic natural language processing methods, particularly methods that are appropriate for a small dataset, are reliant on the raw semantic content. These approaches may not be able to detect the same linguistic observations as a human can. Therefore, **what natural language processing features**

**can capture some of the linguistic differences between people with epilepsy and people with FDS and be used to predict the diagnosis?**

**Research Question 3**

There are two major differences in the spoken descriptions of TLOC that will be collected in this research compared to the previous CA research. First, to automate the collection and analysis of spoken descriptions of TLOC, patients are asked to answer questions posed by a virtual agent, whereas previous research involved dynamic interactions between the patient and a doctor. Secondly, the patients are only asked to describe their most recent experience of losing consciousness because patients first presenting to health services may only have one experience of TLOC, whereas the previous research asked people about multiple seizures. These differences may influence the linguistic profiles for each health condition. Therefore, we are interested in **how people describe their experience of TLOC to a virtual agent?**

**Research Question 4**

The performance of the iPEP was hindered by the particularly challenging differentiation between people with epilepsy and people with FDS, but there are linguistic differences between the spoken seizure descriptions of these two groups. Therefore, **is it possible to improve the predictive performance of the iPEP by incorporating an automated analysis of patient descriptions of TLOC?**

**Research Question 5**

Patients and witnesses must be willing to use the online application in order for it to be effective. People's willingness to use an intervention is influenced by how acceptable the approach is perceived to be. For a new form of technology, this acceptability may also be influenced by the design. Therefore, **do patients and witnesses think the application is acceptable and what changes can be made to improve the acceptability?**

#### 1.3.3.2 Organisation of the thesis

In the current chapter, we have introduced the different methods involved in the differential diagnosis of TLOC and provided a justification for our investigation into whether current clinical decision tools can be improved by incorporating an automated analysis of language. Chapter two will begin by providing an overview of the fundamental principles of machine learning research. It will then provide an overview of the important consideration and methods that are commonly used in medical speech technology research. These two chapters combined will provide the relevant background information for this research.

Having already provided an overview of some types of features that are used for speech processing technology, chapter three will outline an analysis that involves exploring potential features that can be used to differentiate between people with epilepsy and people with FDS. Two groups of features will be assessed: features designed to measure the degree of formulative effort displayed during a single seizure description and features designed to measure semantic differences between people with epilepsy and people with FDS. The objective of the chapter will be to answer the second research question.

Chapter four will provide an overview of the design of the overall research project, the recruitment procedure, and the online web application. The online web application was used to collect the research data that was used in chapters five, six, seven, and eight. Therefore, there was a single recruitment procedure. The application consists of two parts: the data collection front-end that the user interacts with and the machine learning back-end used to make predictions.

Having provided an understanding of how the application works, chapter five will provide an overview of how participants interacted with the virtual agent that was used to collect spoken descriptions of TLOC. This analysis explores whether the front-end of the application can facilitate the production of descriptions that are useful for the differential diagnosis of epilepsy and FDS. Using conversation analysis, we will explore whether the spoken descriptions are qualitatively similar to those observed in previous doctor-patient interactions and consider how people may interact differently with the virtual agent compared to a doctor. This chapter will address research question three.

Chapter six and chapter seven will address how effectively the machine learning back-end can predict the underlying cause of TLOC. Chapter six will explore the effectiveness of predicting the cause of the TLOC using an automated analysis of the TLOC descriptions when spoken descriptions are collected through the online application. Chapter seven will extend the analysis by exploring whether the automated analysis of language can improve the predictive performance of the iPEP. The accuracy of the iPEP will provide the baseline to be improved upon. These chapters will address the research questions one and five.

In chapter eight, we will explore the acceptability of the application from the perspective of patients and witnesses. We will use a mixed methods approach to investigate to what extent people intend to use the application if it were available, what factors influence their likelihood to use the application, and to gain an in-depth understanding of the attitudes of patients and witnesses towards the application. We will record attitudes and intentions towards the application using a questionnaire derived based upon the Technology Acceptance Model (Davis, 1989) and qualitative interviews with people who have used the application. The questionnaire and interviews focus on the application, which encompasses the design of the front-end and the concept behind using the back-end to make diagnostic predictions. Therefore, chapter 8 will address research question 5.

Finally, chapter nine will provide an overview of the main objectives of the thesis and explore the contributions of this thesis towards answering the research questions outlined in the introduction. We will suggest future research directions for the topic.

# Chapter 2

# Background and Related Work

The previous chapter outlined many of the methods used to make the differential diagnosis of TLOC and the motivations for creating a clinical decision tool and exploring the incorporation of an automated analysis of spoken descriptions of TLOC. The process of creating and evaluating modern clinical decision tools uses methods from the discipline of machine learning. Within the discipline of machine learning, there is a sub-field of research that focuses on applications that make predictions using speech and language. Therefore, this first part of this chapter will provide an introduction to machine learning and the methods that are often used to create machine learning models. The second part of the chapter will provide a brief introduction to how automatic speech processing and machine learning have been combined to create and evaluate clinical decision tools for the healthcare industry.

## 2.1 Machine learning

The term machine learning refers to a discipline of research that aims to use computers to solve predictive problems (Jordan and Mitchell, 2015). These methods are considered to "learn" because they are trained on data to improve some measure of performance, for example predicting a medical diagnosis. The most widely used machine learning approaches used in the field of medicine are supervised machine learning models that aim to generate a statistical mapping between a vector of inputs (x) and a select number of target outputs (y) (Jordan and Mitchell, 2015).

### 2.1.1 Feature extraction

Before a machine learning algorithm can be trained, the target data must be transformed into a numerical vector that the algorithm can understand, a method known as feature extraction. The feature extraction methods that are used are largely dependent on the type of target data and objectives of the researcher. For some types of data, it is possible to use a one-to-one mapping between the target data and features, for example each response option in a questionnaire can be allocated unique numerical representations and the responses can then be input into a machine learning model in the form of a vector of responses for each participant (Wardrope et al., 2020a). Other types of data are more complex, for example text and speech data, and must be transformed into a vector of representational features before a machine learning model can make effective predictions. This vector of representational features can be designed using knowledge of the target domain and automatically extracted from the original speech signal, for example exacting features that are designed to measure conversational and interactional properties of speech from individuals with dementia during an interaction with an virtual agent (Mirheidari et al., 2017a). In contrast, there are also methods of automatically generating features that represent the original data using machine learning algorithms trained on a large amounts of data that are able to output vectors that represent the original data and can be inputted into other machine learning algorithms to make predictions, for example BERT can produce vectors that represent textual data (Devlin et al., 2018) and Wav2Vec2 can produce vectors that represent speech data (Baevski et al., 2020). Although the features extracted from these methods are often effective at training highly accurate machine learning models, they are often considered "black boxes" because it is difficult to determine why the algorithm has made a particular prediction.

### 2.1.2 Partitioning the data

Supervised machine learning algorithms generate a statistical mapping between the input (x) and output (y) using a training dataset with known labels. The objective is to train an algorithm to make predictions on future, unseen data, also known as the test set. The ability of an algorithm to successfully predict the output is dependent on whether there is a sufficient relationship between the target features in the training data and the output, whether the statistical methods employed by the algorithm are congruent with the type of data, and whether the training data is sufficiently large to allow the algorithm to detect meaningful patterns within the data that can generalise to unseen data. The performance of algorithms often increases alongside increases in the amount of training data before performance plateaus (Figueroa et al., 2012). Therefore, researchers must partition the dataset in order to maximise the amount of data available for training while also retaining enough data to evaluate whether the model can generalise to unseen data.

The simplest method for partitioning the data is to randomly remove a proportion of the dataset ( 10-30%) for testing (Berrar, 2019). The remaining data is used to train the algorithm. This method requires a large sample size to ensure there is sufficient training data available, which can be problematic in the field of medicine where data is frequently scarce (Latif et al., 2020).

Cross validation is a method that allows researchers to test the generalizability of their machine learning models when they have a small sample size (Berrar, 2019). In K-Fold cross validation, the dataset is segmented into K number of folds of approximately equal size. An iterative analysis is then conducted where a single fold is removed on each iteration to perform as the test dataset and the remaining data is used as the training dataset (Figure 2.1). For each iteration of the cross validation procedure, a machine learning algorithm is trained, and the predictive performance is calculated using the test set. Once each segment of data has been used as the test set, the overall predictive performance of the machine learning algorithm is calculated by averaging the performance across all test sets. This method allows researchers to maximise the size of the training set while still evaluating the generalizability of their model for data external to the training set.

For instances where the amount of data is particularly small, another form of cross validation called Leave-One-Out Cross Validation (LOOCV) can be used (Berrar, 2019). This method follows the same principles as K-Fold cross validation, but rather than segmenting the data into multiple folds, a single datapoint is removed and used as the test on each iteration. Therefore, the total number of iterations is equal to the size of the dataset.

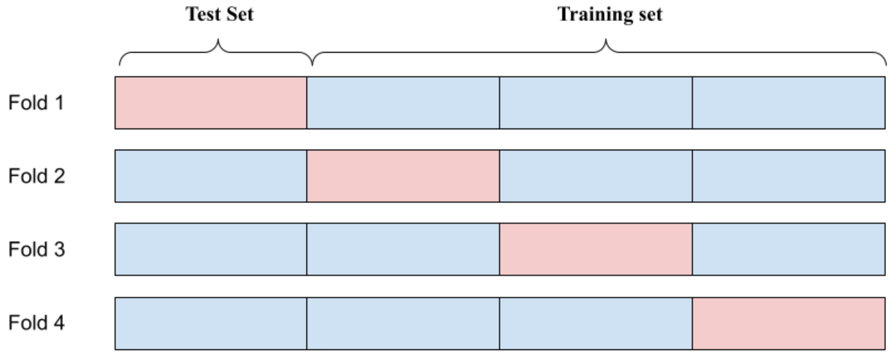


FIGURE 2.1: An example of how the data is segmented during K-Fold cross validation (K=4). The red rectangles represent the fold that is used as the test set for each iteration.

One potential pitfall of cross validation is that a different machine learning model is trained on each iteration. An important part of training a machine learning algorithm is determining the optimum features, also known as feature selection, and hyperparameters for the classification task. Every machine learning model has different hyperparameters, which are parameters that can take different values and influence the performance of the algorithm for a given task (Claesen and De Moor, 2015). Conducting feature selection and hyperparameter tuning outside of the cross-validation procedure can cause over-estimations of the predictive performance of a model (Vabalas et al., 2019). Therefore, it is difficult to generate a single model that can be applied to future data.

### 2.1.3 Training the model

Although there are many different types of supervised machine learning models, some models are more well known than others across disciplines, for example Random Forest, Support Vector Machines, Logistic Regression, K-Nearest Neighbour, and Neural Networks (Ray, 2019). Using well-known models to investigate the performance of new features for predicting the cause of TLOC, namely an automated analysis of spoken descriptions of TLOC, will increase the accessibility of the research between disciplines and provide a benchmark for performance that can be potentially improved upon by using more complex algorithms in the future. Unfortunately, it would not be feasible to evaluate the performance of neural networks for this classification task because neural networks need a large amount of training data (Alwosheel, Cranenburgh, and Chorus, 2018), but this is one approach that could be evaluated in the future if larger dataset are available.

#### 2.1.3.1 Random Forest

Random Forest (Breiman, 2001) is an algorithm that involves training many uncorrelated decision trees and subsequently making predictions based on the majority vote of all decision trees within the forest. A decision tree consists of a hierarchy of nodes where each node asks a question about a single feature and generates “child nodes” that correspond to the possible answers to that question (Figure 2.2). A “child node” can pose an additional question to further segment the responses or become a “leaf

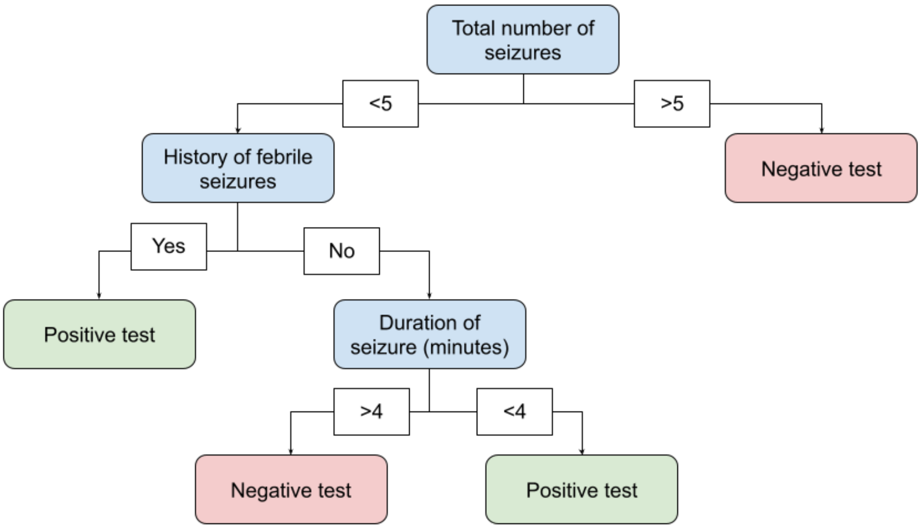


FIGURE 2.2: A basic example of a decision tree. The blue squares represent the features used to train the machine learning classifier. The white squares represent the possible response options from the training data. A cut-off threshold is decided by the algorithm for continuous values. The red (negative prediction) and green (positive prediction) squares represent the outcome nodes where a prediction is made using the portions of outcome values that are present in the node.

node" where the responses are allocated to a given prediction based on the most prominent prediction present within the leaf node from the training data. Predictions about the training set are made by passing each item through the decision tree and making a prediction based upon the resultant leaf node. In Random Forest, the correlation between each decision tree is reduced by using a random sample of training data points to create each decision tree and selecting from a random subsample of features at each node within the decision tree. Reducing the correlation between trees improves the predictive performance of the Random Forest algorithm (Breiman, 2001).

#### 2.1.3.2 Support Vector Machine

Support Vector Machine (SVM) is a method that generates a hyperplane that is able to separate independent classes and uses the hyperplane to predict the class of new data based upon the location of the datapoint in multidimensional space (Noble, 2006). SVM choses the discriminative hyperplane that maximises the distance between the hyperplane and each class. Separating the data using a hyperplane requires the dataset to be linearly separable, which is frequently not the case for complex datasets. SVM can account for linearly inseparable data using a soft margin or kernel function. A soft margin accounts for outliers by allowing some data points to be on the wrong side of the hyperplane without affecting the final result. The kernel function accounts for non-separable data by projecting the data into a higher dimensional space (Kim, Kavuri, and Lee, 2013) and generating a hyperplane.

#### 2.1.3.3 Logistic Regression

Logistic Regression is a statistical method frequently used in binary classification to model the probability of a given class (LaValley, 2008). The algorithm assigns a weight for each feature in the dataset and the weights are updated according to the training data. The input values and weights are combined linearly and passed through a logistic, sigmoidal function that converts the summated value into a value between 0 and 1. The values are assigned to a given class using a linear decision boundary of 0.5. Logistic regression assumes that there is a linear relationship between the input variables and output class; therefore, it can benefit from transformations that highlight the linear relationship.

#### 2.1.3.4 K-Nearest Neighbour

K-Nearest Neighbour (KNN) is a supervised machine learning algorithm that assigns a class based upon the similarity between the features in the test data and those in the training data (Altman, 1992). During training, each data point from the training set is mapped into a multidimensional space based upon the number of features. To test the algorithm, a test data point is projected into the same multidimensional space and the Euclidean distance between the test data point and all other data points is calculated. The algorithm selects a set number of the closest training data points to the test data, a value denoted K, and the most frequent class among these training data points is selected as the predicted class. KNN is a non-linear classification method because it does not use a linear hyperplane.

### 2.1.4 Evaluating the model

Machine learning models are evaluated based upon their ability to accurately predict the target classes when applied to data that was not used to train the algorithm (Tharwat, 2020). In the context of diagnostic technology, these measures aim to provide an indication of how accurately a model can predict a diagnosis in clinical practice. The fundamental underpinning of evaluating a machine learning model involves determining how much the model produces True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). These measures are used to provide insight into how effectively the model is performing. The accuracy of a model reflects the percentage of correct predictions made by the model. Sensitivity or Recall reflects the capacity of the model to correctly identify individuals as belonging to a given class, whereas specificity reflects the ability to correctly identify people that do not belong to a class. Precision refers to the percentage of positive predictions for a class that are correct. Finally, given that a model may be better at predicting the diagnosis of one class over another, the F1 measure provides an estimation of accuracy that takes into consideration any imbalances between the precision and recall of the algorithm. In instances where the model is good at predicting one class but worse at predicting another, the F1 score may be lower than accuracy because this imbalance is considered.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity\, or\, Recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FN}$$

$$F1\ Score = \frac{2\ x\ precision\ x\ sensitivity}{precision + sensitivity}$$

## 2.2 Speech processing and healthcare

Machine learning algorithms can be trained using different types of data. One type of data of particular interest in the healthcare domain is speech data. There is an abundance of research exploring the feasibility of predicting a diagnosis using an automated analysis of speech for a broad range of health conditions, for example cognitive decline (Mirheidari et al., 2016; Mirheidari et al., 2017a; Mirheidari et al., 2017b; Mirheidari et al., 2018; Mirheidari et al., 2019a; Mirheidari et al., 2019b; Pan et al., 2019; Pan et al., 2020; Pan et al., 2021; O'Malley et al., 2021; Petti, Baker, and Korhonen, 2020), Depression (Mundt et al., 2007; Valstar et al., 2013; Gratch et al., 2014; Low, Bentley, and Ghosh, 2020; Rana et al., 2019; Huang, Epps, and Joachim, 2019; Rutowski et al., 2019), Anxiety (Rutowski et al., 2020), Bipolar Disorder (Tondo, H Vazquez, and J Baldessarini, 2017; Ringeval et al., 2018; Matton, McInnis, and Provost, 2019; Pan et al., 2018; Wang et al., 2020), Parkinson's Disease (Sveinbjornsdottir, 2016; Moro-Velazquez et al., 2021), and Amytrophic Lateral Sclerosis (Kühnlein et al., 2008; Bandini et al., 2018; Vashkevich, Petrovsky, and Rushkevich, 2019; Wang et al., 2018; An et al., 2018). Recording speech data allows people to provide a breadth of information. There are a numerous processing steps required to collect speech data and convert a raw audio signal into a machine learning prediction. The remaining sections of this chapter will introduce some of the common processing steps and the common challenges associated with speech technology research using examples from research applying these techniques to the detection of a range of health conditions.

### 2.2.1 Collecting speech data

In order to use a 'healthcare application' in clinical practice, the procedure of recording speech data should be automated to allow the method to be used on a large scale. Although it is possible to investigate the feasibility of predicting a diagnosis by collecting speech recordings by recording human-human interactions (Salekin et al., 2018; Pan et al., 2018; Wang et al., 2020), there may be differences in the speech recordings that are collected through human-computer interactions, and these differences may cause changes to the performance of a system if the method of recruitment is changed as a research project advances. A review of 'healthcare applications' suggested that the success is dependent on three fundamental features of healthcare applications that contributed to their success: interactivity, context-awareness, and adaptiveness (Preum et al., 2021). However, there is a balance between designing the user interface to allow the collection of the necessary information while also managing the users expectations about what the 'healthcare application' is capable of (Mori, MacDorman, and Kageki, 2012).

The CognoSpeak project provides one example of a research project that automatically collected speech data (Mirheidari et al., 2017a). CognoSpeak is an application that is designed to identify the cause of a memory impairment by conducting a cognitive assessment using an automated analysis of speech. The objective was to create an application that could be used to guide referral pathways in Primary Care. People with reported memory impairments completed an interaction with a virtual agent through an online web application. There are other studies that used applications which provided the participant instructions, either in written or verbal format, and recorded spoken responses that were later analysed for the research project (Gratch et al., 2014; Mundt et al., 2007; Vashkevich, Petrovsky, and Rushkevich, 2019; Huang, Epps, and Joachim, 2019; Rutowski et al., 2019; Rutowski et al., 2020). Although some applications were context-aware because they monitored the verbal cues of the participant before asking a subsequent question (Gratch et al., 2014), many of the applications were not responsive to what the patient had said because they did not utilise spoken human understanding. However, the speech recordings collected through these applications were useful for predicting the cause of multiple different health conditions. These findings demonstrate that an automated method of collecting speech recordings does not have to be responsive to the patient to collect useful data.

### 2.2.2 Transcription

A transcript of the speech recording may be required for a machine learning algorithm depending on the features that are used to train the algorithm, for example if the features involve semantic content. Researchers can evaluate the performance of an automated analysis of speech using manual or automatically generated transcriptions. Although the use of manual transcripts provides insight into how effective the model performance on 'clean' research data, the performance metrics from these research studies may not generalise to "real life" where such systems would be reliant on automatic speech recognition (ASR). Although the word-error-rate associated with ASR can reduce the performance of a machine learning model (Mirheidari et al., 2016), these models can still exhibit an acceptable level of performance and the word-error-rate can be reduced when more research data is available to train the ASR system (Mirheidari et al., 2018). Therefore, it is important that feasibility judgements are not reliant on the performance of an ASR system because there is a lot of scope for improvement in future research.

### 2.2.3 Extracting machine learning features from speech

Speech is a continuous signal that contains linguistic and paralinguistic properties (Latif et al., 2020). The linguistic content describes the information or message that the speaker is trying to communicate to the listener. The paralinguistic content transcends beyond the semantic meaning embedded within the speech and contains a breadth of additional information that can be used to make inferences about the speaker or subtle forms of information contained within the semantic content, for example age, gender, identity, and emotional state. In order to support machine learning models to identify the relevant components of speech for a particular classification task, researchers extract sub-components of the speech signal, which are frequently described as features. Researchers can use domain expertise to determine which linguistic and paralinguistic properties of speech may be useful for a particular classification task. Most features that are frequently used can be separated into two categories: acoustic and linguistic features (Latif et al., 2020).

Acoustic features can be further classified into three categories: prosodic, spectral and temporal, and voice quality (Latif et al., 2020). Many acoustic features can be measured using open-source software, such as openSMILE (Eyben, Wöllmer, and Schuller, 2010). Some examples of the features openSMILE can measure include fundamental frequency, formants, Linear Prediction Cepstrum Coefficients (LPCC) (Gupta and Gupta, 2016), Mel-Frequency Cepstrum Coefficients (MFCC) (Chakraborty, Talele, and Upadhya, 2014), and Gammatone Frequency Cepstrum Coefficients (GFCC) (Shao and Wang, 2008). The target features are often calculated for each window of a segmented audio file and descriptive statistics are generated to represent each feature for the entire recording. These features have been effective for emotion recognition, and the detection of stress, suicidal behaviour, distress, Anxiety, Depression, Bipolar Disorder, Parkinson's Disease, and Amyotrophic lateral sclerosis (ALS) and cognitive decline (Mundt et al., 2007; Morales and Levitan, 2016; Sveinbjornsdottir, 2016; Pan et al., 2018; An et al., 2018; Salekin et al., 2018; Vashkevich, Petrovsky, and Rushkevich, 2019; Wang et al., 2018; Rutowski et al., 2019; Latif et al., 2020; Pan et al., 2021).

Linguistic features are often applied to transcriptions of the speech recording. Linguistic features often focus on the type of language used by people with a given health condition, for example measuring the proportion of key words associated with different semantic categories, speech rate, features designed to measure linguistic complexity, and the frequency of particularly part-of-speech labels (Matton, McInnis, and Provost, 2019; Mirheidari et al., 2019a; Wang et al., 2020). Furthermore, many recent approaches have expanded beyond the linguistic content of speech and started exploring the interactional properties of conversations using interactional or dialogue features, for example the number of turns that people take, how much they speak compared to other interactants, the length of turns in the interaction, and the characteristic of pauses (Mirheidari et al., 2017a; Wang et al., 2020). These examples are taken from studies that use speech to assist with the differential diagnosis of Bipolar Disorder (Matton, McInnis, and Provost, 2019) or the differential diagnosis of cognitive impairment (Mirheidari et al., 2017b; Mirheidari et al., 2017a), which demonstrates that similar speech processing features can be used for different classification tasks. Moreover they showcase the breadth of linguistic and interactional features that can be applied to speech processing research.

Many modern state of the art approaches do not rely on hand-craft features based on domain knowledge because there are methods of automatically generated machine learning features based upon contextual knowledge of the data generated by other machine learning models, also known as self-supervised machine learning. One of the most prominent examples is the BERT model (Devlin et al., 2018). BERT can be used to generate a numerical vector representation of individual words such that words with similar meanings have similar numerical vectors. These representations are generated using a masking method where words within a document are hidden from the model and all of the surrounding words are used to predict what the word is. Consequently, the context of language is used to generate the numerical representation of the word. BERT features have been shown to outperform handcrafted features in machine learning models designed to identify health conditions, for example cognitive decline (Liu et al., 2021). Consequently, self-supervised machine learning features have become more prevalent in machine learning research over recent years.

### 2.2.4 Designing a system based on Conversation Analysis findings

CognoSpeak is an application that is designed to identify the cause of a memory impairment by conducting a cognitive assessment using an automated analysis of speech. The methods used to design and create the application are similar to the objectives for this PhD thesis. For example, it is a tool that can be used to guide neurology referral pathways in Primary Care, it involves the differential diagnosis of a functional neurological disorder, and the research was built upon previous conversation analysis research. Therefore, the research outlines a suitable approach to testing the feasibility of using an online application and automated analysis of language for predicting the cause of TLOC.

The early research based on recorded interactions in a memory clinic identified distinct conversational profiles between people with Alzheimer's disease and people with functional memory impairment (Elsey et al., 2015). People with Alzheimer's disease showed a greater reliance on accompanying others to answer questions during the consultation, were less likely to recall recent instances of memory problems, showed difficulty in responding to compound questions, more frequently displayed an inability to answer questions, and were less likely to produce expanded or elaborated responses (Elsey et al., 2015). These findings were translated into features that could be automatically detected using computer software to explore the feasibility of creating a cognitive screening tool (Mirheidari et al., 2016). A linear SVM classifier trained using the same recordings that were used in the conversation analysis research was able to correctly predict the diagnosis with an accuracy of 92% (Mirheidari et al., 2016). The accuracy reduced to 79% when automatic speech recognition was used instead of verbatim transcripts.

The next stage in the project involved creating an application where the questions asked in Elsey et al. (2015) are posed by a digital avatar instead of a neurologist. Furthermore, the number of features was extended to incorporate lexical features that measured the proportions of different parts-of-speech and acoustic properties of the speakers (Mirheidari et al., 2017b). Although the baseline automatic speech recognition algorithm had a high word error rate, the system demonstrated a similarly high level of accuracy when predicting the diagnosis for participants who used the web application (Mirheidari et al., 2017b). Having demonstrated the feasibility of the method, the researchers have been able to improve the performance and utility of the system by using the DementiaBank research corpus to reduce the word-error rate, increasing the number of health conditions related to memory impairment that the application can detect, and exploring different types of data to collect, more complex features to use in the models, and more complex but effective machine learning algorithms (Mirheidari et al., 2018; Pan et al., 2019; Pan et al., 2020; O'Malley et al., 2021). This demonstrates that an initial system does not need to be perfect at conception because there is a large room for improvement once feasibility has been established.

### 2.2.5 Challenges in speech technology and healthcare

The previous section outlines many of the requirements of speech processing and machine learning research. Designing speech processing systems is a complex procedure with many different stages. Consequently, there are numerous challenges associated with this field of research that can impair the utility of the findings. The remaining section of this chapter will discuss some of these challenges.

The performance of machine learning algorithms is largely dependent on the data that is available to train and evaluate the model. Unfortunately, speech processing research in the healthcare domain is often associated with small sample sizes because of the challenges associated with approaching and recruiting participants (Latif et al., 2020). Data scarcity can impair the quality of research because some of the best methods are only feasible with large datasets, for example self-supervised approaches like BERT (Devlin et al., 2018). Furthermore, researchers must rely on cross validation to evaluate the performance of their models, but it is not possible to conduct an exhaustive exploration of potential features using cross validation without potentially introducing bias into the final evaluation metric (Vabalas et al., 2019). Finally, machine learning models tested on a small sample size may not have sufficient data to identify subtle patterns that are useful for making predictions, which can result in an under-estimation of the model performance. This can be especially problematic when conducting novel research because it may discourage future investment in potential solutions that have not been thoroughly tested. Fortunately, a small dataset may be sufficient to demonstrate the feasibility of a research project. For example, one early study exploring the feasibility of predicting depression using speech had a small sample of 35 participants (Mundt et al., 2007), but the advancement of this area of research has resulted in numerous open-source datasets (Rana et al., 2019) that provide researchers with easy access to data and a standardised dataset for comparing the performance of different features and machine learning algorithms. Therefore, early feasibility studies can direct further data collection and research in the future.

Speech processing research for healthcare is an interdisciplinary research field that sits between the disciplines of engineering, linguistics, and medicine. The contributions from all disciplines are vital for the success in this field. An in depth understanding of engineering and computer science is required to design and train robust models. Knowledge of the target health condition, clinical practice, and the patient population is required to design systems that can be integrated into the healthcare system. Research studies that are primarily conducted by one of the disciplines, for example engineers who solely focus on creating the most optimal predicting algorithm, may face barriers to clinical implementation. For example, clinicians reported that they would be less likely to use an algorithm designed to detect suicide risk if they were not able to identify which features were used to make the prediction (Brown et al., 2020). Therefore, clinician involvement in the design of research is important to ensure that it can translate into clinical practice.

There is overlap between the features that are used for different healthcare applications. One feature that is prominently used for the classification of a range of healthcare conditions are MFCC, which have been used for the detection of Depression (Rejaibi et al., 2022), Anxiety (Salekin et al., 2018), Bipolar Disorder (Pan et al., 2018), Parkinson's Disease (Moro-Velazquez et al., 2021), ALS (Vashkevich, Petrovsky, and Rushkevich, 2019) and cognitive decline (Mirheidari et al., 2020). It is not uncommon for individuals to have multiple comorbid health conditions. Although how these features are used for a particular model will depend on the training dataset and the particular classification task, research rarely explores whether the presence of comorbid health conditions is influencing the performance of the model or the presentation of the linguistic or acoustic features.

### 2.2.6 Conclusion

This short review of speech technology in healthcare demonstrates that there are particular characteristics of speech that are indicative of a broad range of health conditions. This is an important consideration when designing future speech applications because comorbidities are common between different health conditions. Although the objective of most of the research studies is to generate more efficient and effective methods of detecting a specific health condition, most of the research projects focus on collecting speech samples rather than on the creation of a speech technology application. Therefore, there is little emphasis on the design principles of the application, for example the interactivity, context-awareness, and adaptiveness (Preum et al., 2021). The Cognospeak project is one example of research that used an application to collect speech data and demonstrates that conversation analysis research can be translated into a system that conducts an automated analysis of spoken responses to detect similar group differences (Mirheidari et al., 2017a; O'Malley et al., 2021). These findings suggest that our objective of translating conversation analysis findings about the cause of TLOC could also be translated into an automated system.

Overall, these findings demonstrate that methods of automatically analysing speech to predict a health diagnosis are powerful and diverse. The breadth of methods and types of features that are available suggest that it may be feasible to apply these methods to spoken descriptions of TLOC. Furthermore, the quality of applications designed to predict a given diagnosis often improve as more research is conducted, suggesting that demonstrating the feasibility of the approach is only the first step to creating an effective system for predicting the diagnosis. The remaining elements of the PhD will explore the feasibility of predicting the cause of TLOC using an automated analysis of spoken descriptions based upon the methods outlined within this section.

# Chapter 3

# Exploring the diagnostic utility of an automated analysis of language

## 3.1 Introduction

The previous chapter provided an overview of the methods commonly used in machine learning and speech processing research and some of the common challenges. Training machine learning models using a small dataset was one of the most prominent challenges across different research projects. Unfortunately, the recruitment procedure for this project was hindered by the coronavirus pandemic. Prior to the pandemic, the recruitment design involved an opportunity to discuss the research project with patients while they were attending a clinic appointment for TLOC at the Royal Hallamshire Hospital. However, it was not possible to speak to patients about the project face-to-face during the pandemic because the TLOC clinic appointments were conducted remotely. Patients could only be approached by letter. It became apparent throughout the project that fewer participants would be recruited than originally anticipated. Choosing automatically detectable features to approximate the qualitatively described profiles for people with epilepsy and FDS requires a training dataset to evaluate the predictive performance of the features. The cross validation method is often used to provide an estimation of model performance when there is a limited amount of data available. However, it is difficult to explore diagnostically relevant features using cross validation without introducing bias into the performance metrics (Vabalas et al., 2019). Therefore, an exploration of potential speech-derived features that could assist the differential diagnosis of epilepsy and FDS was conducted using pre-existing data from previous conversation analysis (CA) research.

This chapter begins by providing an overview of the data from previous research that was used for this analysis (Section 3.2). The data was used to create different datasets designed to explore the two research objectives of the chapter. Firstly, we conducted an exploration of the feasibility of differentiating between epilepsy and FDS using an automated analysis of formulation effort (Section 3.3). Secondly, we explored the utility of multiple semantic categories that measured the proportion of semantically related words (Section 3.4). The findings and limitations from both analyses are then discussed in tandem (Section 3.5) because of the similarities between types of analysis and overlap between the datasets that were used.

### 3.1.1 Datasets

The analysis used data taken from previous conversation analysis research conducted at the Royal Hallamshire Hospital in Sheffield between 2005 and 2013. All of these recordings have been used in previous conversation analysis (CA) research

(Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008; Plug, Sharrack, and Reuber, 2009b; Plug, Sharrack, and Reuber, 2010; Robson et al., 2012; Robson, Drew, and Reuber, 2016; Jenkins et al., 2016). Participants who were currently under examination to determine the cause of their seizures at the Royal Hallamshire Hospital were eligible to participate in the previous CA studies. The final diagnoses for patients were determined using clinical assessment and/or a video-EEG recording of a typical seizure. Participants had not received a final diagnosis at the time of participation. Participants gave consent to be recorded and for the recordings to be used in future studies of communication.

The first dataset was taken from the first CA study that explored the interactional differences in conversations between a clinician and people with epilepsy or FDS in the UK (Schwabe, Howell, and Reuber, 2007). The interactions took place while patients were undergoing monitoring on a video-electroencephalography (EEG) unit. There were no accompanying people and the neurologist followed the conversational guidelines developed in the context of previous similar research in Germany (the Epiling project). These guidelines involved minimising their own input in the interaction, asking a standardised list of questions, and tolerating prolonged silences (Schwabe et al., 2008). Consequently, there are instances of exceptionally long pauses during the interactions and most of the talk is generated by the patient (Table 3.1).

The second dataset was taken from a CA study exploring interactional differences between people with epilepsy and FDS in routine clinical encounters (Robson et al., 2012). Routine clinical encounters follow a different interactional structure to the first dataset because some of the interactions include accompanying others and the neurologist will carry out a broader exploration of the patient's life and medical history, not only focussing on descriptions of their attacks (Cassell, 1985). The neurologists participating in this study did not receive instructions regarding how to communicate with the patients. Therefore, there are more instances of closed questions, for example "Were you confused afterwards?" (Ekberg and Reuber, 2015) (Table 3.1).

The final dataset was taken from a study exploring whether a one-day training course designed to train neurologists to detect the linguistic differences between people with epilepsy and FDS can improve communication and diagnostic predictions (Jenkins and Reuber, 2014). There were recordings from before the training course and after. Therefore, the neurologists followed the interview guidelines similar to those developed for the Epiling project in some recordings, whereas in others they did not. Furthermore, these interactions were still routine clinical encounters, so involved many of the characteristics of the second dataset. Finally, a medical TLOC diagnosis was only available for half of the dataset (Table 3.1).

The studies involved interactions with different clinicians and were conducted using different interview techniques and in slightly different context (for instance routine clinic encounters or research interviews). These differences influence the presentation of the linguistic profile (Ekberg and Reuber, 2015) and the amount of words that the patient spoke (Table 3.1). It was therefore important to consider carefully different methods of combining the datasets so that discrepancies between datasets would be minimised and the sample size maximised.

TABLE 3.1: The CA datasets from previous research. These datasets were used to form the datasets used in this chapter. The table contains descriptive information for each dataset.

| | **Schwabe, Howell and Reuber (2007)** | **Robson et al. (2012)** | **Jenkins and Reuber (2014)** |
|---|---|---|---|
| Average duration of recording (minutes) | 26.18 | 28.38 | 20.87 |
| Average number of words per patient | 2712 | 638 | 1556 |
| Number of patients | | | |
| Epilepsy | 7 | 20 | 17 |
| FDS | 13 | 12 | 9 |

## 3.2 Exploring the diagnostic utility of disfluency features in single seizure descriptions

### 3.2.1 Introduction

One of the most important differentiating features between the speech of people with epilepsy and people with FDS described in the qualitative studies mentioned above is the amount of formulation effort typically expended by patients when they describe their seizure experiences (Schwabe et al., 2008). In this context, formulation effort refers to the number and extent of hesitations, reformulations, restarts, repairs, and changes in grammatical construction (Schwabe, Howell, and Reuber, 2007). Whereas speech produced by people with epilepsy when describing their seizure experiences is characterised by a high level of formulation effort as they struggle to communicate how exactly they experience their seizures, formulation effort is largely absent from the seizure accounts of people with FDS (Schwabe et al., 2008). Hesitations are a prominent aspect of formulation effort and can be detected using automated acoustic language analysis (Liu et al., 2006; Christodoulides and Avanzi, 2015; Mirheidari et al., 2017a). For the first analysis within this chapter, we hypothesised that it is possible to automate the detection of hesitations as a marker of formulation effort in records of clinic conversations with seizures, and that our findings would replicate those previously achieved using qualitative analyses.

Another potentially automatable method for measuring formulation effort involves the identification and analysis of pauses within the interaction. Pauses could be an indicator of formulation effort because they may reflect the difficulties the patient is facing with the accurate description of their complex seizure experiences (Plug, Sharrack, and Reuber, 2009a). The automatic detection of pauses in speech has previously been used as an indicator of dementia (Mirheidari et al., 2017a; Sluis et al., 2020; Yuan et al., 2021). We hypothesised that the inclusion of one or several measures based on pauses would improve the classification performance.

In summary, the present analysis investigates whether features that can be automatically extracted from audio recordings and transcripts of speech as measures of formulation effort can be used to differentiate between epileptic and nonepileptic seizures. We hypothesise that it will be possible to differentiate between seizure accounts provided by people with epilepsy and people with FDS using automatically measurable markers of formulation effort. We will explore the classification performance of a combination of these features using the Random Forest algorithm. Furthermore, we will explore to what extent particular features contribute independently to the classification performance using independent comparisons between groups and exploring the performance of the algorithm using different combinations of features.

### 3.2.2 Method

#### 3.2.2.1 Preprocessing

People with epilepsy display increased formulation effort during descriptions of their subjective symptoms and the unconscious period of a seizure (Schwabe et al., 2008). To investigate whether features designed to measure formulation effort can be used to differentiate between people with epilepsy or FDS, we manually extracted the subsection of each interview in which the neurologist asked the patient to describe their first seizure because this created a dataset where all participants were asked the same question and the focus was on describing what happened during a seizure, which is similar to the questions posed by the virtual agent in our web application. Only extracts where the neurologist asked this question in an open-ended format were included because previous research has observed that the questions that neurologists ask in an outpatient setting can be more restrictive due to the time pressures associated with these interactions, and that this can reduce the presence of CA observations that are important for the differential diagnosis process (Ekberg and Reuber, 2015). Focusing on this question allowed us to create the largest possible corpus of interviews (total sample n=45 - FDS n=24, PWE n=21), while ensuring that patients have been provided with an opportunity to describe this particular seizure experience freely. We defined the end of the target subsection as the point when the neurologist either changed the topic or accepted a change in topic agenda (Fehlenberg, 1986) away from the first seizure by asking questions unrelated to this topic. Changes in topic agenda introduced by patients could be an example of resistance to the question being asked, which is a feature identified by previous CA research as indicative of an FDS description (Schwabe, Howell, and Reuber, 2007).

Audio-recordings were extracted from video recordings of the encounters, transcribed manually and further processed into Extensible Markup Format (XML). XML is a machine and human readable text format that is used to structure information. The transcripts were manually demarcated into individual turns within the conversation and each turn labelled with a speaker identifier, the start time and end time. The start and end time of the target subsection were noted and a new audio file consisting only of the target subsection was created using the AudioSegment function from Pydub (Hu and Wang, 2007). The raw text was converted to lowercase, punctuation and numerical digits were removed, contractions were expanded, and all words were converted to the corresponding lemma through lemmatization using a natural language toolkit (NLTK) in Python (Loper and Bird, 2002).

#### 3.2.2.2 Feature Extraction

Seven features were designed as markers of formulation effort (Table 3.2). Three of the features involved searching for a given word or word pair within the transcript.

TABLE 3.2: the seven formulation effort features and the corresponding description of each feature.

| Features | Definitions |
|---|---|
| Number of hesitations | The frequency of hesitations within the patient speech based on a pre-specified list of possible hesitations ("hm", "um"). |
| Number of Repetitions | The frequency that a word (N) is a repeat of the previous word (N-1) or the word before (N-2). |
| Presence or absence of keywords associated with uncertainty | A list of keywords associated with uncertainty was generated. This feature marked whether any of these words were present in the seizure description or not. |
| Pause Frequency | The frequency of patient pauses that were greater than 30ms in length. |
| Average pause length | The average length of all patient pauses that were greater than 30ms |
| Total pause time | The total time spent pausing when pauses were defined as being longer than 30ms. |
| Average length of between speaker pauses | The average length of all pauses (>30ms) that occurred during a transition between speakers (patient or doctor). |

These features were the total number of hesitations (e.g. "hmm" or "erm"), the total number of repetitions (e.g. "I I don't know") and the presence or absence of words that suggest uncertainty (e.g. "sort of" or "might"). Four features involved measuring pauses within the interaction. Pauses were detected using the WebRTC Voice Activity Detector from Google which checked whether each 10ms window contained speech or not. Only pauses greater than 30 milliseconds were included to minimise the inclusion of plosive phonemes. Pauses in the speech of patients (patient pauses) were identified using a manually created function that aligned each pause with the turn labels on the XML transcript. Between speaker pauses were defined as pauses that crossed the turn allocation boundary. The four pause features were the 'frequency of patient pauses', 'average length of patient pauses', 'total length of patient pauses', and 'average length of between speaker pauses'.

#### 3.2.2.3 Statistical Analysis

Group differences for each feature were compared using an independent t-test, Mann Whitney U test, or chi squared test as appropriate. The alpha level was set at 0.05. A

Bonferroni correction was performed to reduce the risk of a type 1 error and resulted in an adjusted alpha level of 0.007 (0.05/7).

#### 3.2.2.4 Classification

The Random Forest (Breiman, 2001) machine learning algorithm was used to investigate whether the features designed as markers of formulation effort were capable of differentiating between descriptions of epileptic seizures or FDS. Random Forest is an algorithm that involves training many uncorrelated decision trees and subsequently making predictions based on the majority vote of all decision trees within the forest. The correlation between each decision tree is reduced by using a random sample of training data points to create each decision tree and selecting from a random subsample of features at each node within the decision tree. Reducing the correlation between trees improves the performance of the Random Forest algorithm (Breiman, 2001). The Random Forest algorithm was trained by applying the nested "leave-one-out" cross validation method (Vabalas et al., 2019) and using the Scikit-learn toolkit in Python (Pedregosa et al., 2011). A search for the optimum hyperparameters for each cross validation fold was conducted using the "RandomizedSearchCV" function that explores 10 hyperparameter configurations based on the hyperparameters ranges outlined in Appendix A, Table A.1. The best configuration was selected based on the accuracy of the model that was trained using the training data for that specific fold. Only one machine learning classifier was evaluated for simplicity, but the intention to further evaluate other models was planned for future research.

### 3.2.3 Results

#### 3.2.3.1 Participants and seizure descriptions

A chi squared test of independence was performed to examine the relationship between gender and diagnosis. The relationship between these variables was significant, $X^2$ (1, N = 45) = 13, p <0.01. The FDS group included a higher proportion of women than the epilepsy group (women = 82.6% vs 23.8%). A Mann Whitney U test showed that there was no significant difference between the epilepsy and FDS groups in terms of vocabulary size (people with epilepsy median=101 vs. people with FDS median = 100, U=212.5, p=0.187) and word count (people with epilepsy median=257 vs. people with FDS median=210, U=187, p=0.071). A Chi squared test showed that there was no significant difference in word length distribution, $X^2$ (14, N = 45) = 15.2, p = 0.365.

#### 3.2.3.2 Feature Comparison

There were significantly more hesitations and repetitions in the speech of people with epilepsy than that of people with FDS (Table 3.3). There was no significant difference in terms of average pause length, pause frequency, total pause time, average length of between speaker pause and the presence or absence of key words associated with uncertainty (Table 3.3).

#### 3.2.3.3 Random Forest performance

We compared the performance of the Random Forest algorithm using different combinations of the features (Table 3.4). The best performance was achieved when all

TABLE 3.3: The mean (parametric tests) or median (non-parametric tests) for each variable.

| Features | FDS | Epilepsy | Test Statistic | P Value |
|---|---|---|---|---|
| Hesitations † | 2 (7) | 9 (10) | U = 117.5 | 0.001 |
| Repetitions † | 2 (2.25) | 3 (7) | U = 133.0 | 0.003 |
| Pause frequency | 45.4 (22.1) | 46.4 (26.9) | 0.135 | 0.893 |
| Pause average † | 0.996 (0.188) | 0.786 (0.385) | 159.000 | 0.018 |
| Pause total | 45.1 (24.5) | 43.1 (31.6) | -0.238 | 0.813 |
| Between speaker pause average † | 1.15 (0.544) | 0.922 (0.543) | 198.000 | 0.112 |
| Uncertainty keyword § | 13/21 (61.9%) | 10/24 (41.7%) | X2 = 1.115 | 0.291 |

Note: results indicate mean (SDs) unless otherwise indicated. Adjusted alpha set at p<0.007.

t value given unless otherwise specified

† Mann Whitney U, median, and Interquartile range are reported because the variable is not normally distributed

§ Chi squared test, count and percentage because the variable is categorical

formulation effort features were used (accuracy = 71%) (Figure 3.1), followed by hesitations and repetitions alone (accuracy = 68.9%), hesitations, repetitions, and the presence of uncertainty related words (accuracy = 64.5%) and all pause features (accuracy = 48.9%).

## 3.3 An exploration of the utility of semantic categories

Having established the performance of the formulation effort features, this section will now explore the predictive performance of the second feature set. The results of both analyses will be discussed in tandem in the final section.

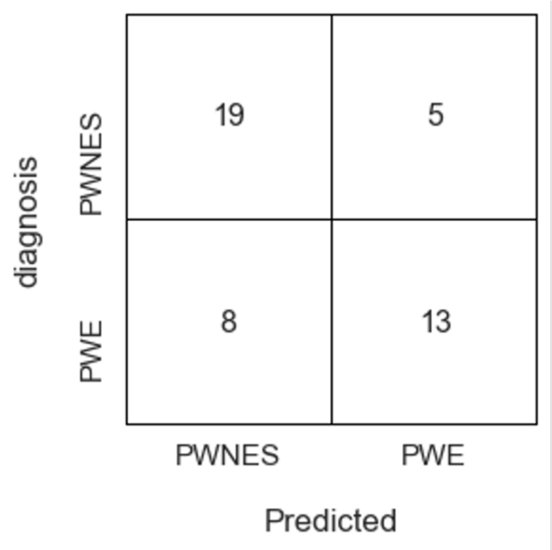
### 3.3.1 Introduction

The previous conversation analysis research identified differences in the semantic content of seizure descriptions between people with epilepsy and FDS. There were many instances where individuals used different words to describe their experience, for example people with FDS were more likely to use the term "blackout" to describe their chief complaint (Plug, Sharrack, and Reuber, 2010), use more catastrophising language when making third party references (Robson et al., 2012), describe their seizure as a "space/place" that they went to (Plug, Sharrack, and Reuber, 2009b), and produce more "complete negations" in their seizure accounts, for example "I do not remember anything" (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008).

TABLE 3.4: The accuracy, sensitivity, and specificity of the Random Forest algorithm trained using Leave-One-Out Cross Validation and different combinations of features

| Features | Accuracy | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|
| All features (7) | 71% | 61.9% | 79.2% | 67% |
| Hesitations & Repetitions (2) | 68.9% | 66.7% | 70.8% | 69% |
| Hesitations, Repetitions & Uncertainty (3) | 64.5% | 52.4% | 75% | 62% |
| Pause features (4) | 48.9% | 42.9% | 54.2% | 49% |

FIGURE 3.1: A confusion matrix for differentiating between people with epilepsy and people with FDS using the Random Forest model trained using all seven formulation effort features.



|  |  | Predicted |  |
|---|---|---|---|
|  |  | PWNES | PWE |
| diagnosis | PWNES | 19 | 5 |
|  | PWE | 8 | 13 |

Furthermore, people with epilepsy were more likely to recall more detailed descriptions of a single seizure experience, including the subjective symptoms that they experienced, whereas individuals with FDS were more likely to provide descriptions of what their seizures are 'generally' like rather than focusing on particular seizure events (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). These differences suggest that there may be semantic differences between the seizure descriptions of each group that can be automatically detected. However, these observations have previously only been detected by trained linguists who can detect these semantic differences in the appropriate context and ignore the semantic differences in other contexts, for example a linguist would avoid coding a complete negation used in a talk not focusing on a seizure description. Therefore, there are no guarantees that semantic differences between the groups will be diagnostically useful when detected using automated methods that do not consider the context.

An automatic analysis of language could include additional features beyond the scope of the original CA research. Previous research suggests that there may be differences in the use of emotive language between people with epilepsy and FDS. Firstly, dissociative seizures are automatic and uncontrolled responses to emotions, thoughts, sensations, or situations perceived as threatening (Brown and Reuber, 2016a). Given that emotions can be an antecedent to FDS, emotion-related words may be more prevalent during the seizure descriptions. Secondly, people with FDS report higher levels of general psychopathology (Brown and Reuber, 2016a), are more likely to experience panic symptoms during a seizure (Rawlings et al., 2017a) and catastrophize life experiences (Whitfield et al., 2020) than people with epilepsy. These findings suggest that emotional semantic content may be prevalent during medical interactions. Therefore, we would like to explore whether differences in the frequency of emotive language could contribute to the differential diagnosis of people with epilepsy and FDS.

One way to explore whether semantic differences between the two groups can assist the differential diagnosis is to measure the frequency of different semantic content within medical interactions. Linguistic Inquiry and Word Count (LIWC) is an application that processes text and measures the proportion of words that correspond to different semantic categories (Pennebaker, Francis, and Booth, 2001). Prior to the inception of the application, the researchers who created LIWC conducted multiple studies that explored relationships between the semantic content of language and different areas of psychology (Pennebaker, Mayne, and Francis, 1997; Pennebaker and Francis, 1996; Pennebaker, 1997; Pennebaker and King, 1999). LIWC was introduced as a tool that can be used by psychologists to conduct similar psychological research by analysing texts to identify psychological dimensions and predict behaviour using the different semantic categories within the application (Chung and Pennebaker, 2012). The words for each semantic category were generated based upon semantic observations from previous research, and a panel of judges rated the inclusion and exclusion of words iteratively until a final selection was achieved for each category (Pennebaker, Francis, and Booth, 2001). The categories were applied to a large corpus of data to identify and eliminate categories that were infrequently activated by the text (Pennebaker, Francis, and Booth, 2001). Furthermore, the external validity of the categories was explored by comparing the ratings with human raters for each category (Pennebaker and Francis, 1996; Alpers et al., 2005).

LIWC has been used to compare differences in language use by people with and without various psychiatric conditions. Anderson et al. (2008) found that people with social anxiety disorder more frequently use first person pronouns and words related to anxiety, sensory/perceptual processes, and physical touch, and make fewer references to other people while writing about a previously distressing social situation. Rosenbach and Renneberg (2015) found that people with borderline personality disorder used more first-person pronouns and words from the semantic categories anger, social, and family during the recollection of autobiographical memories. Multiple studies have reported an increase in first person singular pronoun use by people with depression (Holtzman et al., 2017). Furthermore, Shibata et al. (2016) compared word frequencies between nine people with Alzheimer's disease and nine healthy controls during medical interactions and found that people with Alzheimer's disease used fewer social references and employed the first person pronoun "I", verbs, and words in the present tense more frequently. These findings suggest that LIWC can be used to detect linguistic patterns associated with a specific health condition.

The portion of words for the semantic categories in the LIWC application could be used as diagnostic features to help differentiate between epilepsy and FDS using an automated analysis of language. However, it can be difficult to identify predictive machine learning features because the scarcity of medical speech data (Latif et al., 2020) makes it difficult for research to create a sufficiently large training and test dataset and often leads to an over-reliance on cross validation methods. Therefore, research must identify predictive features using pre-existing datasets that can be used in future research. Cardeña, Pick, and Litwin (2020) explored differences in the LIWC categories between people with epilepsy and FDS during a semi-structured interview and found that people with epilepsy used significantly more instances of "she/he", "we" and family references. Although they only found a small number of group differences, these semantic categories could still make predictive contributions to a machine learning algorithm, for example LIWC features were incorporated into a text classification algorithm for predicting comorbidities in people with epilepsy (Glauser et al., 2020). It is important to explore the predictive performance of features in addition to independent group comparisons because there is no single variable that can reliably separate epileptic and nonepileptic seizures (Reuber et al., 2016; Wardrope et al., 2020a; Avbersek and Sisodiya, 2010).

Therefore, the experimental work presented in this section is to explore the effectiveness of semantic categories from the LIWC application at predicting a diagnosis of epilepsy or FDS when applied to the history-taking phase of routine seizure clinic encounters. The focus was on semantic categories that align with the linguistic differences observed in previous differential diagnostic research.

## 3.3.2 Method

### 3.3.2.1 Preprocessing

In order to explore semantic differences between the two groups, we created a novel dataset that maximised the amount of patient talk and the similarities between the context of the interactions. The dataset was created using the recordings of doctor-patient interactions in routine seizure clinic consultations (Robson et al., 2012; Jenkins et al., 2016). The routine seizure clinic encounters were chosen because the interviews were similar, although not exactly the same, across the two datasets and allowed us to create the largest dataset. In contrast, the original CA studies were not included because they followed a dramatically different interview style, had a smaller number of participants, and solely focussed on the seizures (Schwabe et al., 2008). Two stages of the medical encounters, establishing the reason for the visit and the history taking (Robinson, 2003) were manually extracted from the whole interaction before the doctor started talking about the diagnosis. Although some of the recordings that were used in section 3.2 were included in this dataset, the recordings used in section 3.2 consisted of the response to a single question, whereas the recordings in this dataset consisted of longer interactions.

The dataset consisted of 58 manually transcribed recordings of encounters involving patients and neurologists in a routine seizure clinic setting. The neurologists in one group of the interviews took part in a training program aiming to enhance their ability to pick up interactional and linguistic differential diagnostic features during their clinic interactions with patients. During the training, they were instructed to ask participants about their first, most recent, and worst seizure, and encouraged not to interrupt patients during their narratives (Jenkins et al., 2016). The neurologists in the second group had received no instructions (Robson et al., 2012). Recordings were included in the analysis if a final medical diagnosis had been confirmed by review of all clinical data by an epileptologist or the diagnosis had been confirmed by the video-EEG recording of a typical seizure. Patients were only included if their final diagnosis was one of epilepsy (N=37) or nonepileptic seizures (N=21).

#### 3.3.2.2 Linguistic Inquiry and Word Count

The recordings of the doctor-patient interactions were manually transcribed. A manually created algorithm was used to extract the text that corresponded to all patient turns during the target subsection of the interaction. The most recent version of the LIWC application has a dictionary of almost 6400 words dispersed across 93 different semantic categories (Pennebaker et al., 2015). We used 21 semantic categories to measure the differences observed in previous research and any potential relationships between the independent categories.

The social environment influences the experience and recollection of a seizure differently for people with FDS or epilepsy. Wardrope et al. (2020b) found that the presence of other people at the onset of a seizure resulted in more attempts to alert others about the upcoming seizure, greater intensity, and different post-ictal behaviour for people with FDS compared to people with epilepsy. Previous research identified that people with epilepsy used more words related to the categories "We", "She/He" and "Family Reference" during semi-structured interviews about their seizures compared to people with FDS (Cardeña, Pick, and Litwin, 2020). Moreover, people with FDS are more likely to make catastrophising third party references, whereas people with epilepsy are more likely to make normalising third party references (Robson et al., 2012). Therefore, the first group of semantic categories measuring the frequency of social words were included ("We", "She/He", "Family", "Social", and "Affiliation"). The category "risk" was included to detect differences in the tendency to catastrophize (Whitfield et al., 2020), alongside "Cause" to detect inferences about the consequence of seizures on everyday life and "Reward" as a countermeasure.

There is evidence of differences in emotional experience and expression between people with epilepsy or FDS. People with FDS have increased levels of general psychopathology (Brown and Reuber, 2016a) and many theories suggest that emotional experiences may contribute to the manifestation of FDS (Brown and Reuber, 2016a). Furthermore, a thematic analysis of written accounts of epilepsy and FDS found differences in the "emotional tone" of the accounts (Rawlings et al., 2017c; Rawlings et al., 2017b). People with epilepsy often demonstrated relatively stable moods whereas people with FDS reflected greater anxiety and low mood. Therefore, seven categories were included to measure differences in emotive language ("Emotional tone", "Affect", "Positive emotions", "Negative emotions", "Anxiety", "Anger", and "Sad")

Previous CA research found that people with FDS display an increased tendency to talk about seizures in general rather than focussing on the description of a single past seizure experience and often conflated the unconscious period and the seizure experience by using complete negations to emphasise that they do not know anything about what happened (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). The categories "Focus Present" and "Quantifiers" were included to measure whether people were describing a single experience that happened in the past or multiple seizure experiences in the present tense, for example "I never remember what happened in the seizures" or "I sometimes lose consciousness", instead of "I was walking down the street and began to feel strange". Furthermore, the category

"Certainty" was used to capture holistic statements and complete negations, for example "I never remember anything".

Section 3.2 outlined and demonstrated that people with epilepsy exhibit more formulation effort during descriptions of the seizure experience (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). The semantic category "Tentativeness" was used to capture differences in tentative language associated with formulation effort, for example increased tentativeness during the description of subjective symptoms.

Finally, there is evidence that people with epilepsy or FDS use different metaphoric conceptualisations when describing what happened during a seizure. People with epilepsy more frequently used metaphors that conceptualise the seizure as an agent or force that can be combated, whereas people with FDS more frequently used metaphors that conceptualised the seizure as a space or place that they went to (Plug and Reuber, 2009). The two categories "Space" and "Power" were used to capture some of these differences. A full list of the categories available within the application and those selected for this analysis can be found in Appendix B.

We conducted group comparisons for each semantic category to gain insight into which categories may be the most effective for future research. The Shapiro Wilk test was used to test the normality of each variable, and the Levene test was used to test for homogeneity of variance. Group differences for each semantic category was calculated using an independent T-Test or Mann-Whitney U Test depending on whether the variables were normally distributed and if the samples had equal variances. No correction was made for multiple comparisons because of the exploratory aim of this study. Corrections for multiple comparisons would have increased the risk of making a type 1 error, which could prevent future researchers exploring variables that may improve the predictive performance of machine learning models.

### 3.3.2.3 Demographic and general speech differences

A chi-squared test of independence was performed to investigate whether there was a significant difference in gender between people with epilepsy and people with FDS. There was no significant difference between these variables $X^2$=(1,N=58) = 0.431, p=0.511.

A Mann-Whitney U test was used to evaluate whether there was a difference in the word count and total number of unique words per transcript for people with epilepsy and people with FDS. We found that people with FDS spoke more words in general (median=1445) compared to people with epilepsy (median=1115), U=(1,N=58) = 275, p<0.05. Moreover, people with FDS spoke more unique words (median=351) compared to people with epilepsy (median=271), U=230.5, p<0.01.

### 3.3.2.4 Classification

We explored the classification performance of multiple machine learning classifiers to identify the algorithms most suited to this classification task. The classification performance of the semantic categories was evaluated using five different machine learning models trained using the sci-kit learn toolkit in python (Pedregosa et al., 2011): Random Forest (Breiman, 2001), Support Vector Machine with either a linear or Radial Basis Function (rbf) kernel (Cristianini, Shawe-Taylor, et al., 2000), Logistic Regression, and K-Nearest Neighbor (Altman, 1992). Each model was trained using "leave-one-out" cross validation and a nested search for the optimum hyperparameters to prevent overfitting (Vabalas et al., 2019). A search for the optimum hyperparameters for each cross validation fold was conducted using the "GridSearchCV"

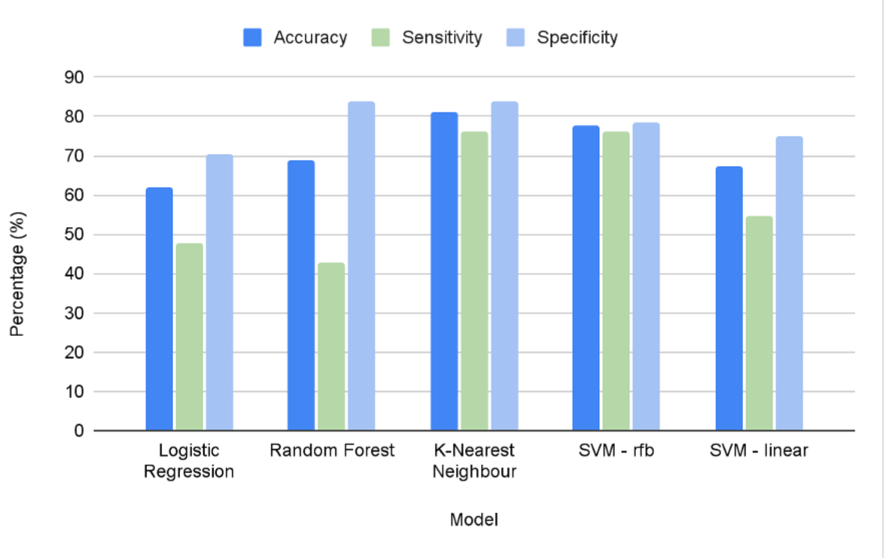


FIGURE 3.2: A comparison of the performance (accuracy, sensitivity, and specificity) of each of the machine learning algorithms using all 21 semantic categories.

function (Pedregosa et al., 2011) that explores all hyperparameter configurations based on the hyperparameters ranges outlined in Appendix A, Tables A.2, A.3, A.4, and A.5. The best configuration was selected based on the accuracy of the model that was trained using the training data for that specific fold.

The importance of each feature for the best performing machine learning models was determined using an ablation analysis where each feature was removed, and the classification accuracy of the model was recalculated. Features that resulted in the largest decrease in classification accuracy were considered the most important.

## 3.3.3 Results

### 3.3.3.1 Comparison of the semantic categories

There was a significant difference in 11 of the 21 LIWC variables (Table 3.5). The semantic categories with a significant group difference were “Negative emotions”, “Emotional tone”, “Quantifiers”, “Focus present”, “Sad”, “Reward”, “Anger”, “Family”, “Power”, “Cause”, and “Affiliation”.

### 3.3.3.2 Classification Performance

The results of the classification analysis demonstrate a large degree of variation between the five classifiers (Figure 3.2). The best performance was demonstrated with the three non-linear classifiers, which were the K-Nearest Neighbour classifier (accuracy = 81%), the Support Vector Machine with a RFB kernel (accuracy = 77.6%) and the Random Forest algorithm (accuracy = 69%). The two classifiers that use a linear operation performed less effectively (Support Vector Machine with a linear kernel, accuracy = 67.2%, and the Logistic Regression algorithm, accuracy = 62.1%). All classifiers demonstrated a greater specificity (70.3-83.8%) than sensitivity (42.9-76.2%).

TABLE 3.5: The mean and standard deviation (parametric) or median and interquartile range (non-parametric) of the percentage of words per semantic category for people with epilepsy and people with FDS. The test statistic and p-value are reported for each group comparison, unless otherwise specified
† - Mann Whitney U, median, and Interquartile range are reported because the variable is not normally distributed or the homogeneity of variance assumption was violated.

| Semantic category | People with epilepsy | People with FDS | Test Statistic | P Value |
|---|---|---|---|---|
| Negative emotions † | 1.04 (1.17) | 1.73 (0.57) | M=210 | P < 0.01 |
| Emotional tone † | 28.8 (28.35) | 21.95 (7.83) | M =216 | P<0.01 |
| Quantifiers † | 1.7 (0.91) | 2.19 (0.4) | M = 229 | P < 0.01 |
| Focus present † | 10.5 (3.54) | 12.94 (2.64) | M = 230.5 | P <0.01 |
| Sad † | 0.21 (0.32) | 0.38 (0.27) | M = 231.5 | P <0.01 |
| Reward | 1.035 (0.61) | 1.399 (0.44) | T = -2.356 | P <0.05 |
| Anger † | 0.05 (0.2) | 0.18 (0.23) | M = 271.5 | P <0.05 |
| Family † | 0.26 (0.37) | 0.35 (0.69) | M = 272 | P < 0.05 |
| Power | 1.189 (0.55) | 1.489 (0.38) | T = -2.181 | P < 0.05 |
| Cause | 1.184 (0.66) | 1.568 (0.59) | T = -2.179 | P < 0.05 |
| Affiliation † | 0.52 (0.52) | 0.71 (0.33) | M = 285.5 | P <0.05 |
| Space | 5.517 (1.74) | 6.254 (1.05) | T = -1.736 | P = 0.08 |
| Social | 5.638 (2.07) | 6.7 (1.97) | T = -1.879 | P = 0.07 |
| Risk † | 0.39 (0.39) | 0.48 (0.41) | M = 286.5 | P = 0.05 |
| We † | 0.06 (0.2) | 0.14 (0.29) | M = 291 | P = 0.05 |
| SheHe † | 0.72 (1.1) | 0.9 (1.09) | M = 289 | P = 0.05 |
| Affect | 2.76 (0.91) | 3.09 (0.64) | T = -1.431 | P = 0.16 |
| Positive Emotions † | 1.54 (0.63) | 1.4 (0.42) | M = 313.5 | P = 0.11 |
| Anxiety † | 0.29 (0.46) | 0.31 (0.38) | M = 318.5 | P = 0.13 |
| Certain | 1.62 (0.77) | 1.489 (0.59) | T = 0.663 | P = 0.51 |
| Tentativeness † | 2.78 (1.34) | 2.65 (1.47) | M = 388 | P = 0.5 |

#### 3.3.3.3 Most Important Features

The most important features were calculated for the K-Nearest Neighbour model and the SVM model with the RFB kernel. The top nine most important features determined from the ablation analysis were: "Focus present tense", "Emotional tone", "Tentativeness", "Quantifiers", "Reward", "Social", "Affect", "We" and "He/She". The subsequent four features ("Positive emotion", "Family", "Cause", and "Affiliation") had the same importance score. Furthermore, four out of the top nine features were among the features within our analysis showing significant group difference in the single item comparisons, excluding "Tentativeness", "Social", "Affect", "We", and "He/She". The Support Vector Machine with a RFB kernel appeared to be the most stable machine learning algorithm because there was no change in the accuracy of the model when 13 out of the 21 features were removed independently, suggesting that this model is less reliant on individual features, whereas K-Nearest Neighbour demonstrated the most and largest changes due to the removal of features (Figure 3.3).

## 3.4 Discussion

Our preliminary results support the hypothesis that it is possible to differentiate between individuals with epilepsy and FDS using an automated analysis of spoken seizure descriptions. Although the features used in this analysis do not take into account the context of what is being said, for example whether the patient is currently describing a subjective symptom (Schwabe et al., 2008), they are still able to predict the diagnosis with a relatively high level of accuracy.

Our first objective was to explore whether features designed to measure formulation effort can be used to differentiate between people with epilepsy and FDS when directly applied to a description of a seizure. In accordance with previous research (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008), we found that people with epilepsy demonstrated significantly more formulation effort as indicated by more hesitations and repetitions while describing their first seizure compared to people with FDS. Furthermore, although the frequency of hesitations and repetitions made the greatest contribution to the overall classification performance, including features that recorded how patient's paused during the seizure description and whether they used key-words that indicated uncertainty improved the overall performance from 68.9% to 71% using the Random Forest algorithm. These findings suggest that measures of formulation effort should be incorporated into an automated analysis of spoken seizure descriptions and can assist the differential diagnosis when applied to the entire seizure description.

Our second objective was to explore the feasibility of differentiating between spoken accounts of epileptic seizures and FDS by comparing semantic differences between the words patients use in routine medical encounters. Using 21 semantic categories measured using the Linguistic Inquiry and Word Count (LIWC) application (Pennebaker, Francis, and Booth, 2001), we were able to accurately predict the diagnosis of epilepsy or FDS in up to 81% of cases. Our ablation analysis demonstrated that some features were making larger contributions to the performance of the algorithm than others, for example verbs in the present tense and words related to the emotional content. These findings provide insight into the most useful semantic differences between the two groups that could be included in our future research.

Our third objective was to explore the performance of different machine learning algorithms (Ray, 2019) to decipher those that appear to be the most effective for this classification task. We compared the performance of multiple machine learning algorithms using the semantic features because this analysis was larger in terms of the number of participants and features. The outcome of this analysis can guide our choice of machine learning algorithm for the automated analysis of language in our application. The largest accuracy was achieved using the K-Nearest Neighbour algorithm (Altman, 1992). However, the algorithm was better at predicting a diagnosis of FDS compared to epilepsy and showed large decreases in performance when individual features were removed using the ablation analysis, suggesting the performance of the algorithm was largely dependent on the features. In contrast, the Support Vector Machine algorithm with a non-linear kernel (Kim, Kavuri, and Lee, 2013) had the second highest accuracy but was equally good at predicting epilepsy and FDS and was less influenced by individual features during the ablation analysis. Therefore, these findings suggest that Support Vector Machine may be a more stable model to use in future research.

Previous qualitative research reported no difference in pause frequency and pause duration between people with epilepsy and people with FDS (Walker et al., 2020). Our findings that there was no significant difference in 'patient pause frequency', 'total pause time', 'average pause length', and 'average length of between speaker pauses' supports this finding. However, we observed an improvement in the performance of the Random Forest algorithm (Breiman, 2001) when the patient pause features were incorporated into the model. Similarly, although we observed a significant group difference for 11 of the 21 semantic categories, only four of these variables were in the top ten performing features for the two best performing models. This demonstrates that features may be effective for predicting the diagnosis without distinct group differences, potentially due to complex relationships between different linguistic and interactional features.

An interesting observation from this analysis is that people with FDS typically said more during the routine clinical consultations than people with epilepsy. This finding was unexpected because previous research has found that people with FDS typically provide less detailed descriptions of their seizures and are more likely to use complete negations instead of describing their seizure experiences more precisely (e.g. statements like "I don't remember anything") (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). One potential explanation is that the wider history taking procedure also involves conversations about other areas of the patient's health, for example conversations about the consequences of seizures, the impact they have on the patient's life, potential causes of the seizure, and information about previous medical interactions and other comorbid health conditions (Cassell, 1985). This finding may be important for designing a fully automated system because it demonstrates that people with FDS can have a lot to say during medical interactions and that questions focusing on what happened during the seizure may not capture all the variations in the responses that allow these semantic categories to perform effectively in the machine learning models.

Our findings provide the basis of a more detailed exploration into possible contributions a fully automated analysis of language could make to the differentiation of epilepsy and FDS. We note that the discrimination we achieved by automated analysis of manually produced transcripts and audio clips was less accurate than the fully manual, qualitative approach based on the analysis of complete interactions and taking account of a wider range of potentially diagnostic features (Reuber et al., 2009; Cornaggia et al., 2012; Papagno et al., 2017; Yao et al., 2017; Biberon et al., 2020). However, our study provides proof of principle that qualitatively described features can be translated into observations which can be made by a computer.

The findings of this study provide encouragement for efforts to develop equivalent methods for other discriminating qualitative observations such as differences in the metaphoric conceptualisations of seizure experiences preferentially used by people with epilepsy and people with FDS (Plug, Sharrack, and Reuber, 2009b), or the extent to which subjective seizure experiences are volunteered, and how periods of unconsciousness are described (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008).

In clinical practice, a TLOC stratification tool would be unlikely to be based on the predictive performance of language features alone as in this chapter. In a clinical system, these features could be used alongside symptom checklists to train a classifier (which may also be more diagnostic if used with a Random Forest Classifier than regression based approaches) (Wardrope et al., 2020a). Future research should therefore explore the performance of a classifier trained using these features in tandem. Another reason for such a combined approach is that a clinical TLOC classification tool should not only be capable of predicting likely diagnoses of epilepsy and FDS but also of syncope (Wardrope, Newberry, and Reuber, 2018). While little is known about the typical linguistic and interactional profile of patient descriptions of syncope, as stated above, this cause of TLOC can be differentiated very well from the two types of seizure based on symptom checklists.

The inclusion of language features into a fully automated clinical decision or stratification tool will require the use of an automatic speech recognition module. Although such systems will generate transcripts that are far less accurate than the manually produced transcripts used in this study, experience with a fully automatic "digital doctor" system, programmed to ask patients questions about memory problems and analyse their answers, suggests that remarkably high correct classification levels can be achieved with erroneous transcripts (Mirheidari et al., 2019b; O'Malley et al., 2021). While the switch from a conversation between clinician and patient to one between a talking head on a computer screen and the patient is likely to have significant consequences on how patients speak about their seizures, there are many similarities between the speech of patients between these two contexts (Walker et al., 2020), so this aspect of automation may actually improve the diagnostic accuracy of a fully automatic classification system.

### 3.4.1 Limitations

Firstly, the features used to approximate formulation effort may not capture all instances of formulation effort within the data. The features used in our analysis may suggest that patients are having difficulty describing their seizures by hesitating more, but another way that people can express formulation effort is by using metatalk (Schiffrin, 1980) to describe their difficulties explaining their seizures (Schwabe, Howell, and Reuber, 2007).

Secondly, the sample size was small and the contexts in which the spoken seizure descriptions were recorded were heterogeneous. Although we used cross validation to demonstrate the machine learning algorithm's ability to generalise to unseen data and to accommodate the analyses to the small dataset available for this analysis (Berrar, 2019), it is difficult to evaluate the variance of a machine learning model using the "leave-one-out" cross validation method. Therefore, a larger sample size would be required before we can be confident that this level of performance will be exhibited across other datasets. Finally, the analysis does not consider the type or severity of the seizures (Fisher, 2017) and future research should explore whether this influences the level of formulation effort that patients exhibit.

Thirdly, the LIWC categories were designed to measure these semantic concepts broadly and are not tailored to seizure consultations. They are not able to measure important semantic categories associated with seizure descriptions, for example the label used for the chief complaint (Plug, Sharrack, and Reuber, 2010) or descriptions of subjective symptoms (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). They contain many keywords unrelated to seizure consultations. It may be possible to generate semantic categories that are customised for seizure consultations that are better able to differentiate between people with epilepsy or people with FDS, but future research would need a larger dataset to detect the broad range of words used for these semantic categories.

Fourthly, the questions that the neurologist asked the patient were not standardised. Although some of the neurologists had received instructions about what questions to ask patients as part of the original research project (Schwabe, Howell, and Reuber, 2007; Jenkins et al., 2016), these instructions did not apply to the whole history taking procedure and the neurologists whose consultations were studied in the other project had received no instructions (Robson et al., 2012). Future research should explore semantic differences in interactions where every participant is asked the same question because this may change what words people use in the interaction.

Fifthly, the semantic analysis focuses on independent words and does not consider the wider context of keywords within the talk. The LIWC may not be as effective at identifying semantic constructs compared to human raters because people are able to label a whole segment of text as corresponding to a construct (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008), whereas LIWC only detects the keywords from that segment (Alpers et al., 2005). There are more complex, non-linear, machine learning algorithms that can process a segment of text rather than a single word, for example recurrent neural networks with long short-term memory (Sundermeyer et al., 2013), that could be used to overcome these limitations. However, these methods typically require larger datasets to be effectively used.

Finally, this analysis used manual transcripts instead of automatic speech recognition. Although this allowed us to test the proof-of-principle of this method, automatic speech recognition would be required for an automatic stratification tool and a small proportion of words will be misidentified due to the associated word-error rate (Mirheidari et al., 2016), which may change the predictive accuracy of this method.

### 3.4.2 Conclusion

The research outlined in this chapter has demonstrated that an automated analysis of TLOC descriptions can be used to differentiate between people with epilepsy or FDS reasonably effectively. The subsequent chapters will explore how effective these features are at predicting the cause of TLOC when applied to spoken descriptions of TLOC that are collected through an online web application, and evaluate whether these features can improve the predictive performance of the iPEP. The application of these features to a novel dataset will provide some insight into the reliability of these findings, but it is important to consider how people communicate with a virtual agent to ensure that their responses are qualitatively similar to those observed during doctor-patient interactions. The subsequent chapter will begin this exploration by introducing the online web application.

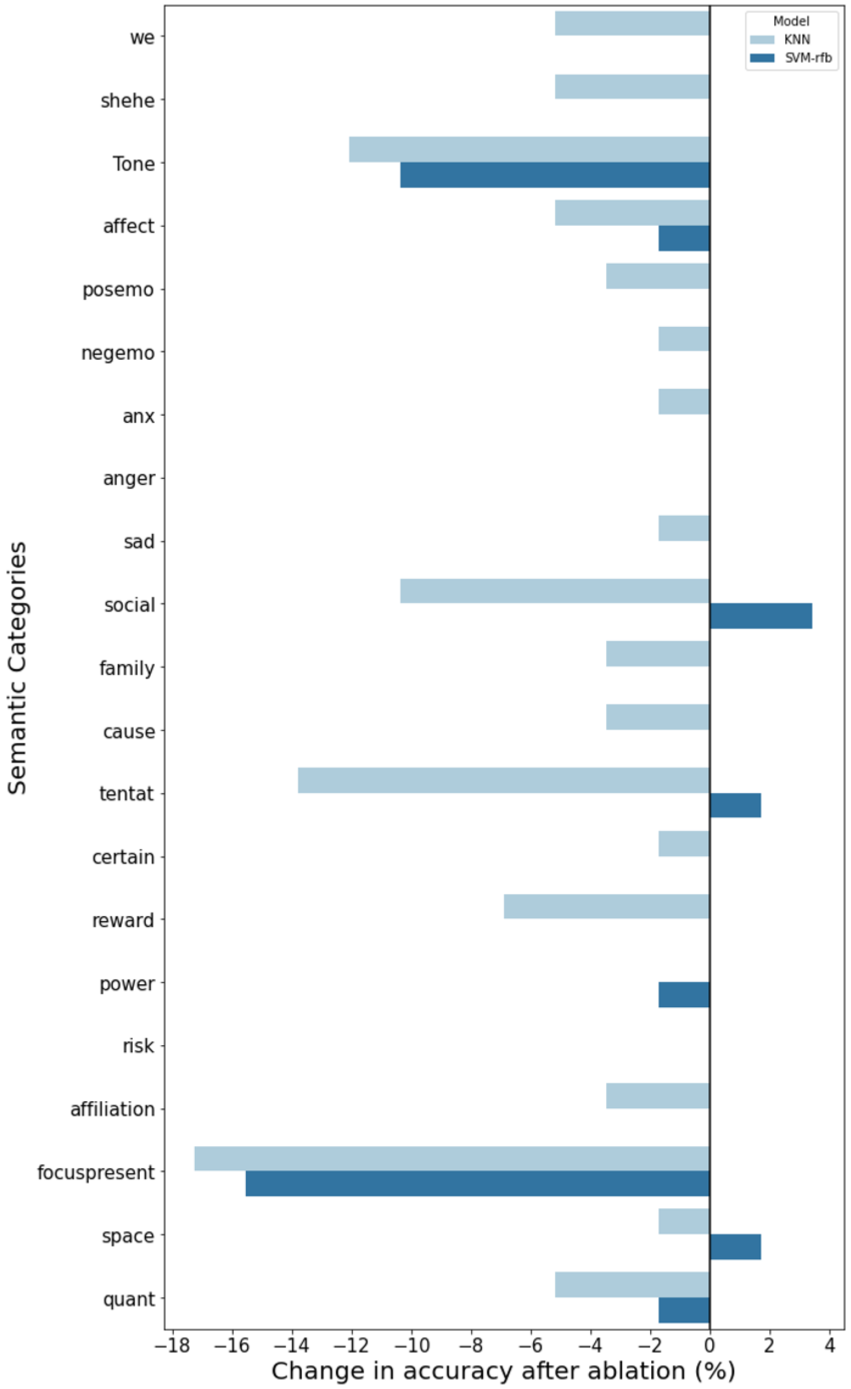


FIGURE 3.3: The change in classification accuracy (x-axis) for each classification model (hue) when each feature (y-axis) is removed from the analysis independently.

# Chapter 4

# The Web Application and Recruitment

The data for this thesis was collected using an online web application created during the first year of the PhD. The web application consists of a NodeJS web server hosted on a virtual machine provided by The University of Sheffield. The application is a prototype which allows the research team to collect the necessary data and receive feedback from the users regarding the functionality of the application to guide future web development projects. The purpose of this chapter is to outline the web application, how it was used to collect the research data, the type of data that was collected, and the participants who were recruited.

## 4.1 Ethics

This research project was approved by the East Midlands - Leicester South Research Ethics Committee on 07/05/2020 and received approval by the Health Research Authority and Health and Care Research Wales on 13/05/2020 (*REC reference: 20/EM/0106*). There were three substantial amendments to the research project to improve recruitment. Firstly, permission was granted (30/11/2021) to incorporate a monthly raffle for a £10 Amazon voucher for participants and to extend our recruitment to patients who had already received a diagnosis previously. Secondly, permission was granted (17/03/2021) to send recruitment letters to patients who had received a "gold-standard" diagnosis from the Royal Hallamshire Hospital between 2012-2016. Thirdly, permission was granted (08/09/2021) to promote and recruit through charities designed to support individuals with TLOC. These participants would be required to self-disclose their diagnosis.

## 4.2 Recruitment

Patients receiving treatment at the Royal Hallamshire Hospital received a letter inviting them to participate in the project, either when they were sent a reminder about a clinic appointment or unrelated to an upcoming appointment for those with a "gold-standard" diagnosis. The letter included a copy of the patient and witness information sheets. Information about the study was also posted through various communication channels by the following charities supporting individuals with TLOC: STARS, Epilepsy Action, FNDHope, FNDAction, Epilepsy Sparks, and the shape network by Epilepsy Research UK. A summary of the project and links to the information sheets and "consent-to-contact" form were hosted on the charities' websites to allow potential participants to sign up. Individuals who were interested in the project were asked to complete an electronic "consent-to-contact" form. They were then contacted to discuss the research project, assess eligibility, and enquire whether there is a witness who would also like to participate in the study. Witnesses were contacted separately to discuss the project. Participants were informed that the objective of the online application was to predict their diagnosis for the purpose of a research study aiming to improve future referral pathways and that their participation would not have any effects on their current care. Those who agreed to participate were sent instructions on how to access the application and the patient (and/or witness) login details via email.

## 4.3 Electronic Consent Form

The consent form for the project was the first web page on the application. Different versions of the consent form were presented depending on the type of participant - patient, witness, recruited through the Royal Hallamshire Hospital or externally. Participants were required to fill in and sign the consent form before completing the procedure. The prompts on the consent form required participants to type their initials, click the appropriate response, and sign using a typed signature. The consent form asked participants if they wanted to provide feedback on the application in the form of a feedback questionnaire, by interview, or both.

## 4.4 Questionnaires

All participants were asked to provide their age, gender, ethnicity, and highest level of education using free typing text boxes. Patients were subsequently presented with an "attack history" questionnaire (Table 4.1). The next stage required participants to complete the iPEP (Wardrope et al., 2020a). The iPEP consists of sets of different binary ("Yes" or "No") questions for patients and witnesses (Figure 4.1). Patients were asked 42 questions about their medical history and symptoms (Table 4.2). Witnesses (if available) were asked 10 questions about what they observed during the attack (Table 4.3). The questions were the same as those used in previous research using the iPEP (Wardrope et al., 2020a).

A different combination of the iPEP questions are used to train the patient only iPEP model and the patient and witness iPEP model (Wardrope et al., 2020a). The patient-only iPEP model uses a dataset of 34 patient directed questions. The patient and witness iPEP model uses a dataset of 26 patient directed questions and 10 witness directed questions. The patient directed questions that are used in each model are different (Table 4.2).

Up until this point, the term "iPEP" has been used to refer to the questionnaire and associated machine learning models from previous research (Wardrope et al., 2020a). This could become a source of confusion upon introduction of the online web application. Therefore, throughout the remaining chapters, the iPEP dataset from the original research will be referenced as "iPEP (original)" and the iPEP dataset collected through the online web application will be referenced as "iPEP (application)". Machine learning models that are trained using the patient only responses or patient and witness responses will be referenced as "patient-only" or "patient and witness", respectively (e.g. patient-only iPEP (application)). Furthermore, the iPEP can be referenced in three ways: questionnaire, responses or dataset, and model. The term "questionnaire" will refer to the questions that were asked, for example the "patient-only iPEP questionnaire" references the 34/42 questions that are relevant for the patient-only analysis. The terms "responses" or "dataset" will reference

TABLE 4.1: The attack history questionnaire presenting to patients

| **Attack History Questionnaire** |
|---|
| How old were you when you had your first seizure? |
| How many years have you been having a seizure for? |
| How many seizures have you had in the last year?<br>None<br>Up to 5<br>Up to 50<br>More than 50 |
| How many times have you been to hospital due to a seizure?<br>Never<br>Once<br>Up to 5<br>More than 5 |
| Have you been to intensive care due to a seizure?<br>No<br>Yes |
| Do you have a family history of seizures?<br>No<br>Yes |

the questionnaire responses, for example the "patient and witness iPEP dataset" references all of the responses to the 26 patient direct questions and 10 witness directed questions. Finally, the term model references the machine learning model that was trained on a particular dataset, for example the "patient-only iPEP model" references a machine learning model trained on the patient-only iPEP dataset.

TABLE 4.2: The iPEP questionnaire presented to patients, taken from Wardrope et al. (2020a). Different questions are used for the patient-only and patient and witness iPEP models. Whether or not a question is included in a given model is marked using an X.

| **Questions** | **Patient** | **Patient/ Witness** |
|---|---|---|
| Did you have Febrile seizures in childhood? | | X |
| Do you suffer from chest pain or tightness? | X | X |
| Do you have a brain tumour? | | X |
| Have you had a head injury with loss of consciousness? | X | |
| Do you get palpitations (sudden fast heart beats)? | X | X |

Continued on next page

**Table 4.2 – continued from previous page**

| **Patient iPEP** | **Patient** | **Patient/ Witness** |
|---|---|---|
| Do your arms and legs jerk briefly when drifting off to sleep? | X | |
| Do your arms and legs jerk at other times? | X | X |
| Do you have spells in which you go light-headed? | X | |
| Do you suffer from poor coordination)? | X | X |
| Do you get breathless unrelated to exercise? | X | |
| My attack came on when I was asleep | X | X |
| The site of blood or needles triggered my attack | | X |
| My attack was associated with sitting or standing for a long time | X | X |
| My attack was associated with emotional stress | X | X |
| In my attack I seemed to be controlled by someone outside of me | X | X |
| In my attack I had a sense of feeling as if I have seen something before when I knew I had not | X | X |
| I felt hot or cold in my attack | X | X |
| During my attack I smelled things that were not really there | X | X |
| During my attack I could see or hear the people near me | | X |
| In my attack I was conscious but could not react to things | X | X |
| I was aware of shaking uncontrollably during the attack | X | X |
| My attack made time go in slow motion | | X |
| During my attack I had memories of a past bad experience which I could not stop | | X |
| During my attack I was frightened that I was going to die | X | X |
| My attack was like a burst of electricity in my brain | X | X |
| My attack was painful like a hammer blow | X | X |
| My attack felt like a knife through the head | X | X |
| I wanted to know what had happened when I had blacked out | | X |
| After my attack I felt relieved | | X |
| My attack built up gradually | X | |
| In my attack I had a sense of feeling as if I'd never seen something before when I knew I had | X | |
| In my attack I felt sick | X | |
| In my attack I experienced tingling or numbness in my skin | X | |
| During my attack I heard things which were not really there | X | |

Continued on next page

**Table 4.2 – continued from previous page**

| Patient iPEP | Patient | Patient/ Witness |
|---|---|---|
| In my attack my mouth went very dry | X | X |
| In my attack I drifted in and out of consciousness | X | |
| During my attack I felt as if I was outside my body | X | |
| In my attack I felt like I was choking or very short of breath | X | |
| I woke from my attack with a cut tongue | X | |
| After my attack my muscles ache | X | |
| After my attack I felt very confused | X | |
| Afterwards I did not know I had had an attack | X | |

## 4.5 The virtual agent

The virtual agent (VA) is a method for participants to provide a spoken description of what happened during the most recent attack. The questions asked by the VA were designed to mirror the questions typically asked during routine epilepsy clinic consultations. Upon loading the page, the application asks for permission to access the microphone and camera. Participants are presented with a video containing a VA (Figure 4.2). Participants are instructed to play the video. Upon completion, the application begins recording the participant and they are instructed to speak their response to the question. Participants are required to press a "next" button when they are finished answering and another question is loaded. There are eight questions for patients (Table 4.4) and four questions for witnesses (Table 4.5). Participant recordings were securely sent to and stored on the web server hosted by the University of Sheffield before they were securely transferred to a shared research storage area.

## 4.6 Feedback questionnaire

A feedback questionnaire for the application was presented after the VA if they selected this option on the consent form (Table 4.6). Participants responded to each question by selecting the appropriate answer and pressing "next". The questionnaire is based upon the Technology Acceptance Model (TAM) (Davis, 1989), which postulates that a range of factors influence a user's decision to use a new form of technology, for example usefulness, ease of use, and attitude. The model has been validated for many different forms of health technology (Holden and Karsh, 2009). The questions were modified from previous research that aimed to evaluate participants' acceptance of hypothetical diagnostic technology (Lanseng and Andreassen, 2007). However, the additional construct "trust towards provider" that was introduced in the original paper was not incorporated because the application is purely theoretical and trust would depend upon who the service provider was in the future. Further information about TAM will be discussed in chapter seven.

TABLE 4.3: The iPEP questionnaire presented to witnesses, taken from Wardrope et al., 2020a

| **Witness iPEP** |
|---|
| The attack involved chewing, smacking or licking movements of the mouth and lips |
| The attack involved scratching or bicycling movements of the legs |
| The attack involved fiddling, picking, or fumbling movements of the hands |
| In the attack the head moved rapidly from side to side |
| The attack involved violent movements of arms and legs |
| During the attack the arms and legs were limp |
| During the attack the arms and legs were rigid |
| Shaking of the arms and legs went on for over 1 minute |
| The attack involved movements into unusual positions |
| The skin or lips looked pale during the attack |

TABLE 4.4: The virtual agent questions presented to patients.

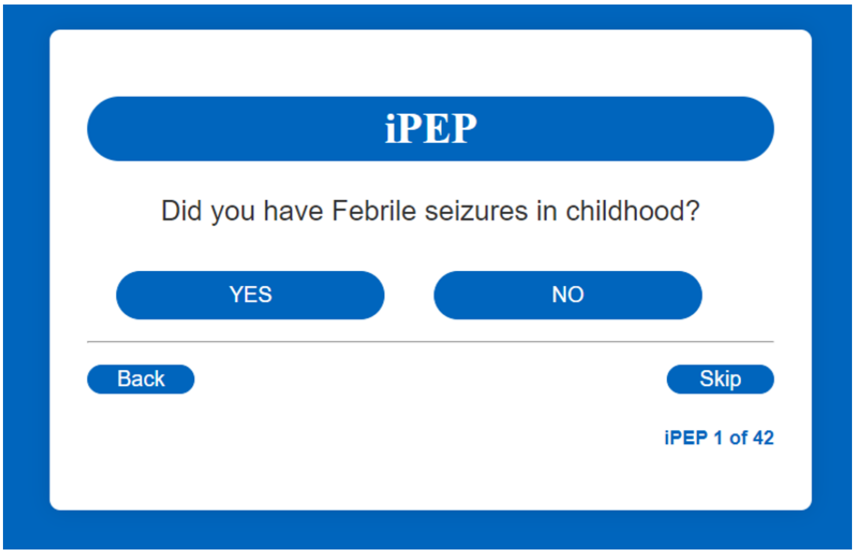
| **Virtual agent questions for patients** |
|---|
| Please tell me in as much detail as you can remember what happened during the most recent attack that caused you to lose consciousness? |
| What were you doing when the attack started and how were you feeling? |
| Do you think there was any trigger for the attack? |
| What was the first sign of the attack? |
| How did you feel during the attack? |
| How did the attack end? |
| How did you feel after the attack? |
| Did you injure yourself during the attack? |

TABLE 4.5: The virtual agent questions presented to witnesses.

| **Virtual agent questions for witnesses** |
|---|
| Please tell me in as much detail as you can remember what happened during the attack that you witnessed |
| What was happening before the attack? |
| How responsive were they during the attack? |
| What were they like when the attack had finished? |

TABLE 4.6: The Technology Acceptance Model feedback questionnaire taken from Lanseng and Andreassen (2007). All questions were asked on a 7-point Likert scale ranging between “strongly agree” to “strongly disagree”. The questionnaire consists of 4 subscales that measure the usefulness (1-6), ease of use (7-12), attitude (13-16), and intention to use (17).

| **Feedback questionnaire** |
|---|
| The digital doctor is useful because it would save me time |
| The digital doctor is useful because it would save me effort |
| Using the digital doctor is more convenient than booking and attending a medical appointment |
| Using the digital doctor is easier than booking and attending a medical appointment |
| The digital doctor will help me to be referred to the right service |
| The digital doctor will help me to receive the correct diagnosis |
| I found the ‘digital doctor’ confusing to use |
| I found the ‘digital doctor’ time consuming |
| I found that the ‘digital doctor’ takes a lot of effort to use |
| I found the ‘digital doctor’ complicated to use |
| I found that the procedure required little work to use |
| I found the ‘digital doctor’ easy to talk to |
| I think the ‘digital doctor’ is good |
| I think the ‘digital doctor’ is pleasant |
| I think the ‘digital doctor’ is beneficial |
| I think the digital doctor is favourable |
| If this technology was available in the future, I would use it |



iPEP

Did you have Febrile seizures in childhood?

iPEP 1 of 42

FIGURE 4.1: A screenshot of the iPEP questionnaire from the online web application. Participants responded to each question by clicking on the appropriate response option.

## 4.7 Feedback Interview

Participants who expressed an interest in providing feedback on the application by interview were presented with another consent form. A member of the research team contacted them to arrange a telephone interview. Interviews were conducted on average 11 days after completing the application (SD = 7.4). The telephone interviews were recorded and stored in the shared research storage area.

## 4.8 Transcription

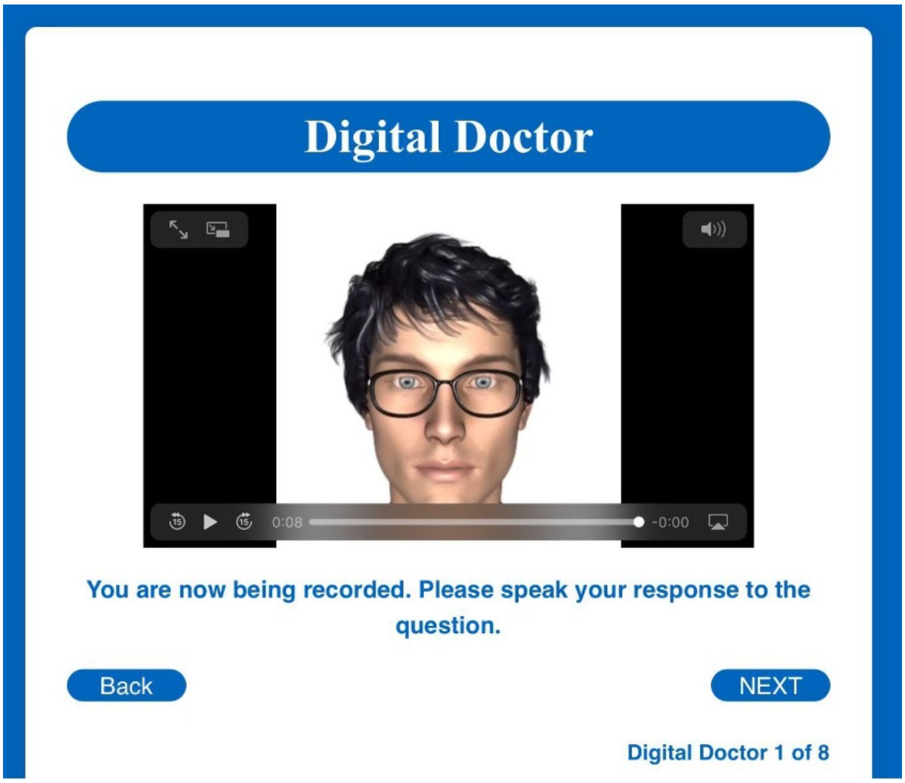
The audio files were encrypted and shared with an external transcription service. The transcripts were verbatim and included timed pauses that were greater than 1 second. More detailed timed pauses were manually calculated using Praat (version 6.1.34, 1992-2020, produced by Paul Boersma and David Weenink) and added to the transcripts for the qualitative research.

## 4.9 Confirming the diagnosis

Most of the participants who were recruited through the Royal Hallamshire Hospital did not have a diagnosis at the time of recruitment. Participants received a diagnosis through the seizure or syncope clinic and the Royal Hallamshire Hospital. The exact methods used for the diagnosis were not recorded. The diagnosis was confirmed by Professor Markus Reuber through an examination of the hospital medical records six months after participation.

## 4.10 Conclusion

This chapter provides an outline of the online web application that was used to collect the research data that will be analysed in the subsequent chapters of this thesis.

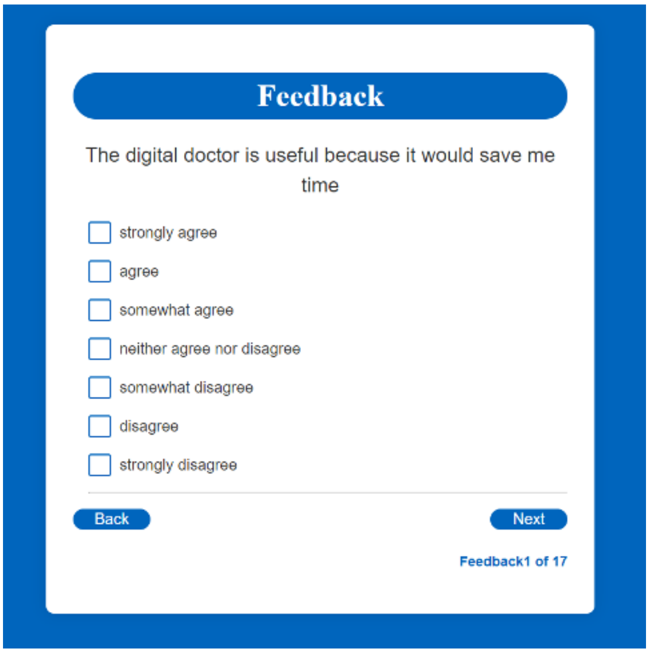


Digital Doctor

You are now being recorded. Please speak your response to the question.

FIGURE 4.2: A screenshot of the virtual agent from the web application.

The application collected symptom and medical history information from patients and witnesses using a questionnaire and by recording an interaction with a VA. Participants were given the option to provide feedback on the application using a feedback questionnaire or telephone interview, which will be used to assess the acceptability of the approach from the perspective of patients and witnesses and guide the future development of the application. Henceforth, the thesis will explore the data that was collected using this method throughout the remaining chapters.



Feedback

The digital doctor is useful because it would save me time

□ strongly agree

□ agree

□ somewhat agree

□ neither agree nor disagree

□ somewhat disagree

□ disagree

□ strongly disagree

FIGURE 4.3: A screenshot of the feedback questionnaire presented to all participants.

TABLE 4.7: The questions that were used for the semi-structured feedback interview.

| **Interview questions** |
|---|
| What do you think about the application? |
| How did your experience using the application compare with your prior expectations? |
| What do you think are the advantages of using the application? |
| What do you think are the disadvantages of using the application? |
| What would you change about the application? |
| Were you able to share all the information that you wanted to share when speaking with the 'digital doctor'? |

# Chapter 5

# Conversation analysis of interactions with a digital avatar

The objective of this chapter is to address the third research question of this thesis - how do people describe their experience of TLOC to a VA. The aim is to explore whether the conversational profiles for people with epilepsy or FDS from previous doctor-patient interactions are present during interactions with a VA using conversation analysis (CA). The presence of these profiles would support the feasibility of detecting these differences using an automated analysis of language.

## 5.1 Introduction

CA is a rigorous and empirically-based exploration of the structure of human interaction. The approach originated in the field of sociology and was first outlined by Harvey Sacks, Emanuel Schegloff and Gail Jefferson. In a seminal lecture in the early phase of the conception of this methodological approach, Harvey Sacks described interaction using the metaphor of machinery (Sacks, 1992). What people say during an interaction is caused by underlying machinery that responds to the content of the interaction and produces an appropriate output based upon certain principles of interaction that people adhere to. Adherents of CA posit that an interaction is not best understood based upon guesses at an individual's internal cognitive representation, but rather that the elements of the interaction both cause and are caused by what is happening in the interaction (Sidnell, 2011, p. 2). The objective of CA is to understand the machinery that governs interaction by conducting an analysis of interactions on a turn-by-turn basis to understand the relationships between a given input (what was previously said in the interaction) and the corresponding output (what is being said now) based solely upon the information that is available to those in the interaction.

Although interactions are predominantly organised on a micro level based on the content of the interaction, the macro aspects of the social world can influence the structure of interactions. These macro influences are often labelled as the context of the interaction and can include broader societal structures, such as social institutions (law and medicine) and social stratification (class and race) (Schegloff, 1992). There are an enormous amount of contextual factors that could influence a conversation at any given moment, so it is important that context is only considered in instances where participants in an interaction have made a given context relevant (Schegloff, 1992). Context can be made relevant by something a speaker has said or by performing particular sequences that are restricted to or repeated in a particular setting that demonstrates that all participants in the interaction are orienting to the normality of the behaviour for this particular context. Robinson (2003) demonstrated that, having established that the patient is visiting the doctor with a new medical concern, doctors and patients mutually orient to an "interactional project" that involves advancing through four stages of the medical interaction: establishing the reason for the visit, gathering information through verbal and/or physical examination, delivering a diagnosis, and recommending treatment. These stages have been extended for secondary care interactions where requesting further tests can be an additional stage (Toerien, Jackson, and Reuber, 2020). The doctor and patient have different roles and responsibilities throughout the medical interaction that can influence the trajectory of the interaction and may be made relevant by one or both parties. The patient is responsible for presenting the information about their health concern and answering the doctors questions, whereas the doctor is responsible for asking the relevant questions, making a diagnosis, and providing treatment suggestions (Heritage and Maynard, 2006). These expectations are shaped by the context of this particular encounter. Therefore, context can have an impact on interactions, particularly those occurring in an institutional setting.

Understanding how the context of an interaction can affect the talk that takes place within it is relevant for the research conducted in this thesis because we are investigating whether we can detect linguistic and communicative differences between the way that people with epilepsy and FDS describe their seizure experience, which have been outlined in previous doctor-patient interactions, during an interaction with a VA. The utilisation of machine learning features designed to measure particular interactional differences would presume the presence of such interactional differences during the conversation with the VA. The performance would be hindered if these features were not present. Exploring how patients interact with the VA and whether these interactions resemble the doctor-patient interactions from previous studies would confirm the theoretical feasibility of this analysis and aid the interpretation of the performance of the model.

Previous research has compared how patients presenting clinically with memory impairment speak to a VA or to a doctor (Walker et al., 2020). During doctor-patient interactions, the doctor will often change the format of a question in response to the information the patient has provided to preceding questions. This behaviour can be problematic in medical interactions where the doctor has been instructed to ask particular questions in a particular format in order to evaluate the communicative behaviour of the patient. Relatively minor changes to the way they ask a question can have clear effects on the patient's answer, potentially diminishing its differential diagnostic value. Walker et al. (2020) found that patients may use the question changes to avoid answering a question, particularly if the question can be interpreted as a polar question instead of a request to share more detailed information. These findings suggested that a standardised method of asking questions, like a VA programmed to ask a set of preformulated questions, could result in greater consistency in the responses that are collected. In the analysis of VA or doctor interactions with patients with memory concerns, this effect was indeed observed while the differential diagnostic features previously described in doctor-patient interactions could still be recognised (Walker et al., 2020).

While the questions asked by the VA and doctors in the memory project were quite similar, there were important differences between the questions asked by the VA in our research project involving patients with seizures and those from the previous CA research (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008) based on doctor-patient interactions. The original interview guide started with an open question not mentioning the patient's complaint (i.e. "What were you expecting to get out of this admission?" or "What can I do for you?"). The interview guide then went on to propose questions about three specific, memorable seizure events (the first, most recent and worst seizure) (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). The topical agenda of the questions used in this study was restricted to the most recent attack because patients who are first presenting to the seizure clinic may have only had one episode of TLOC (Tables 4.4 and 4.5). Furthermore, the question "what can I do for you" may not have been appropriate for interactions that were part of a research study rather than clinical care because people may not have the expectation that the VA would do anything for their medical care. The differences in the questions may have had an effect on observable interactional, topical or linguistic features of potential differential diagnostic value.

Importantly, the linguistic representation of seizures could also be different during interactions with the VA because the avatar is not interactive. The original CA research was conducted in the context of a research study in which the doctor was expected to follow an interview guide (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). These instructions involved giving participants time to talk by avoiding interruptions or asking subsequent questions too quickly. Furthermore, the prompts about the first, last, and worst seizures were conceived as challenges that allowed patients to provide more detailed accounts (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). Although the interviewers were instructed to be unusually passive in these interactions, they frequently responded to patients with back-channelling responses (e.g. "hm"). They are also likely to have affected patients' contributions by tolerating long pauses in the interaction. Refusing to take a turn at a Transition Relevance Place (Sacks, 1974) may indicate to the patient that the neurologist cannot respond until they receive more information, which may prompt the patient to continue talking in instances where the turn was designed to be complete at an earlier point. In contrast, in our VA application, the patient has full control over when they are finished providing their response, potentially leading to patients producing concise responses. Shorter contributions from patients (such as those typically elicited in routine face-to-face encounters involving frequent interruptions by the doctor and long series of closed questions), could limit the diagnostic potential of speech analysis (Ekberg and Reuber, 2015).

### 5.1.1 Objective

The objective of this section was to examine the spoken seizure descriptions of people with epilepsy and people with FDS during an interaction with an VA. Using conversation analytic techniques, we focussed on whether the linguistic and interactional features previously shown to differentiate between accounts of people with epilepsy and people with FDS can also be identified in patient interactions with a VA. The communicative profiles for people with FDS or epilepsy are extensive and multi-faceted (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). However, recent research has shown that it is possible to make accurate predictions about the diagnosis using a scoring tool consisting of fewer items (Biberon et al., 2020). The most prominent items on this condensed scoring tool related to description of the unconscious period and subjective symptoms and the associated formulation effort (Biberon et al., 2020). Therefore, we focussed on whether and how people describe the symptoms that they experience during their seizure and how much detail people provide about the situation and their experiences before and after they lose consciousness. Restricting our analysis to a smaller portion of the communicative profiles that are highly prevalent across the Diagnostic Scoring Aid used to make predictions about the diagnosis (Reuber et al., 2009), we were able to produce a more detailed analysis of the interactional behaviour of the participants for these items.

## 5.2 Method

### 5.2.1 Participants

A subset of the overall sample consisting of 20 patient recordings was used for this qualitative analysis (Table 5.1). Further information about the recruitment procedure was outlined in chapter 4. The average age was 39 years old (people with FDS = 33.75, people with epilepsy = 42.58) and 60% of participants were female (people with FDS = 75%, people with epilepsy = 50%). Responses from people with syncope were not incorporated in this analysis because there is no previous CA research exploring how they describe their experience of TLOC. Some participants had already received a diagnosis of either epilepsy or FDS, whereas others were attending their first appointment.

Although the responses are contrasted with the previous CA research throughout this chapter, extracts from the original research are not included in the analysis. An overview of the recordings used in the original CA research can be found in section 3.1.

### 5.2.2 Analytic Approach

This analysis uses the micro-analytic approach of CA to explore how people with epilepsy or FDS describe their most recent seizure experience when interacting with a VA. The interactions with the VA are different to the human-human interactions that are analysed using CA because we cannot interpret the contributions of the VA to demonstrate what sense the VA has made of what was previously said. Unlike human-human interactions, the VA cannot understand what the participants have said and does not tailor the responses based upon what has previously been said. Consequently, we cannot apply the full methodology of CA to our analysis of this data. However, previous research has demonstrated that CA can provide meaningful insights into how people interact with a VA in this context (Walker et al., 2020). Therefore, this analysis uses CA to analyse the responses to the questions posed by the VA. The majority of the analysis focused on the responses to the first question, which asked participants to provide as much information as they could remember from their most recent attack (Table 4.4). This question allowed patients to shape their own account of what happened without any additional direction from the VA, in contrast to the more direct follow-up questions. The sequential analytic approach used in CA was applied to the analysis of the interaction in two ways: how the participants' response was influenced by the question asked by the VA, and how participant responses were shaped by their responses to the preceding questions.

Another non-conventional component of this analysis is that we contrasted the communicative behaviour of the participants who were speaking with the VA to the communication profile for seizures identified in previous research. An overview of these communicative profiles was provided in section 1.3.2. The analysts in the previous research were blinded to the diagnosis of the patients. However, this was not

TABLE 5.1: The demographic, seizure history information, and diagnosis for the participants. All participants were assigned a pseudonym.

| Pseudonym | Sex | Age | Duration of seizure (years) | Seizure frequency per year | Diagnosis status at time of participation | Final diagnosis |
|---|---|---|---|---|---|---|
| Margaret | Female | 40 | 17 | 0-50 | Epilepsy | Epilepsy |
| Julia | Female | 40 | 10 | 0-50 | Epilepsy | Epilepsy |
| Rachael | Female | 25 | 1 | 0-50 | N/A | FDS |
| Jonathan | Male | 33 | 0 | 0-5 | N/A | Epilepsy |
| Eleanor | Female | 30 | 7 | 0-50 | Epilepsy | FDS |
| Richard | Male | 64 | 1 | 0-5 | N/A | Epilepsy |
| Phillip | Male | 16 | 5 | 50+ | FDS | FDS |
| Sophie | Female | 33 | 2 | 50+ | FDS | FDS |
| George | Male | 80 | 0 | 0-5 | FDS | FDS |
| Fred | Male | 19 | 1 | 50+ | Epilepsy | Epilepsy |
| Anthony | Male | 43 | 0 | 0-5 | N/A | Epilepsy |
| Luke | Male | 77 | 1 | 0-5 | N/A | Epilepsy |
| Edward | Male | 44 | 25 | 0-5 | Epilepsy | Epilepsy |
| Lucy | Female | 21 | 2 | 0-50 | FDS | FDS |
| Mary | Female | 23 | 7 | 0-50 | Epilepsy | Epilepsy |
| Cindy | Female | 59 | 3 | 0-50 | Epilepsy | Epilepsy |
| Victoria | Female | 33 | 20 | 50+ | Epilepsy | Epilepsy |
| Michelle | Female | 32 | 16 | 50+ | FDS | FDS |
| Angela | Female | 36 | 22 | 50+ | Epilepsy | Epilepsy |
| Olivia | Female | 33 | 18 | 50+ | FDS | FDS |

possible for this analysis because the analyst was also involved in recruitment where discussions of diagnosis happened. Although this could introduce bias to the analysis, the presentation of each extract should support the arguments that are made by displaying the communicative behaviours that are discussed. This component of the analysis highlights the interdisciplinary nature of the research and illustrates the usefulness of applied CA for the purpose of differential diagnosis. This analytic approach allows us to evaluate the feasibility of automating the detection of the two conversational profiles for interactions with the VA.

## 5.3 Results

The results for this analysis are structured into two separate sections. The two sections will illustrate the conversational, interactional, and linguistic presentations of people with epilepsy and people with FDS during interactions with the VA. Multiple extracts containing responses to the VA from participants will be examined. Extracts from the verbatim English transcripts will be presented that highlight the relevant features from the Diagnostic Scoring Aid (Table 1.1) for people with each diagnosis. Each section will finish with a summary of the relevant linguistic presentations that were identified in previous CA research for the diagnostic group that is being examined within the section.

### 5.3.1 Responses from people with epilepsy

The first extract that will be examined comes from a participant with the pseudonym Richard. It contains Richard's response to the first question asked by the VA.

Extract 1 - Richard (person with epilepsy)

```
VA:   Please tell me in as much detail as you can remember what happened
      during the most recent attack that caused you to lose consciousness?
Pat:  (0.5) .hhh I wa::s having dinner with my partner (0.4) sitting at the
      (1.1) table (.) .h (.) a::nd (0.5) I began to fee:l stra:nge (0.9) and
      (1.1) um (1.8) was apparently staring (1.8) at (.) her (1.1) I got up
      (0.5) because I felt slightly s:ick (0.6) and went over to the sink
      but I'm not exactly sure I couldn't exactly remember (0.9) that (0.8)
      a:nd it lasted for about (0.4) lasted for (.) .hh (.0) you know, five
      or ten seconds I think but it ff I thought I'd been out for half an
      hour (1) um (0.5) and there was a strange sense of (0.3) sort of
      hallucinations or something like that like (.) .h almost like dream
      images but I couldn't describe them (6.2)
VA:   What were you doing when the attack started and how were you feeling?
```

After a short description of the situation, Richard begins to describe an unusual and difficult to understand experience through his choice of adjective "I began to feel strange". Richard then begins a description of what happened. He got up, felt slightly sick, and went over to the sink. However, his ability to recall the series of events is marked as uncertain using the juxtaposition of "apparently staring" and "but I'm not exactly sure I couldn't exactly remember". Richard states that something lasted for five to ten seconds, which is again contrasted with his own personal experience by stating "but it I thought I'd been out for half an hour". The term "I'd been out" is an example of a conceptualised metaphor that describes an unconscious period as a space/place that someone goes to (Plug, Sharrack, and Reuber, 2009b).

The use of the coordinating conjunction 'but' and the sequential organisation of the duration estimation and conceptualised metaphor suggests that the duration is referring to a period of reduced awareness. After pausing for one second, Richard returns to describing what he experienced during the seizure "there was a sense of sort of hallucinations" where Richard exhibits examples of formulation effort (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008) because he reformulates the description "almost like dream images", uses multiple hedges, for example "sort of" and "I think", and emphases the challenges associated with producing the description "but I couldn't describe them". Richard's description of the symptoms that he personally experienced during his seizure and his attempts to accurately convey his experience of these symptoms by reformulating the descriptions aligns with multiple components of the linguistic profile for people with epilepsy (Reuber et al., 2009).

In the next examples, Margaret provides a description of the most recent attack in response to question one. The responses from Margaret were extensive: therefore, shorter extracts have been removed from the wider response. Margaret begins by providing details about the situation, namely that it was stressful, upsetting and worrying. She then states that she experienced signs of frontal temporal epilepsy, which is portrayed as a known experience through the use of the adjective 'usual' and by stating that it has already been diagnosed. However, this experience of frontal temporal lobe epilepsy is marked as different to the previous experiences. Margaret formulates and reformulates descriptions about how the experience was different from her previous experiences, for example 'more symptoms', 'less awareness', and 'I felt, I don't know, different, more (sick)'. After an extended sequence that highlights that these novel experiences are different to the previous experiences of frontal temporal lobe epilepsy, Margaret continues to describe the events that preceded the loss of consciousness - she was sitting on the floor after completing a Skype call with her boyfriend. Margarets description of the sequence of events that preceded the loss of consciousness is finished by reiterating the final action that she can recall "I just remember sitting there". The loss of consciousness is then represented as a transition to the first available memory afterwards "next thing I remember I was sort of waking up on the floor" and her mental state is described as being confused. Therefore, Margaret has contoured the unconscious period with her memories of what happened immediately before and after (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). Similarly to the example in extract 1, Margaret's description focuses on her personal experience of the seizure because she describes and reformulates descriptions of her symptoms and details her memories of the events that surrounded the seizure. Furthermore, Margaret displays a willingness to provide information about the seizure that continues throughout her interaction with the VA.

Extract 2 - Margaret (person with epilepsy)

```
VA:   Please tell me in as much detail as you can remember what happened
      during the most recent attack that caused you to lose consciousness?
Pat:  Pat: (24 seconds) Oh I remember being in a very stressful and
      upsetting and worrying situation and building up a lot of anxiety and
      thoughts to do with that situation, um, and then (2 seconds) I
      remember (1 second) having my usual signs of frontal temporal lobe
      epilepsy that I've had since 2003, er diagnosed since 2010, um (3
      seconds) but it felt different, it felt like there was more symptoms,
      there was more er, there was less awareness, um, and I felt, I don't
      know, different, more sick and I was aware of some sort of gibberish
      that sort of came out of my mouth, and er one of my best friends er
```

said the last three that she'd witnessed over previous months (2 seconds) had been like that, and I'd almost lost some consciousness, like I wasn't in the room properly, er which I wasn't aware of afterwards. I had not experienced these symptoms with temporal lobe before; I'd always remained fully conscious and even been able to maintain a conversation. Um (1 second) so then er this, the loss of consciousness one I, so I was having this exaggerated version of the frontal temporal lobe one, um, and then that seemed to pass; I was sat on the floor actually, luckily, um cos we'd just been doing a Skype, a Skype call, me and my boyfriend, um, and then I (1 second) um (1 second) remember, I just remember being sat there and then the next thing I remember (1 second) was sort of waking up on the floor confused and, and (Boyfriend name) was on, I think (Boyfriend name) was on the phone to an ambulance

After further talk in response to the same question, Margaret returns to describing what happened when she lost consciousness by drawing upon a third person account of the event (Extract 3). The account includes descriptions of the signs of the seizure, for example "went down on the floor", "shaking", "twitching", and "a sort of gargling noise". As evidenced in extract two, Margaret has no recollection of what happened while she was unconscious, but has attempted to produce a description of the events by drawing upon the information that is available (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). Attempts to reconstruct what happened during the unconscious period using their own memories or information they have received from a witness are components of the linguistic profile for epilepsy and further demonstrate attempts to provide detailed descriptions of what happened during a single seizure experience.

Extract 3 - Margaret (person with epilepsy)

Pat: Um, my boyfriend said that I was, oh, and I've got a video of it, he said I was er (1 second) went down on the floor and was shaking and twitching arms and legs, head was moving about, um, and I was making a sort of gargling noise er, which I think might have been the sort of production of sputum and blood that you get.

The subsequent questions that are asked by the VA are designed to encourage people to elaborate on the descriptions they provided in response to question one. How people respond to prompts or challenges posed by the interview are considered in the Diagnostic Scoring Aid (Table 1.1) because people with FDS may only provide information when prompted or not provide the information requested by a prompt. However, these follow-up questions were challenging for individuals with epilepsy who had already provided a detailed description of their seizure. The second question asks participants to describe what happened before the attack and how they were feeling. Margaret begins her response by stating that she had already answered this question "Oh sorry I've like covered most of that in the last answer" and providing some additional contextual information about what happened before the seizure (example not shown). After addressing the topic agenda set by the question, Margaret transitions to a description of her experience of epileptic seizures in general (Extract 4). The first symptom that Margaret reports is racing thoughts, which is expressed in two different formats. Thoughts are described as happening too fast, and Margaret refers to her brain becoming too hectic and malfunctioning "it's like it trips over itself". Margaret then reports her experience of auras where she suddenly recalls vivid memories from past experiences. Similarly to the example in extract one, these descriptions are marked as difficult to express through the use of words like "strange" and "weird" and statements that highlight the challenges associated with the description "I don't know how else to describe it". Given that Margaret stated in Extract 2 that her experiences of losing consciousness are preceded by her symptoms of frontal temporal lobe epilepsy, providing a more elaborate description of how she feels during these seizures provides relevant information about how she feels before losing consciousness.

Extract 4 - Margaret (person with epilepsy)

Pat: I, I, I feel like when I do have epilepsy seizures my (1 second) thoughts have got almost so fast or I'm thinking about too many different things; er I doesn't always have to be negative, it could be some project I'm excited about or positive, or some idea I've got, or I'm trying to um, I don't know, do the washing up and a workout and take some phone calls and send some emails and get out to the shop and er, er I get, it's almost like my brain gets too hectic and then (1 second) it's like it trips over itself and becomes too much and that's when I feel like it's misfiring and it goes into this strange thing. But one of the things I experienced prior to epilepsy is auras, um, and sometimes I get auras without the epilepsy developing as well, but that can, that usually consists of a sudden slight absence of mind um where I'll suddenly remember er a whole scene from any point in my life, childhood, adulthood, um and also it happens with memories from dreams as well um (1 second) where I can (1 second) remember, I'm not actually physically smelling anything but I can remember the smell and ambience and atmosphere of every detail of that moment, whether it's positive, negative, completely insignificant, you know, in a supermarket when I was, I don't know, eighteen, or on a beach when I was seven, it could be absolutely anything, um (2 seconds) it's so weird, I don't know (laughs) how else to describe it.

Extract 5 contains the final extract from Margaret's response to question 2 where she is describing her personal experience of seizures. Margaret is describing her experience and response upon recognising that a seizure is about to happen. There are two statements in her response that portray the seizure as a moving object: "part of me wants to let go and go with it" and "sometimes it catches me unawares and does creep up". This type of description aligns with the metaphoric conceptualisation that a seizure is an external agent (Plug, Sharrack, and Reuber, 2009b). Furthermore, the use of the statement "mind over matter" implies an attempt to prevent the seizure from happening, which portrays the seizure as something that can be combated (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). Therefore, transitioning into a description of the general seizure experience allows Margaret to adhere to the question posed by the VA while avoiding solely repeating information that she has already provided. Not only does this type of response allow Margaret to provide more information that is relevant for the linguistic profile for epilepsy, for example further describing the symptoms that she experiences, but this contrasts with how people with FDS respond to the follow-up questions, which will be demonstrated in the subsequent section.

Extract 5 - Margaret (person with epilepsy)

Pat: part of me wants to let go and go with it, and part, sometimes I feel like I can almost mind over matter it but sometimes I haven't and er,

```
          but sometimes it catches me more unawares and does creep up, but most
          of the time I have these sort of strange feelings that are an
          indication, um yeah.
```

The linguistic profiles for epilepsy and FDS have multiple features of seizure descriptions that may be indicative of a particular diagnosis, but individuals do not have to exhibit all of the features of a given profile for a prediction to be made (Reuber et al., 2009). The remaining examples outline responses from people that contain some features of the linguistic profile for FDS, but the analysis will highlight the relevant components of the linguistic profile for epilepsy that are similar to the examples outlined previously.

In extract 6, Fred provides a description of the symptoms that he experiences during his attacks in general instead of providing a description of the most recent attack, which is evident by the use of the present and future tense ("I get" and "I will"). In doing so, Fred is exhibiting interactional resistance by not focusing on a single seizure experience, which was the topical agenda introduced by the question. This type of response is characteristic of the linguistic profile for FDS (Reuber et al., 2009). However, Fred's response is focusing on his personal experience of symptoms across his seizures. The symptoms include a feeling in his chest, a pounding in the head, déjà vu, recognising smells that aren't there, and feeling tired. Fred displays uncertainty about what he is describing on numerous occasions through repetitions "che che che chest", hedging "kinda a pounding", and reformulations "I smell I recognise smells that aren't there". Fred ends by describing his personal experience of impaired consciousness whereby he feels removed from the situation but is able to respond if required. Fred's description of his symptoms and his display of the challenges associated with producing the description are similar to those observed in the previous responses.

Extract 6 - Fred (person with epilepsy)

```
    VA:   Please tell me in as much detail as you can remember what happened
          during the most recent attack that caused you to lose consciousness?
    Pat:  erm so I get a build up of (1.5) a feeling in my che che che chest and
          that will then (lipsmack) (1) develop in my head to kinda a pounding
          (.) I feel very ill (1) er a sense of dêjà vu erm (0.8) smells I smell
          I recognise smells (0.8) that aren't there erm (3.4) right afterwards
          I'll feel very tired erm I'm (0.8) aware of what's going on around me
          and I (1) c:an respond if I need to like (3) but I: (1) am very (0.8)
          removed from the situation (1.5)
    VA:   What were you doing when the attack started and how were you feeling?
```

The examples provided so far all include descriptions of the subjective symptoms associated with a seizure. However, not all individuals with epilepsy experience these symptoms, and patients who do not experience symptoms prior to losing consciousness cannot provide descriptions of the symptoms, which impacts on the score that they would receive on the Diagnostic Scoring Aid because subjective symptoms are considered in multiple items on the scale (Reuber et al., 2009). The next extract will examine the responses from Jonthan who has a diagnosis of epilepsy but does not report experiencing an aura prior to losing consciousness.

Extract 7 - Jonathan (person with epilepsy)

```
 1   VA:   Please tell me in as much detail as you can remember what happened
 2         during the most recent attack that caused you to lose consciousness?
 3   Pat:  (1.9) .hh er (.) most recent seizu:re was (0.7) one Thursday afternoon
 4         if (.) but I believe (0.3) about just after three o'clock I was sat at
 5         my desk working cos I work from home .hhh I'd just eaten (0.3) just
 6         eaten some crackers (0.3) um (0.) next thing I know (.) I'm laid on
 7         the floor (.) my office chair (0.2) sort of broken next to me (0.4) er
 8         I woke up very (0.3) disorientated (0.6) um stood up and all I wanted
 9         to do was get somewhere safe↑ (0.6) um (.) but I knew I was at home so
10         my safe place was my bedroom (.) so I managed to get to my bedroom and
11         fell asleep (0.6) um (1) the initial seizure I remember looking at my
12         (0.6) clock (.) computer clock just beforehand (0.4) and it was about
13         three-thirty in the afternoon (0.3) and then when I got (0.2) sort of
14         came round from it I recall seeing the time (0.2) about ten past four
15         .hh ish (0.5) and then (0.7) sort of slept for another like
16         forty/fifty minutes (0.5) um before seeking advice=and then when I
17         came too I felt (0.8) fine as thought i'd just woken up from a nap↑ in
18         essence .hh (.) erm (0.6) then sought advice from (0.3) erm (.) 111
19         (.) online (0.3) and then subsequently rang and they adviced me to go
20         to (0.4) A and E (0.2) erm n went to ((hospital name)) (0.5) er went
21         for=had the CT scan (0.3) ECG (0.5) stayed over night (0.6) erm (0.2)
22         discharged Friday morning when I went for an MRI (.) at ((hospital
23         name)) (2.6)
24   VA:   What were you doing when the attack started and how were you feeling?
```

From line 3 to 5 Jonathan starts his response by describing what happened prior to the seizure onset: "I was sat at my desk working. . . I'd just eaten just eaten some crackers". This is then cast as the last thing to have happened before the seizure. The phrase "next thing I know " is used to represent a shift in time that is followed by a description of the first memory that follows this period "I'm laid on the floor, my office chair sort of broken next to me". Jonathan therefore provides a contour of the events that surrounded the unconscious period (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008) and orients to the question by claiming to provide as much detail as he can remember from his own perspective. Jonathan provides an extensive description of what happened after the seizure that includes how he felt and the actions that he took "I woke up very disoriented, um stood up and all I wanted to do was get somewhere safe". Furthermore, Jonathan tries to fill in the missing time in lines 8-10 by estimating the duration of the seizure using his memory of the time prior to the seizure onset, which he states at the beginning of his description on line 4 and reiterates in line 11. Although Jonathan does not report the experience of subjective symptoms, Jonathan's description focuses on what he can recall about the events that surrounded the seizure rather than what he does not know. Therefore, there are parallels between this description, earlier descriptions (for example Margaret's description of her most recent seizure), and the linguistic profile for people with epilepsy because Jonathan contours his memories before and after he lost consciousness and attempts to reconstruct what happened using his own memories.

Given that Jonathan is unable to recall what happened during the attack, his response to the fifth question about how he felt during the attack (Table 4.4) posed by the VA will now be explored (Extract 8). Jonathan begins with a complete negation that states that he cannot remember any details of the seizure and that this applied to both of the seizures which he has had. The response is followed by a one second pause in line 3 at a potential Transition Relevance Place (Sacks, Schegloff, and Jefferson, 1978). Although Jonathan's response covers his epistemic access to the topic of the question, a transition to the next question at this point would result in an "I don't know" response that could be evaluated as insufficient by the recipient (Sacks, 1992) and be considered a dispreferred response (Stivers and Robinson, 2006). Jonathan appears to overcome this bind by introducing a third party account of what happened during his first seizure "I re regained consciousness and panicked". Considering that Jonathan was alone during his most recent seizure and is claiming no memories of what happened, the witness report that he panicked allows him to provide information about how he was feeling during the attack. The introduction of information about other seizures demonstrates a willingness to provide as much detail as possible in response to the questions asked by the VA. Furthermore, the introduction of information about other seizures in order to introduce relevant information is similar to the responses from Margaret.

Extract 8 - Jonathan (people with epilepsy)

```
VA:   How did you feel during the attack?
Pat:  (2.1) During the seizure I don't remember (0.2) it at all (.) um
      neither seizure (.) I don't remember any details (1) er the first
      seizure that I came (.) apparently the witness (0.5) that found me
      (0.2) first time .h (.) said I (.) re regained consciousness and
      panicked↑ (.) at the sight of them being there .h (.) second seizure
      (0.4) no recollection or anything other than waking up (0.5) sort of
      coming round in my spare room and with the urge to (0.3) get somewhere
      safe (1.5)
VA:   How did the attack end?
```

The final two extracts display responses from Julia who reports that she is unable to describe what happened during the last seizure where she lost consciousness because she was asleep. Given that many individuals with FDS are unable to recall what happened during a seizure (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008), exploring the responses from Julia will showcase how individuals with epilepsy or FDS may exhibit different linguistic profiles even though they both cannot recall a seizure.

Julia starts by reporting that she does not usually lose consciousness with her seizures (Extract 9). In response to question one, Julia states that the last time she lost consciousness was eight year ago. She states that the seizure happened when she was asleep, that it was quite a big seizure, and that her husband was unable to wake her up afterwards. Given that Julia was asleep when the seizure began, it is unsurprising that she makes a complete negation about her ability to remember what happened "I can't remember anything about it", which is a feature typically observed in the responses of individuals with FDS (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008; Reuber et al., 2009). Although every question focuses on the most recent attack, Julia's responses to the follow up questions are similar to those observed in the previous extracts because she transitions to describing her seizures more generally to overcome the challenges associated with describing a seizure that she cannot recall (Extract 10). In response to question 4 about the first sign of the seizure (Table 4.4), Julia provides an elaborate description of the subjective symptoms that she typically experiences during her seizures. Many of the symptom descriptions are repeated and reformulated " I get erm (.) like I'm trying to remember something like I've got a memory in my head and I'm trying really hard to try and remember" and are marked by hesitations and hedging statements "kind of (.) erm". Therefore, although Julia is unable to provide a detailed description based upon the confines of the topic that was set by the questions, she displays other features of the linguistic profile for people with epilepsy in response to the follow-up questions by introducing information that is beyond the scope of the questions, for example describing the subjective symptoms that she experiences during a typical seizure and displaying formulation effort.

Extract 9 - Julia (people with epilepsy)

```
VA:   Please tell me in as much detail as you can remember what happened
      during the most recent attack that caused you to lose consciousness?
Pat:  erm I don't usually lose consciousness with my seizures er the last
      time that I did was erm (.) eight years ago when I was pregnant with
      my son (.) erm I was in bed asleep and I had er quite a big seizure
      and then er my husband couldn't wake me up afterwards so (.) I can't
      really remember anything about it
VA:   How did the attack end?
```

Extract 10 - Julia (people with epilepsy)

```
VA:   What was the first sign of the attack?
Pat:  .hh erm I get: er a feeling it's either like a déjà vu (.) feeling
      like a really strong feeling that all of this has happened before
      (0.5) or I get erm (.) like I'm trying to remember something like I've
      got a memory in my head and I'm trying really hard to try and remember
      (.) what happened (0.7) or I get confused like there's something wrong
      (0.7) and then once I've had that feeling everything kind of (.) erm
      (.) I know it's going to happen I always go "oh no" and everything
      kind of closes in (.) erm I get this feeling in my chest like
      everything's sort of been sucked into my chest (0.5) and er then
      everything goes dark for a while (2.8) erm I can hear people talking
      at the time but erm I can't respond for (.) just a few minutes (1)
VA:   How did you feel during the attack?
```

#### 5.3.1.1 Summary

The responses from people with epilepsy demonstrated many of the characteristics of the linguistic profile from previous research. There was extensive evidence that people with epilepsy produced descriptions that focused on their subjective experience, including descriptions of their subjective symptoms, because people typically volunteered descriptions of subjective symptoms, provided more detailed descriptions of the symptoms, and treated them as a central component of the seizure (Reuber et al., 2009). There were numerous examples of people displaying extensive discursive work as they attempt to adequately communicate their experience to another individual (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008) in the form of an increase in hesitations, repetitions, hedging, and statements that highlight the challenges associated with the description. The format that people presented descriptions of the unconscious period where characteristic of the linguistic profile for epilepsy because people with epilepsy often treated the unconscious period as one of several elements of the seizure, contoured the unconscious period by detailing their memories immediately before and after, and attempted to reconstruct what happened while they were unconscious using the information available to them (Reuber et al., 2009). Furthermore, we outlined an example of someone with epilepsy who volunteered information about a seizure suppression attempt and conceptualised the seizure as an external agent and a conflict.

Overall, our analysis of the responses from people with epilepsy demonstrated a tendency to provide detailed responses to the questions. Although there were instances when patients were unable to provide much detail while also adhering to the topic agenda set by the question, for example Julia's inability to recall the last seizure where she lost consciousness, a sequential analysis across the responses to different questions demonstrated that people with epilepsy often overcome the constraints of the questions by introducing information about different seizure experiences or transitioning to talking about seizures more generally, which is interestingly a characteristic of the linguistic profile for people with FDS (Reuber et al., 2009), in order to provide relevant responses to the questions. The next section will transition to a comparative analysis of the responses from people with FDS.

### 5.3.2 Responses from people with FDS

The first extract (Extract 11) contains the response that follows the first question asked by the VA from a patient with FDS.

Extract 11 - Sophie (person with FDS)

```
VA:   What was the first sign of the attack?
Pat:  .hh erm I get: er a feeling it's either like a déjà vu (.) feeling
      like a really strong feeling that all of this has happened before
      (0.5) or I get erm (.) like I'm trying to remember something like I've
      got a memory in my head and I'm trying really hard to try and remember
      (.) what happened (0.7) or I get confused like there's something wrong
      (0.7) and then once I've had that feeling everything kind of (.) erm
      (.) I know it's going to happen I always go ''oh no'' and everything
      kind of closes in (.) erm I get this feeling in my chest like
      everything's sort of been sucked into my chest (0.5) and er then
      everything goes dark for a while (2.8) erm I can hear people talking
      at the time but erm I can't respond for (.) just a few minutes (1)
VA:   How did you feel during the attack?
```

After a short period of hesitancy that is indicated by the elongation of "erm" and "was" and long gaps of silence between the turn constructional units, the only contextual information that Sophie provides is that the most recent attack was yesterday. Although the start of Sophie's response appears to focus on a single seizure "it", Sophie transitions into describing multiple attacks by saying "I had several (.) yesterday". These attacks are described as solely consisting of a loss of consciousness "I just (2.1) blanked out", and Sophie provides no information about the events that surround this period of unconsciousness. The additional details that are provided about the attack are attributed to what was observed by others "and erm stopped (0.7) breathing so I've been told". These details make it difficult to differentiate which of the multiple seizures that Sophie experienced on that day is being referenced in the response. Furthermore, other than demonstrating the recognition of losing consciousness, the description provides very little information about Sophie's personal experience or any subjective symptoms that were experienced during the attack. Therefore, this example displays many characteristics of the linguistic profile for people with FDS, for example treating the unconscious period as the defining features of the seizure, no contouring of the gap, and the use of complete negations (Reuber et al., 2009).

In extract 12, Sophie is responding to one of the follow-up questions that is described to prompt the patient to describe the events that preceded the period of unconsciousness and any potential symptoms that they experience. Sophie's states that the first sign of an attack is something observable by the third party "my partner notices it tends to be my eyes rolling back in my head". When Sophie later elaborates on this description by introducing an additional early sign of an attack, the sign is still reported as something observable by another "I will go really quiet and then somebody will look at me". This response contrasts with the responses from people with epilepsy who typically detail their own memories and recollections before losing consciousness rather than the recollections of others, for example the description of symptoms that was provided by Julia in response to the same question (Extract 10) where Julia reports experiencing "like a déjà vu (.) feeling", " I've got a memory in my head and I'm trying really hard to try and remember", and "I get this feeling in my chest". In contrast, it is not clear whether Sophie experiences similar symptoms from the description provided.

Extract 12 - Sophie (person with FDS)

```
VA:   What was the first sign of the attack?
Pat:  (1) the first sign tends erm well my partner notices it tends to be my
      eyes rolling back in my head (.) erm and my head sloping off to one
      side (.) erm it can then be a combination of things after that (.) but
      the first sign is I will go really quiet and then somebody will look
      at me .h and my eyes will be rolling back in my head and I will be
      slumped usually to one side (0.7)
VA:   How did you feel during the attack?
```

In the next extract from a patient with FDS (Extract 13), upon hearing the question by the VA, Michelle begins with an exclamation "WHA:T" and repeats the question. This exclamation displays a problem associated with the question that has been asked, which prompts the involvement of an accompanying person (speech not transcribed). Upon completion of the accompanying person's turn, Michelle subsequently seeks advice on how to answer the question. On receiving a response from the accompanying person, Michelle makes a complete negation (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008) by stating that she has no recollection of the attack, which is subsequently downgraded somewhat by the inclusion of some additional information "except for waking up with a slight memory loss". In doing so, Michelle's response equates the attack and the unconscious period (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). Not only does the use of a complete negation align with the linguistic profile for people with FDS (Reuber et al., 2009), it suggests that the problem with the question was caused by the inability to recall what happened during a period of unconsciousness. Michelle provides some contextual information about the seizure in response to the first follow-up question, but it is important to note that this information was provided upon being prompted.

Extract 13 - Michelle (person with FDS)

```
VA:   Please tell me in as much detail as you can remember what happened
      during the most recent attack that caused you to lose consciousness
Pat:  WHA:T (3.7) PLEASE TELL ME WHAT HAPPENED THAT YOU CAN REMEMBER WHEN
      YOU LAST LOST CONSCIOUSNESS (2.8) with this
Oth:  (not transcribed)
Pat:  (0.8) ah (1.0) what do I say to that?
```

```
Oth:  (not transcribed)
Pat:  (1.5) right (10) .h (3.6) I can’t remember anything except for (0.8)
      waking (0.4) up and having a slight memory loss
VA:   What were you doing when the seizure started and how were you feeling?
Pat:  (6) I was having a nap (0.5) and (1) felt (.) tired (0.6)
VA:   What were you doing when the attack started and how were you feeling?
```

The next example comes from Lucy (Extract 14). Lucy’s response starts with ”most of the time”, referencing what usually happens when she has an attack. Rather than reporting what happens before the attacks, Lucy uses a complete negation (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008) and reports ”feeling nothing before I collapse (0.4) until I come round”. Upon being asked a follow-up question about the circumstances of the attack, an accompanying person joins the interaction and produces a response instead of Lucy. Lucy does not answer questions two, three, and four. In doing so, Lucy has not provided an elaboration about what she knows and what she does not know about her seizures. Lucy continues to respond to the questions again at question 5. Upon being asked how she felt during the attack, Lucy states that she is unable to provide a description of how she felt because she was unconscious during the attack. Therefore, Lucy’s responses appear to conflate the unconscious period and the attack, thereby rendering the seizure as indescribable.

Extract 14 - Lucy (person with FDS)

```
VA:   Please tell me in as much detail as you can remember what happened
      during the most recent attack that caused you to lose consciousness
Pat:  (2.1) Most of the time I feel nothing before I collapse (0.4) until I
      come round (0.6)
VA:   What were you doing when the attack started and how were you feeling?
Oth:  (not transcribed)
VA:   Do you think there was a trigger for the attack?
Pat:
VA:   What was the first sign of the seizure?
Pat:
VA:   How did you feel during the seizure?
Pat:  (1.2) .h I was unconscious so I can’t say (0.2)
```

Extract 15 consists of an interaction with George and an accompanying other in response to the first question asked by the VA. In contrast to the previous examples, the speech of the accompanying other is transcribed and forms part of the analysis because they were also a participant in the study.

Extract 15 - George (people with FDS)

```
 1    VA:   Please tell me in as much detail as you can remember what happened
 2          during the most recent attack that caused you to lose consciousness
 3    Oth:  You have to speak now
 4    Pat:  I can’t remember can I
 5    Oth:  well you don’t (1.5) always lose consciousness do you
 6    Pat:  no no
 7    Oth:  (3) er what do you want to say (2) you have to answer it *name* (4)
 8    Pat:  what was the question
 9    Oth:  You’re being recorded so
10    Pat:  yeah please speak your response well what was the question
11    Oth:  The question was about losing consciousness during your attack
12    Pat:  (incomprehensible)
13    Oth:  You can’t recall it can you
14    Pat:  No
15    Oth:  Can’t recall so you have to say can’t recall
16    Pat:  CAN’T RECALL
```

Upon completion of the question asked by the VA, there is a three second pause that allows the absence of a response from George to be noticeable, which prompts the accompanying person to join the interaction and instruct George to answer the question “you have to speak now”. Rather than adhering to the instruction and speaking with the VA, George challenges the assertion based on the grounds that he has no knowledge of the attack and makes relevant a further response from the accompanying person using a tag question. The accompanying person challenges the assertion using a negative statement (Heritage, 2002) and a well-prefaced response (Heritage, 2015) by asserting that he does not always lose consciousness, which implicitly asserts that there are seizures where he is conscious and therefore should be able to talk about. Moreover, by making an assertion about George’s seizures, the accompanying person makes a claim of epistemic access with regards to the seizures. This is followed by another instruction on line 7 that reiterates the obligation of George to answer the question and a reminder about the recording “you are being recorded”. Upon completion of an insert sequence (Schegloff, 1972) between lines 8-12, the accompanying person accepts that the response to the question is that George cannot recall what happened and therefore instructs him to produce a response to the initial question that was asked by the VA “you have to say can’t recall”, which is repeated by George with an increase in loudness on line 16.

Although producing a response that does not answer the question is dispreferred (Stivers and Robinson, 2006), the accompanying person does not volunteer any information about what happened in the presence of a complete negation from George, even though they have already indicated they have some knowledge about what happened. This pattern of involvement has been similarly observed during family interactions with Amazon Alexa where family members use a range of discourse scaffolding methods to support others during the interaction, for example direct instructions, modelling, redirection, and expansion (Beneteau et al., 2019). Furthermore, the increased involvement of an accompanying other mirrors the finding from doctor-patient interactions where individuals with FDS were more likely to invite an accompanying other to engage in the interaction or an accompanying other self-initiated their involvement (Robson, Drew, and Reuber, 2016). It is of particular interest that the accompanying other in extract 15 implies that George should talk about the seizures that he does remember "you don’t always lose consciousness" because they are suggesting that George should provide information beyond the topical agenda of the question, similarly to the response pattern observed by people with epilepsy, but George does not provide the information.

Extract 16 provides the final example from Phillip who has a diagnosis of FDS. Phillip outlines the context of the attack “I was out for a walk” before inviting the involvement of an accompanying other using a tag question “weren’t I”. After the accompanying other has taken a turn (not transcribed), Phillip expands on his previous turn by stating that he was with the dog. Although Phillip does not make an explicit reference to a seizure, he implies the presence of a seizure by stating that the postman had to pick him up. The juxtaposition of walking and needing to be picked up suggests that something has happened between these two actions. Phillip expresses uncertainty about what happened and whether it was the postman who picked him up or someone else by saying “or summat” and “someone had to”. Phillips response to question one provides details about the events that preceded and followed the attack, but is different to the examples by people with epilepsy that were previously outlined because Phillips description doesn’t contain information about the proximity of the events to the attack, whereas the memories detailed in extracts 2 and 7 mark the memory as being close to the unconscious period using the word “just”. In this example, it is not clear whether Phillip is detailing his own memory or whether it is a recollection of information shared by another person.

Upon receipt of the second question about the events that preceded the attack, Phillip produces a well-prefaced turn, which have been shown to link the current turn with a previous turn and can act as an alert that a response is going to be rejected, dispreferred, or not straightforward (Heritage, 2015). Phillip states that he was having a seizure and signifies that his turn is finished by a long pause. The turn is hearable as complete by the accompanying other who subsequently takes a turn. Therefore, Phillip’s dispreferred response suggests that he cannot answer the question and the term “seizure” is used as a justification - thereby implying that seizures cannot be recalled and conflating the seizure and the unconscious period. After further talk by the accompanying other, Phillip produces an oh-prefaced turn that suggests that the prior turn has caused a shift in attention and induced a “change of state” (Heritage, 1998). Phillip’s subsequent description that he was sitting down implies that the change either relates to the recall of what was happening before his seizure or the realisation about the topical agenda that was set by the question. However, Phillip finishes his turn by stating that he cannot really remember. It appears that Phillip is stating that he cannot remember the seizure as a whole because Phillip does not specify what he can and cannot recall. The accompanying other appears to accept that Phillip cannot remember anything about the seizure because they provide the response to question three.

Extract 16 - Phillip (people with FDS)

```
VA:   Please tell me in as much detail as you can remember what happened
      during the most recent attack that caused you to lose consciousness
Pat:  (2.3) Right I was I was out for a walk (0.7) weren’t I
Oth:  (Not transcribed)
Pat:  With the dog (.) and the (0.5) like postman had to pick me up or
      summat (2.0) (laughing) Someone had to (0.6)
Oth:  (Not transcribed)
VA:   What were you doing when the attack started and how were you feeling?
Pat:  (1.14) Well I was having a seizure (2.16)
Oth:  (Not transcribed)
Pat:  Oh yeah because I was sat down when I when I (0.8) I can’t really
      remember
```

```
13    Oth:  (Not transcribed)
14    Pat:  I don’t like the face
15    VA:   Do you think there was a trigger for the attack?
16    Oth:  (Not transcribed)
17    Int:  What was the first sign of the attack?
```

#### 5.3.2.1 Summary

The responses from people with FDS demonstrated many of the interactional features that were reported in previous research (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008; Reuber et al., 2009). People often displayed challenges associated with describing what happened during the most recent attack and reported that they were unable to recall what happened. This inability to recall and frequent use of complete negations was a prominent feature of the linguistic profile for people with FDS (Reuber et al., 2009). Compared to people with epilepsy, the responses by people with FDS had fewer instances of descriptions of the individuals personal experience of the attack, subjective symptoms, or details about the events that immediately preceded and followed the unconscious period. Furthermore, there was an increase in the involvement of accompanying others during the interaction. Although in most instances the speech of the accompanying others was not transcribed, there was evidence that accompanying others joined the interaction to support the patient to answer the questions. Accompanying others often joined the interaction because the patient spoke to them or because the patient was exhibiting difficulties answering the questions posed by the VA. This pattern of involvement has previously been associated with a diagnosis of FDS during doctor-patient interactions (Robson, Drew, and Reuber, 2016). Overall, the responses demonstrate many of the patterns observed in previous CA research and contrast dramatically with the responses from people with epilepsy.

## 5.4 Discussion

The objective of this chapter was to explore whether the linguistic, conversational, and interactional profile for people with epilepsy or FDS that were first identified during doctor-patient interactions (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008) are observable during interactions with a VA. The results showed that people with epilepsy provided more detailed descriptions that focused on their personal experience and subjective symptoms of the seizure. These descriptions included more instances of formulation effort, and people with epilepsy were more likely to incorporate references to other seizure experiences in their answers to the questions posed by the VA. In contrast, people with FDS provided limited or no descriptions of the seizures, were more likely to conflate the seizure and unconscious period, and displayed an increased reliance on accompanying others when answering the questions posed by the VA. These findings demonstrate that there are differences in how people with epilepsy and people with FDS talk about their most recent seizure to a VA and that these differences align with the findings from previous research (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008; Robinson, 2003).

A VA provides a more standardised method of asking questions in contrast to human interviewers because previous research has shown that doctors vary how they ask a question based upon the previous responses by a patient (Walker et al., 2020). By removing the variability in how the questions are asked, we are able to demonstrate that the communicational differences between the two groups are not co-produced by changes in the communication style of the doctor during interactions between doctors and patients. The patient responses to the VA can be compared more readily (Walker et al., 2020). By demonstrating that the communication differences are still evident during interactions with the VA, future research can explore how effectively the spoken descriptions can be automatically transcribed and analysed to make predictions about the cause of TLOC, which may help to guide referral pathways and reduce the waiting time for a diagnosis (Wardrope, Newberry, and Reuber, 2018).

A single feature from the linguistic profile can be described in two ways based on the principles of CA: the action that someone is performing during the conversation (e.g. displaying an inability to answer the question) and the linguistic method that has been used to perform the action Sidnell, 2011) (e.g. directing questions towards an accompanying other, long pauses after the virtual agent has asked a question, skipping questions, and making complete negations). An automated analysis of language must have sufficient training data to detect a statistical mapping between the different linguistic methods and the associated action. Unlike humans who have an abundance of linguistic and interactional knowledge based on a lifetime of conversing (Sidnell, 2011), machine learning models can only develop knowledge based upon the features that are inputted into the model Jordan and Mitchell, 2015). Machine learning models may require many instances of the different linguistic methods to reliably detect each action that is relevant for the linguistic profiles. Therefore, it is important that a large training dataset is acquired to increase the frequency of different linguistic presentations of a given feature. Otherwise, the model may fail to detect meaningful differences that are observable by humans.

The web application used in this research project is designed to be a stratification tool for patients newly presenting with a TLOC experience (Wardrope, Newberry, and Reuber, 2018). Therefore, the questions were different from previous research because they focused on a single seizure experience to account for individuals who have only had one seizure. Although the interviews from previous research also asked questions about the most recent seizure, the questions were situated in a wider interview context that involved asking other questions and allowing the participants to speak freely (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). It may be possible to encourage people to speak more to the VA by incorporating questions about clinically relevant information relating to other topics (Cassell, 1985) because we found that people with FDS typically said more during the routine clinical consultations analysed in chapter 3. Furthermore, allowing people to adjust to speaking with a VA before they produce their seizure descriptions might help them feel more comfortable with the interaction, which could increase how much they say. This may support participants who have only had one seizure that they are unable to recall, for example if the seizure occurred during sleep (Extract 9), because it provides them with questions that they are able to answer. However, the inclusion of additional questions might require further qualitatively analysis to determine useful linguistic differences that can be incorporated into an automated analysis of language.

### 5.4.1 Limitations

One limitation of this analysis is that it uses a small sample. Although the sample size is the same as one of the original research studies that identified the linguistic differences between people with epilepsy or FDS (Schwabe, Howell, and Reuber,

2007) and may be sufficiently large to demonstrate that the relevant linguistic features are present during the interactions with the VA, using a small sample size may not capture many of the variations in spoken descriptions that may be important for an automated analysis of language. Exploring the linguistic features present in a larger sample size may help to guide the development of features that can automatically detect the relevant linguistic presentations.

One considerable difference between this study and the previous research is that the patients in this study did not all receive a "gold-standard" diagnosis using video-EEG (Noachtar and Rémi, 2009; Kinney, Kovac, and Diehl, 2019). Patients were recruited when they were referred to the seizure or syncope clinic at the Royal Hallamshire Hospital and the diagnosis was confirmed after six months by reviewing their medical record. Many patients will have received a diagnosis solely based on clinical interviews and their medical history (Plug and Reuber, 2009; Malmgren, Reuber, and Appleton, 2012). Therefore, the diagnoses may not be as reliable as the gold-standard diagnosis used in previous research.

Although the analysis demonstrates that these linguistic profiles are evident during interactions with a VA, there is no indication how effective these linguistic differences are for predicting the diagnosis. The utility of the linguistic profiles for predicting the diagnosis during doctor-patient interactions has been shown by asking linguists who are unaware of the diagnosis to apply the Diagnostic Scoring Aid to recordings of the interaction and generate a score (Reuber et al., 2009; Cornaggia et al., 2012; Papagno et al., 2017; Yao et al., 2017; Biberon et al., 2020). Applying the Diagnostic Scoring Aid to interactions with the VA would allow an estimation of the predictive capabilities of the linguistic profiles for these interactions, which would later be compared to the predictive performance of the automated analysis of language. However, the validity of this approach is reliant on the linguistic scorers being blinded to the diagnosis of the patients (Reuber et al., 2009). Unfortunately, it was not possible for the main researcher to be blinded to the diagnosis because they were responsible for recruitment and consenting, which included a discussion about pre-existing diagnoses.

Another important consideration is that the interactions with the VA were separate to the care that patients were receiving for TLOC. This may influence who participates and how people interact with the VA. Participants with an unfavourable attitude towards a clinical decision tool for TLOC may not choose to participate in the study, but how they interact with the VA may be influenced by their attitude. Furthermore, people may interact differently with the VA if the responses they provide will be used to guide their referral pathway. Some participants in the study did not provide answers to some of the questions asked by the VA, but this pattern of responding may be different if the application was integrated into the care pathway.

The questions that were asked by the VA are different to those used in previous research because they solely focused on the most recent attack, whereas previous studies asked participants about their expectations from the consultation and asked them to describe their first, worst, and most recent attack (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). Although focusing on a single attack allows individuals who have only experienced one episode of TLOC to answer all of the questions, this can influence the presentation of the linguistic profile for individuals who have different types of seizures because their ability to recall information can vary depending on the seizure in question, which was demonstrated by the analysis conducted for extract 9. Future research should consider asking questions about other seizure experiences for individuals who have had multiple seizures to ensure that the topical agenda of the questions (Heritage and Maynard, 2006) does not constrain the responses that individuals provide, thus allowing individuals who want to volunteer additional information to do so within the constraints of the interaction.

### 5.4.2 Conclusion

This chapter demonstrates that there are differences in the spoken seizure responses between people with epilepsy and people with FDS during interactions with a VA. The VA provides a standardised method of collecting spoken descriptions of TLOC where the differences in patient responses are not influenced by changes in the way that the question is asked. This standardisation may make the differences in the responses more readily apparent and support the automatic transcription and analysis of spoken descriptions, which could be used to make diagnostic predictions for an automatic patient stratification tool. Although these differences may be useful for making diagnostic predictions, future research should explore how changes to the VA and the questions that are asked can improve the amount of detail in the spoken responses.

# Chapter 6

# Predicting the cause of transient loss of consciousness using web recorded descriptions

## 6.1 Introduction

The qualitative analysis detailed in the previous chapter has demonstrated that individuals with a diagnosis of epilepsy or FDS display many of the interactional, linguistic, and conversational differences outlined in previous research (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008; Robson et al., 2012; Robson, Drew, and Reuber, 2016) during an interaction with a VA. Given that the spoken descriptions of TLOC contained features that have previously been shown to be useful for predicting the diagnosis, these findings suggest that an automated analysis of the recordings collected through the online application could be useful for predicting the diagnosis.

The analysis outlined in chapter three demonstrated that it may be feasible to differentiate between individuals with a diagnosis of epilepsy or FDS by conducting an automated analysis of seizure descriptions. More specifically, features designed to measure formulation effort and features designed to measure the proportion of words corresponding to relevant semantic categories were able to predict the diagnosis with an accuracy of 71% and 81%, respectively. However, it is unclear how well these features will generalise to recordings from a novel patient group that were collected under a different research paradigm, with a different interviewer, and using different interview questions.

Two of the most prominent differences between the spoken descriptions of TLOC for individuals with a diagnosis of epilepsy or FDS outlined in the previous chapter was the extent to which individuals produced descriptions of their subjective symptoms and how much information people provided about the events that surrounded the unconscious period. People with epilepsy were more likely to provide descriptions of subjective symptoms and details about their memories immediately before and after they lost consciousness, which aligned with the findings from the original conversation analysis research (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008). These two interactional differences are important for predicting the diagnosis using the Diagnostic Scoring Aid (Reuber et al., 2009) and are a prominent component of a condensed version of the DSA from previous research (Biberon et al., 2020), but they are not effectively measured using the features outlined in chapter three. Semantic features that are restricted to particular word types, for example verbs, adjectives, and adverbs, might be able to detect these important communicative differences by capturing differences in the actions and appraisals that people incorporate, or do not incorporate, in their description. Exploring these additional features may help to improve the predictive performance.

The analyses of language that have been conducted so far in this thesis have focused on the differentiation between epilepsy and FDS, but a clinical decision tool for TLOC should also correctly identify individuals with syncope (Wardrope, Newberry, and Reuber, 2018). Therefore, it is important to consider how useful these features are for detecting individuals with syncope given that there is no research exploring the linguistic and communicative profile for syncope.

The features that were outlined in chapter three mostly measure the semantic content of the description of TLOC. However, it is important to consider potential confounding variables that could influence the predictions of the machine learning model Mukherjee et al., 2022 because the semantic content of the description may be influenced by factors other than the diagnosis. For example, vocabulary can be influenced by age, education, and multilingualism (Keuleers et al., 2015). People who are more educated may produce more verbose descriptions of what they experienced. Therefore, it is necessary to consider the relationship between education and the amount of detail that individuals provide.

The online web application must include automated speech recognition in order for the system to be fully automatic. Although modern ASR systems are not without error, an automated analysis of language that uses ASR can still demonstrate good predictive capabilities in the presence of ASR errors (Mirheidari et al., 2016; Mirheidari et al., 2017a; Mirheidari et al., 2018). It is necessary to explore the impact of ASR on the predictive performance of this automated analysis of language because the identification of challenges associated with ASR can drive future research.

### 6.1.1 Aims

The objective of this chapter is to explore the predictive performance of an automated analysis of descriptions of TLOC. The first aim was to further evaluate the formulation effort and semantic features outlined and evaluated in chapter three by applying these features to the descriptions of TLOC collected through the online web application. The second aim was to evaluate a different feature set that compared the usage of verbs, adjectives, and adverbs between each diagnostic group. The objective was to evaluate these features for the binary classification between epilepsy and FDS and the threeway classification that incorporated individuals with a diagnosis of syncope. The third aim was to further explore the utility of these features by evaluating their predictive performance after the incorporation of ASR. The final aim was to explore whether the amount of detail that individuals provide is influenced by their educational background.

## 6.2 Method

### 6.2.1 Data

The analysis used the audio recordings collected through the online web application. An overview of the recruitment procedure and the interaction with the VA can be found in chapter four. The responses to all of the questions asked by the VA were combined and utilised.

### 6.2.2 Automatic Speech Recognition (ASR)

An ASR system is required to make the system fully-automatic because the feature extraction module of the application requires a transcript. An open-source software application called Kaldi (Povey et al., 2011) was used to train the ASR system. The Kaldi documentation contains programming scripts, also known as recipes, that are designed to train an ASR system from scratch using a pre-specified dataset. This ASR system was trained using the Librispeech recipe that uses the Librispeech corpus consisting of approximately 1000 hours of read English speech sampled at 16kHz (Panayotov et al., 2015). Given that the Librispeech recipe is a standardised pipeline that provides a blueprint for training an ASR system using Kaldi and the Librispeech corpus, a brief overview will be provided of the major steps of the analysis instead of detailing every component. The data is segmented into a training and testing dataset. The audio files are separated into independent 25ms frames with a sliding window of 10ms and MFCC's and i-vectors (Dehak et al., 2011) are extracted from each frame. The features that are extracted from the waveforms of each frame are used to predict phonemes. For this recipe, the predictions are made using a Time Delay Neural Network (Povey et al., 2011). A separate language model, which can be considered an out-of-domain language model because it was trained using the Librispeech corpus and applied to the dataset used in this research project, was trained using a 3- and 4-gram pruned lattice Recurrent Neural Network Language Model (Xu et al., 2016). The language model is used to predict words and word sequences based upon the predicted phonemes.

The trained Librispeech model was used to generate transcripts for the VA dataset. The online web application saved the responses to each question as a separate audio file. Manual speaker diarisation was performed on audio files with more than one speaker. Although a fully automated system would require a separate module within the ASR system to perform speaker diarisation, this processing step was not performed for the analysis due to the low number of audio recordings with multiple speakers. The audio recordings were converted to 16kz PCM files and passed into the ASR system for transcription. The word-error rate was calculated by comparing the ASR transcripts with high quality manual transcripts produced by a professional transcription service. The manual transcripts and ASR transcripts were both evaluated separately in the automated analysis of language to explore the effect of the ASR system on the overall predictive performance of the model.

### 6.2.3 Data Augmentation

The overall number of participants who interacted with the VA was small and the number of participants with each diagnosis was imbalanced. Machine learning algorithms that are trained on imbalanced datasets can learn to accurately identify the majority class but fail to reliably identify instances from the minority classes (Guo et al., 2008). We used a data augmentation method called Adaptive Synthetic sampling (ADASYN) (He et al., 2008) to upsample the number of samples in the epilepsy and syncope groups to mitigate this to some extent. ADASYN was applied to the training data for each fold of the cross validation progress to increase the number of samples available for training without influencing the test dataset.

ADASYN starts by calculating the ratio of minority examples, d, where $M_s$ and $M_l$ represent the minority and majority samples, respectively.

$$d = \frac{M_s}{M_l}$$

The total number of synthetic minority data to generate is calculated, G, where $\beta$ represents the ratio of minority and majority data that is desired. $\beta$ was set at one for this analysis to balance the classes.

$$G = (M_l - M_s)\beta$$

The algorithm identifies the data points that are closest to each minority data point using K-Nearest Neighbour. K was set to seven for every model to ensure that neighbourhoods were large enough to contain a minimum of two data points for each minority class. $r_i$ is then calculated for each neighbourhood, where #majority is the number of majority data points within the neighbourhood.

$$r_i = \frac{\#majority}{K}$$

All values of $r_i$ are normalised to generate $\hat{r}_i$.

$$\hat{r}_i = \frac{r_i}{\sum r_i}$$

The number of synthetic examples to calculate per neighbour, $G_i$ is calculated. More synthetic samples are generated for neighbourhoods that are considered harder to learn because the neighbourhood contains more data points from the majority class (higher $r_i$ values).

$$G_i = G * \hat{r}_i$$

Finally, the samples, $S_i$ , are generated for each neighbourhood. Two minority samples are selected, $X_i$ and $X_zi$ , and the difference between the two samples is multiplied by a random number ranging between 0 and 1, $\lambda$.

$$S_i = X_i + (X_zi - X_i)\lambda$$

### 6.2.4 Features Extraction

The machine learning features were extracted from the raw audio recording and the transcripts. The first set of features were designed to measure formulation effort (Table 3.2). Using all seven features resulted in the best performance for differentiating between people with epilepsy and people with FDS in chapter three. However, we did not include the average length of between speaker pauses for this analysis because it is likely to reflect people's understanding of how the recording element of the application works rather than any communicative difference between the two groups in this context. Furthermore, the keywords measured for the feature "keywords associated with uncertainty" in the analysis in chapter 3 were defined by the research team. The LIWC application contains a semantic category that contains the keywords associated with uncertainty but is more extensive (Pennebaker, Francis, and Booth, 2001). Therefore, this LIWC category was used instead of the list of keywords that were used in chapter three to capture more instances of uncertainty and to reduce the complexity of the analysis, given that the LIWC is already being used. The final six features that were used to measure formulation effort were number of hesitations, the number of repetitions, keywords associated with uncertainty, frequency of patient pauses, average length of patient pauses, and total time the patient spent pausing. Pauses were measured using the Google Web RTC Voice Activity Detector (VAD). The presence or absence of speech in each 10ms window was identified and used to calculate the duration of each pause or speech segment. Pauses less than 300ms were removed because short pauses may happen within independent words.

The second feature set contained semantic categories measured using the Linguistic Inquiry and Word Count (LIWC) application (Pennebaker, Francis, and Booth, 2001). The application estimates the proportion of words that correspond to a given semantic category within a document. The categories that had the largest impact on accuracy from the analysis in chapter three were selected for this stage of the analysis (Table 3.3). The objective was to select the top 10 features, but the 10th, 11th, 12th, and 13th best forming features all had the same impact on accuracy. Therefore, the top 9 best performing LIWC categories were selected. These categories measured how people referred to others and their social life ('We', 'He/She', and 'Social'), the emotional content of the description ('Emotional Tone' and 'Affect'), the extent to which people displayed tentativeness regarding what they were describing ('Tentativeness'), the amount of present tense verbs that people used ('Focus Present'), how much the description referenced quantities or the description of their experience was quantified ('Quantifiers'), and the use of words associated with reward ('Reward'). A full list of the categories available within the application and those selected for this analysis can be found in Appendix B.

Finally, we created a feature set designed to capture differences in the description of symptoms and actions that surrounded the period of unconsciousness. These features measured the frequency of specific adjectives, adverbs, and verbs that were found to be effective predictors of the diagnosis. The text was lemmatised and the corresponding part-of-speech (POS) label was identified using Spacy (Honnibal and Montani, 2017). The POS labels were used to extract all verbs, adjectives, and adverbs. It is difficult to define a comprehensive and generalisable list of words that are diagnostically relevant without a large sample size because there are a broad range of words that can be used to describe similar actions and experiences. Therefore, we used the Term Frequency Inverse Document Frequency (TFIDF) vectoriser from Scikit Learn library in python (Pedregosa et al., 2011) to convert the words into vector representations of the verbs, adjectives, and adverbs. TFIDF is a simple and efficient method of representing a document as a set of terms that can be easily interpreted (Alsmadi and Gan, 2019). Term Frequency describes the number of times a word is in a document divided by the total number of words in the document. Inverse Document Frequency describes the logarithm of the number of documents divided by number of documents that contain the word.

$$W_x, y = tf_x, yxlog(\frac{N}{df_x}$$

$$tf_x, y = Frequency\ of\ x\ in\ y$$

$$df_x = number\ of\ documentscontainingx$$

$$N = total\ number\ of\ documents$$

There are multiple parameters that are specified in the TFIDF vectorizer that can facilitate the selection of the most discriminative words and improve the predictive performance of the selected features. This analysis only focused on single words that were included in a minimum of three documents and no more than seven. One of the parameters, max features, restricts the number of words included in the final vector by selecting N number of words with the highest term frequency across the whole dataset. The algorithm was applied to the training data for each fold of the leave-one-out cross validation procedure. The maximum number of features (N) was determined by evaluating the predictive performance of different values (10, 20, 50, 100) using a "nested" five fold cross validation that was restricted to the training data. The TFIDF vectorizer then identified all verbs, adjectives, and adverbs within the test data that were selected by the algorithm and generated a vector that was equal to the value of N and contained the TFIDF values for each word. Therefore, the number of words and selected words was different for each fold.

### 6.2.5 Classification

The classification performance of each feature set was evaluated separately using a Support Vector Machine with a RFB kernel (Cristianini, Shawe-Taylor, et al., 2000). Each model was first evaluated for the binary classification between people with epilepsy or FDS before the analysis was repeated with the inclusion of people with syncope. This model was chosen because it had the second highest classification accuracy in chapter three but was the least influenced by changes in the features that were used to train the model when evaluated using the LIWC categories. The models were trained using the nested leave-one-out cross validation method (Vabalas et al., 2019). A search for the optimum hyperparameters for each cross validation fold was conducted using the "GridSearchCV" function (Pedregosa et al., 2011) that explores all hyperparameter configurations based on the hyperparameters ranges outlined in Appendix A, Table A.6. The best configuration was selected based on the accuracy of the model that was trained using the training data for that specific fold.

The predictive performance of each feature set was subsequently evaluated again using the ASR generated transcripts. This analysis was restricted to the binary classification between epilepsy and FDS because exploring the changes for one analysis was considered sufficient to explore the impact of the inclusion of ASR.

### 6.2.6 Correlation Analysis

We conducted an analysis to explore whether there is a significant relationship between the level of education and how much information people provide while speaking with the VA. The highest level of education was converted into an ordinal scale ranging between 0 (no education) and 6 (people with a PhD). This analysis only focused on patients because witnesses were asked fewer, different questions by the VA. Spearman's Rho was chosen because educational attainment was an ordinal variable (Spearman, 1904).

## 6.3 Results

### 6.3.1 Automated Analysis of Language

A total of 76 patients participated in the study. Twenty-six (34%) were recruited through the Royal Hallamshire Hospital and 50 (66%) patients were recruited through independent charities. Most patients already had a diagnosis at the time of participation (71%). A breakdown of the demographic and seizure history information can be found in the subsequent chapter (Table 7.1). Out of all the patients that completed the iPEP (application), 61 (78%) also completed the interaction with the VA. This group included 20 people with epilepsy, 29 people with FDS, and 12 people with

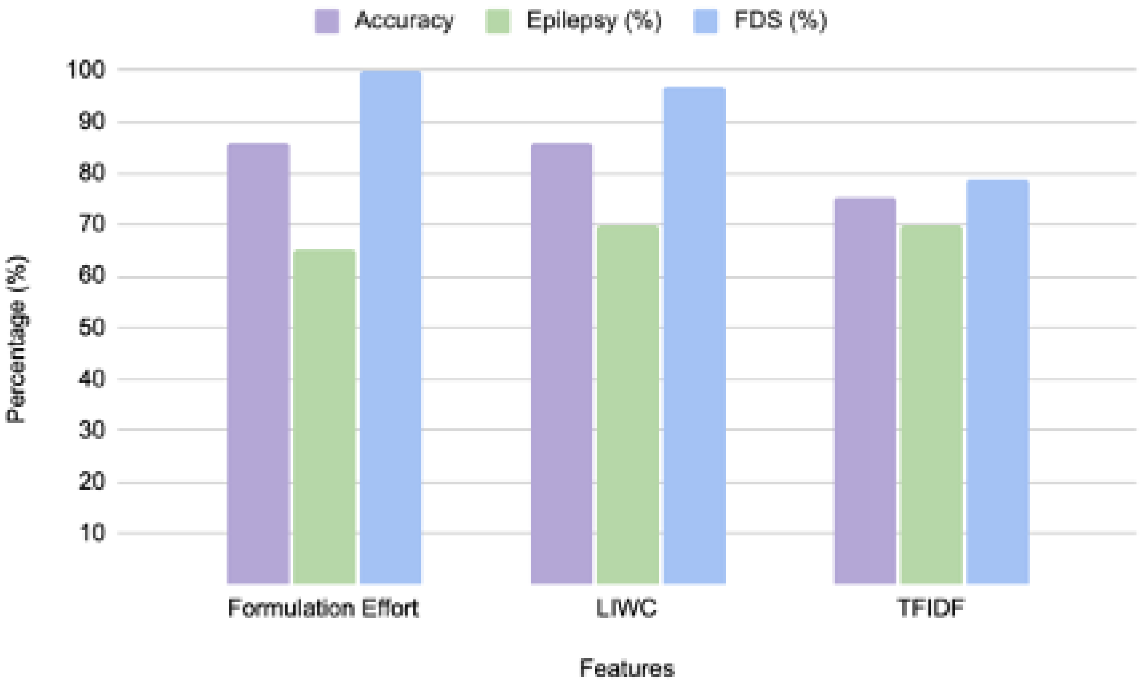


FIGURE 6.1: A bar chart showing the overall accuracy and percentage of people with epilepsy and FDS that were correctly identified by the formulation effort features, LIWC semantic categories, and TFIDF features using a Support Vector Machine with an RFB kernel.

syncope. Some individuals were unable to complete the interaction with the VA due to technological difficulties and time constraints. There were three participants who had an inconclusive diagnosis, of which none completed the VA interaction. Additionally, 26 witnesses completed the interaction with the VA, but these responses were not incorporated into this analysis due to the insufficient sample size.

### 6.3.2 Differentiation between epilepsy and FDS

The performance of the three language feature sets were evaluated for the binary classification between the clinical diagnoses of epilepsy or FDS. All feature sets were good at differentiating between epilepsy and FDS. The formulation effort features and semantic categories extracted from the LIWC application achieved an accuracy of 85.7%, and the TFIDF features had an accuracy of 75.5% (Figure 6.1). All feature sets were better at identifying individuals with FDS compared to epilepsy, and the formulation effort features successfully identified all cases of FDS.

### 6.3.3 Differentiating between epilepsy, FDS, and syncope

The accuracy of each feature set was dramatically reduced by the inclusion of people with syncope. The accuracy of the formulation effort features, LIWC semantic categories, and TFIDF features was 59%, 32.8%, and 52.5%, respectively (Figure 6.2). The inclusion of people with syncope reduced the ability of the model trained using the formulation effort features to identify people with epilepsy by 45%, but the number of people with FDS that were correctly identified remained at 100%. In contrast, the

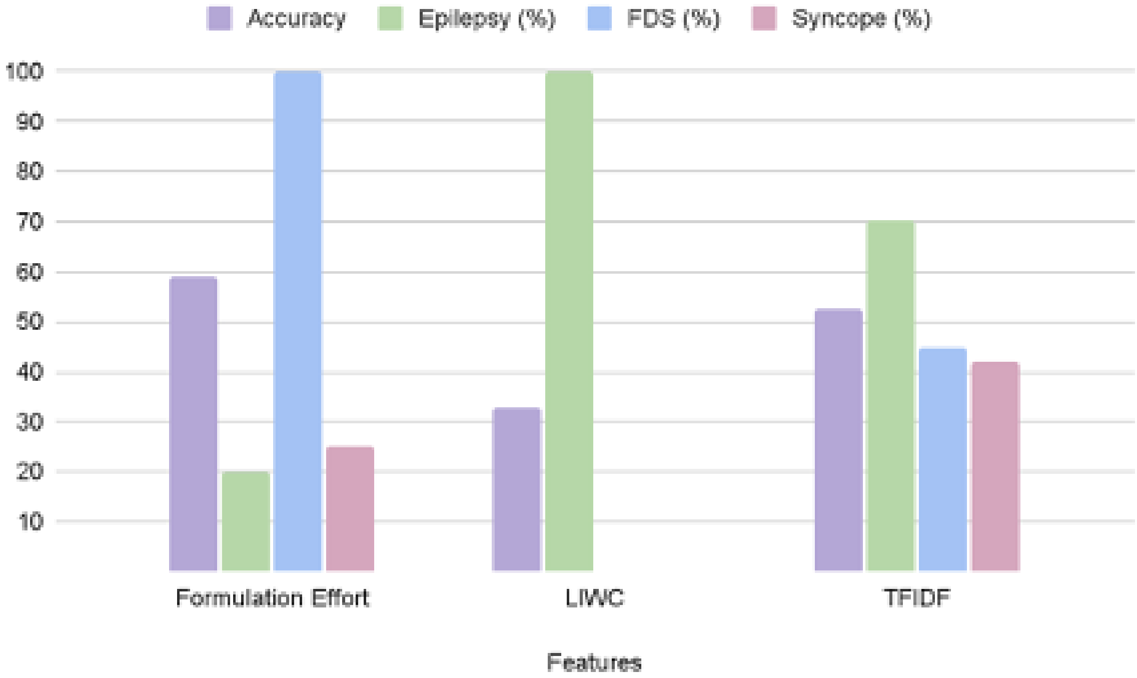


FIGURE 6.2: A bar chart showing the overall accuracy and percentage of people with epilepsy, FDS, and syncope that were correctly identified by the formulation effort features, LIWC semantic categories, and TFIDF features using a support vector machine with an RFB kernel.

model trained with the LIWC semantic categories changed from being more effective at identifying FDS to correctly identifying all people with epilepsy and nobody with FDS. All models performed poorly at identifying people with syncope, but the best performance was exhibited by the TFIDF model, which identified 42% of people with syncope.

### 6.3.4 Evaluating the performance using automatic speech recognition

The ASR algorithm had a word error rate of 39.28%. As expected, the performance of the language analysis was impaired by the inclusion of ASR (Figure 6.3). The accuracy of the binary classification between epilepsy and FDS decreased by 24.5% for the formulation effort features and LIWC semantic categories and by 12.2% for the TFIDF features.

### 6.3.5 Exploring the relationship between education and speech

The number of spoken words for 60 participants who spoke with the VA were included in this analysis. One participant could not be included because they did not provide information about their education level. Spearman's Rho correlation coefficient was computed to calculate the correlation between word count and education level. There was a significant weak correlation between the two variables ($r_s$=0.33, p< 0.05, n=60) (Figure 6.4).

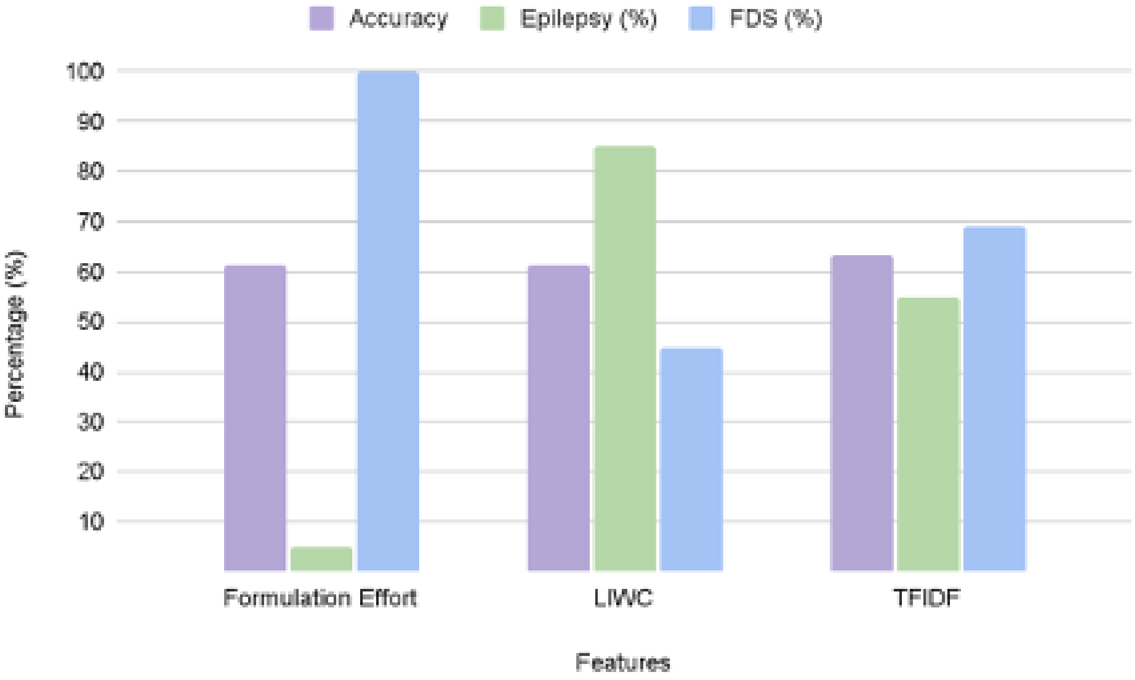


FIGURE 6.3: A bar chart showing the overall accuracy and percentage of people with epilepsy, FDS, and syncope that were correctly identified by the formulation effort features, LIWC semantic categories, and TFIDF features using a Support Vector Machine with an RFB kernel when the model was trained using the ASR output transcripts.

## 6.4 Discussion

The objective of this chapter was to explore whether an automated analysis of spoken descriptions of TLOC collected through the online web application can be used to predict the diagnosis. We explored the effectiveness of two sets of features that were previously tested on doctor-patient interactions, features designed to measure formulation effort and semantic categories. Furthermore, we tested a novel type of feature designed to capture differences in the descriptions of subjective symptoms and the reported actions that surrounded the unconscious period, changes in the predictive performance after the inclusion of ASR, and the relationship between how much people say and their educational background.

The automated analysis of language features were effective at differentiating between individuals with epilepsy or FDS. The performance of the formulation effort features and semantic categories extracted by the LIWC application exceeded the performance observed in chapter three by 14.7% and 8.1%, respectively, suggesting that these features may be reliable predictors for epilepsy or FDS. Although the TFIDF features (Alsmadi and Gan, 2019) were also effective predictors with an accuracy of 75.5%, they were less effective than the remaining two sets of features because the accuracy of these features was lower by 10.2%. Given that we observed in chapter three that features with a limited discriminative capacity can still improve the performance of a predictive model, these features may still make a valuable contribution to future models.

The predictive performance of each independent feature set was reduced by including individuals with syncope into the model. The accuracy of the models

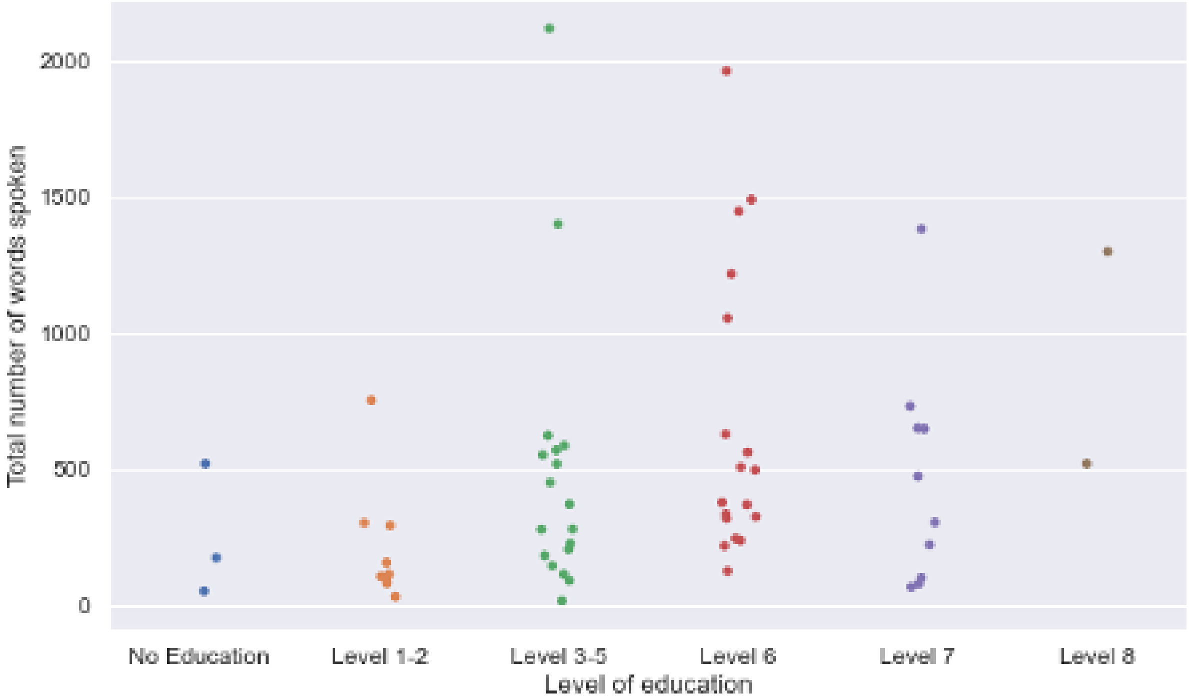


FIGURE 6.4: A stripplot displaying the total number of words spoken by each patient and the corresponding level of education. The levels of education represent the UK equivalent of GCSE (Level 1-2), A-Level (Level 3-5), Bachelor's degree (Level 6), Master's degree (Level 7), and PhD (Level 8).

trained using formulation effort features, LIWC semantic categories (Pennebaker, Francis, and Booth, 2001), and TFIDF features (Alsmadi and Gan, 2019) was reduced by 26.7%, 52.9%, 23%, respectively. The reduction in accuracy may be because individuals with syncope produce spoken descriptions that are similar to the descriptions from individuals with epilepsy or FDS, which makes it difficult for the model to identify patterns in the features that are a reliable indicator of a single diagnosis. Interestingly, it appears that the degree of overlap may be dependent on which features are used. The model trained with formulation effort features became worse at identify epilepsy and the model trained using LIWC lost the ability to identify FDS and became excellent at identifying epilepsy. This may be because people with syncope displayed formulation effort similarly to people with epilepsy and used similar words to people with FDS. It is unsurprising that the performance of these two models was reduced given that these features were selected to discriminate between epilepsy and FDS and did not change for this analysis.

The capacity of an automated analysis of language to reliably identify syncope may increase with the inclusion of features tailored to the communication profile of people with syncope. Unfortunately, there is no previous CA research detailing how individuals with syncope describe their experience of TLOC and contrasting this communication style with epilepsy or FDS. In this analysis, the model trained using TFIDF features was the most effective at identifying syncope. The procedure used to train the TFIDF model contained nested feature selection because the number of features (and therefore number of words to include in the model) was selected using the training data for each fold of the cross validation process (Vabalas et al., 2019). This may have increased the models ability to identify syncope because the model could increase the number of features used until there were words in the training data set that were indicative of syncope. Although the percentage of people with syncope who were reliably identified is still low, the number of individuals with syncope included in the dataset was also low (20%). Increasing the size of the dataset and the number of individuals with syncope may improve the models capacity to identify patterns indicative of syncope and improve the overall predictive performance.

ASR is a fundamental component of an application that conducts an automated analysis of language because manual transcription is a labour and time intensive process. The ASR module that was used in this research had a word error rate of 39.28%. The accuracy is typical of ASR in the medical domain (Kodish-Wachs et al., 2018). Unfortunately, the introduction of ASR negated the predictive performance of the automated analysis of language. There are two fundamental components to a traditional ASR system: the acoustic model and the language model (O'Shaughnessy, 2008). The acoustic model is a statistical model representing and detecting the sounds that make up words. The language model is a probability distribution of words and word sequences. Although the LibriSpeech corpus can be used to train an effective acoustic model of the English language, the language in the model was generated from read English speech (Panayotov et al., 2015) and therefore the language model may not be tailored to detect words frequently observed in medical consultations about TLOC. Furthermore, there was a large reduction in the number of people with epilepsy who were correctly classified by the formulation effort features when ASR was used, which may be because the ASR system was not effective at detecting hesitations and repetitions because the language model was not trained on spontaneous speech. Fortunately, there are many methods that can be employed to create an ASR system that is specialised at detecting the language used in interactions about TLOC, for example, designing ASR systems that are specialised in the detection of disfluencies (Liu et al., 2006), training a model using data that has been identified as being more information for an ASR system designed for a particular task (Wu and Wu, 2007), using additional text data to improve the language model and therefore the automatic speech recognition model (Toshniwal et al., 2018), and fine-tuning a pretrained model using domain specific speech data (Yu, Deng, and Dahl, 2010). Future research should explore these methods for improving the ASR system for TLOC interactions, and methods of tailoring accessible pretrained models to speech technology that is applied to other health conditions in order to reduce the redundancy involved with training novel ASR models for each medical domain.

It is important to consider potential confounding variables that can influence the predictions from an automated analysis of language. One potential confound is the education level of participants. Vocabulary size has been shown to increase as people gain higher level qualifications (Keuleers et al., 2015). A larger vocabulary may support people to produce more verbose descriptions of what happened during their experience of TLOC. We found a significant, weak correlation between education level and the number of words spoken to the VA. Although these findings suggest that education has a minimal impact on how much the patient says to the VA, the strength of the correlation could be attenuated by the small sample size used in the study. Therefore, future research should explore this relationship further using a larger sample size.

### 6.4.1 Limitations

Training an ASR system is a complex research endeavor that involves exploring a range of methods to improve the system (Latif et al., 2020). This analysis acts as a starting point by setting the benchmark for how well these features perform with a pretrained system, but this area of research needs exploring further to create an ASR system that is specialised for TLOC descriptions.

The machine learning features used in the automated analysis of language were not current state of the art methods. These features were chosen because they are easily interpretable by clinicians who may be interested in understanding how the application makes diagnostic predictions, an important consideration that can influence the likelihood that a clinician will use an application (Brown et al., 2020), but the predictive performance of these features may be lower than what can be achieved using more advanced methods. For example, the performance of the semantic categories and TFIDF features has demonstrated that semantic differences are useful for predicting the diagnosis, and there is a natural language processing technique called BERT that frequently outperforms the TFIDF method (González-Carvajal and Garrido-Merchán, 2020). Although predictive performance is very important, the ability for clinicians to interpret a model is also vital (Brown et al., 2020). Therefore, future research should explore the predictive performance of more advanced natural language processing methods for this classification task, but also collect feedback from clinicians that specialise in TLOC about the acceptability of different classification approaches for this type of clinical decision tool.

The features used in this analysis are not designed to identify individuals with syncope. The three-way classification may be improved by incorporating additional features that focus on the speech of people with syncope. Unfortunately, a larger sample of patients with syncope would be required to allow a machine learning model to adequately detect these patterns.

The CA analysis conducted in chapter five identified interactional differences between individuals with epilepsy or FDS. Individuals with FDS were more likely to rely on the contributions of accompanying others. These features were not incorporated because the research paradigm did not encourage the involvement of accompanying others during the interaction with the VA. However, this pattern of involvement has also been observed in routine clinical encounters (Robson, Drew, and Reuber, 2016). Therefore, future research should consider instructing participants that they can engage with the VA alongside an accompanying other to allow and explore the predictive capability of interactional features using novel research data. Furthermore, detecting these features in new data would demonstrate that this interactional behaviour is not limited to the sample collected in this PhD.

It is important that a diagnostic pathway is not bias towards particular individuals. There is always a risk that machine learning models may make predictions based upon confounding variables, which could influence the accuracy of the models predictions across different groups of individuals (AlHasan, 2021). Although we explored the relationship between educational background and how much people said, there are a broad range of additional confounding variables that could influence the model, for example ethnicity and language (Latif et al., 2020). As this area of research develops, future research should explore additional potential confounding variables using larger sample sizes to ensure that the effects are accurately estimated.

## 6.5 Conclusion

The research outlined in this chapter demonstrates that an automated analysis of spoken descriptions of TLOC collected through an online web application can help to differentiate between individuals with epilepsy or FDS. These findings support the reliability of the findings from chapter three by demonstrating that the previously tested features generalise to a novel sample. Future research is required to extend the findings from this analysis using a larger sample size. For example, exploring language features that can improve the identification of individuals with syncope, testing the performance of more advanced machine learning methods, creating a more effective ASR model, and exploring the predictive performance for individuals from a broad range of demographic backgrounds. In the next chapter, we will investigate whether these models can improve the accuracy of the iPEP.

# Chapter 7

# Integrating the iPEP and automated analysis of TLOC descriptions

## 7.1 Introduction

A clinical decision tool that is capable of stratifying people who have experienced TLOC could speed up appropriate investigations and decrease the number of people who are initially misdiagnosed in Primary and Emergency Care Services. Xu et al. (2016) conducted a review of 26 studies to identify the number of people referred for clinical care in a tertiary setting with a diagnosis of epilepsy who were misdiagnosed. The misdiagnosis rate was between 2-71% with a median of 20%. These findings suggest that a tool capable of predicting the diagnosis with an accuracy above 80% could improve clinical practice, particularly in areas with a higher misdiagnosis rate.

As described in the introduction and section 4.4, the iPEP (original) has been developed as a clinical decision tool intended to stratify people who have experienced Transient Loss of Consciousness (TLOC) into the three most likely diagnoses to cause TLOC (Wardrope et al., 2020a). The iPEP (original) combines two sets of questions: a patient symptom/medical history questionnaire and a witness observation questionnaire. A Random Forest algorithm trained using the patient-only iPEP (original) responses or the patient and witness iPEP (original) responses from the dichotomised questionnaire was capable of correctly identifying a diagnosis of epilepsy, FDS, or syncope with an accuracy of 78.3% and 86%, respectively (Wardrope et al., 2020a). In both instances, the questionnaire was most effective at identifying cases of syncope: in fact, modelling of the diagnostic performance of the proposed stratification tool suggested that all patients with syncope would be correctly identified using the patient and witness responses (Wardrope et al., 2020a). In the modelling, it proved more challenging to differentiate between the diagnoses of epilepsy and FDS. Although modelling of the symptom and medical history questionnaire demonstrated considerable promise in terms of its ability to differentiate between people with epilepsy, FDS, or syncope (Reuber et al., 2016; Chen et al., 2019; Wardrope et al., 2020a), there is no research exploring the performance when the questions are administered as a binary questionnaire and when the sample includes individuals who have only experience one of few experiences of TLOC, which will be relevant for individuals first presenting to Primary and Emergency Care Services due to TLOC.

The previous chapters have demonstrated that an automated analysis of spoken descriptions of TLOC can differentiate between people with epilepsy or FDS. Therefore, incorporating these features into a machine learning model trained using the iPEP may improve the model's discriminative capacity. However, it is important to consider how the iPEP (application) and language features are integrated into a machine learning model because this may influence performance. The most common method of integrating features is to train a single machine learning model using all features. The language features investigated throughout this thesis are designed to distinguish between talk of people with epilepsy and that of people with FDS. We do not know the linguistic profile of individuals with syncope. Furthermore, incorporating people with syncope reduced the performance of the automated analysis of language in chapter six. Therefore, training a single model with all participants and all features may result in a machine learning model that performs poorly.

A second method of integrating the machine learning features is stacking. Stacking is an ensemble machine learning method where multiple machine learning models are trained and the predictions are used to train a meta-learning algorithm that makes the final diagnostic prediction (Pavlyshenko, 2018). The iPEP (original) was an effective predictor of syncope in the previous research because it had a sensitivity and specificity of 83.8% and 94.6% for the patient only analysis and 100% and 91.7% for the patient and witness (Wardrope et al., 2020a). These findings suggest that the iPEP (original) has the capacity to detect most people with syncope. Using the model stacking approach, the predictions from the iPEP (application) could be used to filter out individuals who may have a diagnosis of syncope and retain individuals for whom the iPEP (application) suggests a diagnosis of epilepsy or FDS for the automated analysis of language in order to improve the overall differentiation between these two conditions. Although the iPEP (application) may incorrectly classify some individuals (e.g. predicting syncope when the diagnosis is epilepsy/FDS or predicting epilepsy/FDS when the diagnosis is syncope), restricting the automated analysis of language to the predictions of epilepsy and FDS may still improve the overall classification performance of the iPEP (application).

### 7.1.1 Aims

The overarching aim of this chapter is to evaluate the predictive performance of the automated analysis of the data collected by the online web application outlined in chapter 4. The first objective was to evaluate the predictive performance of the iPEP (application) alone. Firstly, we were interested in whether the iPEP (original) could be used to predict the diagnosis from the iPEP (application), which contains of individuals in an earlier stage of their diagnostic journey. This analysis provides us with insight into the similarity between the two different datasets. Given that there may be differences between the two datasets, we also explored the performance of machine learning models trained and evaluated using the iPEP (application). The second objective was to explore whether we can improve the predictive performance of the iPEP (application) by incorporating an automated analysis of recordings and transcripts of the spoken descriptions of what happened. We explored two different methods of integrating the iPEP (application) and language analysis: training a single model using all features and all participants or the model stacking approach outlined above.

## 7.2 Method

### 7.2.1 Data

Three datasets were used for this analysis: iPEP (original), iPEP (application), and recordings of patient interactions with the VA (patient VA). Although we also collected recordings from witnesses, there were too few recordings to use this data for our analysis. An overview of the iPEP (original) and the recruitment procedure for this dataset can be found in section 1.3.1 and in the previous research papers (Reuber et al., 2016; Chen et al., 2019; Wardrope et al., 2020a). An overview of the recruitment procedure and methods for the iPEP (application) and VA datasets can be found in chapter 4.

### 7.2.2 Machine learning models

#### 7.2.2.1 Evaluating the iPEP (application) using the iPEP (original)

The first analysis evaluated how effectively the iPEP (original) dataset can predict the diagnosis of the iPEP (application) dataset. The iPEP (original) is based on gold-standard diagnoses and contains a larger number of participants. If a machine learning model trained using the iPEP (original) dataset is effective at predicting the diagnosis for the iPEP (application) dataset, the results suggest that there are similarities between the two datasets and that the model will generalise to individuals who are first presenting. Furthermore, the iPEP (original) model can be used to make predictions for this analysis to overcome the small sample size for this research.

The iPEP (original) models in the previous research studies were trained in Matlab (Wardrope et al., 2020a). Given that this project was conducted in Python, two baseline Random Forest algorithms were trained for this analysis using the patient-only and patient and witness iPEP (original) datasets but following the same training pipeline. The hyperparameters for the Random Forest algorithm were selected to match those used in the original study (Wardrope et al., 2020a). The data was segmented into training ($\frac{2}{3}$) and a validation ($\frac{1}{3}$) datasets. The validation data provided a comparison of how well the algorithm performed on data that had been collected in the same format as the training data.

The patient-only iPEP (application) and patient and witness iPEP (application) were used to form two test datasets. The performance on the test datasets provides insight into whether there are differences in the response profiles between the iPEP (original) and iPEP (application) datasets. All missing values were imputed using K-Nearest Neighbour and a nested gridsearch (Vabalas et al., 2019) for the optimum hyperparameter value for K for each variable.

#### 7.2.2.2 Training and evaluating the iPEP (application) using cross validation

Given that the iPEP (application) was administered in binary format and to a different sample of participants, the performance of a model trained and evaluated using this dataset alone allows us to compare the performance of this approach with that based on a model trained using the iPEP (original) dataset. Two models were trained using the iPEP (application) dataset and leave-one-out cross validation. The first model used the patient-only iPEP (application) dataset. The second model used the patient and witness iPEP (application) dataset. We investigated the performance of two machine learning models: Random Forest and Support Vector Machine (SVM) with an RFB kernel. The hyperparameters for SVM were selected using a nested gridsearch of different options (Vabalas et al., 2019). A search for the optimum hyperparameters for each cross validation fold was conducted using the "GridSearchCV" function (Pedregosa et al., 2011) that explores all hyperparameter configurations based on the hyperparameters ranges outlined in Appendix A, Table A.6. The best configuration was selected based on the accuracy of the model that was trained using the training data for that specific fold. SVM was chosen because it was the second highest performing model in chapter three but was the least perturbed by changes in the features.

#### 7.2.2.3 Integrating the iPEP and language features

The iPEP (application) and language analysis features were integrated using two different methods: training a single machine learning algorithm using all features and all diagnostic groups and the model stacking approach outlined in the introduction. In the model stacking approach (Figure 7.1), the diagnostic prediction process was started with the patient-only iPEP (application) model, trained using leave-one-out cross validation. If the prediction was syncope or the participant had only completed the iPEP (application) but not interacted with the VA, the participants were removed from the model and the predictions from the iPEP (application) model were used for the final evaluation. The second stage used the three machine learning models trained using each independent feature set from the language analysis outlined in chapter 6. These models used leave-one-out cross validation to make the binary classification between epilepsy and FDS. The predictions from all four machine learning models were used to train a meta-model (SVM with RFB kernel) using leave-one-out cross validation to generate the final predictions of either epilepsy or FDS. All diagnostic predictions (iPEP only or iPEP and language analysis) were combined for the final evaluation.

## 7.3 Results

### 7.3.1 iPEP

The iPEP (application) consists of responses from 76 patients and 26 witnesses. Out of these participants, 26 (34%) patients were recruited through the Royal Hallamshire Hospital in Sheffield and 50 (66%) were recruited online through independent charities. Most patients already had a diagnosis at the time of participation (71%). There were three participants whose diagnosis was inconclusive and who were therefore excluded from the analysis. All relevant demographic and medical information can be found in table 7.1. A thorough overview of the research paradigm and recruitment strategy has been provided in chapter four.

TABLE 7.1: A breakdown of the seizure history and demographic for participants who completed the iPEP (application).

| | | Epilepsy | FDS | Syncope |
|---|---|---|---|---|
| N | | | | |
| | Patient | 24 | 36 | 16 |
| | Witness | 12 | 9 | 5 |
| **Recruitment Arm** | | | | |
| | Sheffield Teaching | 10 | 8 | 8 |

Continued on next page

**Table 7.1 – continued from previous page**

| | Epilepsy | FDS | Syncope |
|---|---|---|---|
| Hospital | | | |
| Charities and online | 14 | 28 | 8 |
| **Diagnosis status upon participation** | | | |
| Diagnosed | 18 | 32 | 3 |
| Undiagnosed | 6 | 4 | 13 |
| **Age** (years) | 43 (15.5) | 36 (27.1) | 55 (23.2) |
| **Age at onset** | 32.4 (20.7) | 31 (15.6) | 49.8 (25) |
| **Duration** | 10.6 (11.5) | 6.5 (6.1) | 5.4 (6.8) |
| **TLOC in the last year** | | | |
| None | 5 | 0 | 4 |
| Up to 5 | 7 | 9 | 20 |
| Up to 50 | 6 | 12 | 2 |
| 50+ | 6 | 15 | 0 |
| **Hospitalisation** | | | |
| Never | 6 | 7 | 7 |
| Once | 2 | 9 | 8 |
| Up to 5 | 11 | 15 | 1 |
| More than 5 | 4 | 5 | 0 |
| **Intensive care** | | | |
| No | 1 | 4 | 0 |
| Yes | 23 | 32 | 16 |
| **Family History** | | | |
| No | 4 | 3 | 3 |
| Yes | 20 | 33 | 13 |
| **Gender** | | | |
| Male | 8 | 5 | 6 |
| Female | 16 | 31 | 10 |
| **Ethnicity** | | | |
| White British | 17 | 21 | 13 |
| Black British Caribbean | 0 | 1 | 0 |
| White | 3 | 6 | 2 |
| Australian and Italian | 0 | 1 | 0 |
| English/German | 0 | 1 | 0 |
| German | 0 | 1 | 0 |
| British | 3 | 2 | 1 |
| Indonesia | 1 | 0 | 0 |
| Any other asian | 0 | 1 | 0 |
| Mixed | 0 | 2 | 0 |
| **Education** | | | |
| No education | 1 | 1 | 0 |
| Secondary education (GCSE or equivalent) | 0 | 7 | 3 |
| Further education (A-Level or equivalent) | 8 | 11 | 4 |
| Higher education (Undergraduate or equivalent) | 8 | 10 | 5 |

Continued on next page

**Table 7.1 – continued from previous page**

| | Epilepsy | FDS | Syncope |
|---|---|---|---|
| Higher Education (MSc or equivalent) | 5 | 5 | 3 |
| Higher Education (PhD or equivalent) | 2 | 1 | 0 |

A Random Forest algorithm was trained using the patient only responses from the original research. It demonstrated a similar level of accuracy for the validation dataset (78.8%) as the performance observed in the original research (78.3%) (Wardrope et al., 2020a). However, the model did not perform as effectively at predicting the cause of TLOC when it was tested using the iPEP test data collected through the online web application (63.2%) (Table 7.2). In addition, the diagnostic capability of the model differed for the three conditions in the training compared to the testing datasets (Table 7.2), for example the model correctly identified 35/37 (95%) cases of syncope on the validation dataset, but only 22/29 (76%) cases of epilepsy and 21/33 (64%) cases of functional seizures. On the test dataset, the model still performed best at identifying syncope because it identified 12/16 (75%) cases, but identified more cases of FDS (67%) compared to epilepsy (50%).

A Random Forest model trained using the patient and witness responses from the original research demonstrated a similar, yet slightly lower, level of accuracy for the validation dataset (83.1%) as the performance observed in the original research (86%) (Wardrope et al., 2020a). However, the accuracy on the testing dataset was dramatically reduced (to 46.2%). The validation dataset identified all cases of syncope (100%), most cases of epilepsy (87%) and fewer cases of FDS (62%). In the test dataset, most cases of syncope were accurately identified (80%), almost half of all cases of FDS (56%), and very few cases of epilepsy (25%). The inclusion of witness responses improved the performance of the patient-only model for the validation data by 4.8%, but reduced the performance for the test data by 19.9%.

Leave-one-out cross validation was used to evaluate the predictive performance of the patient iPEP data when it was trained solely using the responses collected through the online web application. The overall accuracy was 63.2% for the Random Forest algorithm and 65.8% for the Support Vector Machine. Compared to the performance of the model trained on the dataset from previous research, the leave-one-out cross validation model correctly identified more individuals with FDS and fewer individuals with syncope (Table 7.3).

The performance of the leave-one-out cross validation models using the patient and witness iPEP data had an accuracy of 38.5% for the Random Forest algorithm and 50% for the Support Vector Machine (Table 7.3). In contrast to the patient iPEP results, more individuals with epilepsy and fewer individuals with FDS and syncope were correctly identified.

### 7.3.2 Combining the iPEP and language analysis

Out of all the patients that completed the iPEP (application), 61 (78%) also completed the interaction with the VA. This group included 20 people with epilepsy, 29 people with FDS, and 12 people with syncope. Some individuals were unable to complete the interaction with the VA due to technological difficulties and time constraints. The three participants with an inconclusive diagnosis did not complete the VA interaction. Additionally, 26 witnesses completed the interaction with the VA, but these responses were not incorporated into this analysis due to the insufficient

TABLE 7.2: The performance of a Random Forest algorithm trained on the iPEP data used in previous research (Wardrope et al., 2020). A sample from the original research was used to create patient-only (N=99) and patient and witness (N=83) validation datasets. This model was evaluated on a patient only (N = 76) and patient and witness (N = 26) test dataset from the online web application. The table includes the overall accuracy and a breakdown of the accuracy per condition (sensitivity). The witness questionnaire is described as the "Paroxysmal Event Observer" (PEO).

| Dataset | Accuracy (%) | Epilepsy (%) | FDS(%) | Syncope(%) |
|---|---|---|---|---|
| Original iPEP<br>(Validation) | 78.8 | 75.9 | 63.6 | 94.6 |
| iPEP (Test) | 63.2 | 50 | 66.7 | 75 |
| Original iPEP & PEO<br>(Validation) | 86.7 | 61.5 | 100 | 81.1 |
| iPEP & PEO<br>(Test) | 46.2 | 25 | 55.6 | 80 |

TABLE 7.3: The performance of a Random Forest algorithm and SVM with an RFB kernel trained using the responses recorded through the online web application and leave-one-out cross validation. The table includes the overall accuracy and number of participants in each dataset alongside the percentage of people who were correctly classified for each health condition.

| Features | Accuracy (%) | Epilepsy (%) | FDS(%) | Syncope(%) |
|---|---|---|---|---|
| iPEP<br>(Random Forest) | 63.2 | 50 | 75 | 56.3 |
| iPEP<br>(SVM | 65.8 | 54.2 | 72.2 | 68.8 |
| iPEP & PEO<br>(Random Forest) | 38.5 | 66.6 | 22.2 | 0 |
| iPEP & PEO<br>(SVM) | 50 | 83.3 | 33.3 | 0 |

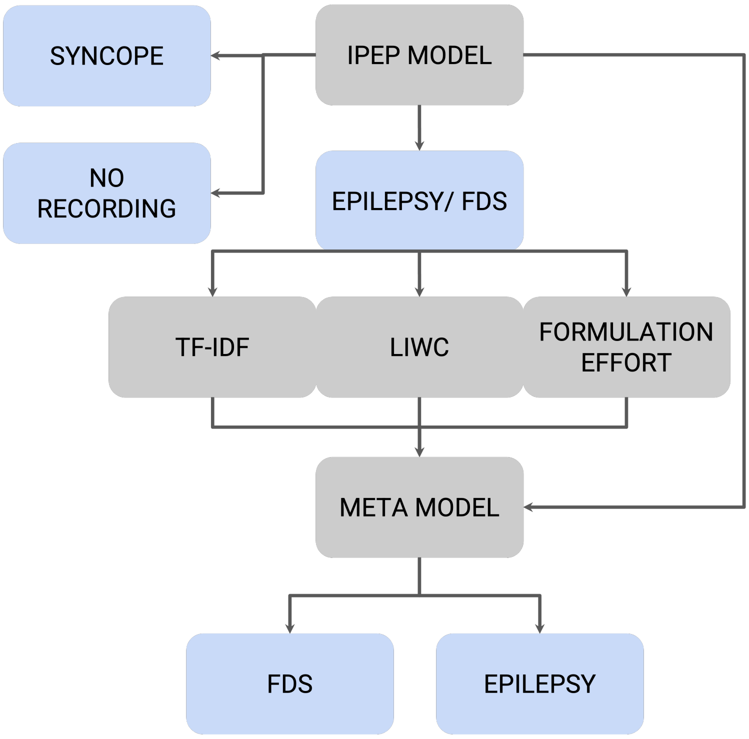


FIGURE 7.1: A representation of the model stacking algorithm. The grey boxes represent the different machine learning models. The blue boxes represent diagnostic predictions. All predictions were used to evaluate.

sample size. A Support Vector Machine with RFB kernel trained using the combination of the iPEP responses and the three language-based feature sets for the 61 participants, who completed the interaction with the VA, achieved an accuracy of 59% (Figure 7.2). The model correctly predicted 20% of cases with epilepsy, 100% of cases with FDS, and 25% of cases with syncope. The accuracy of the model was 6.8% less than the patient-only iPEP model trained using all 78 iPEP responses.

Integrating the features using a model stacking approach outperformed the model trained using all features (Table 7.2). During the first stage of the model stacking algorithm, the iPEP predicted that 14/76 (18%) people had syncope (the actual diagnosis of these predictions were: syncope = 11; epilepsy = 3). The accuracy of this subsample was 78.6%. Of the participants with iPEP predictions of epilepsy and FDS, a further 16 had not completed the interaction with the VA (epilepsy = 4; FDS = 7; syncope = 5). The accuracy of those without a VA interaction was 62.5%. The baseline accuracy of the iPEP predictions for the remaining 46 people who had completed the interaction with the VA and who had an iPEP prediction of either epilepsy or FDS was 63%. The meta-model trained using the predictions from the iPEP, formulation effort, LIWC, and TFIDF models for these 46 participants was used to predict the diagnosis with an accuracy of 95.7%. Combining the predictions from each stage of the stacking approach resulted in an overall accuracy of 85.5%. The model correctly identified 83% of people with epilepsy, 94% of people with FDS, and 69% of people with syncope.

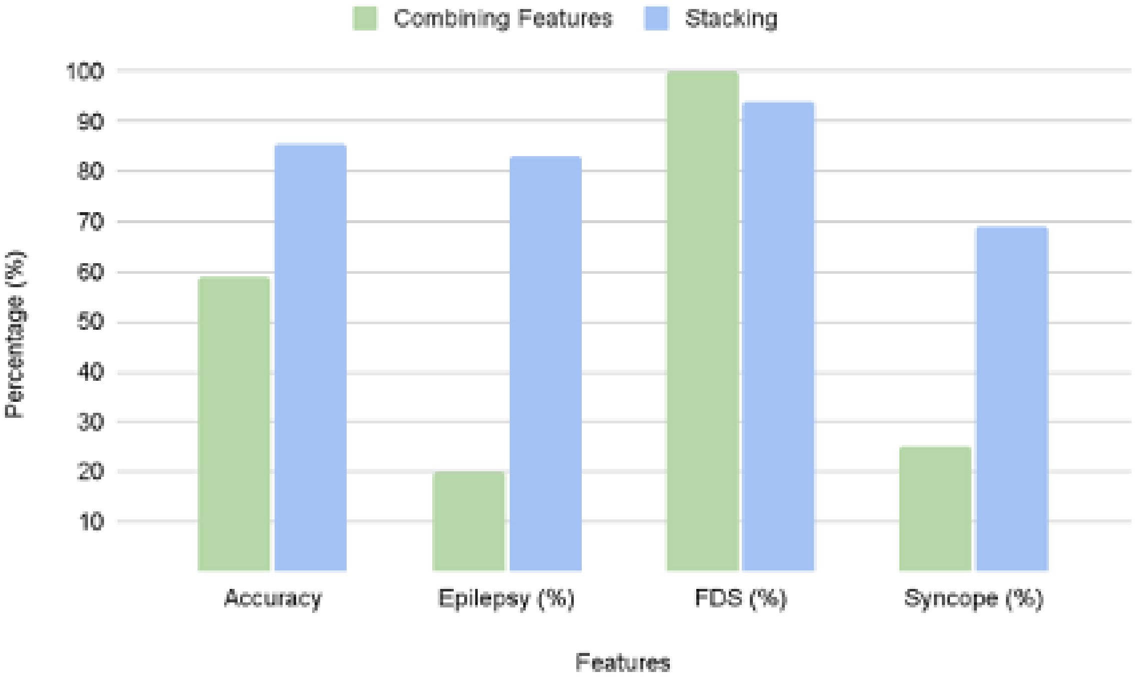


FIGURE 7.2: A bar chart showing the overall accuracy and percentage of people with epilepsy and FDS that were correctly identified by the formulation effort features, LIWC semantic categories, and TFIDF features using a support vector machine with an RFB kernel.

## 7.4 Discussion

The first objective of this chapter was to explore whether the predictive performance of the iPEP demonstrated in previous research (Wardrope, Newberry, and Reuber, 2018) is maintained when the questionnaire is administered in a binary format using an online web application. We found that the iPEP was not as effective at predicting the diagnosis for the responses collected using the online web application. The accuracy of a Random Forest algorithm trained using the iPEP responses from previous research showed a reduction in accuracy of 15.6% for the patient only questionnaire and of 36.9% for the patient and witness questionnaire when applied to the responses collected through the online web application compared to a validation dataset extracted from the questionnaire responses provided in the original iPEP study (Wardrope et al., 2020a). These findings suggest that the response profiles collected through the online web application are different to the responses collected using the 5-point Likert scale in previous research. One potential explanation is that the sample collected in this study is coincidentally different from the previous research. A second potential explanation is that administering the questionnaire in binary format changes the response profile of participants and therefore reduces the concordance with the original iPEP questionnaire.

If administering the iPEP as a binary questionnaire influences the responses that participants provide, training a separate machine learning algorithm using the responses collected through the iPEP (application) may improve the predictive performance because the model will be trained to detect the patterns that are present in the new dataset (Jordan and Mitchell, 2015). However, we did not observe a large increase in performance when the iPEP was trained using leave-one-out cross validation. The accuracy of a Random Forest model trained using the binary patient-only iPEP and leave-one-out cross validation was equal to the accuracy of the model trained using answers provided by patients in the previous research study on a five point Likert scale (Wardrope et al., 2020a). Furthermore, there was a decrease in the performance when the stratification was based on the combination of patient and witness responses provided on the iPEP (application). One potential explanation for this finding is that there was an insufficient number of participants in the training data of the online dataset to allow the model to accurately and reliably detect the group differences needed to achieve a better distinction. This may explain why the patient-only iPEP trained using leave-one-out cross validation had the greatest accuracy for people with FDS because this was the largest cohort within the sample. Future research could investigate this further by exploring the performance of the iPEP using a larger sample size.

Improving the predictive performance of the iPEP is dependent upon how the features are integrated into the model. Training a single machine learning classifier using the iPEP and language analysis features caused an overall decrease in performance of the iPEP by 6.8%. The language features were designed to differentiate between epilepsy and FDS and therefore did not perform effectively when applied to individuals with syncope. These findings do not suggest that an automated analysis of spoken descriptions of TLOC is not effective for detecting syncope, but it does suggest that additional features that are designed to detect syncope should be considered in the future. For example, features could be designed to detect the situational triggers (Lempert, Bauer, and Schmidt, 1994; Colman et al., 2004) and frequently observed symptoms of syncope (Malmgren, Reuber, and Appleton, 2012). Furthermore, the potential of our model was restricted because there were very few cases of syncope within the dataset, which would make it difficult to detect patterns in the responses for people with syncope. Therefore, future research should explore whether there are additional features that can improve the differential diagnosis of syncope using a larger dataset.

The model stacking approach could be a flexible and appropriate method to use within a clinical decision tool in the future because it does not require the same information from all participants. Although witness observations have previously been shown to improve the performance of clinical decision tools (Chen et al., 2019; Wardrope et al., 2020a), not all patients will have witnesses who are able to complete these questionnaires. Furthermore, not all patients within our study completed the interaction with the VA, and a similar trend may be observed if the application were to be used in clinical practice. Model stacking would allow predictions to be made on the data that is available, but also allow the generation of more accurate predictions using any additional information. Based upon this flexibility, future research could investigate additional methods of improving the accuracy of the approach. For example, our patient and witness iPEP model was not effective at predicting the diagnosis, but the model was able to identify 100% cases of syncope in previous research (Wardrope et al., 2020a). The predictive performance of the witness questionnaire may increase if there was more data available to train the leave-one-out cross validation model, and this model could be used to further stratify the predictions of syncope to reduce the number of people with epilepsy or FDS who are incorrectly predicted as having syncope using the patient-only iPEP. The model could incorporate an analysis stage where iPEP predictions for syncope are analysed to further stratify individuals with syncope and transfer the predictions of epilepsy or FDS that were not identify by the iPEP back into the automated analysis of language for epilepsy and FDS. Moreover, we have not explored the feasibility of predicting the cause of TLOC using an automated analysis of witness descriptions of what happened. Future research using CA could use the data that we have collected to identify potential patterns in witness descriptions that could be detected using automated methods to improve the detection of syncope, epilepsy, and/or FDS. Model stacking could allow these methods to be incorporated into the model while still maintaining a baseline prediction using the patient-only iPEP.

### 7.4.1 Limitations

Many of the limitations discussed in chapter six also apply to this analysis given that the analysis uses the same data. However, there are two additional points that are more relevant to highlight in this chapter rather than chapter six.

The sample size in this study is very small for a machine learning research project. The recruitment for the project was hindered due to the coronavirus pandemic because most people who were referred to the seizure and syncope clinics at the Royal Hallamshire Hospital attended their appointments remotely, which made it more difficult to approach potential participants and engage them in the research project. We attempted to overcome the reduced sample size using nested leave-one-out cross validation to evaluate the effectiveness of the model at predicting the diagnosis of unseen data, but the consequence of this approach is that we were unable to create a single machine learning model, which reduces the utility of the research because the model cannot be applied to future samples. Fortunately, the objective of the research was to explore the feasibility of the approach, which provides a justification for continuing this research in the future. Therefore, future research should further validate the method using a larger sample size.

The sample used in this study is not ethnically diverse because most participants were white and British. The data used to train an ASR system often uses speech from individuals who are native speakers of the target language, but these models can perform less effectively for individuals who are non-native speakers of the target language (Cumbal et al., 2021). Therefore, ethnicity can have an impact on the performance of an automated analysis of language (Latif et al., 2020), and these confound variables, alongside additional confounds, should be explored more extensively in future research.

### 7.4.2 Conclusion

This chapter has explored the feasibility of predicting the cause of TLOC using an online patient symptoms and witness observation questionnaire (iPEP) and an automated analysis of spoken descriptions of TLOC. We found that the predictive performance of the iPEP was reduced when applied to responses collected through an online web application. Furthermore, we demonstrated that it is possible to improve the challenging differentiation between people with epilepsy or FDS using an automated analysis of seizure descriptions. However, increases in performance were only achieved when the iPEP was used as a first stage stratification tool and the automated analysis of language was restricted to people with epilepsy and FDS. These findings demonstrate the feasibility of using this method to improve the differential diagnosis, but future research can improve upon this research by exploring whether the predictive performance of the version of the iPEP that was administered through the online web application can be improved by training a machine learning model using a larger sample size, identifying linguistic features that are useful for identifying individuals with syncope, creating an ASR system that is tailored towards descriptions of TLOC, and identifying and mitigating confounding variables. Finally, it is important to evaluate the acceptability of the approach from the perspective of users to ensure this is a clinical decision tool that patients and witnesses would be prepared to use.

# Chapter 8

# Evaluating the acceptability of the online application

## 8.1 Introduction

## 8.2 Acceptability

The 2021 framework for developing and evaluating complex interventions recommends that a feasibility study should evaluate the acceptability of a complex intervention (Skivington et al., 2021). It is important that patients with TLOC accept the application because people must be willing to use the application and motivated to thoroughly answer all of the questions for the implementation of the application to be successful. Acceptability is a construct that has been defined and measured for healthcare interventions in many different ways with little standardisation (Sekhon, Cartwright, and Francis, 2017). For instance, it has been defined as "a multi-faceted construct that reflects the extent to which people delivering or receiving a healthcare intervention consider it to be appropriate, based on anticipated or experienced cognitive and emotional responses to the intervention" (Sekhon, Cartwright, and Francis, 2017). This definition is made up of seven different constructs, and the authors of the definition recommend measuring the different constructs at different time points for an intervention (Figure 8.1).

An important first step towards measuring the acceptability of a healthcare intervention is to decide the most appropriate method. Many previous studies have opted to assess acceptability using objective measures, such as withdrawal rates and self-report measures that focus on satisfaction, attitudes, perceptions and experiences (Sekhon, Cartwright, and Francis, 2017). Although these measures do provide insight into the acceptability of an intervention, a large range of measures are required to capture all the component constructs for acceptability, and objective measures may not provide insight into what improvements to the complex intervention can increase the acceptability. Therefore, we have chosen to take a multi-modal approach to evaluating acceptability using quantitative and qualitative methodologies. Given that there is no clear decision boundary about what constitutes an acceptable intervention, we will qualitatively evaluate the findings of our analysis in accordance with each of the component constructs of acceptability to gain insight into what the users consider important determinants of acceptability for the application and potential areas of improvement that will improve the overall acceptability.

### 8.2.1 Technology Acceptance Model

The Technology Acceptance Model (TAM) is a theory that attempts to explain the process and factors that influence a person's acceptance of a new technology (Davis,

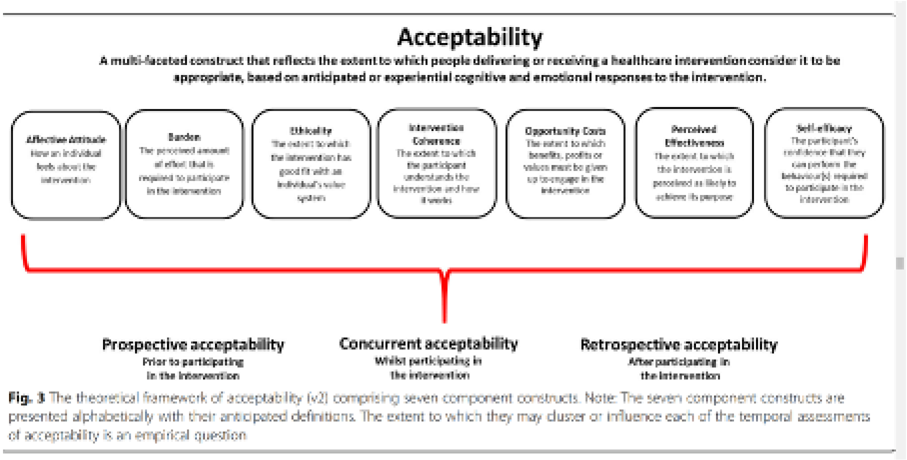

Acceptability
A multi-faceted construct that reflects the extent to which people delivering or receiving a healthcare intervention consider it to be appropriate, based on anticipated or experiential cognitive and emotional responses to the intervention.
Affective Attitude
How an individual feels about the intervention
Burden
The perceived amount of effort that is required to participate in the intervention
Ethicality
The extent to which the intervention has good fit with an individual's value system
Intervention Coherence
The extent to which the participant understands the intervention and how it works
Opportunity Costs
The extent to which benefits, profits or values must be given up to engage in the intervention
Perceived Effectiveness
The extent to which the intervention is perceived as likely to achieve its purpose
Self-efficacy
The participant's confidence that they can perform the behaviour(s) required to participate in the intervention
Prospective acceptability
Prior to participating in the intervention
Concurrent acceptability
Whilst participating in the intervention
Retrospective acceptability
After participating in the intervention

**Fig. 3** The theoretical framework of acceptability (v2) comprising seven component constructs. Note: The seven component constructs are presented alphabetically with their anticipated definitions. The extent to which they may cluster or influence each of the temporal assessments of acceptability is an empirical question

FIGURE 8.1: The seven different constructs that define acceptability. The image is taken from Sekhon, Cartwright, and Francis (2017)·

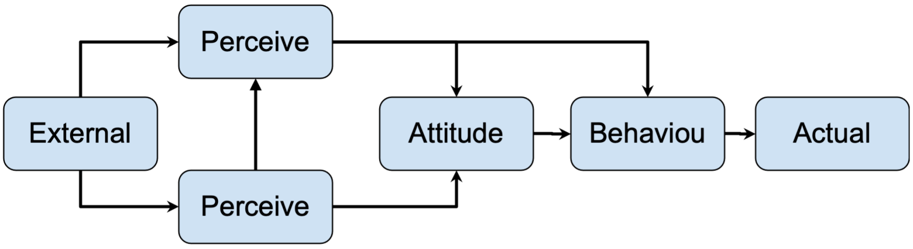


FIGURE 8.2: A schematic of the related constructs in the Technology Acceptance Model.

1989; Davis, Bagozzi, and Warshaw, 1989). The theory is an extension of the theory of reasoned actions (Ajzen and Fishbein, 1969). According to TAM, the adoption of new technology is mediated by the behavioural intention to use the technology, which is in turn influenced by the users attitude (Figure 8.2). There are two notable mediators of attitude within the mode, the perceived usefulness and perceived ease of use of the technology. The model theorises that having positive perceptions of these factors will result in increased usage of the technology. However, these factors can be influenced by external social factors that must be considered. The relevant external factors may vary depending on the type and purpose of a technology, but prominent examples may include demographic information such as age.

TAM has been extensively applied to technology inside and outside of the healthcare domain. A review of the validity of TAM across studies exploring medical professional adoption of health technology found that many of the relationships in the model are frequently validated across studies with a large effect size, although the relationship between ease of use and acceptance was less frequently demonstrated (Holden and Karsh, 2009). Similar findings have been reported for studies applying TAM to patient acceptance of medical technology. Razmak and Bélanger (2018) found that TAM was able to predict patient intention to access and use medical records, although they also did not report a significant relationship between perceived ease of use and attitude. El-Wajeeh, Galal-Edeen, and Mokhtar (2014) validated the relationship between the fundamental constructs in the model when predicting the intention to use mobile health technology. Finally, Lanseng and Andreassen (2007) validated the model for predicting the intention to use self-service diagnostic technology, but were unable to detect a direct relationship between expected usefulness and behavioural intention. Furthermore, their model was extended to explore the relationship between the construct 'trust' and two of the original constructs from the model: expected usefulness and expected ease of use. Trust was found to influence the two constructs. These findings demonstrate that TAM can be an effective model for understanding the factors that influence an individual's acceptance of medical technology and can provide insight into the factors that may influence the future intention to use the technology, which can help shape the future design and implementation of the technology.

### 8.2.2 Thematic Analysis

A qualitative analysis may be effective at capturing the independent perspective of users of the healthcare intervention by measuring the meanings, experiences, and views of the participants to establish what people think about the intervention and why they think it (Pope and Mays, 1995). The depth associated with a qualitative analysis can help to determine how acceptable people think the application is and what changes can be made in the future to improve the acceptability as the development of the application advances.

Thematic analysis is a method that may be useful for evaluating the acceptability of a healthcare intervention. Thematic analysis is a qualitative method that involves identifying, analysing, and reporting patterns within data (Braun and Clarke, 2006). The patterns are organised into different themes that are used to produce a detailed description of the data. Previous research has used thematic analysis to understand the attitudes towards healthcare interventions that can account for intervention adherence. Jørgensen et al. (2019) demonstrated that various themes extracted from interviews with mental health nurses can explain attitudes and adherence towards the "Guided Self Determination" intervention, such as whether people view themselves as a novice with regards to the intervention and understand the theoretical underpinnings of the intervention. Furthermore, Valley and Stallones (2018) explored how perceived benefits, perceived barriers, and self-efficacy can describe adherence to a mindfulness intervention for health care workers, which could be used to influence the adoption of mindfulness practices in healthcare workers. These studies demonstrate that thematic analysis can provide insights into attitudes and usage of healthcare interventions that could be used to make evaluations and improvements.

Evaluating the acceptability of a healthcare intervention from the patients perspective requires consideration of the individual experiences of having a specific diagnosis and the corresponding healthcare that is associated with the condition. Thematic analysis has been used to understand the lived experiences of people with epilepsy and FDS. People with epilepsy have reported that the onset of seizures marks a significant moment in their life that has an impact on their family and friends, finances, employment, independence, and self-esteem because of the stigma and consequences of having a seizure and taking anti-epileptic medication (Rawlings et al., 2017b). Many individuals with epilepsy report having positive experiences with health professionals and view medication as a means of controlling seizures, although there were forms of trial and error associated with identifying optimum treatments (Rawlings et al., 2017c). People with FDS have reported a lack of understanding about their health condition by themselves, others, and health professionals, negative attitudes by health professionals, and limited access to treatment, even though individuals report increased psychiatric comorbidities, isolation, and a reduction in the ability to cope (Rawlings et al., 2017c). The process of receiving a diagnosis is often long and many of the tests for identifying FDS often produce "normal" results, which may explain why many individuals are resistant to the diagnosis (Rawlings et al., 2017c). Using a qualitative approach will allow us to understand how people draw upon knowledge and experience while evaluating the acceptability of an automated method of predicting the cause of TLOC.

It is equally important to consider the experiences of individuals supporting people with a diagnosis of TLOC. Witness accounts of what happened during an episode of TLOC can be vital for making the correct diagnosis (Chen et al., 2019; Wardrope et al., 2020a). Therefore, it is important that an automated method for predicting the cause of TLOC is acceptable to witnesses as well as patients. Many witnesses, for example family members, are consistently involved in the patient care pathway and experience their own trials and tribulations throughout the transition between recognising a problem, receiving treatment, and recovery (Pieters et al., 2016). These experiences may be influential in the evaluation of acceptability.

### 8.2.3 Aim

This analysis will use a mixed-methods approach to evaluate the acceptability of the online application from the perspective of the patients and witnesses that have used it. Firstly, we will explore the general attitudes towards the application using the responses to a closed questionnaire. The questionnaire is based upon the Technology Acceptance Model, a theory that provides insight into people's intention to use new technology. Secondly, we will use the methodology of thematic analysis to explore the attitudes of patients and witnesses towards the application. This analysis will use an inductive approach to increase understanding of how the application is perceived by the users based upon their lived experience. We hope that the thematic analysis will provide a more in depth understanding that can guide the interpretation of user attitudes that were measured using the Technology Acceptance Model. Thirdly, we will evaluate how people have used the application, for example dropout rates, because these findings may provide further insight into the perceived acceptability of the application. Finally, we will evaluate each of the component constructs of acceptability in accordance with the outcome of each stage of this analysis to formulate conclusions about the perceived acceptability of the application and potential areas of improvement.

## 8.3 Method

### 8.3.1 Mixed Methods Approach

The mixed methods research approach involved combining quantitative and qualitative research methods to increase the breadth and depth of understanding and corroboration, thereby strengthening the conclusions that stem from the research project (Johnson, Onwuegbuzie, and Turner, 2007). There are many different approaches that can be used in mixed methods research and the most appropriate research design for a given research question should be constructed based upon a series of decisions (Schoonenboom and Johnson, 2017). One of the first important decisions is determining the objective of the mixed methods research (Greene, Caracelli, and Graham, 1989; Bryman, 2006). Our objective is to use the qualitative method to provide contextual information about the broader feedback questionnaire (Bryman, 2006). The questionnaire could be more generalisable because more people will provide feedback through this method. Furthermore, given previous research has demonstrated that the underlying factors measured by the questionnaire are predictive of intention to use the application, the feedback can help determine potential changes to improve intention to use. The qualitative feedback on the application can provide context about the quantitative feedback and help guide future improvements or changes to improve acceptability. The second important design decision regards whether the different components of the project are conducted sequentially or concurrently (Guest, 2013). Although all participants completed the feedback questionnaire before the interviews were conducted, the interviewer did not look at the feedback questionnaire responses before interviewing participants. The objective was to collect feedback that was guided by what the participants wanted to share, rather than the objective of the researchers. The third important design decision involves determining the "point of integration" where each of the independent research components are mixed or connected in some way (Schoonenboom and Johnson, 2017). Our findings were integrated during the inferential stage of the research project (Tashakkori, 2009) where we explore the insights into acceptability that each component of the project can provide. Therefore, the final inferential stage that evaluates the acceptability of the application will utilise a deductive approach by applying the construct of acceptability to the findings from each of the components of the mixed-methods approach.

### 8.3.2 Sample

Participants who took part in the wider research project chose whether they wanted to provide feedback in the form of a questionnaire, telephone interview, or both. Therefore, there are different participants for each component of this project. Table 8.1 provides a breakdown of the demographic information for each component of the analysis.

A total of 54 participants completed the feedback questionnaire (59.3% of whom were patients). The respondents were aged 16-82 $\bar{x}$=44.5, std=25). More of the respondents were female than male (66.7%) and most were white and British. The most frequent level of education was a university undergraduate degree or above (57.4%).

The feedback interviews were conducted with 24 participants, from which 66.7% were patients. Interviewees were aged 19-80 ($\bar{x}$=52.3, std=19.5). More of the interviewees were female (58.3%) and were predominantly white and British. The most frequent level of education was a university undergraduate degree or above (66.7%).

TABLE 8.1: A breakdown of the demographic information for the participants who provided feedback on the online application by completing the feedback questionnaire or participating in the feedback interview.

| | | Questionnaire | Interview |
|---|---|---|---|
| N | | 81 | 24 |
| **Participant Type** | | | |
| | Patient | 58 | 16 |
| | Witness | 23 | 8 |
| **Age** | | | |
| | 16-25 | 14 | 3 |
| | 26-40 | 21 | 5 |
| | 41-65 | 31 | 10 |

Continued on next page

**Table 8.1 – continued from previous page**

| | | Questionnaire | Interview |
|---|---|---|---|
| | 66+ | 15 | 6 |
| **Gender** | | | |
| | Male | 27 | 10 |
| | Female | 54 | 14 |
| **Ethnicity** | | | |
| | White British | 57 | 16 |
| | White European | 1 | 1 |
| | Black British Caribbean | 0 | 1 |
| | White | 12 | 3 |
| | German | 1 | 0 |
| | White mixed | 1 | 0 |
| | White other | 1 | 0 |
| | British | 5 | 2 |
| | Indonesia | 1 | 0 |
| | Any other asian | 1 | 0 |
| | Mixed | 1 | 0 |
| **Education** | | | |
| | No education | 3 | 0 |
| | Secondary education (GCSE or equivalent) | 11 | 2 |
| | Further education (A-Level or equivalent) | 27 | 6 |
| | Higher education (Undergraduate or equivalent) | 24 | 7 |
| | Higher education (postgraduate degree or equivalent) | 11 | 6 |
| | Higher Education (PhD) | 4 | 2 |
| | Information not available | 1 | 1 |

### 8.3.3 Thematic Analysis Analytic Approach

Semi-structured qualitative research interviews were conducted on average 11 days (std = 7.4) after participants completed the procedure. The interviews were conducted over the phone, recorded, transcribed verbatim, and subsequently checked for any errors. Participants were informed that although the interviewer had a short list of questions (Table 4.7), the purpose of the interview was to hear all of their thoughts regarding the application and they were free to talk about anything that came to mind. To prevent the feedback becoming too constrained by the questions, the interviewer encouraged participants to openly share thoughts during the initial stage of the interview and then subsequently asked more specific questions that could prompt further thoughts once participants indicated they were ready to move on. On average the interviews lasted 22.8 minutes (SD = 13.5).

The interviews were analysed using the methodology of thematic analysis outlined by Braun and Clarke (2006). Firstly, the main author read through the interviews repeatedly and made notes to become familiar with the data and review the quality of the transcripts. Secondly, the first author used an inductive approach to coding the data based on the features present within the data. Thirdly, the codes were organised into four initial themes with sub-themes. The early themes followed the format of "bucket" themes (Braun and Clarke, 2022). All authors met to discuss the codes and themes. Subsequently, NP & TW reviewed the data in greater detail and explored how the themes could be developed to transition from a "bucket" to a story, leading to the generation of the three themes that transitioned from purely describing the data to including some interpretation. Fourthly, the first author began reviewing the themes by revisiting the original data and codes to explore whether the themes align with the data and to ensure the themes have internal and external homogeneity (Patton, 1990). During stages three and four, it became apparent that the codes were predominantly semantic and much of the latent content was nested under semantic codes. Therefore, some of the codes were split into multiple smaller codes to allow the latent content to be assessed independently. The themes were then written into a research report to provide an extensive account alongside the supporting evidence.

### 8.3.4 Questionnaire Preprocessing

The questions were grouped based upon the latent construct they were designed to measure by the Technology Acceptance Model. The scoring was reversed on the negatively framed questions to ensure the scores were consistent for each question. We calculated descriptive statistics for each latent construct for the ease of reporting the findings, but independent questions were explored where appropriate for providing further information regarding the acceptability of the application. Missing values were removed from the analysis using a pairwise deletion.

## 8.4 Results

### 8.4.1 Responses to the closed questionnaire

The internal consistency of each of the component constructs from the Technology Acceptance Model were measured using Cronbach's alpha. The internal consistency was considered "acceptable" based on the recommended guidelines (Nunnally, 1978) for usefulness (0.74), attitude (0.77), and ease of use (0.74).

When the average score across all participants for each construct is rounded to one significant figure, people reported that they somewhat agree that the application is useful ($\bar{x}$=4.99, std=0.95), easy to use ($\bar{x}$=4.48, std=1.06), and for the attitudinal measures regarding the application ($\bar{x}$ = 4.84, std=0.98). For the single question about whether people would use the application if it were available in the future, on average people selected somewhat agree ($\bar{x}$=5.13, std=1.57).

### 8.4.2 Themes from the thematic analysis

There were three themes identified by the thematic analysis. The first theme centered around the importance of providing the right information in order to receive an accurate diagnosis, and the influence of the web application on the information that people can provide. In the second theme, people often described the medical pathway as a journey and discussed the impact the application might have on the progression through this journey. The final theme centered around the importance of control and the ability to make choices about the application. Control over the application design was thought to influence the information that people provided, which related back to the first theme.

#### 8.4.2.1 Providing information is crucial for making the diagnosis

TABLE 8.2: A breakdown of the percentage of responses for each of the component constructs of the Technology Acceptance Model. Missing responses were removed using a pairwise deletion. Therefore, the percentages on each row do not always sum to 100%.

| | **Strongly disagree** | **Disagree** | **Somewhat agree** | **Neither agree or disagree** | **Somewhat agree** | **Agree** | **Strongly agree** |
|---|---|---|---|---|---|---|---|
| Usefulness | 0% | 2.47% | 1.23% | 23.46% | 41.98% | 27.16% | 3.7% |
| Ease of use | 0% | 2.47% | 13.58% | 34.57% | 28.4% | 19.75% | 1.23% |
| Attitude | 1.23% | 1.23% | 6.17% | 22.22% | 46.91% | 18.52% | 2.47% |
| Intention to use | 3.7% | 6.17% | 3.7% | 9.88% | 30.86% | 23.46% | 19.75% |

TABLE 8.3: A breakdown of the responses for each of the questions on the 17 item questionnaire. The 7pt Likert scale was converted into binary format for ease of interpreting the findings. The percentages were calculated before the pairwise deletion of missing values. Therefore, the percentages on each row do not always sum to 100%.

| Question | Disagree | Unsure | Agree |
|---|---|---|---|
| Save time | 7.4% | 17.3% | 75.3% |
| Save effort | 14.8% | 17.3% | 64.2% |
| More convenient than booking and attending a medical appointment | 19.8% | 16.1% | 61.7% |
| Easier than booking and attending a medical appointment | 11.1% | 13.6% | 71.6% |
| Help me to be referred to the right service | 4.9% | 24.7% | 66.7% |
| Help me to receive the correct diagnosis | 11.1% | 34.6% | 50.6% |
| Confusing to use | 66.7% | 9.9% | 22.2% |
| Time consuming | 81.5% | 8.6% | 7.4% |
| Takes a lot of effort to use | 77.8% | 8.6% | 11.1% |
| Complicated to use | 81.5% | 6.2% | 8.6% |
| Required little work to use | 11.1% | 9.9% | 75.3% |
| Easy to talk to | 17.3% | 8.6% | 72.8% |
| Good | 9.9% | 16.1% | 71.6% |
| Pleasant | 9.9% | 18.5% | 70.4% |
| Beneficial | 8.7% | 17.3% | 71.6% |
| Favourable | 21% | 28.4% | 49.4% |
| I would use it in the future | 13.6% | 9.9% | 74.1% |

TABLE 8.4: Quotes from the feedback interviews that relate to the theme “providing information is crucial for making the diagnosis”.

| label | Quote |
|---|---|
| T1A | it has to rely on collecting information that I give it, and if I don’t give it then it potentially misses out on information that would be relevant in making a diagnosis |
| T1B | I answered all the questions to the best of my ability and I think I gave, and I gave everything that I thought was right for that question. But I don’t know if it was because it can’t tell me whether I’ve given it enough information, so I don’t know whether that’s, and I don’t know how you would ever know. |
| T1C | If I didn’t feel that I’d given enough information or the system didn’t allow, I would probably follow it up and say, I don’t think I was able to say everything and I need to speak to somebody about it. |
| T1D | if you’re not getting any prompts and you’re not getting anything, even any kind of eye contact, any sorta smiles or raised eyes or ooh really, that’s interesting from the clinician, you might not say as much as would be useful to help your diagnosis. |
| T1E | A lot of the symptoms were kind of really odd things to explain. Like some of the (1 second) just the things I felt, and so (1 second) just having the, you know, doctor just being able to just nod and, OK, yeah, I get it, almost. (1 second) Just helps. |
| T1F | Um, even when you go to the doctor you like forget stuff until, because they’re very well trained (laughter) until somebody prompts you (1 second) and says, well tell me more about that, or what do you mean when you say fainting and, you know, that kinda thing. So I felt like I’d only really touched the surface. |
| T1G | Cos it’s kind of, it’s kind of linear and so you don’t get the chance to say, oh well when I said that at the beginning what I meant was (1 second) do you, do you see what I mean? |
| T1H | I felt it was just a bit restricted in terms of just having a yes/no option; I wanted kind of a maybe option to then be able to say, well my, my case was, was this, if that makes sense? |
| T1I | And the other one was that there were some questions asked and then they were almost yes/no answers but I’d got background coming off, like I’ve got a pacemaker. And it didn’t give me the opportunity to, to input that |
| T1J | Yeah, I think for me it enables you to articulate, I think I might have touched on this earlier to be fair, but I think it, it enables you to articulate what happened more freely than you would be being asked questions on a, on a form or something as well. |
| T1K | It’s cos you’re able to go into a bit more detail as well. And I think, sorry, if you can go into more detail then I think that’ll (1 second) help diagnose it a bit quicker as well. |
| T1L | I found that more reassuring and gave me more confidence |

Continued on next page

**Table 8.4 – continued from previous page**

| Label | Quote |
|---|---|
| | in the application because it made my, made me think oh it must be quite sophisticated if it can take information from tha |
| T1M | perhaps more open ques, asking more open questions like, you know, do you (1 second) do you have a range um; what is it? Do you have a range of experiences relating to blackouts? Um, something whi, which opens up the ability for you to, for you to give infor, more information |
| T1N | I was taking blood pressure medication, which, you know, a common side effect of that is you can feel, um you can feel dizzy or light-head, headed and, and both the medication, in fact there were three medications I had all had that as a potential side effect. So my GP and in fact other consultants I've seen thought it was probably to do with um (1 second) with, with medication, um but I stopped taking the medication altogether about three months ago and the, and the problem, er while that general dizziness that I, well in fact it's not dizziness but something light-headed um has (1 second) um has got less |
| T1O | I mean um (1 second) I think specifically in, in, in the um case of, of my daughter, of (Daughter name) um that I, I, I thought that I hadn't had enough time to um go through the, the sort of the, the social psychological er or emotional context er (1 second) you know, of her situation at the moment um, and um, you know, what she'd experienced in the, in the last days, weeks, er months before she had the seizure. Er, I, I didn't find that there, there was an opportunity to do this in sufficient detail er during the exchange with, with the er, with the um AI. |
| T1P | I would feel more content personally that, ooh, I've done a good job, giving the yes/no answers seems a bit er (1 second) when you think there's context to go with it, if you've not put the context in you feel what, not fulfilled. |
| T1Q | that's sort of guiding me in one direction a bit and I'm, I'm not sure that's gonna guide, guide this process in the right direction for me. But that was more to do with the questions themselves again, as I said, cos they didn't cover the scope of what I was experiencing. |
| T1R | But you really do think, need to think about, about the questions almost from a patient's perspective rather than just a doctor's perspective. |
| T1S | Yeah, exactly, yeah. And I think even if, even if there was sort of an opportunity to say, is there anything you want to add, I think for both of us (Pat name) could have, could, could have kinda contextualised why he, he said no to some of the questions. |

This theme centres around the ability of users to provide the necessary information to the application to make a diagnosis. Providing information, particularly the relevant information, is considered a vital preceding step towards making a diagnosis and any hindrance to this task is understood to hinder the diagnostic process [T1A]. Although there were some participants who felt they were able to give all the information that they wanted, many people reported that they could not and outlined a range of factors that caused this restriction when compared to routine doctor-patient interactions.

People expressed uncertainty about whether the information they had provided was sufficient to make the diagnosis [T1B]. During routine consultations, the doctor, not the patient, has the knowledge of what information is required to make the diagnosis. In the absence of a doctor, the patient is responsible for making this judgement but must do so without the knowledge and expertise to make a diagnosis. Users expressed concern that they may not share diagnostically relevant information, which may influence their trust in the application and prompt them to seek out further support from medical professionals [T1C]. Some participants requested prompts alongside the questions to aid them in answering the questions. Prompts would remove some of the responsibility from the patient because they are less responsible for deciding what information is relevant and can tailor their responses based upon what the professionals have decided is important.

In addition to knowing which information is relevant for making a diagnosis, the verbal and non-verbal feedback that doctors provide was considered important for facilitating and encouraging patients and witnesses to share information. Users orientated towards a continuous interaction between the patient and doctor where non-verbal cues prompt patients to continue talking [T1D], which was absent from interactions with the VA and consequently resulted in shorter and less detailed responses. This was considered particularly troublesome during descriptions of symptoms that were considered difficult to explain, for example subjective symptoms that are experienced during a seizure [T1E]. The absence of interactivity during the application did not encourage users to elaborate their responses based on the recipient's interest [T1F] and caused users to perceive the interaction in a linear fashion where they were unable to draw upon and elaborate their previous responses [T1G], which often resulted in superficial responses.

The questions were another element of the application that influenced the information that people were able to provide because they made relevant certain responses. People reported feeling restricted by the binary questionnaire because they were unable to give the response that they wanted, for example some people did not have any memory of the seizure so would have preferred to provide a neutral "maybe" response [T1H] and others felt there was important contextual information that should have been provided alongside their response, for example that palpitations were associated with a pacemaker [T1I]. In contrast, many people reported a preference for providing spoken descriptions of their attacks because they were able to provide more information [T1J] and that this detail would help make a diagnosis [T1K]. Providing spoken descriptions gave the impression that the application was more sophisticated and increased user confidence [T1L]. However, many users still felt that these questions restricted the amount of information that they could provide because there was information that they believed was important for the diagnosis that was beyond the scope of the topical agenda of the questions (Heritage and Maynard, 2006), for example providing more detail about the different type of attacks that they experience [T1M] and contextual information that may be related to symptoms and the cause of the attacks, for example the side effects of medication [T1N] and social, psychological, and emotional factors [T1O]. The questions appear to presuppose that users can answer the questions and that the response they give is sufficient to make a diagnosis, but instances where this presupposition does not align with the users expectations or experience can leave users feeling unfulfilled

[T1P] and concerned about how effective the application will understand their experience [T1Q]. Changing the design of the questions, for example generating questions that allow patients to share all the information that they feel is important [T1R] and providing people with the opportunity to add relevant information at the end of the procedure [T1S], may increase the trust and satisfaction for users of the application.

#### 8.4.2.2 Progression through the medical pathway

TABLE 8.5: Quotes from the feedback interview that correspond to the theme "progression through the medical pathway"

| label | Quote |
|---|---|
| T2A | speed up process of people getting diagnosed a lot quicker |
| T2B | er the doctor's gonna call you back (1 second) on Thursday (laughs) and it's now Monday (laughs) you know, um you can speak to a doctor as such straight away and know then that your situation has immediately gone into the, into the system |
| T2C | Cos it took 'em about four/five years before they finally diagnosed me. (laughs) |
| T2D | I wouldn't want is, is it to be used by the NHS as, as a sort of a, a delaying tactic |
| T2E | there was like a huge waiting list to even be seen, and I think it could massively reduce people's anxiety if, if they kinda like had this intermediate step of like a digital consultation |
| T2F | follow-up must be good and quick. |
| T2G | I don't know. When things disappear into cyber space I'm, I'm always thinking well does, does, does it go somewhere or has it got, has it got lost in cyber space, do you know what I mean? So, you know, you kind of need to know that it's gone somewhere for a start |
| T2H | I want to see um something come up (1 second) which tells me, thank you, um (1 second) thank you, um we will be back in touch with you, somebody will call you or what have you in (1 second) X amount of (2 seconds) time |
| T2I | I wouldn't want it to be the only tool, I would still want somebody to sit in front of me and to probe. |
| T2J | I'd hate to think that the computer diagnosed me and the neurologist went along with the computer and didn't explore |
| T2K | I think I believe that er I myself, it's a personal opinion, um, you know, I, I believe that the, the doctor's er empathy and understanding is very important |
| T2L | I found that quite (1 second) comforting cos it were just like speaking to a normal doctor |
| T2M | it feels a lot nicer than you, you know, if, if you, you might be kind of do a, on triage or you might be given a questionnaire, so it's a lot nicer than doing sort of some big kind of yes/no tick box questionnaire because you've got that |

Continued on next page

**Table 8.5 – continued from previous page**

| Label | Quote |
|---|---|
| | opportunity to, to, to kinda speak |
| T2N | I didn't feel connected to it, let's put it that way. I, I kind of felt like it was an exercise I was doing, I didn't, didn't feel like it was er anything I was connected to on a personal or medical level. |
| T2O | it's fine as long as it's presented to you as, as the very early stages, this is what we do in the very beginning because it will help us direct you in the right direction. You know, it's just, as I say, I wouldn't mind if I knew that there was more to come after it |
| T2P | It would just be really useful (laughs) if you could tell one person or like one computer application and then everybody got the information; I think that's a frustration for people |

This theme describes how the medical care pathway is often conceptualised like a journey because users often describe the importance of making progress along the pathway and the defining milestones along the way. Many areas of feedback about the application focus on the impact that the application will have on progress along the pathway. Many users thought that the application would lead to faster progression along the pathway [T2A] and referenced the importance of the information being in the 'system' [T2B]. Having information available in the 'system' may represent a milestone along the medical care pathway where information about the health problem is visible to medical professionals and the professionals can begin to make progress in their responsibilities, for example making referrals. Many individuals reported long diagnostic delays [T2C], an unpleasant experience where progress along the medical pathway is delayed, and it was important that the application improved progress rather than delaying it [T2D]. Furthermore, one user suggested that completing the application may give rise to the perception that they are making progress along the medical pathway while they are waiting for upcoming appointments, which would reduce some negative emotions associated with the delay [T2E]. In order for this to be effective though, there should be a quick follow-up [T2F] and progress should be observable to the user in the form of knowing who the information has been sent to [T2G] and when the follow-up will be [T2H].

A second facet to the theme was about whether the application would change the current medical pathway. It was important to users that the application did not replace consultations with doctors [T2I] and that doctors did not solely rely on the output of the application to make a diagnosis [T2J]. One reason for this is that the doctor's empathy and understanding were an important part of the medical pathway [T2K] and although some people thought speaking with the VA was pleasant [T2L], particularly more pleasant than completing a questionnaire [T2M], the VA lacked the human connection that is present in human interaction [T2N]. Therefore, the application should be used to complement rather than change the pre-existing medical pathway [T2O], although one user did speculate whether the spoken descriptions of what happened during an attack could be shared with multiple medical professionals to reduce the requirement to repeat the story on several occasions [T2P].

#### 8.4.2.3 Making choices while completing the application

TABLE 8.6: Quotes from the feedback interviews that correspond to the theme “making choices while completing the application”.

| label | Quote |
|---|---|
| T3A | just give the user more control over what’s actually happening on screen. |
| T3B | So it is like do you wanna see a male or female. (1 second) Person? So then you click and you get the, either the female or the (2 seconds) male one, yeah, all stuff like that, and if you wanna see some (1 second) you, um related to your ethnicity or something like that, so that you kinda feel; er, er what do you call it now? Comfortable |
| T3C | Because of those inbuilt um (1 second) experiences that you have and you relate to, um someone in the medical profession. I still have the prejudice of, I’m just being honest here, inbuilt, this is not what I agree with logically and rationally, this is my feelings, um I saw an inbuilt thing of um (1 second) (1) men being dismissive and all that |
| T3D | Because for me I was concentrating on telling you what was happening with me and how it happened; and I’m having a conversation literally in the same way I’m having a conversation with you now. So I’m not watching the technology |
| T3E | Other than that I found it really, it was pretty much (?) (laughs) but literally you gave your questions and gave, you gave the question, you gave the answer, you went to your next question, bish, bash, bosh, job’s a good ‘un; it weren’t, it didn’t take an exorbitant amount of time, it were like relatively quick and easy |
| T3F | Um, I thought, I thought it worked really well; it was easy to use. |
| T3G | actually thought about it, it was actually pretty good and it was actually nice to talk to a face rather than just having a voice recorded. |
| T3H | Personally I would, I would just scrap the whole animation thing. Just ha, have a voice instead (1 second) to make, that would have made it far easier to concentrate on the question |
| T3I | I was kind of expecting to see a doctor in a doctors’ surgery looking, looking like a doctor |
| T3J | So I don’t know if it might be worth giving an option to give people chance to think of their answer first, start the recording to give their answer. |
| T3K | And you were able to re-answer if you thought you hadn’t re-answered something |
| T3L | It’d be easier if you could cos you’d actually look over what you said and see if you’d missed anything yerself |
| T3M | I have to stop. So if that happened halfway through (2 seconds) um is there a button you could have that says (1 second) er (1 second) I don’t know, er exit, save, come back to |

The freedom to make choices was a prominent area of feedback. Many people expressed a preference for more freedom to make choices about the design of the application [T3A], for example choosing the type of avatar [T3B]. Having choice can allow individuals to tailor their experience to their own needs and preferences to make the experience more pleasant, for example one individual reported that they had experienced discrimination by male medical professionals in the past and explained that they would feel more comfortable if the avatar was female [T3C].

Choosing the design of the application may allow users to self-manage their expectations about the application because making these choices at the start will forewarn people of what is going to happen further into the application. One user reported that some design elements of the application acted as a distraction because they drew their attention to the application itself and away from the activity that they were trying to focus on, which was providing information about what happened during their experience of TLOC [T3D]. Given that previous research has demonstrated that people orient to the normative rules in medical interactions (Heritage and Maynard, 2006), it is clear that people know what to expect during these interactions, which is absent when interacting with the application because this is a novel activity. Therefore, choice may allow people to overcome unexpected and unusual experiences.

Having more choice about the design of the application may change how easy the application is to use. Although many individuals reported that the application was easy to use [T3E-F], individuals have different preferences and having greater control of the design of the application would allow people to choose designs that they personally feel are easier to use. For example, people expressed different opinions about the design of the avatar [T3G-I] and the design of the avatar was considered important for concentrating on the task.

Users also expressed an interest to have greater control over how they provide their responses, for example choosing when they start the recording so that they have time to plan their response and the opportunity to review and edit the responses that they have already given [T3J-L]. This would allow people to provide information that they may have forgotten and to revisit the questions if they felt they were not providing the best responses under the current circumstances [T3M]. Although people do not have time to rewrite their responses during interactions with doctors, the interaction is fluid because the patient can ask the doctor to reformulate questions, the doctor can prompt the patient for more relevant information if it was not provided, and the patient can add in more information at a later stage during the interaction. However, this is not possible with the application - the absence of the doctor leaves all the responsibility to the patient. Therefore, users may want more control over how responses are given to allow them to provide responses that they feel are most effective for the task in the absence of the fluid nature of human-human interactions, thus responding to the responsibility to provide the necessary information that the application instils.

#### 8.4.2.4 Field notes on the use of the application

Out of all participants who took part in the experiment, only 78.2% (61/78) completed the interaction with the VA section of the application, indicating a drop-out rate of 22% for participants throughout the procedure. Although there is little feedback about why these participants dropped out of the experiment, some participants indicated that the procedure was taking them longer than they had anticipated and had hoped to return to complete the remaining section at a later date.

There were a range of technological issues associated with completing the application. One patient logged into the application using the witness login details rather than the patient login details. One participant skipped the first question and only started answering the second. One participant reported that they thought the application was recording and spoke their response, but they subsequently had to repeat their response because the application was not recording. The recordings produced by one participant have a high level of distortion, potentially indicating that the recording device was covered during the interaction with the VA. There was background noise in many of the recordings, for example the TV. The final audio file was not received by the server for one participant, which may have been caused by a connectivity issue on the participants device. Finally, given that the application was initially designed to be used on a single device at the Royal Hallamshire Hospital, the application was not vigorously tested for cross platform devices. There was a device compatibility issue associated with the recording element of the application that interfered with performance on apple mobile devices. Consequently, a couple of participants who only had access to these devices encountered issues providing spoken descriptions through the application.

### 8.4.3 A deductive analysis of acceptability

TABLE 8.7: Quotes that directly relate to the evaluation of acceptability but were not used in the description of themes in the previous sections

| label | Quote |
|---|---|
| T3A | just give the user more control over what's actually happening on screen. |
| T3B | So it is like do you wanna see a male or female. (1 second) Person? So then you click and you get the, either the female or the (2 seconds) male one, yeah, all stuff like that, and if you wanna see some (1 second) you, um related to your ethnicity or something like that, so that you kinda feel; er, er what do you call it now? Comfortable |
| A1 | But um like, as I say, I love the concept but I just don't (1 second) think it would, unless you've got like dead specific questions for people with different symptoms. So you could like separate them and categorise them, I don't think it would work. |
| A2 | I, I think if my scepticism, shall I call it, er is wrong then it may be helpful because I know these days most people, either rightly or wrongly, do look into, in er YouTube and ask about, shall we say, medication or illnesses, er what's the treatment, what are they for, things like that. So providing the answers are accurate, or the choices given (1 second) are accurate (1 second) no, I don't think I've any other problem, it is this just (1 second) can we feed in the information to get the accurate output, as it were. |
| A3 | I just rushed through it because I felt like I had, I almost felt like there was a deadline, as if there was gonna be a line going across and I had to give the answer within a |

Continued on next page

**Table 8.7 – continued from previous page**

| Label | Quote |
|---|---|
| | timeframe. |
| A4 | I did sort of, the first (1 second) er screen I sort of sat there and looked at it before I realised I had to press play; I don't know why, I just thought it would automatically start playing at me |
| A5 | I was kind of expecting to see a doctor in a doctors' surgery looking, looking like a doctor (1 second) but not (1 second) a real person, but looking very much like the one |
| A6 | I've got very used to using Zoom it, it did my head in that there wasn't any um sort of like (1 second) er visual feedback |

#### 8.4.3.1 Affective attitude

Affective attitude refers to how someone feels about taking part in an intervention. The findings from the affective component of the Technology Acceptance Model suggest that users are largely undecided about their attitude towards the application. Although 67.9% of users agreed that the application had positive characteristics, most of the respondents indicated that they somewhat agree with this (46.91%), in accordance with the average rating of somewhat agree. Furthermore, 22.2% reported that they were undecided about the application. This suggests there is room for improvement for the application.

Individuals may be undecided about their attitude towards the application because there is a conflict between their attitude towards the concept and the design of the actual application. One participant reported that they like the concept of the application but suggested that the application may not work in the current design [A1]. Given that the Technology Acceptance Model has shown that perceived usefulness and ease of use both influence affective attitude, it's unsurprising that people may be uncertain about how they feel about the application given that they are uncertain about whether the information that they have provided is sufficient to receive the correct diagnosis, which would be exacerbated if people encountered issues with the application that influence the ease of use. One user displayed this understanding because they explained that they have a scepticism about whether the information that the application collected could be used to predict the right diagnosis, but if this scepticism was demonstrated to be unfounded, they would not have a problem with the application because it aligns with people's current behaviour of accessing medical information through the internet, although they expressed uncertainty about whether this behaviour is the right or wrong approach to medical care [A2]. Therefore, improving the design of the application or improving trust in the design of the application by demonstrating that the design of the application is sufficient to accurately predict the diagnosis may be required to improve the users' attitudes.

#### 8.4.3.2 Burden

Burden describes the perceived amount of effort required to participate in an intervention, for example time, expense, and cognitive effort. The subscale "usefulness" should provide insight into the perceived burden of the application because the application is designed to guide referral pathways and reduce burden, so a useful application should reduce burden. 72.8% of users reported that the application was useful and 23.5% were unsure. An exploration of the responses to independent questions on the subscale can provide insight into where effort is required. Most people agreed that the application would save them time (75.3%), but fewer people thought that the application would save them effort (64.2%). Many people thought that the application was easier than booking a medical appointment (71.6%), but fewer reported that it was more convenient (61.7%). People may perceive the application as requiring more effort because the interaction with the VA requires them to determine what information is relevant and important for making the diagnosis, whereas the doctor would typically have this responsibility during medical interactions. Therefore, patients have more responsibility while completing the application, which could be construed as increased cognitive effort required for the application compared to routine clinical interactions.

The subscale "ease of use" also provides insight into the burden of the application. Only 49.4% of people reported that the application was easy to use across all questions on the subscale and 34.6% were unsure. The score for ease of use appears to be influenced by particular elements of the application, for example only 7.4% of users reported that the application was time consuming, 8.6% reported that it was complicated to use, and only 11.1% agreed that it required a lot of effort, but 22.2% agreed that the application was confusing and 17.3% disagreed that the application was easy to talk to. These findings suggest that the ease of use is largely influenced by how much people understand the intervention and how easily people can provide spoken descriptions of what happened to the VA. These findings were corroborated by the thematic analysis because many people reported difficulty providing all the information that they felt was relevant because the lack of interactivity by the VA reduced the amount that people elaborated their descriptions, and the topic agenda set by the questions reduced the scope of what people thought they could talk about and made it difficult to describe different types of information that they thought were relevant. These findings suggest that increasing the scope of topics that people can talk about and including cues to guide the topic could reduce the burden for users.

#### 8.4.3.3 Opportunity Costs

Opportunity costs refers to the extent to which benefits, profits, or values must be given up to engage in the intervention. The most relevant cost that was discussed during the feedback interviews was the potential loss of future interactions with medical professionals. Doctors were regarded for their knowledge, empathy, and understanding, and many users expressed a concern that future interactions with doctors would be reduced or the quality of these interactions hindered because of the reliance on an online application. Although the application is not designed to replace medical professionals, a future provider would have to be cautious that such an application did not influence interactions with medical professionals and ensure that users understood this to reduce the impact on the perceived acceptability of the application.

#### 8.4.3.4 Ethicality

Ethicality describes the extent to which an intervention aligns with an individual's moral system. Unfortunately, ethicality is not covered by the Technology Acceptance

Model questionnaire and was not a prominent part of the thematic analysis. Therefore, this area of acceptability would need to be assessed further in future research.

#### 8.4.3.5 Self-Efficacy

Self-Efficacy refers to people's confidence that they can perform the behaviours required for the intervention. Although some participants encountered technological barriers to completing the application and some design elements of the application sometimes reduced how easy the application was to use, these factors are not related to the individuals confidence in their own abilities to complete the activity. The most prominent area of feedback that related to self-efficacy was whether people felt they had the knowledge required to determine the appropriate information required by the questions in order to make the correct diagnosis. Many people reported uncertainty about whether they could do this. Therefore, future implementations of the application should provide guidance about what information is required to increase people's confidence that they can complete the task.

#### 8.4.3.6 Perceived Effectiveness

The extent to which the intervention is perceived as likely to achieve its purpose. An effective online application would successfully predict the underlying cause of TLOC and refer them to the appropriate service. Most people thought that the application might help them be referred to the right service (66.7%), but a large proportion were uncertain (24.7%). Fewer people thought the application would assist them to receive the right diagnosis (50.6%) and more people were uncertain about this (34.6%). These findings suggest that many people are uncertain about the potential effectiveness of the application, which is unsurprising given that the application is currently only being tested. The most probable reason that people are uncertain about the effectiveness of the application is that they do not believe the information collected by the application is sufficient to predict the diagnosis, both because the closed questionnaire and spoken descriptions does not take into consideration important contextual information, particularly when compared to the breadth of information that is considered during routine medical consultations.

#### 8.4.3.7 Intervention Coherence

Intervention coherence refers to how well people understand the intervention and how it works. The intervention that we are researching is not currently being used and we do not know how it would be used in clinical practice. However, we can evaluate intervention coherence based upon our theoretical definition of how the application would be used and the misconceptions that were present during the feedback interviews. The feedback interviews revealed that people often develop a preconceived idea of what the application was going to consist of and how it would be used in clinical practice. These preconceptions can influence how they engage with the application. Examples of the misconceptions include perceiving that the recording elements of the application have a time limit [A3], expecting the video to play automatically [A4], expecting the avatar to resemble a doctor in a doctor's surgery [A5], and expecting to see a video of themselves on the screen too to resemble video conferencing [A6]. People reported being disappointed when the application did not align with their prior expectations, which demonstrates the importance of managing expectations on perceived acceptability. We found that people wanted to make decisions about the presentation of the application themselves, which may allow people to self-manage their expectations about the application and increase the acceptability.

Users frequently displayed a misunderstanding about how the application would be used in clinical practice. The theoretical idea of the application is that it would be a standalone application that patients (and witnesses) would be instructed to complete by a medical professional in Primary Care. The application would provide a predicted diagnosis and, the medical professional would use their own expertise and the prediction from the application to guide the referral. In contrast, the users displayed a different understanding of how the application might be used in practice under the theme "progression through the medical pathway". People thought that the information collected through the application would be shared with multiple medical professionals, stored in a central system that all medical professionals can access, and that medical professionals involved in their future care will look at the spoken descriptions collected by the application before future appointments, which would influence the information they need to share during the appointment. These observations are important for understanding the acceptability of the application because it demonstrates that users will formulate an understanding of an intervention that may be different from what the intervention was designed for or how it is used in clinical practice, which could reduce the perceived acceptability of an intervention if a discrepancy becomes apparent.

## 8.5 Discussion

The objective of this analysis was to explore the acceptability of the online web application from the perspective of the users. The analysis used a mixed-methods approach combining the responses from a questionnaire inspired by the Technology Acceptance Model (Lanseng and Andreassen, 2007) and a thematic analysis of the feedback that users provided during telephone interviews. Users reported that they somewhat agree that the application is positive and useful, but there was uncertainty regarding whether the application was easy to use. The thematic analysis resulted in the identification of three prominent themes surrounding the application. Firstly, the application was described as a method of collecting information. The design of the application influenced the users ability to provide the necessary information that was perceived as required for making an accurate diagnosis, which consequently influenced the users trust about the usefulness of the application. Secondly, medical care was frequently represented as a journey for which progression and satisfaction were intertwined. The appraisal of the application can be dependent upon whether it is perceived to improve or hinder progress, or whether it interferes with other preferred elements of the medical journey, such as interactions with medical professionals. Finally, users frequently expressed a desire for more choice about the design elements of the application, and control over the application was related to comfort and improvements in the ease of use. These findings demonstrate the multi-faceted nature of acceptability (Sekhon, Cartwright, and Francis, 2017) and the challenge associated with trying to determine whether an intervention is considered acceptable because it is difficult to integrate the findings from each of the independent constructs identified in the analysis.

We attempted to understand how our findings relate to the concept of acceptability (Sekhon, Cartwright, and Francis, 2017) by contrasting our findings with each of the component constructs in the theoretical framework. The analysis revealed that users consistently regarded the application as "somewhat" or "potentially" acceptable but the acceptability was dependent on the effectiveness of the application. Regardless of the actual effectiveness, the most prominent influence on perceived effectiveness was people's perception of what information was required to make an accurate diagnosis and whether they were able to provide all the necessary information using the application. People frequently reported barriers to providing the information, for example being restricted to binary responses during the questionnaire, only being asked about a single attack, and uncertainty about what information is required. We noted that this area of feedback was related to many of the component constructs of acceptability, for example affective attitude, burden, self-efficacy, and perceived effectiveness. This relationship between the ability to provide information, perceived effectiveness, and acceptability is in line with the findings from the Technology Acceptance Model (Davis, 1989; Davis, Bagozzi, and Warshaw, 1989). Providing information is a primary objective of the application and therefore relates to the "ease of use" construct. This construct subsequently influences all elements of the model, for example perceived usefulness, attitude, intention to use, and behaviour (Figure 8.2). Therefore, it makes sense that this would have a large impact on acceptability, and the technology acceptance model would suggest making changes to this construct would subsequently improve the users perception on many of the other constructs.

The acceptability of the application is also dependent on how the application is integrated into the pre-existing medical pathway. The perceived acceptability of the application is likely to decrease if the application reduces the amount of time or quality of interactions with medical professionals. This area of feedback appeared to depend on whether users trusted that the provider would use the application for its designed purposes or whether it would be used for other purposes, for example delaying referrals due to long waiting lists. Given that individuals with FDS often experience long diagnostic delays, a lack of understanding about their condition, prejudice, and limited access to treatment (Rawlings et al., 2017c), trust in the service provider may be more important. Furthermore, previous research has demonstrated that trust influences the perceived ease of use and usefulness of theoretical diagnostic technology (Lanseng and Andreassen, 2007), so future research should consider evaluating the trust of a provider of this technology before it is implemented. Furthermore, users displayed some misunderstandings about how the application may be used in practice, which suggests that the intervention coherence would need to be monitored to ensure that misunderstandings about the application do not reduce acceptability.

### 8.5.1 Improvements to the application that may improve acceptability

The application could be improved by broadening the scope of the questions that are asked by the VA. Currently, the application focuses on the most recent attack. We originally chose to focus on a single attack because we hypothesised that this would result in the greatest difference in the responses from people with epilepsy and FDS. Many users felt that these questions were too restrictive because it did not allow them to talk about the different types of attacks that they experience or provide other contextual information, for example other potentially related symptoms, the influence of their medication, and psychosocial factors that may be related to the attacks. Users could specify the number of different types of attacks that they have or the number of attacks that they wish to describe and be asked to produce a description of each attack independently. This approach would allow users to record and report multiple attacks across the time period between referral to a specialist service and the initial appointment. Furthermore, users could be given the option to answer nonessential questions that cover many of the topics addressed in routine medical interactions, for example medication, other medical symptoms, and psychosocial factors that may influence their health (Cassell, 1985), but are not required for all users to complete because it may exceed the amount of time that people are prepared to give to the application. This approach would satisfy some users' desire to provide more information and make more information available for the automated analysis of language.

Broadening the scope of the questions to represent the different topics discussed in routine medical encounters may encourage users to speak more. During the preliminary analysis that we conducted in chapter three using the routine medical interactions, people with FDS spoke more than people with epilepsy. This is the opposite of what would be expected based on the previous CA research (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008) and is contradictory to the response patterns we observed during interactions with the VA in chapter five. Therefore, broadening the questions may increase how much people say, and it could lead to the identification of additional group differences that could improve the predictive performance of the automated analysis of language based upon the semantic measures.

The application could be updated to provide users with more choice and greater control. One change could be to provide users with a choice about how the questions are presented, for example choosing the type of avatar or opting to have the questions in a written format. This may allow some users to feel more comfortable and would alert users about what to expect as they advance through the application.

The application could change how the questions are asked too. The application could become nonlinear so that participants do not have to complete all questions in a single sitting. Users could be presented with a dashboard that contains separate sections for all of the questions. Moreover, the follow-up questions could be removed from the application to prevent repetition and reduce the time taken to answer the questions. However, once users have produced a single description of the attack, they could be given a list of prompts and asked to provide any further information that is then subsequently recorded separately. This would allow the application to differentiate between the freely shared information and prompted information without asking participants to provide the same information multiple times. This method could allow people to add more information to their response at a later date.

Finally, users could be given the option to start the recording once the question has been asked, rather than the recording starting automatically. We originally designed the application like this because individuals are not given time to plan their answers during human-human interactions and this may influence the presentation of formulation effort.

### 8.5.2 Future Research

Future research could expand on these findings by exploring the acceptability of the application from the perspective of clinicians. Clinicians could provide greater insight into how an intervention like this could be integrated into the pre-existing care pathway and how clinicians would use the application. Furthermore, the analysis could provide greater insight into what information clinicians would like the application to collect, which could guide future improvements of the application, for example the incorporation of additional questions. Given that users frequently speculated about how clinicians could use the application, the feedback from clinicians would provide greater clarity about how it would be used in practice, which could be used to improve the intervention coherence.

### 8.5.3 Limitations

It is important to consider whether the information reported in this analysis has been influenced by socially desirable answers because the analysis is mostly reliant on self-report measures. The socially desirability bias refers to the tendency for people to provide responses that are perceived as appropriate or socially acceptable to others (Grimm, 2010). All areas of this research project were developed and conducted by a single researcher, for example the online web application, signing participants up to the study, and the feedback interviews. It was possible for participants to recognise that the application was designed by the researcher who was conducting the interviews because the participant information sheet informed the participants that the research project was part of a PhD program and the main researcher's voice was used to record the questions that the avatar asked. This may have increased the likelihood that participants produced socially desirable responses because they did not want to critique someone's work. One participant explicitly said this during one of the feedback interviews. One way that people may have overcome the bind between wanting to critique the application without being rude to the creator is to provide constructive criticism towards the design of the application, which was frequently observed throughout the feedback interviews. This emphasises that the absence of direct negative feedback is not evidence that individuals do not have a negative attitude towards the application and that constructive criticism should provide insight into the acceptability of the application to the users - a lot of constructive criticism is indicative of lower acceptability.

The study is also subject to a recruitment bias because we only sought feedback from participants who signed up to complete the online web application. There may be many people who have a negative attitude towards health technology and perceive the general approach as unacceptable. These individuals may have chosen not to participate in the research project. Moreover, people who perceive themselves as lacking the necessary technological skills to complete the online application may not sign up to the study. It is important to collect feedback from the wider population to thoroughly understand the acceptability of the approach. Therefore, future research should explore the acceptability independent of using the application.

Another important consideration is that people may be providing feedback about different things while completing the questionnaire. For example, people could be evaluating the iPEP questionnaire or the VA, or they may be evaluating the application as a theoretical construct rather than the current design of the application because they recognise that it is currently only a prototype design. Participants often made this distinction during the feedback interviews because they explained that they were positive about the concept but had uncertainties about the particular design. This makes it difficult to decipher what the questionnaire responses are referring to, and future research should consider collecting feedback on these different elements independently to better understand what a broader range of people think about each component of the application and concept.

Ethicality is an important sub-component of acceptability. Unfortunately, ethicality as a construct is not part of TAM and was not a prominent area of discussion throughout the feedback interviews. People's perception of the ethnical ramification of the application are likely to extend beyond this particular research project and encapsulate attitudes towards the use of machine learning in medical care in general (Latif et al., 2020). Considering the ethical ramifications of artificial intelligence research is incredibly important (Vollmer et al., 2020; Zhang et al., 2021) and researchers in the field have recommended that research regarding the ethical implementation of artificial intelligence should be domain-specific (Starke et al., 2022). Therefore, future research should explore attitudes towards the ethicality of this area of research using a multidisciplinary approach.

### 8.5.4 Conclusion

This analysis aimed to explore the acceptability of the online web application from the perspective of the patients and witnesses who have used it. Based upon the responses to the feedback questionnaire and a thematic analysis of feedback interviews, it appeared that the application is currently considered somewhat acceptable, but many users expressed dissatisfaction towards different components of the application through the constructive criticism that they provided. The application should be updated in future to improve these design elements and hopefully the overall acceptability of the application. The suggested improvements included broadening the scope of information that people can provide to the application so that it includes many of the elements frequently discussed in doctor-patient interactions and allowing people to have more control over the design of the application and the interaction with the virtual agent. Future research should expand on the findings from this analysis by seeking the views of medical professionals and a more diverse range of patients without the requirement that patients complete the online web application, for example seeking feedback on the concept rather than the specific application and the ethical ramifications of this type of technology. These studies would help to further our understanding of the acceptability of this concept and guide the future design of the application.

# Chapter 9

# Summary and scope for future research

This thesis has investigated the feasibility of predicting the cause of TLOC using an online web application. Approximately 20% of individuals who experience TLOC are initially misdiagnosed in Primary and Emergency Care Services (Xu et al., 2016). Although there are a range of potential clinical decision tools that could assist in the differential diagnosis, many of these tools are not designed for the three-way differentiation between epilepsy, FDS, and syncope or they under-perform in the differentiation of two of the most common causes of TLOC: epilepsy and FDS (Wardrope, Newberry, and Reuber, 2018). Therefore, this thesis investigated alternative methods to improve the differential diagnosis of a pre-existing, high performing clinical decision tool (Wardrope et al., 2020a).

Analysing the spoken descriptions of TLOC is one of the most important methods in the diagnostic process (Plug and Reuber, 2009). Not only do spoken descriptions include information about the symptoms associated with the experience of TLOC (Malmgren, Reuber, and Appleton, 2012), there are also a broad range of communicative behaviours associated with the descriptions of epilepsy or FDS that can assist in the differential diagnosis (Reuber et al., 2009). The use of speech technology for the identification of various health conditions is vastly developing field (Latif et al., 2020). However, to date, there has been no research utilising these methodologies for the identification of the cause of TLOC. Therefore, the major contribution of this thesis is the exploration and evaluation of this area of research. This chapter will provide an overview of the key research findings and outline what future research can do to improve and extend this research.

## 9.1 Research Questions

### 9.1.1 Research question 1

Machine learning research using a small dataset is reliant on cross validation to estimate the capacity of a model to make predictions for unseen data (Berrar, 2019). Research paradigms that perform feature selection and cross validation on the same dataset often produce overly optimistic performance estimates (Vabalas et al., 2019). This research project was reliant on cross validation because the number of people recruited to this research project was hindered by the Coronavirus pandemic. Given that there is no previous research automating the analysis of TLOC descriptions, the first objective was to identify useful acoustic and semantic features for differentiating between people with epilepsy and FDS using seizure descriptions collected from previous CA research.

Chapter three investigated the performance of two feature sets: features designed to measure formulation effort and features measuring the proportion of words corresponding to 21 semantic categories. The two feature sets were effective at differentiating between people with epilepsy and people with FDS with an accuracy of 71-81%. Although this was not an exhaustive exploration of potential features, the performance suggested that these features could contribute to machine learning models trained using other effective features. Furthermore, we identified that a SVM model with an RFB kernel outperformed Random Forest - the model used to evaluate the iPEP in previous studies, which guided our choice of machine learning model in the research conducted in subsequent chapters.

### 9.1.2 Research question 2

A questionnaire designed to predict the cause of TLOC using patient endorsed symptoms and witness observations called the iPEP has been previously tested on individuals with gold-standard diagnoses (Wardrope et al., 2020a). The questionnaire was previously administered as a 5-point likert scale, but the responses were dichtomised before the final model was trained. It was not clear whether these results would generalise to novel research samples when the questionnaire was administered in binary format or to individuals first presenting with TLOC.

The research outlined in chapter six demonstrates that the iPEP was not as effective at predicting the cause of TLOC for the respondents who completed the online web application. A Random Forest algorithm was trained using the dichotomised responses from the original research project (Wardrope et al., 2020a) and subsequently tested on two additional data-sets: a sub-sample of the dichotomised responses collected from the original research study and the responses that were collected through the online web application. The model displayed a similar level of performance to that observed in the original research project when evaluated on the sub-sample that was collected in the original study, but the overall accuracy reduced by 15.6% for the patient-only responses and 36.9% for the patient and witness responses when the model was evaluated using the data collected through the online web application. Furthermore, there was little improvement in the overall accuracy of the iPEP when a model was trained and evaluated using the responses collected through the online web application using leave-one-out cross validation and a SVM, for which there was an accuracy of 65.8%. These findings suggest that the iPEP model from previous research does not demonstrate a similar level of performance when administered as a binary questionnaire to a novel research sample, but the performance of the leave-one-out cross validation model trained using the sample collected through the online web application may improve with increases in the sample size. These improvements may be particularly evident for individuals with syncope who only made up 21% of the total sample.

### 9.1.3 Research question 3

Although there is extensive research demonstrating conversational, interactional, and linguistic differences between how people with epilepsy or FDS communicate their seizure experience to a doctor (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008; Plug, Sharrack, and Reuber, 2009b; Robson et al., 2012; Robson, Drew, and Reuber, 2016), there are no guarantees that people will display this communicative behaviour while conversing with a VA. Given that the automated analysis of language will presume the presence of these linguistic profiles, we explored whether these communicative behaviours are still evident during interactions with the VA using a sub-sample from the overall study. Confirming the presence of these behaviours supports the validity and feasibility of an automated analysis of language.

The analysis outlined in chapter five used conversational analytic techniques to contrast the responses between people with epilepsy or FDS during interactions with the VA. We found that people with epilepsy provided descriptions that predominantly focused on their personal and subjective experience of TLOC, which resulted in more detailed descriptions overall. These descriptions contained extensive formulation effort, for example hesitations, repetitions, meta-discursive comments about the challenges associated with the description, and hedging statements. Furthermore, although the questions restricted the scope of responses by only focusing on the most recent experience of TLOC, individuals with epilepsy were more likely to display a willingness to provide information by describing additional TLOC experiences that allowed them to outline symptoms that were relevant for the question but were not evident in the most recent attack. In contrast, people with FDS provided very limited descriptions of their most recent TLOC experience. They were more likely to make complete negations that conflate the unconscious period and overall seizure, for example "I cannot remember anything". Moreover, they exhibited an increased reliance on third parties to answer the questions posed by the VA. These findings demonstrate differences in how people with epilepsy or FDS describe their TLOC experience to a VA. Many of these communicative behaviours are similar to those observed during doctor-patient interactions (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008; Robson, Drew, and Reuber, 2016). Therefore, we concluded that an automated analysis based upon the linguistic profiles outlined in previous research may be able to differentiate between these two patient groups using interactions with a VA.

### 9.1.4 Research question 4

The analysis in chapter three demonstrated that the formulation effort and LIWC features were effective at differentiating between people with epilepsy or FDS using doctor-patient interactions. The analysis in chapter six explored whether these features were equally effective when applied to VA-patient interactions. Both feature sets exhibited a high classification accuracy of 86% for the binary classification. The consistency between the two analyses suggests that these features are useful for differentiating between epilepsy and FDS.

One limitation of these two feature sets is that they do not capture any differences in people's descriptions of subjective symptoms or the events that preceded and followed the unconscious period. We explored the predictive performance of additional semantic features based on the verbs, adjectives, and adverbs used by the patient. These features were also effective at differentiating between people with epilepsy or FDS with an accuracy of 75.5%.

Up until chapter six, the automated analysis of language had not included individuals with syncope. We explored how the performance of these feature sets changed when individuals with syncope were included. Unfortunately, there was a large decrease in the accuracy of each feature set when people with syncope were included. These findings suggest that the selected features are not effective predictors of syncope.

The final analysis outlined in chapter seven explored whether the automated analysis of spoken TLOC descriptions can improve the baseline performance of the iPEP. Training a SVM using the patient only iPEP responses, the three feature sets from the automated analysis of language, and all three diagnoses reduced the overall accuracy of the iPEP from 65.8% to 59%. However, the use of a model stacking approach that allowed us to restrict the automated analysis of language to individuals for whom the iPEP predicted a diagnosis of either epilepsy or FDS increased the overall performance of the iPEP to 85.5%. The results suggest that it is possible to improve the predictive performance of the iPEP using an automated analysis of TLOC, but the automated analysis of TLOC may be better suited for the challenging differentiation between individuals with epilepsy or FDS.

### 9.1.5 Research question five

Chapter eight collated and analysed the application feedback from the patients and witnesses who used the application based on the construct of acceptability (Sekhon, Cartwright, and Francis, 2017). The application was often appraised as "somewhat acceptable" across a range of areas of feedback. One particular noteworthy piece of feedback was that individuals preferred speaking than completing the binary questionnaire because they were able to provide important contextual information. The feedback seemed to suggest that people positively appraised the concept because it could potentially improve the current care pathway and speed up the time that it takes to receive the correct diagnosis, but there were some challenges associated with the design of the current application that reduced the overall acceptability. Most areas of feedback centred around the usability and capacity of the application to collect a sufficient amount of high quality and diagnostically relevant information in order to make an accurate clinical judgement. This was often contrasted with people's previous clinical encounters, which were considered extensive. Providing all the relevant information was considered vital for receiving the right diagnosis, so aspects of the application that reduced the information that people could provide also reduced their perception of trust. Users provided extensive feedback about how the application design could be improved, which should be utilised to improve the application for future research.

## 9.2 Future Research

The limitations of the research conducted in this thesis have been discussed throughout. Rather than repeating these limitations, this section will provide an overview of the potential avenues of future research that have been discussed because the suggested areas of future research incorporates and extends the limitations from this thesis.

### 9.2.1 More data

The coronavirus pandemic hindered the data collection for this thesis. The limited sample size has likely caused an under-estimation of performance for the iPEP because an insufficient amount of data makes it difficult for the machine learning model to detect relationships between the input features and diagnoses. This may explain why the model was less effective at detecting syncope compared to previous research (Wardrope, Newberry, and Reuber, 2018). The effectiveness of the model stacking algorithm that we outlined in chapter seven is dependent on the capacity to detect individuals with syncope. Therefore, future research should collect more data to investigate whether increasing the sample size, particularly participants with a diagnosis of syncope, can increase the baseline performance of the iPEP.

### 9.2.2 Feature engineering

The review in chapter two demonstrated the breadth of features that can be used in an automated analysis of language or speech. Given the limited amount of data available during this research, it has been difficult to conduct a vigorous exploration of potential features that can improve the predictive performance of the automated analysis of language because feature selection should be conducted using a separate train, validation, and test dataset (Vabalas et al., 2019). Future research should explore additional features that can be incorporated into the analysis, particular features that may better measure the linguistic differences identified in previous CA research. Furthermore, self-supervised machine learning methods that can convert textual data into a feature vector, for example BERT (Devlin et al., 2018), may automatically generate useful machine learning features, especially considering that semantic differences were effective predictors of the diagnosis. However, more data will be required before these methods can be implemented.

### 9.2.3 Linguistic profile for syncope

Previous CA research has focused on the differential diagnosis of epilepsy or FDS (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008), but there are no CA studies exploring how individuals with syncope describe their experience of TLOC and how this potential linguistic profile contrasts with the profiles from the other two diagnostic groups. Future CA research could help to increase our understanding about how people with syncope describe their experience of TLOC. An outcome of this analysis may help to design features that can reliably separate individuals with syncope from individuals with epilepsy or FDS. These features could be incorporated into the model stacking algorithm to provide a "second stage" analysis that precedes the automated analysis of language for predictions of epilepsy or FDS to improve the identification of syncope and potentially increase the number of people with epilepsy or FDS that are incorporated in the appropriate language analysis. The stacking analysis would be more effective if the sensitivity for syncope was higher.

### 9.2.4 Incorporating witness contributions

One of the benefits of the model stacking approach is that the analysis can be segmented into stages and the earlier stages can be used to make a prediction (albeit perhaps less accurately) when the information is not available for the later stage analysis (e.g. the individuals without a VA interaction). Previous research has demonstrated that witness contributions can dramatically improve the predictive performance of the iPEP, especially for the identification of syncope (Wardrope et al., 2020a). Therefore, future research should recruit more witnesses and explore methods of incorporating the witness responses into the model stacking algorithm. Aside from research exploring the interactional contributions of accompanying others during routine clinical encounters (Robinson, 2003) or description of seizure manifestations (Malmgren, Reuber, and Appleton, 2012), there is no researching exploring how the communicative behaviour of witnesses can assist the differential diagnosis, which would help guide an automated analysis of witness interactions with the VA. Therefore, future research should identify and collate the interactional and communication profiles of witnesses for the differential diagnosis of TLOC and explore how these behaviours can be automatically analysed using machine learning methodology. This analysis could explore the contributions of accompanying others during the patient's interaction with the VA and the witnesses independent interaction with the VA.

### 9.2.5 Broadening the topical agenda of the questions asked by the VA

The thematic analysis outlined in chapter eight identified that people thought providing information was vital for receiving the correct diagnosis. However, people reported that some design elements of the application reduced their capacity to provide all the information they thought was relevant. One component was that the questions solely focused on the most recent attack, whereas people wanted the opportunity to provide information about other attacks and other contextual information about their life that may be relevant for the diagnosis. The history taking procedure during doctor-patient interactions often involves discussions beyond the immediate medical concern (Cassell, 1985). Including questions about the patients health, lifestyle, and other medical concerns may encourage patients to talk more. We observed in chapter three that people with FDS typically said more during the routine history taking procedure. However, previous research has shown that they say less when describing a seizure (Schwabe, Howell, and Reuber, 2007; Schwabe et al., 2008), a finding that was corroborated in our CA analysis of interactions with the VA. Including additional questions may prompt further talk from this patient group, and these contributions may be diagnostically relevant because people with FDS have increased general psychopathology which may relate to their manifestation of FDS (Brown and Reuber, 2016a).

It would be important to consider how these questions are integrated into the automated analysis to ensure that the contribution to these additional questions does not interfere with the features that are designed specifically for talk about the attack; for example, some of the additional questions could be analysed independently and incorporated into the model stacking algorithm as a separate machine learning model. Therefore, broadening the scope of the questions may provide additional information that can improve the predictions made by the automated analysis of language and improve the perceived acceptability of the application because people feel that they have been able to provide all the information that they consider relevant.

### 9.2.6 Broadening the analysis of acceptability

The analysis of acceptability conducted in chapter eight was restricted to people who participated in the research project. Therefore, the findings may be affected by a sampling bias because the attitudes of individuals who chose not to participate because they negatively appraised this type of technology or who felt that they did not have the necessary technical skills to participate were not measured. Given that the introduction of a tool like this would be a dramatic change to the clinical care pathway, future research should explore the acceptability of this approach using a more diverse population. This research should also explore the attitudes of clinicians involved in the care pathway for TLOC and the potential ethical considerations of this approach, which is a fundamental component of the acceptability construct but was not considered in sufficient detail during this research.

### 9.2.7 Automatic Speech Recognition

Successful ASR is vital for an application that conducts an automated analysis of language. Unfortunately, the research conducted in this PhD did not focus on creating a fine-tuned ASR system for the particular task because there were multiple additional commitments within the PhD that reduced the time available for this task, for example developing the online web application, recruiting participants, and conducting the qualitative analyses. The review in chapter two demonstrated that there are many methods that can be used to improve an ASR system once there is more research data available for training and testing. Therefore, future research should explore methods of fine-tuning the language and acoustic models of the ASR system.

### 9.2.8 Confounding variables

Clinical decision tools and machine learning algorithms can be influenced by confounding variables (AlHasan, 2021). The analysis in chapter seven found a weak correlation between educational level and how much people talked during the interaction with the VA. However, there are many more confounding variables that can influence the performance of the iPEP and the automated analysis of language (Mukherjee et al., 2022; Keuleers et al., 2015). Identifying these confounds is vital for understanding how the clinical decision tool works and ensuring that it does not inadvertently discriminate towards particular minority groups if it is used in clinical practice. For example, the sample used in this research is not ethically diverse, so it is not possible to conclude that these findings would generalise to the wider populations. Therefore, future research should aim to identify and mitigate potential confounds.

# Appendix A

# Model Configurations

TABLE A.1: The hyperparameters that were used for the "RandomisedSearchCV" in conjunction with the Random Forest Classifier in Chapter 3 section 3.2.

| Hyperparameter | Value Range |
|---|---|
| n estimators | 100,200,300,400,500 |
| max features | auto, sqrt |
| min samples split | 2,5,10 |
| criterion | Gini, Entropy |
| bootstrap | True, False |
| min samples leaf | 1,2,4 |

TABLE A.2: The hyperparameters that were used for the "GridSearchCV" in conjunction with the Logistic Regression Classifier in Chapter 3 section 3.3.

| Hyperparameter | Value Range |
|---|---|
| Penalty | l1, l2 |
| C | 0.001,.009,0.01,.09,1,5,10,25 |

TABLE A.3: The hyperparameters that were used for the "GridSearchCV" in conjunction with the K-Nearest Neighbours Classifier in Chapter 3 section 3.3.

| Hyperparameter | Value Range |
|---|---|
| n neighbors | 3,5,11,19 |
| weights | uniform,distance |
| metric | euclidean,manhattan |

TABLE A.4: The hyperparameters that were used for the "GridSearchCV" in conjunction with the K-Nearest Neighbours Classifier in Chapter 3 section 3.3. Two separate models were implemented, one using a linear kernal and one using the RBF kerbel

| **Hyperparameter** | **Value Range** |
|---|---|
| C | 0.1,1, 10, 100 |
| gamma | 1,0.1,0.01,0.001 |

TABLE A.5: The hyperparameters that were used for the "GridSearchCV" in conjunction with the Random Forest Classifier in Chapter 3 section 3.3.

| **Hyperparameter** | **Value Range** |
|---|---|
| Bootstrap | True, False |
| max depth | 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, None |
| max features | auto, sqrt |
| min samples leaf | 1, 2, 4 |
| min samples split | 2, 5, 10 |
| n estimators | 200, 400, 600, 800, 1000,<br>1200, 1400, 1600, 1800, 2000 |

TABLE A.6: The hyperparameters that were used for the "GridSearchCV" in conjunction with the Support Vector Machine with an RBF kernel in Chapters 6 and 7.

| **Hyperparameter** | **Value Range** |
|---|---|
| C | 1, 10, 100, 1000 |
| gamma | 0.001, 0.0001 |

# Appendix B

# LIWC Categories

TABLE B.1: A full list of the LIWC categories. The categories incorporated in the analyses in chapters, 3, 6 and 7 are marked with an X

| LIWC Category | Chapter 3 | Chapters 6 and 7 |
|---|---|---|
| Function | | |
| Pronoun | | |
| Ppron | | |
| I | | |
| We | X | X |
| You | | |
| She/He | X | X |
| They | | |
| Ipron | | |
| Article | | |
| Prep | | |
| Auxverb | | |
| Adverb | | |
| Conj | | |
| Negate | | |
| Verb | | |
| Adjective | | |
| Compare | | |
| Interogative | | |
| Number | | |
| Quantifiers | X | X |
| Emotional Tone | X | X |
| Affect | X | X |
| Positive Emotions | X | |
| Negative Emotions | X | |
| Anxiety | X | |
| Anger | X | |
| Sadness | X | |
| Social | X | X |
| Family | X | |
| Friend | | |
| Female | | |

Continued on next page

**Table B.1 – continued from previous page**

| LIWC Category | Chapter 3 | Chapters 6 and 7 |
|---|---|---|
| Male | | |
| Cognitive Processes | | |
| Insight | | |
| Cause | X | |
| Discrepancy | | |
| Tentativeness | X | X |
| Certain | X | |
| Differ | | |
| Percept | | |
| See | | |
| Hear | | |
| Feel | | |
| Biological | | |
| Body | | |
| Health | | |
| Sexual | | |
| Ingest | | |
| Drives | | |
| Affilitation | X | |
| Achieve | | |
| Power | X | |
| Reward | X | X |
| Risk | X | |
| Focus Past | | |
| Focus Present | X | X |
| Focus Future | | |
| Relative | | |
| Motion | | |
| Space | X | |
| Time | | |
| Work | | |
| Leisure | | |
| Home | | |
| Money | | |
| Religion | | |
| Death | | |
| Informal | | |
| Swear | | |
| Netspeak | | |
| Assent | | |
| NonFlu | | X |
| Filler | | |

# Bibliography

Adkisson, Wayne O and David G Benditt (2017). "Pathophysiology of reflex syncope: A review." In: *Journal of cardiovascular electrophysiology* 28.9, pp. 1088–1097.

Ajzen, Icek and Martin Fishbein (1969). "The prediction of behavioral intentions in a choice situation". In: *Journal of experimental social psychology* 5.4, pp. 400–416.

AlHasan, AJMS (2021). "Bias in medical artificial intelligence". In: *The Bulletin of the Royal College of Surgeons of England* 103.6, pp. 302–305.

Alpers, Georg W et al. (2005). "Evaluation of computerized text analysis in an Internet breast cancer support group". In: *Computers in Human Behavior* 21.2, pp. 361–376.

Alsmadi, Issa and Keng Hoon Gan (2019). "Review of short-text classification". In: *International Journal of Web Information Systems*.

Altman, Naomi S (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". In: *The American Statistician* 46.3, pp. 175–185.

Alwosheel, Ahmad, Sander van Cranenburgh, and Caspar G Chorus (2018). "Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis". In: *Journal of choice modelling* 28, pp. 167–182.

An, KwangHoon et al. (2018). "Automatic Early Detection of Amyotrophic Lateral Sclerosis from Intelligible Speech Using Convolutional Neural Networks." In: *Interspeech*, pp. 1913–1917.

Anderson, Barrett et al. (2008). "Self-representation in social anxiety disorder: Linguistic analysis of autobiographical narratives". In: *Behaviour Research and Therapy* 46.10, pp. 1119–1125.

Angus-Leppan, Heather (2008). "Diagnosing epilepsy in neurology clinics: a prospective study". In: *Seizure* 17.5, pp. 431–436.

Asadi-Pooya, Ali A (2019). "Semiological classification of psychogenic nonepileptic seizures: a systematic review and a new proposal". In: *Epilepsy & Behavior* 100, p. 106412.

Avbersek, Andreja and Sanjay Sisodiya (2010). "Does the primary literature provide support for clinical signs used to distinguish psychogenic nonepileptic seizures from epileptic seizures?" In: *Journal of Neurology, Neurosurgery & Psychiatry* 81.7, pp. 719–725.

Baevski, Alexei et al. (2020). "wav2vec 2.0: A framework for self-supervised learning of speech representations". In: *Advances in Neural Information Processing Systems* 33, pp. 12449–12460.

Bandini, Andrea et al. (2018). "Automatic detection of amyotrophic lateral sclerosis (ALS) from video-based analysis of facial movements: speech and non-speech tasks". In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, pp. 150–157.

Baroni, Gislaine et al. (2021). "A novel scale for suspicion of psychogenic nonepileptic seizures: development and accuracy". In: *Seizure* 89, pp. 65–72.

Beghi, Massimiliano et al. (2020). "The semantics of epileptic and psychogenic nonepileptic seizures and their differential diagnosis". In: *Epilepsy & Behavior* 111, p. 107250.

Benbadis, Selim R and W Allen Hauser (2000). “An estimate of the prevalence of psychogenic non-epileptic seizures”. In: *Seizure* 9.4, pp. 280–281.

Benbadis, Selim R et al. (1995). “Value of tongue biting in the diagnosis of seizures”. In: *Archives of Internal Medicine* 155.21, pp. 2346–2349.

Beneteau, Erin et al. (2019). “Communication breakdowns between families and Alexa”. In: *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–13.

Berrar, Daniel (2019). *Cross-Validation.*

Biberon, Julien et al. (2020). “Differentiating PNES from epileptic seizures using conversational analysis on French patients: A prospective blinded study”. In: *Epilepsy & Behavior* 111, p. 107239.

Bien, Christian G et al. (2000). “Localizing value of epileptic visual auras”. In: *Brain* 123.2, pp. 244–253.

Bisignani, Antonio et al. (2019). “Implantable loop recorder in clinical practice”. In: *Journal of arrhythmia* 35.1, pp. 25–32.

Bradley, John G and Kathy A Davis (2003). “Orthostatic hypotension”. In: *American family physician* 68.12, pp. 2393–2398.

Braun, Virginia and Victoria Clarke (2006). “Using thematic analysis in psychology”. In: *Qualitative research in psychology* 3.2, pp. 77–101.

— (2022). “Conceptual and design thinking for thematic analysis.” In: *Qualitative Psychology* 9.1, p. 3.

Breiman, Leo (2001). “Random forests”. In: *Machine learning* 45.1, pp. 5–32.

Briggs, R et al. (2017). “Investigation and diagnostic formulation in patients admitted with transient loss of consciousness”. In.

Brignole, Michele et al. (2018). “2018 ESC Guidelines for the diagnosis and management of syncope”. In: *Kardiologia Polska (Polish Heart Journal)* 76.8, pp. 1119–1198.

Brigo, Francesco et al. (2015). “Postictal serum creatine kinase for the differential diagnosis of epileptic seizures and psychogenic non-epileptic seizures: a systematic review”. In: *Journal of neurology* 262.2, pp. 251–257.

Brown, Lily A et al. (2020). “Machine learning algorithms in suicide prevention: clinician interpretations as barriers to implementation”. In: *The Journal of Clinical Psychiatry* 81.3, p. 10951.

Brown, Richard J and Markus Reuber (2016a). “Psychological and psychiatric aspects of psychogenic non-epileptic seizures (PNES): a systematic review”. In: *Clinical Psychology Review* 45, pp. 157–182.

— (2016b). “Towards an integrative theory of psychogenic non-epileptic seizures (PNES)”. In: *Clinical psychology review* 47, pp. 55–70.

Bryman, Alan (2006). “Integrating quantitative and qualitative research: how is it done?” In: *Qualitative research* 6.1, pp. 97–113.

Cardeña, Etzel, Susannah Pick, and Richard Litwin (2020). “Differentiating psychogenic nonepileptic from epileptic seizures: A mixed-methods, content analysis study”. In: *Epilepsy & Behavior* 109, p. 107121.

Cascino, Gregory D (2002). “Video-EEG monitoring in adults”. In: *Epilepsia* 43, pp. 80–93.

Cassell, Eric J (1985). *Talking with patients, Volume 1: The theory of doctor-patient communication*. MIT Press.

Chakraborty, Koustav, Asmita Talele, and Savitha Upadhya (2014). “Voice recognition using MFCC algorithm”. In: *International Journal of Innovative Research in Advanced Engineering (IJIRAE)* 1.10, pp. 2349–2163.

Chen, Min et al. (2019). “Value of witness observations in the differential diagnosis of transient loss of consciousness”. In: *Neurology* 92.9, e895–e904.

Chowdhury, Fahmida A et al. (2021). “Localisation in focal epilepsy: a practical guide”. In: *Practical Neurology* 21.6, pp. 481–491.

Christodoulides, George and Mathieu Avanzi (2015). “Automatic detection and annotation of disfluencies in spoken French corpora”. In: *Sixteenth Annual Conference of the International Speech Communication Association*.

Chung, Cindy K and James W Pennebaker (2012). “Linguistic inquiry and word count (LIWC): pronounced “Luke,”... and other useful facts”. In: *Applied natural language processing: Identification, investigation and resolution*. IGI Global, pp. 206–229.

Claesen, Marc and Bart De Moor (2015). “Hyperparameter search in machine learning”. In: *arXiv preprint arXiv:1502.02127*.

Colman, N et al. (2004). “Diagnostic value of history taking in reflex syncope”. In: *Clinical Autonomic Research* 14.1, pp. i37–i44.

Cornaggia, Cesare Maria et al. (2012). “Conversation analysis in the differential diagnosis of Italian patients with epileptic or psychogenic non-epileptic seizures: a blind prospective study”. In: *Epilepsy & Behavior* 25.4, pp. 598–604.

Cragar, Dona E et al. (2002). “A review of diagnostic techniques in the differential diagnosis of epileptic and nonepileptic seizures”. In: *Neuropsychology review* 12.1, pp. 31–64.

Cristianini, Nello, John Shawe-Taylor, et al. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Cumbal, Ronald et al. (2021). ““You don’t understand me!”: Comparing ASR results for L1 and L2 speakers of Swedish”. In: *Interspeech 2021*.

Davis, Fred D (1989). “Perceived usefulness, perceived ease of use, and user acceptance of information technology”. In: *MIS quarterly*, pp. 319–340.

Davis, Fred D, Richard P Bagozzi, and Paul R Warshaw (1989). “User acceptance of computer technology: A comparison of two theoretical models”. In: *Management science* 35.8, pp. 982–1003.

Dehak, Najim et al. (2011). “Language recognition via i-vectors and dimensionality reduction”. In: *Twelfth annual conference of the international speech communication association*. Citeseer.

Devlin, Jacob et al. (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805*.

Dovgalyuk, Jacqueline et al. (2007). “The electrocardiogram in the patient with syncope”. In: *The American journal of emergency medicine* 25.6, pp. 688–701.

Ekberg, Katie and Markus Reuber (2015). “Can conversation analytic findings help with differential diagnosis in routine seizure clinic interactions?” In: *Communication & medicine* 12.1, pp. 13–24.

El-Wajeeh, Marwan, G Galal-Edeen, and Hoda Mokhtar (2014). “Technology acceptance model for mobile health systems”. In: *IOSR Journal of Mobile Computing and Acceptance* 1.1, pp. 21–33.

Elsey, Christopher et al. (2015). “Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics”. In: *Patient Education and Counseling* 98.9, pp. 1071–1077.

Eyben, Florian, Martin Wöllmer, and Björn Schuller (2010). “Opensmile: the munich versatile and fast open-source audio feature extractor”. In: *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462.

Fehlenberg, Dirk (1986). “The Discourse of Medicine: Dialectics of Medical Interviews (Book).” In: *Sociology of Health & Illness* 8.2, pp. 195–197.

Fiest, Kirsten M et al. (2017). "Prevalence and incidence of epilepsy: a systematic review and meta-analysis of international studies". In: *Neurology* 88.3, pp. 296–303.

Figueroa, Rosa L et al. (2012). "Predicting sample size required for classification performance". In: *BMC medical informatics and decision making* 12.1, pp. 1–10.

Fisher, Robert S (2017). "The new classification of seizures by the International League Against Epilepsy 2017". In: *Current neurology and neuroscience reports* 17.6, pp. 1–6.

Fitsiori, Aikaterini et al. (2019). "Morphological and advanced imaging of epilepsy: beyond the basics". In: *Children* 6.3, p. 43.

Frucht, Michael M et al. (2000). "Distribution of seizure precipitants among epilepsy syndromes." In: *Epilepsia* 41.12, pp. 1534–1539.

Glauser, Tracy et al. (2020). "Identifying epilepsy psychiatric comorbidities with machine learning". In: *Acta Neurologica Scandinavica* 141.5, pp. 388–396.

González-Carvajal, Santiago and Eduardo C Garrido-Merchán (2020). "Comparing BERT against traditional machine learning text classification". In: *arXiv preprint arXiv:2005.13012*.

Gratch, Jonathan et al. (2014). *The distress analysis interview corpus of human and computer interviews*. Tech. rep. UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES.

Greene, Jennifer C, Valerie J Caracelli, and Wendy F Graham (1989). "Toward a conceptual framework for mixed-method evaluation designs". In: *Educational evaluation and policy analysis* 11.3, pp. 255–274.

Grimm, Pamela (2010). "Social desirability bias". In: *Wiley international encyclopedia of marketing*.

Guest, Greg (2013). "Describing mixed methods research: An alternative to typologies". In: *Journal of mixed methods research* 7.2, pp. 141–151.

Guo, Xinjian et al. (2008). "On the class imbalance problem". In: *2008 Fourth international conference on natural computation*. Vol. 4. IEEE, pp. 192–201.

Gupta, Harshita and Divya Gupta (2016). "LPC and LPCC method of feature extraction in Speech Recognition System". In: *2016 6th international conference-cloud system and big data engineering (confluence)*. IEEE, pp. 498–502.

He, Haibo et al. (2008). "ADASYN: Adaptive synthetic sampling approach for imbalanced learning". In: *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, pp. 1322–1328.

Heritage, John (1998). "Oh-prefaced responses to inquiry". In: *Language in society* 27.3, pp. 291–334.

— (2002). "The limits of questioning: Negative interrogatives and hostile question content". In: *Journal of pragmatics* 34.10-11, pp. 1427–1446.

— (2015). "Well-prefaced turns in English conversation: A conversation analytic perspective". In: *Journal of Pragmatics* 88, pp. 88–104.

Heritage, John and Douglas W Maynard (2006). *Communication in medical care: Interaction between primary care physicians and patients*. Vol. 20. Cambridge University Press.

Hingray, C et al. (2016). "Psychogenic non-epileptic seizures (PNES)". In: *Revue neurologique* 172.4-5, pp. 263–269.

Holden, Richard J and Ben-Tzion Karsh (2009). "A theoretical model of health information technology usage behaviour with implications for patient safety". In: *Behaviour & Information Technology* 28.1, pp. 21–38.

Holtzman, Nicholas S et al. (2017). "A meta-analysis of correlations between depression and first person singular pronoun use". In: *Journal of Research in Personality* 68, pp. 63–68.

Honnibal, Matthew and Ines Montani (2017). "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". In: *To appear* 7.1, pp. 411–420.

Hu, Guoning and DeLiang Wang (2007). "Auditory segmentation based on onset and offset analysis". In: *IEEE transactions on audio, speech, and language processing* 15.2, pp. 396–405.

Huang, Zhaocheng, Julien Epps, and Dale Joachim (2019). "Investigation of speech landmark patterns for depression detection". In: *IEEE Transactions on Affective Computing*.

Iglesias, Juan F et al. (2009). "Stepwise evaluation of unexplained syncope in a large ambulatory population". In: *Pacing and Clinical Electrophysiology* 32, S202–S206.

Jefferson, Gail et al. (1983). *An exercise in the transcription and analysis of laughter*. Tilburg Univ., Department of Language and Literature.

Jenkins, Laura et al. (2016). "Neurologists can identify diagnostic linguistic features during routine seizure clinic interactions: results of a one-day teaching intervention". In: *Epilepsy & Behavior* 64, pp. 257–261.

Johnson, R Burke, Anthony J Onwuegbuzie, and Lisa A Turner (2007). "Toward a definition of mixed methods research". In: *Journal of mixed methods research* 1.2, pp. 112–133.

Jordan, Michael I and Tom M Mitchell (2015). "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245, pp. 255–260.

Jørgensen, Rikke et al. (2019). "The deadlock of saying "That is what we already do!" A thematic analysis of mental healthcare professionals' reactions to using an evidence-based intervention". In: *Journal of Psychiatric and Mental Health Nursing* 26.1-2, pp. 39–48.

Josephson, Colin B, Susan Rahey, and R Mark Sadler (2007). "Neurocardiogenic syncope: frequency and consequences of its misdiagnosis as epilepsy". In: *Canadian journal of neurological sciences* 34.2, pp. 221–224.

Kalkan, Asım et al. (2022). "A new biomarker in the differential diagnosis of epileptic seizure: Neurogranin". In: *The American Journal of Emergency Medicine* 54, pp. 147–150.

Kerr, Wesley T et al. (2020). "Objective score from initial interview identifies patients with probable dissociative seizures". In: *Epilepsy & Behavior* 113, p. 107525.

Keuleers, Emmanuel et al. (2015). "Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment". In: *The Quarterly Journal of Experimental Psychology* 68.8, pp. 1665–1692.

Kim, Sangwook, Swathi Kavuri, and Minho Lee (2013). "Deep network with support vector machines". In: *International Conference on Neural Information Processing*. Springer, pp. 458–465.

Kinney, Michael Owen, Stjepana Kovac, and Beate Diehl (2019). "Structured testing during seizures: a practical guide for assessing and interpreting ictal and postictal signs during video EEG long term monitoring". In: *Seizure* 72, pp. 13–22.

Kodish-Wachs, Jodi et al. (2018). "A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech". In: *AMIA Annual Symposium Proceedings*. Vol. 2018. American Medical Informatics Association, p. 683.

Koene, Ryan J, Wayne O Adkisson, and David G Benditt (2017). “Syncope and the risk of sudden cardiac death: Evaluation, management, and prevention”. In: *Journal of arrhythmia* 33.6, pp. 533–544.

Kohno, Ritsuko, Wayne O Adkisson, and David G Benditt (2018). “Tilt table testing for syncope and collapse”. In: *Herzschrittmachertherapie+ Elektrophysiologie* 29.2, pp. 187–192.

Kotsopoulos, Irene AW et al. (2003). “The diagnosis of epileptic and non-epileptic seizures”. In: *Epilepsy research* 57.1, pp. 59–67.

Kühnlein, Peter et al. (2008). “Diagnosis and treatment of bulbar symptoms in amyotrophic lateral sclerosis”. In: *Nature clinical practice Neurology* 4.7, pp. 366–374.

LaFrance Jr, W Curt et al. (2013). “Minimum requirements for the diagnosis of psychogenic nonepileptic seizures: a staged approach: a report from the International League Against Epilepsy Nonepileptic Seizures Task Force”. In: *Epilepsia* 54.11, pp. 2005–2018.

Lanseng, Even J and Tor W Andreassen (2007). “Electronic healthcare: a study of people’s readiness and attitude toward performing self-diagnosis”. In: *International Journal of Service Industry Management* 18.4, pp. 394–417.

Latif, Siddique et al. (2020). “Speech technology for healthcare: Opportunities, challenges, and state of the art”. In: *IEEE Reviews in Biomedical Engineering* 14, pp. 342–356.

LaValley, Michael P (2008). “Logistic regression”. In: *Circulation* 117.18, pp. 2395–2399.

Lempert, T (1996). “Recognizing syncope: pitfalls and surprises”. In: *Journal of the Royal Society of Medicine* 89.7, pp. 372–375.

Lempert, T, M Bauer, and D Schmidt (1994). “Syncope: a videometric analysis of 56 episodes of transient cerebral hypoxia”. In: *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 36.2, pp. 233–237.

Liu, Yang et al. (2006). “Enriching speech recognition with automatic detection of sentence boundaries and disfluencies”. In: *IEEE Transactions on audio, speech, and language processing* 14.5, pp. 1526–1540.

Liu, Ziming et al. (2021). “Automatic diagnosis and prediction of cognitive decline associated with alzheimer’s dementia through spontaneous speech”. In: *2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. IEEE, pp. 39–43.

Loper, Edward and Steven Bird (2002). “Nltk: The natural language toolkit”. In: *arXiv preprint cs/0205028*.

Low, Daniel M, Kate H Bentley, and Satrajit S Ghosh (2020). “Automated assessment of psychiatric disorders using speech: A systematic review”. In: *Laryngoscope Investigative Otolaryngology* 5.1, pp. 96–116.

Malmgren, Kristina, Markus Reuber, and Richard Appleton (2012). “Differential diagnosis of epilepsy”. In: *Oxford textbook of epilepsy and epileptic seizures*, pp. 81–94.

Matton, Katie, Melvin G McInnis, and Emily Mower Provost (2019). “Into the wild: Transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder”. In: *Interspeech*.

Mirheidari, Bahman et al. (2016). “Diagnosing people with dementia using automatic conversation analysis”. In: *Proceedings of interspeech*. ISCA, pp. 1220–1224.

Mirheidari, Bahman et al. (2017a). “An avatar-based system for identifying individuals likely to develop dementia”. In: *Interspeech 2017*. ISCA, pp. 3147–3151.

Mirheidari, Bahman et al. (2017b). "Toward the automation of diagnostic conversation analysis in patients with memory complaints". In: *Journal of Alzheimer's Disease* 58.2, pp. 373–387.

Mirheidari, Bahman et al. (2018). "Detecting Signs of Dementia Using Word Vector Representations." In: *Interspeech*, pp. 1893–1897.

Mirheidari, Bahman et al. (2019a). "Computational cognitive assessment: Investigating the use of an intelligent virtual agent for the detection of early signs of dementia". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2732–2736.

Mirheidari, Bahman et al. (2019b). "Dementia detection using automatic analysis of conversations". In: *Computer Speech & Language* 53, pp. 65–79.

Mirheidari, Bahman et al. (2020). "Improving Cognitive Impairment Classification by Generative Neural Network-Based Feature Augmentation." In: *INTERSPEECH*, pp. 2527–2531.

Mizrachi, Esther M and Kranthi K Sitammagari (2018). "Cardiac Syncope". In.

Mohan, KK, ON Markand, and V Salanov (1996). "Diagnostic utility of video EEG monitoring in paroxysmal events". In: *Acta Neurologica Scandinavica* 94.5, pp. 320–325.

Morales, Michelle Renee and Rivka Levitan (2016). "Speech vs. text: A comparative analysis of features for depression detection systems". In: *2016 IEEE spoken language technology workshop (SLT)*. IEEE, pp. 136–143.

Mori, Masahiro, Karl F MacDorman, and Norri Kageki (2012). "The uncanny valley [from the field]". In: *IEEE Robotics & automation magazine* 19.2, pp. 98–100.

Moro-Velazquez, Laureano et al. (2021). "Advances in Parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects". In: *Biomedical Signal Processing and Control* 66, p. 102418.

Mukherjee, Pritam et al. (2022). "Confounding factors need to be accounted for in assessing bias by machine learning algorithms". In: *Nature Medicine* 28.6, pp. 1159–1160.

Mundt, James C et al. (2007). "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology". In: *Journal of neurolinguistics* 20.1, pp. 50–64.

NICE (2022). "Epilepsies in children, young people and adults NG217] https://www.nice.org.uk/guidance In.

Noachtar, Soheyl and Jan Rémi (2009). "The role of EEG in epilepsy: a critical review". In: *Epilepsy & Behavior* 15.1, pp. 22–33.

Noble, William S (2006). "What is a support vector machine?" In: *Nature biotechnology* 24.12, pp. 1565–1567.

Nunnally, Jum C (1978). "An overview of psychological measurement". In: *Clinical diagnosis of mental disorders*, pp. 97–146.

Oliva, Megan et al. (2008). "The diagnostic value of oral lacerations and incontinence during convulsive "seizures"". In: *Epilepsia* 49.6, pp. 962–967.

O'Malley, Ronan Peter Daniel et al. (2021). "Fully automated cognitive screening tool based on assessment of speech and language". In: *Journal of Neurology, Neurosurgery & Psychiatry* 92.1, pp. 12–15.

O'Shaughnessy, Douglas (2008). "Automatic speech recognition: History, methods and challenges". In: *Pattern Recognition* 41.10, pp. 2965–2979.

Pan, Yilin et al. (2019). "Automatic hierarchical attention neural network for detecting AD". In: *Proceedings of Interspeech 2019*. International Speech Communication Association (ISCA), pp. 4105–4109.

Pan, Yilin et al. (2020). "Acoustic feature extraction with interpretable deep neural network for neurodegenerative related disorder classification". In: *Proceedings of Interspeech 2020*. International Speech Communication Association (ISCA), pp. 4806–4810.

Pan, Yilin et al. (2021). "Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic-and BERT-Based Alzheimer's Dementia Detection Through Spontaneous Speech." In: *Interspeech*, pp. 3810–3814.

Pan, Zhongde et al. (2018). "Detecting manic state of bipolar disorder based on support vector machine and gaussian mixture model using spontaneous speech". In: *Psychiatry investigation* 15.7, p. 695.

Panayotov, Vassil et al. (2015). "Librispeech: an asr corpus based on public domain audio books". In: *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 5206–5210.

Papagno, Costanza et al. (2017). "Differentiating PNES from epileptic seizures using conversational analysis". In: *Epilepsy & Behavior* 76, pp. 46–50.

Patton, Michael Quinn (1990). *Qualitative evaluation and research methods*. SAGE Publications, inc.

Pavlyshenko, Bohdan (2018). "Using stacking approaches for machine learning models". In: *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*. IEEE, pp. 255–258.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12, pp. 2825–2830.

Pennebaker, James W (1997). "Writing about emotional experiences as a therapeutic process". In: *Psychological science* 8.3, pp. 162–166.

Pennebaker, James W and Martha E Francis (1996). "Cognitive, emotional, and language processes in disclosure". In: *Cognition & emotion* 10.6, pp. 601–626.

Pennebaker, James W, Martha E Francis, and Roger J Booth (2001). "Linguistic inquiry and word count: LIWC 2001". In: *Mahway: Lawrence Erlbaum Associates* 71.2001, p. 2001.

Pennebaker, James W and Laura A King (1999). "Linguistic styles: language use as an individual difference." In: *Journal of personality and social psychology* 77.6, p. 1296.

Pennebaker, James W, Tracy J Mayne, and Martha E Francis (1997). "Linguistic predictors of adaptive bereavement." In: *Journal of personality and social psychology* 72.4, p. 863.

Pennebaker, James W et al. (2015). *The development and psychometric properties of LIWC2015*. Tech. rep.

Petkar, Sanjiv et al. (2011). "Initial experience with a rapid access blackouts triage clinic". In: *Clinical medicine* 11.1, p. 11.

Petti, Ulla, Simon Baker, and Anna Korhonen (2020). "A systematic literature review of automatic Alzheimer's disease detection from speech and language". In: *Journal of the American Medical Informatics Association* 27.11, pp. 1784–1797.

Pieters, Huibrie C et al. (2016). ""It was five years of hell": Parental experiences of navigating and processing the slow and arduous time to pediatric resective epilepsy surgery". In: *Epilepsy & Behavior* 62, pp. 276–284.

Pillai, Jyoti and Michael R Sperling (2006). "Interictal EEG and the diagnosis of epilepsy". In: *Epilepsia* 47, pp. 14–22.

Plug, Leendert and Markus Reuber (2009). "Making the diagnosis in patients with blackouts: it's all in the history". In: *Practical Neurology* 9.1, pp. 4–15.

Plug, Leendert, Basil Sharrack, and Markus Reuber (2009a). "Conversation analysis can help to distinguish between epilepsy and non-epileptic seizure disorders: a case comparison". In: *Seizure* 18.1, pp. 43–50.

— (2009b). “Seizure metaphors differ in patients’ accounts of epileptic and psychogenic nonepileptic seizures”. In: *Epilepsia* 50.5, pp. 994–1000.

— (2010). “Seizure, fit or attack? The use of diagnostic labels by patients with epileptic or non-epileptic seizures”. In: *Applied Linguistics* 31.1, pp. 94–114.

Pope, Catherine and Nick Mays (1995). “Qualitative research: reaching the parts other methods cannot reach: an introduction to qualitative methods in health and health services research”. In: *bmj* 311.6996, pp. 42–45.

Povey, Daniel et al. (2011). “The Kaldi speech recognition toolkit”. In: *IEEE 2011 workshop on automatic speech recognition and understanding*. CONF. IEEE Signal Processing Society.

Preum, Sarah Masud et al. (2021). “A review of cognitive assistants for healthcare: Trends, prospects, and future directions”. In: *ACM Computing Surveys (CSUR)* 53.6, pp. 1–37.

Raj, Vidya et al. (2014). “Psychogenic pseudosyncope: diagnosis and management”. In: *Autonomic Neuroscience* 184, pp. 66–72.

Rana, Rajib et al. (2019). “Automated screening for distress: A perspective for the future”. In: *European journal of cancer care* 28.4, e13033.

Rawlings, GH et al. (2017a). “Panic symptoms in transient loss of consciousness: frequency and diagnostic value in psychogenic nonepileptic seizures, epilepsy and syncope”. In: *Seizure* 48, pp. 22–27.

Rawlings, Gregg H et al. (2017b). “Written accounts of living with epilepsy: A thematic analysis”. In: *Epilepsy & Behavior* 72, pp. 63–70.

— (2017c). “Written accounts of living with psychogenic nonepileptic seizures: A thematic analysis”. In: *Seizure* 50, pp. 83–91.

Ray, Susmita (2019). “A quick review of machine learning algorithms”. In: *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*. IEEE, pp. 35–39.

Razmak, Jamil and Charles Bélanger (2018). “Using the technology acceptance model to predict patient attitude toward personal health records in regional communities”. In: *Information Technology & People* 31.2, pp. 306–326.

Rejaibi, Emna et al. (2022). “MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech”. In: *Biomedical Signal Processing and Control* 71, p. 103107.

Reuber, Markus et al. (2009). “Using interactional and linguistic analysis to distinguish between epileptic and psychogenic nonepileptic seizures: a prospective, blinded multirater study”. In: *Epilepsy & Behavior* 16.1, pp. 139–144.

Reuber, Markus et al. (2011). “Psychogenic nonepileptic seizure manifestations reported by patients and witnesses”. In: *Epilepsia* 52.11, pp. 2028–2035.

Reuber, Markus et al. (2016). “Value of patient-reported symptoms in the diagnosis of transient loss of consciousness”. In: *Neurology* 87.6, pp. 625–633.

Ringeval, Fabien et al. (2018). “AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition”. In: *Proceedings of the 2018 on audio/visual emotion challenge and workshop*, pp. 3–13.

Robinson, Jeffrey D (2003). “An interactional structure of medical activities during acute visits and its implications for patients’ participation”. In: *Health Communication* 15.1, pp. 27–59.

Robson, Catherine, Paul Drew, and Markus Reuber (2016). “The role of companions in outpatient seizure clinic interactions: a pilot study”. In: *Epilepsy & Behavior* 60, pp. 86–93.

Robson, Catherine et al. (2012). “Catastrophising and normalising in patient’s accounts of their seizure experiences”. In: *Seizure* 21.10, pp. 795–801.

Rosenbach, Charlotte and Babette Renneberg (2015). "Remembering rejection: Specificity and linguistic styles of autobiographical memories in borderline personality disorder and depression". In: *Journal of behavior therapy and experimental psychiatry* 46, pp. 85–92.

Rutowski, Tomasz et al. (2019). "Optimizing Speech-Input Length for Speaker-Independent Depression Classification." In: *Interspeech*, pp. 3023–3027.

Rutowski, Tomasz et al. (2020). "Depression and anxiety prediction using deep language models and transfer learning". In: *2020 7th International Conference on Behavioural and Social Computing (BESC)*. IEEE, pp. 1–6.

Sacks, Harvey (1974). "Schegloff, and Jefferson, GA Simplest Systematic for the Turn Taking for Conversation". In: *Language* 33.4.

— (1992). "Lectures on conversation (2 vols.; G. Jefferson, Ed.)" In: *Cambridge, MA*.

Sacks, Harvey, Emanuel A Schegloff, and Gail Jefferson (1978). "A simplest systematics for the organization of turn taking for conversation". In: *Studies in the organization of conversational interaction*. Elsevier, pp. 7–55.

Salekin, Asif et al. (2018). "A weakly supervised learning framework for detecting social anxiety and depression". In: *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2.2, pp. 1–26.

Scheffer, Ingrid E et al. (2017). "ILAE classification of the epilepsies: position paper of the ILAE Commission for Classification and Terminology". In: *Epilepsia* 58.4, pp. 512–521.

Schegloff, Emanuel A (1972). "Notes on a conversational practice: Formulating place". In.

— (1992). "On talk and its institutional occasions". In: *Talk at work: Interaction in institutional settings* 101, p. 34.

Schiffrin, Deborah (1980). "Meta-talk: Organizational and evaluative brackets in discourse". In: *Sociological inquiry* 50.3-4, pp. 199–236.

Schoonenboom, Judith and R Burke Johnson (2017). "How to construct a mixed methods research design". In: *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie* 69.2, pp. 107–131.

Schwabe, Meike, Stephen J Howell, and Markus Reuber (2007). "Differential diagnosis of seizure disorders: a conversation analytic approach". In: *Social science & medicine* 65.4, pp. 712–724.

Schwabe, Meike et al. (2008). "Listening to people with seizures: how can linguistic analysis help in the differential diagnosis of seizure disorders?" In: *Communication & medicine* 5.1, p. 59.

Sekhon, Mandeep, Martin Cartwright, and Jill J Francis (2017). "Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework". In: *BMC health services research* 17.1, pp. 1–13.

Shao, Yang and DeLiang Wang (2008). "Robust speaker identification using auditory features and computational auditory scene analysis". In: *2008 IEEE international conference on acoustics, speech and signal processing*. IEEE, pp. 1589–1592.

Shibata, Daisaku et al. (2016). "Detecting Japanese patients with Alzheimer's disease based on word category frequencies". In: *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pp. 78–85.

Shorvon, Simon D (2011). "The etiologic classification of epilepsy". In: *Epilepsia* 52.6, pp. 1052–1057.

Sidnell, Jack (2011). *Conversation analysis: An introduction*. John Wiley & Sons.

Skivington, Kathryn et al. (2021). "A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance". In: *bmj* 374.

Sluis, Rachel A et al. (2020). "An automated approach to examining pausing in the speech of people with dementia". In: *American Journal of Alzheimer's Disease & Other Dementias®* 35, p. 1533317520939773.

Spearman, Charles (1904). "The Proof and Measurement of Association Between Two Things." In: *The American Journal of Psychology* 15.1, pp. 72–101.

Staley, Kevin (2015). "Molecular mechanisms of epilepsy". In: *Nature neuroscience* 18.3, pp. 367–372.

Starke, Georg et al. (2022). "Explainability as fig leaf? An exploration of experts' ethical expectations towards machine learning in psychiatry". In: *AI and Ethics*, pp. 1–12.

Stiell, Ian G and Carol Bennett (2007). "Implementation of clinical decision rules in the emergency department". In: *Academic Emergency Medicine* 14.11, pp. 955–959.

Stivers, Tanya (2002). "Presenting the problem in pediatric encounters:" symptoms only" versus" candidate diagnosis" presentations". In: *Health Communication* 14.3, pp. 299–338.

Stivers, Tanya and Jeffrey D Robinson (2006). "A preference for progressivity in interaction". In: *Language in society* 35.3, pp. 367–392.

Sueri, Chiara et al. (2018). "Diagnostic biomarkers of epilepsy". In: *Current Pharmaceutical Biotechnology* 19.6, pp. 440–450.

Sundermeyer, Martin et al. (2013). "Comparison of feedforward and recurrent neural network language models". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 8430–8434.

Sveinbjornsdottir, Sigurlaug (2016). "The clinical symptoms of Parkinson's disease". In: *Journal of neurochemistry* 139, pp. 318–324.

Tannemaat, Martijn R et al. (2013). "The semiology of tilt-induced psychogenic pseudosyncope". In: *Neurology* 81.8, pp. 752–758.

Tashakkori, Abbas (2009). *Are we there yet? The state of the mixed methods community*.

Tharwat, Alaa (2020). "Classification assessment methods". In: *Applied Computing and Informatics*.

Toerien, Merran, Clare Jackson, and Markus Reuber (2020). "The normativity of medical tests: test ordering as a routine activity in "new problem" consultations in secondary care". In: *Research on Language and Social Interaction* 53.4, pp. 405–424.

Tondo, Leonardo, Gustavo H Vazquez, and Ross J Baldessarini (2017). "Depression and mania in bipolar disorder". In: *Current neuropharmacology* 15.3, pp. 353–358.

Toshniwal, Shubham et al. (2018). "A comparison of techniques for language model integration in encoder-decoder speech recognition". In: *2018 IEEE spoken language technology workshop (SLT)*. IEEE, pp. 369–375.

Vabalas, Andrius et al. (2019). "Machine learning algorithm validation with a limited sample size". In: *PloS one* 14.11, e0224365.

Valley, Morgan and Lorann Stallones (2018). "A thematic analysis of health care workers' adoption of mindfulness practices". In: *Workplace health & safety* 66.11, pp. 538–544.

Valstar, Michel et al. (2013). "Avec 2013: the continuous audio/visual emotion and depression recognition challenge". In: *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pp. 3–10.

Vashkevich, Maxim, Alexander Petrovsky, and Yuliya Rushkevich (2019). "Bulbar ALS detection based on analysis of voice perturbation and vibrato". In: *2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. IEEE, pp. 267–272.

Vollmer, Sebastian et al. (2020). “Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness”. In: *bmj* 368.

Walker, Traci et al. (2020). “Developing an intelligent virtual agent to stratify people with cognitive complaints: A comparison of human–patient and intelligent virtual agent–patient interaction”. In: *Dementia* 19.4, pp. 1173–1188.

Wang, Bo et al. (2020). “Learning to detect bipolar disorder and borderline personality disorder with language and speech in non-clinical interviews”. In: *arXiv preprint arXiv:2008.03408*.

Wang, Jun et al. (2018). “Automatic prediction of intelligible speaking rate for individuals with ALS from speech acoustic and articulatory samples”. In: *International journal of speech-language pathology* 20.6, pp. 669–679.

Wang, Yan-Qiu et al. (2021). “Prolactin levels as a criterion to differentiate between psychogenic non-epileptic seizures and epileptic seizures: a systematic review”. In: *Epilepsy Research* 169, p. 106508.

Wardrope, Alistair, Ellen Newberry, and Markus Reuber (2018). “Diagnostic criteria to aid the differential diagnosis of patients presenting with transient loss of consciousness: A systematic review”. In: *Seizure* 61, pp. 139–148.

Wardrope, Alistair et al. (2020a). “Machine learning as a diagnostic decision aid for patients with transient loss of consciousness”. In: *Neurology: Clinical Practice* 10.2, pp. 96–105.

Wardrope, Alistair et al. (2020b). “Peri-ictal responsiveness to the social environment is greater in psychogenic nonepileptic than epileptic seizures”. In: *Epilepsia* 61.4, pp. 758–765.

Waytz, Josh, Adam S Cifu, and Scott DC Stern (2018). “Evaluation and management of patients with syncope”. In: *JAMA* 319.21, pp. 2227–2228.

Whitfield, Andrew et al. (2020). “Catastrophising and repetitive negative thinking tendencies in patients with psychogenic non-epileptic seizures or epilepsy”. In: *Seizure* 83, pp. 57–62.

Wieling, Wouter et al. (2009). “Symptoms and signs of syncope: a review of the link between physiology and clinical clues”. In: *Brain* 132.10, pp. 2630–2642.

Wu, Hao and Xihong Wu (2007). “Context dependent syllable acoustic model for continuous Chinese speech recognition”. In: *Eighth Annual Conference of the International Speech Communication Association*.

Xu, Ying et al. (2016). “Frequency of a false positive diagnosis of epilepsy: a systematic review of observational studies”. In: *Seizure* 41, pp. 167–174.

Xue, Lanny Y and Anthony L Ritaccio (2006). “Reflex seizures and reflex epilepsy”. In: *American journal of electroneurodiagnostic technology* 46.1, pp. 39–48.

Yao, Yuan et al. (2017). “Conversation analysis in differential diagnosis between epileptic seizure and psychogenic nonepileptic seizure”. In: *Chinese Journal of Neurology*, pp. 266–270.

Yu, Dong, Li Deng, and George Dahl (2010). “Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition”. In: *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*. sn.

Yuan, Jiahong et al. (2021). “Pauses for detection of Alzheimer’s disease”. In: *Frontiers in Computer Science* 2, p. 624488.

Zhang, Baobao et al. (2021). “Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers”. In: *Journal of Artificial Intelligence Research* 71, pp. 591–666.