Phenotype and aetiology of ALS investigated using genetic, environmental, and clinical data analysis

Martin, Sarah

Awarding institution:
King's College London

Phenotype and aetiology of ALS investigated using genetic, environmental, and clinical data analysis

Sarah Opie-Martin

Institute of Psychiatry, Psychology and Neuroscience

Thesis submitted for the degree of Ph.D.

2021

Table of contents

Table of figures

Table of tables

Dedication

This thesis is dedicated to my family and friends.

Abstract

The work presented in this thesis uses large-scale data collection and analysis techniques to understand the factors modifying ALS risk and phenotype. I test the hypotheses that overall incidence of amyotrophic lateral sclerosis (ALS) in the UK is similar to previously reported values, smoking is a modifier of disease risk in ALS and *SOD1* mutations are modifiers of ALS phenotype.

Amyotrophic lateral sclerosis is a neurodegenerative disease characterised by death or loss of function of upper and or lower motor neurons leading to paralysis. There is currently no cure for ALS, and symptoms are relentlessly progressive with most people dying within 2-3 years of diagnosis, generally of respiratory failure.

ALS has a variable phenotype, with age of onset and progression differing between people. Additionally, although genetic modelling shows that up to 40% of variance in liability could be attributable to environmental factors, these have yet to be established.

Disease risk and phenotype modifiers can only be understood by analysing detailed genetic, environmental and clinical information on large numbers of people, using multiple methodologies.

In this thesis I report an updated incidence estimate for the UK based on data from the newly established MND Register for England, Wales and Northern Ireland. I also present work on whether people in the UK who smoke are at risk increased of ALS by using a novel to ALS methodology to quantify smoking intensity, and follow up with a Mendelian Randomisation study to provide corroboration with our findings. Finally, I present a global study of clinical phenotype in people with ALS who have tested positive for mutations in the *SOD1* gene.

We find that ALS incidence in the UK is similar to previously reported values. Smoking is unlikely to be a risk factor for ALS, and that people with *SOD1* ALS have a distinct phenotype from those with sporadic ALS.

Chapter 1 Introduction

1.1 Clinical summary of Amyotrophic Lateral Sclerosis

Amyotrophic lateral sclerosis (also referred to as motor neuron disease in the UK) is the umbrella term for a group of related diseases characterised primarily by progressive loss of function or death, of upper and or lower motor neurons. The first known descriptions of ALS cases are from reports written in 1824 by Charles Bell, however Jean-Marie Charcot associated clinical symptoms with corticospinal tract pathology, and first coined the term Amyotrophic Lateral Sclerosis (Charcot, 1874).

The cellular pathological process leads to clinical symptoms of progressive paralysis with death often from respiratory failure (Brown and Al-Chalabi, 2017). Onset of motor symptoms is usually focal and spreads, either within the same body region for example, starting in the foot and spreading to the whole leg, or through neurologically connected areas for example rostro-caudal (head to tail) or contra-lateral (eg right side central nervous system (CNS) damage affecting left side) (Dharmadasa, Matamala, Howells, Vucic, & Kiernan, 2020; Ravits and La Spada, 2009). Any voluntary muscle can be affected, and symptoms could start in any area the disease can affect. Although exact proportions vary between geographic regions, onset of motor symptoms is most commonly observed in the limbs followed by onset in the bulbar region (changes affecting speech or swallowing) and in rare cases onset of motor symptoms is in the thoracic region (affecting respiratory function) (Marin et al., 2016). The sphincter, ocular muscles and hearing associated musculature are not usually affected until very late in the disease and are muscles that are not paralysed in rapid eye movement sleep, which is consistent with the idea of shared motor networks (Turner and Al-Chalabi, 2020).

Proportion of upper and lower motor neuron degeneration is not uniform between people and under the ALS umbrella are three main diagnostic groups based on motor neuron involvement. Classic ALS (referred to as ALS in the UK) involves degeneration of both upper and lower motor neurons; Primary Lateral Sclerosis (PLS) describes when only upper motor neurons are affected, and this remains the case over time; Progressive Muscular Atrophy (PMA) is the diagnosis given if only lower motor neurons are affected, and again this remains the case over time (Al-Chalabi et al., 2016; Swinnen and Robberecht, 2014). Between the extremes of PLS and PMA there are upper and lower motor neuron predominant forms of ALS, although some regard all patterns, including PLS and PMA, as forms of ALS. Motor neuron loss may be confined to certain body regions in combination with restricted involvement of upper and or lower motor neurons. For example, flail arm variant, involves lower motor neuron degeneration in the upper limbs with possible upper motor neuron involvement in the lower limbs, and progressive bulbar palsy affects lower motor neurons in the bulbar region

(Al-Chalabi et al., 2016; Swinnen and Robberecht, 2014). Diagnoses require monitoring over time as symptoms in second or third CNS regions (regions mapping to upper limbs, lower limbs, bulbar and thoracic), or signs of motor involvement may appear much later in the disease course.

The clinical symptoms of lower motor neuron loss are fasciculations, muscle wasting, weakness and hyporeflexia, and of upper motor neuron loss include hyperreflexia and spasticity (de Carvalho, Kiernan, & Swash, 2017; Huynh et al., 2016). Needle-point electrophysiology is used in addition to clinical assessment to measure lower motor neuron loss (de Carvalho et al., 2008). 'Split' syndromes caused by dissociated muscular atrophy are a feature of ALS motor symptoms. The first to be described, and most widely recognised, is split hand syndrome where thenar muscles and the lateral interosseus muscles are affected but hypothenar muscles are not (Wilbourn, 2000). The split hand index, developed to quantify the phenomenon, is both sensitive and specific for ALS, so can be used diagnostically (Menon, Kiernan, Yiannikas, Stroud, & Vucic, 2013). More recently split elbow (where the biceps are more affected than triceps) and split leg syndrome (ankle plantar flexors are more affected compared to dorsiflexors) have also been described – although these have not been validated as diagnostic tools (Khalaf et al., 2019; Simon et al., 2015). It has been suggested these syndromes provide evidence for a positive link between cortical representation and risk of degeneration (Vucic, 2019).

Many non-motor symptoms of ALS have been reported with common symptoms including cognitive and behavioural changes, and weight loss (T. Fang, Jozsa, & Al-Chalabi, 2017). These symptoms may appear before, contemporaneously, or after appearance of motor symptoms. Cognitive impairment is present in 50% of people with ALS and as many as 15% of these people will reach criteria for a diagnosis of frontotemporal dementia (FTD) (Raaphorst, de Visser, Linssen, de Haan, & Schmand, 2010; Ringholz et al., 2005). There are three main forms of cognitive impairment that are seen in people with ALS; these are: executive dysfunction where people struggle with planning, problem-solving, organisation and time management; problems with working memory, language impairment, social cognition; and behaviour changes such as ego centric or selfish behaviours and apathy. There is some genetic overlap between FTD and ALS, in that risk genes such as *C9orf72, FUS, TBK1, TARDBP* and *VCP* are associated with both diseases (Abramzon, Fratta, Traynor, & Chia, 2020). There is cellular pathological overlap in the form of TDP-43 protein aggregates and clinical overlap where some people with FTD also have motor symptoms and go on to develop ALS (Ferrari, Kapogiannis, Huey, & Momeni, 2011; Scotter, Chen, & Shaw, 2015).

Another behavioural change is pseudobulbar affect (sometimes called pseudobulbar palsy), which is characterised by frequent, spontaneous involuntary laughing and or crying that is discordant to how

a person is feeling or the underlying situation and is observed in 15-60% of people with ALS (Brooks, Crumpacker, Fellus, Kantor, & Kaye, 2013). Severe weight loss is a frequent occurrence in people with ALS and is associated with worse prognosis (Desport et al., 1999; Moglia et al., 2019). It is not completely clear why people lose weight and is probably due to malnutrition, hypermetabolism, cachexia (wasting of the body due to severe chronic illness) and loss of appetite (T. Fang et al., 2017).

Due to the variability of symptoms and the spread of disease, people need multidisciplinary care, the nature of which will change over time (Balendra, Al Khleifat, Fang, & Al-Chalabi, 2019). Specialist multidisciplinary clinics have been shown to improve outcomes (Hardiman et al., 2017; Martin et al., 2017; Rooney et al., 2015).

1.1.1. Diagnosis of ALS

Diagnosis of ALS takes on average a year, likely due to a combination of time-consuming factors such as referral to specialist centres, exclusion of ALS mimics and the need to show the disease is progressing (Richards, Morren, & Pioro, 2020). This is compounded by the lack of a diagnostic test, and the lack of an effective therapy for what is a devastating diagnosis, so that doctors wait until they are very certain of the diagnosis before referral. The diagnosis is clinical, meaning it is based on symptoms of motor neuron involvement, patient history and negative tests for other motor disorders. In some cases, genetic testing may be undertaken as part of diagnosis, but this is usually for people with a reported positive family history and does not replace clinical diagnostic workup (Shefner et al., 2020).

The El Escorial criteria, and their Arlie House and Awaji revisions are used as a diagnostic framework for ALS, and are detailed in table 1.1 (Brooks, 1994; Brooks, Miller, Swash, & Munsat, 2000; de Carvalho et al., 2008). They have now been superseded by the Gold Coast Criteria (Shefner et al., 2020). A simplification of these criteria is that, the body is divided into central nervous system regions (upper limb, lower limb, bulbar and thoracic) and depending on the number of regions with presence of upper and or lower motor neuron symptoms people are assigned diagnostic categories. A recent revision, the Awaji criteria, categorises people as having Possible, Probable or Definite ALS (de Carvalho et al., 2008). The revisions have clarified acceptable lower motor neuron signs and improved on diagnostic accuracy (both sensitivity and specificity) (Boekestein, Kleine, Hageman, Schelhaas, & Zwarts, 2010; Johnsen et al., 2019). However, the category nomenclature does not reflect diagnostic certainty, interrater reliability is low, and people with PLS may be included (Johnsen et al., 2019; Shefner et al., 2020). To address these limitations, and to simplify the diagnostic categories into a minimum criterion needed to be diagnosed with ALS, the ‘Gold Coast

Criteria' have been developed (Shefner et al., 2020). If someone has progressive motor impairment, upper and lower motor neuron symptoms in one body region or lower motor neuron symptoms in two body regions and other causes have been excluded then they meet minimum diagnostic criteria. The Gold Coast Criteria represent a consensus of experts and the diagnostic category is not yet clinically validated. The current frameworks do not include presence of non-motor changes, and further revision may be required with the development of serological and radiological biomarkers (Verber et al., 2019).

Criteria	Definite ALS	Probable ALS	Laboratory supported probable ALS	Possible ALS	Suspected ALS
El Escorial criteria (1994)	Three CNS regions with UMN and LMN signs	Two CNS regions with upper and LMN signs, with UMN signs rostral to LMN signs	n/a	One CNS region with UMN and LMN signs, UMN signs only in two regions or LMN signs rostral to UMN signs	Two or more CNS regions with LMN signs only
Arlie house criteria (2000) (**Awaji-Shima criteria 2008)**	**Clinical and electrophysiological evidence of** UMN and LMN signs in the bulbar region and at least two spinal regions and LMN signs in three spinal regions	**Clinical or electrophysiological evidence** of UMN and LMN in at least two regions with some UMN signs rostral to lower motor neuron signs	Clinical evidence of UMN and LMN signs in only one region or UMN signs alone in one region and evidence of LMN signs in at least two regions (**removed for Awaji-Shima revision)**	**Clinical or electrophysiological evidence** of UMN and LMN signs in only one region or UMN signs alone on two or more regions or LMN signs rostral to UMN signs	n/a
Gold coast criteria	Minimum criteria for diagnosis of ALS (diagnosis is either ALS or not ALS)				
	(1) progressive motor impairment documented by history or repeated clinical assessment, preceded by normal motor function; (2) presence of UMN and LMN signs in at least 1 body region (with UMN and LMN dysfunction noted in the same body region if only one body region is involved) or LMN dysfunction in at least 2 body regions; and (3) investigations excluding other disease processes				

Table 1-1 ALS diagnostic criteria. CNS = central nervous system, UMN = upper motor neuron, LMN = lower motor neuron

Phenotypic traits such as age and site of onset of disease symptoms, presence of family history, severity of symptoms, degree of involvement of motor neurons and presence of cognitive changes have been used in diagnostic description (Al-Chalabi et al., 2016). Some of these traits have clinical implications for disease progression and care needs, such as site of onset. Many of these features are continuous traits with non-standardised definitions which can make comparisons and pooling data from different sources challenging, although standardisation efforts are underway due to initiatives such as TRICALS (https://www.tricals.org).

1.1.2. Disease progression

The simplest way of quantifying disease progression in ALS is by measuring time from onset of motor symptoms to death, and mortality is a common endpoint in research studies. As there is currently no cure for ALS, disease prognosis is an important part of information given at diagnosis. Many models that use data collected at first presentation to predict disease duration have been developed and the most comprehensively validated model of European survival is the ENCALS model based on the records of 11,475 people with ALS from 9 countries (Westeneng et al., 2018). From this dataset, time from onset of symptoms to either death, tracheostomy or use of non-invasive ventilation for >23 hours/day is accurately modelled as being in one of 5 categories that range from a median survival of 17 months to a median survival time of 7 years, demonstrating the variability of survival in ALS (Westeneng et al., 2018).

Simple measures of survival time do not account for disability progression as muscles lose function. The ALS functional rating scale - revised (ALSFRS-R) measures a range of daily living activities and respiratory measures, focussing on those activities affected by motor symptoms (Cedarbaum et al., 1999). It is possible for patients to self-report their score online and the score is used as an outcome in many clinical trials (Maier et al., 2012). The rate of change in score can be used as an indication of how quickly disease is progressing and is included in some prediction models (Steinbach et al., 2020; Westeneng et al., 2018). The scale has been criticized, both for its statistical properties and in particular, the poor assessment of respiratory function (van Eijk and van Den Berg, 2020). What constitutes a clinically meaningful change in the ALSFRS-R scale is hard to determine and the scale is ordinal so parametric statistics may not be appropriate without conversion to an interval or ratio scale.

Staging systems are clinical tools that put people into groups depending on the severity of disease, in a way that is therapeutically and prognostically meaningful. There are two widely used staging systems in ALS. The King's staging system allocates people into categories ranging from 1-5, where 1 is symptom onset and 5 is death (Balendra et al., 2019). Stages 2 and 3 indicate involvement of two

or three CNS regions respectively while stage four is needing gastrostomy and or non-invasive ventilation (Balendra et al., 2019). The stage at which a patient is at on the King's staging system can be estimated from ALSFRS-R scores and has shown to have good interrater reliability (Balendra et al., 2014). The Milano and Torino Staging (MITOS) system directly uses the ALSFRS-R to define domains (for example respiratory or speech) and the stages are based on domains where function is lost (Chiò, Hammond, Mora, Bonito, & Filippini, 2015). There are six stages ranging from 0-5 where 0 is no functional domains lost and 5 is death (Chiò et al., 2015).

ALS is in almost all cases progressive, although short plateaus in progression as detected by repeat measurements of the ALSFRS-R are not uncommon (Bedlack et al., 2016). In very rare cases, people diagnosed with ALS have been recorded to make full recoveries. It is not yet clear why this is, the cases recorded are not typical of the normally progressive population – and are often associated with concurrent autoimmune diseases such as myasthenia gravis (Harrison et al., 2018).

1.1.3. Treatments and clinical trials

There are two approved pharmacological treatments that slow ALS progression, riluzole and edaravone (Abe et al., 2017; Bensimon, Lacomblez, & Meininger, 1994). Additionally, nuedexta has been approved to treat symptoms of pseudobulbar palsy (Cruz, 2013), and anecdotally may also improve bulbar function generally. Non-invasive ventilation and gastrostomy are physical interventions undertaken in the late stages of disease that treat symptoms of breathlessness and dysphagia and also prolong life (Bourke et al., 2006; Group, 2015).

Riluzole is a glutamate blocker and was investigated because glutamate-induced excitotoxicity is a pathological consequence of ALS. Riluzole was shown to be effective in a clinical trial in 1994, is safe (provided blood count and liver enzymes are monitored regularly) and is now widely available to patients (Bensimon et al., 1994). Meta-analysis of clinical trial evidence has shown that riluzole extends median survival by approximately 3 months (on average in a trial with a 12 month end-point), although reviews considering data from clinic databases and population registers suggest that the survival benefit is more than this (Andrews et al., 2020; Hinchcliffe and Smith, 2017; R. G. Miller, Mitchell, & Moore, 2012). Post-hoc analysis of the original riluzole trial shows that most of the survival benefit can be attributed to extending a later clinical stage of ALS, which may be less desirable for patients as this is a stage of the disease where there is more disability (T. Fang et al., 2018); this study did not have people in Stage 1, and subsequent study of other data showed that there is also a benefit in Stage 1 disease (de Jongh, van Eijk, & van den Berg, 2019).

Edaravone, is a neuroprotective drug that may reduce oxidative stress. Although an initial trial did not find a positive effect of drug treatment, post-hoc analysis identified a group of people with rapid

disease progression who did respond (Yoshida et al., 2006). Another phase 3 trial was undertaken to include people who matched the clinical phenotype of responders and Edaravone was shown to reduce speed of decline of ALSFRS-R score in these patients (Abe et al., 2017). A replication trial in Italy did not show a positive effect and so far the drug is only licenced in Japan, the USA and Canada (Lunetta et al., 2020).

Antisense oligonucleotides are an emerging molecular therapy that target and block the RNA of genes that cause disease, reducing their translation into damaging protein products (Klim, Vance, & Scotter, 2019). In human clinical trials for antisense oligonucleotide therapies are being conducted, and include the targets *SOD1, C9ORF72, ATXN2* and *FUS* (Amado and Davidson, 2021). The *SOD1* antisense oligonucleotide therapy, torfersen, has been tested in phase II trials, and has shown a bigger effect in subgroups of people with faster progressing *SOD1* ALS, although the trials were not designed to test efficacy, and phase III trials are ongoing. (T. Miller et al., 2020). Clinical trials to find effective treatments for ALS are ongoing alongside efforts to develop solutions to common trial design problems (Kiernan et al., 2020). Large platform trials, where multiple compounds are tested under one master protocol can improve efficiency by streamlining movement from phase two to phase 3 trials and reducing the proportion of people recruited to placebo arms (Hirakawa, Asano, Sato, & Teramukai, 2018; Saville and Berry, 2016). Patient stratification by clinical or genetic characteristics may increase power through reduction of heterogeneity, with the trade-off of a loss of power of ALS through smaller numbers and increased time to recruit eligible patients. Getting the balance of stratification right is key as stringent stratification can lead to a lack of generalisability, a criticism of the Edaravone phase 3 trial; but lack of stratification may lead to missing subgroups of people who benefit from a therapy, for example the post-hoc finding that people with *UNC13A* variants benefit from lithium treatment while the general ALS population do not (Hardiman and van den Berg, 2017; van Eijk et al., 2017). Reform of patient eligibility criteria from lists of univariate criteria designed to reflect likely progression to multivariate models of predicted disease course may also help increase power (van Eijk et al., 2019). Reducing measurement error in trial outcomes through harmonising reporting standards and training on progression measures like the ALSFRS-R and clinical staging is ongoing work of the TRICALS project.

Alternative and off-label treatments are commonly marketed to people with ALS, many being recommended to people with ALS based on anecdotal evidence. ALSUntanged is an initiative where these treatments are reviewed and rated by groups of experts to inform the ALS community of likely effectiveness (Bedlack and Hardiman, 2009).

1.1.4. *Clinical subgroups of ALS*

Heterogeneity of aetiology, presentation and progression necessitates the identification of clinically meaningful subgroups of ALS. Many different phenotype descriptions are used by ALS clinicians and researchers, which can make pooling data challenging (Al-Chalabi et al., 2016).

Subgroups may be defined based on individual clinical features obvious to physicians at presentation such as site of onset, this is meaningful as it has care and prognostic implications and is easy to identify but it does not fully account for variation. Subgroups have also been defined using statistical learning techniques such as latent class clustering analysis but these groups have not been validated with external datasets (Ganesalingam et al., 2009).

More broadly, neurodegenerative diseases involve death or loss of function of a particular type of neuron alongside predominant pathological hallmarks, such as intracellular inclusions and protein aggregates. It may be appropriate to group diseases with distinct clinical characteristics but shared neuropathology, for example TDP-43 proteinopathy occurs in most cases of ALS and some cases of frontotemporal dementia in addition to there being shared genetic risk factors (Neumann et al., 2006; Scotter et al., 2015). Some disease-causing genetic variants can cause different neurodegenerative conditions as in the case of *C9orf72*, which has been associated with frontotemporal dementia and schizophrenia as well as ALS (McLaughlin et al., 2017). Basket studies, such as those in cancer research, that group a disease by common biology rather than clinical features may be appropriate in ALS, although some work would be needed for find appropriate outcomes in the trials - the ALSFRS-R as an outcome measure would not be appropriate for someone with TDP-43 FTD but no motor symptoms (van Es, Goedee, Westeneng, Nijboer, & van den Berg, 2020).

1.2 Epidemiology of Amyotrophic Lateral Sclerosis

Global average estimated ALS incidence is 1-2 per 100,000 person-years, and the prevalence is approximately 5 per 100,000 persons (Chiò et al., 2013; Logroscino et al., 2018; Marin et al., 2017; Xu et al., 2020). When incidence and prevalence are estimated at sub-continent levels the variation is more like 1-3 per 100,000 person years incidence and prevalence of 2-10 per 100,000 person years, the highest incidence being recorded in Western Europe and the lowest in South Asia (Xu et al., 2020). Although ALS incidence is low, the lifetime risk of ALS is approximately 1 in 350 in men and 1 in 340 in women, with the risk increasing steadily, until it levels off at about 80 (Clare A Johnston et al., 2006; Ryan, Heverin, McLaughlin, & Hardiman, 2019).

Peak age of incidence and proportion of people with different sites of focal onset vary between countries (Marin et al., 2018). In analysis based on data from clinic registers, ALS is diagnosed more

frequently in men than women, with a male: female ratio of about 3:2, the difference reducing as age increases. In population registers, although the proportion of men is still higher, the ratio may be closer to 1:1, possibly due to the greater capture of older people with ALS (Chiò et al., 2013). Additionally, the proportion of people with different sites of onset varies between countries in which this is studied. Differences in the presentation and progression of ALS between countries probably reflects an interplay between genetics, healthcare system, behavioural differences in uptake of interventions and possibly environmental factors (Chiò et al., 2010; Zou et al., 2017).

Clinical and demographic data about people with ALS comes from clinic databases at specialist centres, population registers and other research studies, including clinical trials. Population-based registers collect data on all cases in a defined geographical area and provide unbiased insights into the epidemiology and aetiology of ALS (Hardiman et al., 2017; Logroscino et al., 2008). Clinic databases and data from research studies provide the most detailed level of diagnostic and progression information but on a biased subset of the population. Population registers will suffer ascertainment bias in very elderly people with multiple comorbidities, because the condition may go undetected for example, symptoms such as weakness may be considered a normal consequence of ageing, or people may be too ill to attend a hospital-based clinic for a formal diagnosis. Inclusion of hospices as data collection sites may ameliorate the second problem if consultant neurologists are able to run clinics in hospices, but the first will be more challenging to address. The MND Register dataset includes data collection on comorbidities, and detailed analysis of these comorbidities may elucidate disease networks, so it may be possible to advise checking for symptoms of ALS in people with other diseases that co-occur with ALS, for example FTD.

Healthcare system and geography informs the design of population-based registers. In some countries there are single centres that see most people with ALS, in these cases a single clinic database can also act as the population register, with some supplementary data capture from other care providers. In other places, where there is well organised, centralised data sources national healthcare statistics can be used. Insurance data may be a good source of healthcare information, web-based systems such as self-registration websites or centralised care planning databases have also been designed in some areas. In the UK there are 22 specialist ALS centres. In addition, people are diagnosed in general neurology clinics and treated by clinical nurse specialists and in palliative care settings such as hospices. All sites should be included to obtain the most complete case ascertainment. There are some population registers already in existence and care planning and recording of patient data varies locally. An efficient approach in the UK is to allow data collection to be organised locally and collecting and cleaning the data centrally.

The aim of the MND Register for England, Wales and Northern Ireland is to act as a central source of data for all people with ALS in the area defined. Although there are other motor neuron diseases, the aim of the database was determined by funding constraints and because of an unmet need for an ALS population register.

1.2.1. Genetic risk factors for Amyotrophic Lateral Sclerosis

Approximately five to 20 percent of people with ALS have a positive family history of ALS, according to prospective, population-based studies (Byrne et al., 2011; Ryan et al., 2018). What constitutes a positive family history in terms of which relatives and diseases to include is not standardised (Al-Chalabi, 2017; Byrne, Elamin, Bede, & Hardiman, 2012). All genetic variants that are causal for ALS that have been identified in people with a family history of ALS have been found in sporadic cohorts and clinically and pathologically there is no difference between people with a family history and those without. Additionally, although age of onset is lower in familial ALS compared with apparently sporadic ALS, this is not the case when compared with those people with apparently sporadic ALS but harbouring a variant in an ALS risk gene of Mendelian inheritance (Mehta et al., 2019). In other words, the presence of a Mendelian gene is driving the lower age of onset, not the presence of a family history. Familial ALS, which may be thought of as a subtype of ALS describes those people with a high liability for ALS, likely due to a genetic cause. Due to small family sizes and incomplete penetrance this category will not be sensitive enough include everyone with a high genetic liability, and disagreements about eligibility of positive family history lead to measurement error, so it is flawed. Despite its flaws, it can be used meaningfully in genetic risk factor research and may be useful to identify people in clinic who are likely to benefit from genetic testing before all people are able to be genotyped (Kenna et al., 2016).

Through genetic linkage analysis, *SOD1* was the first gene in which variants were found to be associated with ALS, in 1993 (Rosen et al., 1993). Since then, about 150 genes have been recorded as associated with ALS; approximately 20 are considered to have enough evidence to be classified as having variants that can cause an ALS phenotype, with many more having their association replicated in independent studies (Brown and Al-Chalabi, 2017) (www.alsod.ac.uk).

Estimates of heritability, the amount of phenotypic variance attributable to genetic variation, vary by methodology. Twin-based studies have the highest estimation of approximately 60%, concordance rates amongst parent-offspring pairs estimate heritability at 50% and common SNP-based heritability is estimated at 10-20% (Al-Chalabi et al., 2010; McLaughlin, Vajda, & Hardiman, 2015; Wingo, Cutler, Yarab, Kelly, & Glass, 2011). The difference between estimates, or 'missing heritability' may be explained by rare, private mutations that are not sampled in traditional GWAS

studies, or by the presence of structural variants (McLaughlin et al., 2015). Another possibility is the presence of somatic mutations, not generally detected by peripheral blood based genetic testing. Large whole genome sequencing association studies of people with ALS show that it is likely private mutations with medium effects that cause ALS in a lot of cases (van Rheenen et al., 2016).

The genetic architecture of ALS seems to comprise a mixture of simple Mendelian inheritance, oligogenic inheritance and polygenic inheritance patterns seen in common complex diseases (Veldink, 2017). Genograms of people with a family history of ALS show Mendelian inheritance patterns of disease, and the most common, large effect genes such as *SOD1*, *TARDBP*, *FUS* and *C9orf72* have Mendelian inheritance. Oligogenic inheritance, where there are variants in more than one disease causing gene, is seen in 0.4-1% of people. Estimates from GWAS summary statistics show polygenic risk accounts for approximately 5% of heritability (McLaughlin et al., 2017; van Rheenen et al., 2016). Sporadic cases with a genetic background are found in ALS as would be expected by a polygenic model of disease risk (Yang, Visscher, & Wray, 2010).

Another source of genetic variability may be found in somatic mutations accumulating in cells during replication (D'Gama and Walsh, 2018). A study of C9orf72 ALS post-mortem tissue samples does not support this finding, although there could be attrition due to cell death (Ross et al., 2019).

1.2.2. Environmental risk factors for ALS

Heritability estimates of less than 1 (or 100%), indicate the possibility that the remainder of the variability is due to sampling or measurement error, random noise, or environmental factors. Environmental risk factors in ALS have proven difficult to identify, in part because of problems common to all environmental studies such as an infinite exposome but also ALS-specific problems of disease heterogeneity and long disease lead in time (Al-Chalabi and Hardiman, 2013). For some risk factors that are being tested, such as smoking, a finding of a positive association would not change public health advice but confirming such a causal link would be invaluable in elucidating disease mechanism and informing drug development.

Smoking may be associated with increased risk of ALS, although the data are not conclusive. Meta-analysis of case-control and cohort studies show a slight increased risk of smoking, with a stronger effect shown in cohort studies (Alonso, Logroscino, & Hernán, 2010a). Since then there have been a mix of positive and negative studies, and there may be a higher risk in some subgroups (Alonso, Logroscino, Jick, & Hernán, 2010b). Individual studies have shown dose dependent effects by age of smoking initiation and time since smoking cessation reducing risk (Peters et al., 2019; H. Wang et al., 2011). Mendelian randomisation studies to investigate causality have reported positive and negative associations (Bandres‐Ciga et al., 2019; Zhan and Fang, 2019).

There are several high-profile sports people who have been diagnosed with ALS. Additionally, many people with ALS have higher levels of voluntary sports participation and a low BMI on presentation (Scarmeas, Shih, Stern, Ottman, & Rowland, 2002). It is not clear whether participating in high levels of physical activity raises the risk of ALS and, if it does, whether it is directly a consequence of the exercise or some form of shared genetical predisposal between sporting prowess and ALS (Lacorte et al., 2016). Additionally, low BMI may be confounded with the pathological processes of disease. Trauma, including head injury which may occur as a result of sporting activity, also appears to be a risk factor according to meta-analysis (M. D. Wang, Little, Gomes, Cashman, & Krewski, 2017).

It is not clear whether there are occupational exposures that cause ALS, and there is mixed evidence for associations between occupational exposures such as pesticides, heavy metals, solvents, electromagnetic fields, cyanotoxins, electric shock and diesel exhaust fumes (Abhinav, Al-Chalabi, Hortobagyi, & Leigh, 2007a; Bozzoni et al., 2016; Delzor et al., 2014; Dickerson et al., 2018; Fischer et al., 2015; Koeman et al., 2017; Malek et al., 2015; Pamphlett and Rikard-Bell, 2013; Rooney et al., 2016; Sutedja et al., 2009).

Specific occupations that have replicated associated with ALS are military service with deployment and football (Beard and Kamel, 2015; Blecher et al., 2019; Chiò et al., 2009a; Pupillo et al., 2020; Tai et al., 2017). Both occupations represent a wide range of exposures including toxins, physical activity and psychological stress and it is not clear what the causative factor would be.

1.2.3. Observational study biases

Many environmental risk factors that are the subject of analysis such as smoking, head injuries, electrical shock and chemical exposure are investigated because they are known to cause other diseases. It is therefore unethical to investigate the effect of these kinds of exposures on an outcome using a randomised controlled trial to experimentally demonstrate causality. In these cases, observational research is undertaken where the proportions of people who have encountered an exposure with and without developing the disease are compared. While this is ethically necessary, there are several methodological pitfalls that make determining causality more difficult.

Observational studies have the potential to suffer from unmeasured confounder bias due to the unrandomized nature of the design. Unmeasured factors that are associated with the factors under investigation in the study may be driving the association, rather than the measured factor (Greenland and Neutra, 1980). Multivariate analysis, where outcomes can be adjusted for the presence of confounding is common practice to mitigate some of the effects of confounders (Skelly, Dettori, & Brodt, 2012). Adjusting for all possible confounders is almost impossible, and confounders

could be important at stages of life much earlier than when the disease develops (Lawlor et al., 2005).

If an exposure and an outcome both affect how likely someone is to be sampled for a study this can lead to spurious causal associations driven by collider bias (Cole et al., 2010). For example, moderate exercise may slow disease course in people with ALS (McCrate and Kaspar, 2008). Additionally, people who are enrolled in clinical studies tend to have a slower disease course to have enough time to be included in the study. Therefore, sample selection bias could produce a correlation between physical activity and ALS where there isn't one in the ALS population more generally – so population level data are preferred. A recent, large population-based study using data from the Netherlands, Ireland and Italy found a small but significant increased risk of ALS with different levels of physical activity, although the project was population based, the percentage of responders, and their clinical phenotype compared to non-responders, was not reported (Visser et al., 2018).

Retrospective observational studies that reply on questionnaire data may suffer from self-reporting biases (Althubaiti, 2016). One such bias is recall bias, which causes people to over or underestimate an exposure (Neugebauer and Ng, 1990). Another bias is social desirability bias – people may over-estimate their participation in socially acceptable or desirable behaviours, such as healthy eating and underestimate undesirable behaviours such as drinking alcohol. Despite this, the retrospective design is generally favoured in ALS due to the rarity and older age prevalence of the disease meaning prospective cohort studies would suffer from problems of scale (Al-Chalabi and Hardiman, 2013). Prospective studies investigating other diseases, or studies such as the UK Biobank can be used to look for associations and this will help attenuate recall bias.

Measurement error, although not confined to observational studies, can affect both independent and dependent variables and lead to regression dilution. If there is measurement error in the independent variable this will bias the effect size to zero; if the error is in the dependent variable, the effect size will not be biased, but the test is less likely to show a strong correlation between the variables (K. Liu, 1988). Measurement error may derive from situations such as people interpreting the same questions differently in a survey, variables that measure more than one exposure, human error in recording, and the variable not accurately measuring the exposure (Coggon, Rose, & Barker, 2003). ALS as an outcome could suffer measurement bias through misdiagnosis (O'Reilly, Brazis, & Rubino, 1982; Traynor et al., 2000).

Replication of studies is essential to strengthen findings, particularly as gene-environment effects may mean in some populations effects between exposure and risk are not seen. However, as bias can be replicated, later studies must also address different sources of bias and not be simple

replication of previous methods. This kind of replication, to vary the source of bias has been termed triangulation or consistency (Lawlor, Tilling, & Davey Smith, 2017).

1.2.4. Phenotype modifiers of Amyotrophic Lateral Sclerosis

As well as predisposing people to increased risk of ALS, genetic, demographic, and environmental factors can affect disease course. The factors that affect prognosis are not necessarily the same as those that are associated with increased risk (Chiò et al., 2009b). While therapeutics are confined to treating the disease rather than primary prevention, targeting factors that are phenotype modifiers will have the most effect on disease course (Shatunov and Al-Chalabi, 2020).

Age and sex strongly affect phenotype characteristics such as site of onset of symptoms, proportion of upper and lower motor neuron involvement and cognitive impairment, with genetic variation having a smaller effect (Chiò et al., 2020).

Smoking at time of diagnosis has also been found to affect progression rate in ALS, and this has some dose dependency in that people who used to smoke have a slightly faster progression than those who have never smoked (Calvo et al., 2016).

1.2.5. Models of ALS aetiology

The multistep hypothesis of ALS is a model that conceptualises the pathological development of ALS as a sequence of steps (Al-Chalabi et al., 2014). Consistent with this hypothesis, if the log incidence of ALS is plotted against the log age of onset the resulting slope is linear, and the regression coefficient can be used to quantify the number of steps needed for disease development, with the number of steps being regression coefficient (b)+1. Analysis of data from Australian, Japanese and South Korean registers have replicated the initial finding from multiple European registers that on average, people with ALS will take 5-6 steps to reach ALS (Vucic et al., 2020; Vucic et al., 2019). Analysis of incidence in people with ALS that have variants in ALS causing genes show that in people with *SOD1* variants on average 2 steps are needed to develop disease, *C9orf72* it is 3 steps and *TARDBP* it is 4 steps (Chiò et al., 2018). A study comparing risk of prior diagnosis of schizophrenia and prior diagnosis of cardiovascular disease found that after controlling for prior death as a competing risk, people with prior cardiovascular disease have 3 steps needed to develop ALS (Garton, Trabjerg, Wray, & Agerbo, 2020). Men also have on average, half a step less than women to develop ALS. Each step is not thought to be exposure to an individual risk factor, but likely represents a biological process, possibly happening at a cellular level. Population-level data is required to test whether incidence as a function of age is mathematically consistent with a multistep process. To find subgroups with different numbers of steps required for disease onset, detailed genotype and phenotype information needs to be available on all people within a population. People

develop ALS after encountering fewer additional exposures, such as those with *SOD1*-mediated ALS may provide subgroups with a smaller exposome to investigate potential environmental triggers of ALS. Establishing a population register in for ALS in the UK will lay the foundation for data collection needed for ALS subgroup analysis.

The multistep model is also consistent with the liability-threshold model of disease risk (Al-Chalabi and Hardiman, 2013). Liability is defined as the burden of exposures, which may be a mixture of genetic or environmental factors, and everyone will have a different liability, assumed to be normally distributed, and if an individual's liability is above the disease threshold then disease develops (Falconer, 1996; Read, 2018).If liability is based on inherited genetic risk then it is determined at birth and does not change, however the multistep model conceptualises it as being time-associated so differs from the liability threshold model in this way. . Features of ALS, such as lower age of onset in people with a family history or Mendelian variant of ALS, support this model because people with genetic risk variants of ALS have a higher liability (Mehta et al., 2019).

1.3 Conclusions

ALS is likely to be the result of complex interactions between genetic and environmental risk factors, which can affect both risk of developing the disease as well as the clinical course once someone is affected. There have been significant advances in our understanding of ALS through the discovery of genetic risk factors. The optimal phenotype subgroups of ALS and whether environmental risk factors have a role to play in disease risk, is not yet fully clear.

Disease risk and phenotype modifiers can only be understood by analysing detailed genetic, environmental and clinical information on large numbers of people, using multiple methodologies. Data collection efforts and technologies are increasingly able to provide this level of detail on people with ALS.

Chapter 2 Summary of thesis objectives

It is not known how many people in England, Wales and Northern Ireland are diagnosed with ALS every year, all estimates are extrapolated from local population registers. In this thesis I report the set-up of the MND Register for England, Wales and Northern Ireland (the MND Register), a project that aims to collect clinical information on everyone with ALS in England, Wales and Northern Ireland. I also report the first results of updated incidence estimate for the UK based on an area of England where there has not been data reported previously, collected as part of the MND Register. In this project we test the hypothesis that incidence in this area of England is like estimates previously reported in the UK and in other European databases.

To test the hypothesis that people who smoke are at risk increased of ALS, I have analysed a dataset of questionnaires, collected from centres in England. I used a novel to ALS methodology to quantify a continuous measure of smoking and investigate dose-dependency. I followed this up with a Mendelian Randomisation study to provide corroboration with our findings using a different epidemiological method to assess causality of exposure.

Finally, I present a global study of clinical phenotype in people with ALS who have tested positive for mutations in the *SOD1* gene. The dataset is the largest dataset of people with *SOD1* ALS and the genotype-phenotype correlations identified show there may be subtypes of *SOD1* ALS defined by clinical variant.

Chapter 3 Methods

This section expands on the main statistical methods used in chapters 4-6, particularly where journal space restricted provision of more detailed explanation. The methods are addressed in order of appearance in each chapter.

3.1 Incidence

Incidence measures the rate of novel cases of a disease that occurs in a defined population of disease-free individuals in a specific timeframe. It can be conceptualised using equation 1 (Critchley, 2004).

$$Incidence\ rate = \frac{Number\ of\ new\ cases\ of\ disease}{Population\ at\ risk}\ in\ a\ period\ of\ time \quad (1)$$

ALS is a relatively rare event and large populations must be observed over a period of several years to estimate incidence. When incidence is reported in ALS, the 'period of time' is usually a yearly rate and the population at risk is typically measured in summed person-years of observation using the population size from a census. When population estimates are used, there will be a few people not at risk because they have the disease. As ALS is a rare disease, this is unlikely to affect the estimate greatly.

Crude incidence rates are the proportion of people that develop a disease, divided by the overall population, usually multiplied by 100,000 for readability. When using census data, if the period of observation is two years, the number of person-years is the number of people in one year multiplied by two. As ALS risk varies by age and sex, age and sex-specific rates are calculated by dividing the population into age groups by sex and separately calculating incidence rates. To compare incidence rates between different geographical populations the crude rate must be standardised to a reference population. This is because even if the risk of disease is the same in different age groups, if there are different age structures in two populations, the incidence will be different. The age-standardized rate can be a theoretical rate that would occur if the rates occur in a fabricated population structure (for example the European standard population) or could be taken from census data from a particular country.

The standard error for the direct standardisation method can be calculated using the binomial or Poisson approximations – they are similar and either can be used (Bowden et al., 2016b; Keyfitz, 1966; Ulm, 1990). From the standard error, the variance and confidence intervals can be calculated using the appropriate standardised normal deviate.

3.2 Multiple imputation

Imputation is a method to estimate data where it is missing using information gleaned from other, complete cases in the dataset (Rubin, 1987). Multiple imputation describes when this process is repeated to generate multiple possible datasets by introduction of an element of randomness in generating dummy values in place of missing values. Multiple imputation is preferred to mean or median value imputation, because mean and median imputation decrease the variance in the dataset.

In the MND Register dataset, there were many cases of missingness of date of diagnosis. Although date of onset of symptoms could arguably be used instead, due to lag times between onset and diagnosis, it is possible we would miss new cases. Instead, we decided to impute diagnostic delay from the other values in the sample and add that to date of onset to estimate date of diagnosis. This was repeated 20 times, as there were 20% of date of diagnosis values missing (Bodner, 2008; White, Royston, & Wood, 2011). Predictive mean matching was used to estimate diagnostic delay values. Predictive mean matching is where a regression is used to estimate values for a particular record and then records with full data that closely match the predicted value are used to pick an imputed value from (White et al., 2011). Pooled parameter and standard error estimates are calculated using Rubin's rules (Rubin, 1987). To calculate the pooled parameter estimate, the mean of all imputed incidence calculations were taken (Enders, 2010). To calculate the pooled standard error, the within imputation variance (squared standard error of each dataset divided by the number of datasets) and the between imputation variance (variance of each imputed incidence calculation) are combined (Enders, 2010). The variance derived from these estimates are not directly comparable to historical estimates but confidence intervals from unimputed data were provided as well as pooled incidence estimates to give an indication of data spread. Multiple imputation was performed in R, using the 'mice' package and incidence estimates were calculated in MS Excel (Groothuis-Oudshoorn, 2011; Team, 2018).

3.3 Measures of smoking exposure

Smoking is a multidimensional exposure that may be measured in a variety of ways, using both categorical and continuous variables. Categorical variables include whether someone has ever smoked, or whether they are currently smoking and are relatively easy to collect and these measures are common in ALS smoking literature (Alonso et al., 2010a). Continuous variables of smoking exposure include, age started smoking, years of smoking (duration), cigarettes smoked per day (intensity) and time since cessation. Using these variables, it is possible to investigate dose dependency – which provides further evidence towards causality in observational research.

Dose dependency is commonly assessed using cigarette pack years, a model that assumes that risk increases linearly with increased smoking exposure and that intensity and duration are equally important to risk (Prignot, 1987). Additionally, time since cessation is absorbed into this measure but may be separately important. Including time since cessation as well as age of initiation as separate variables in a regression is inappropriate because of variable co-dependency. Cigarette pack years is calculated as described in equation 2:

$$Pack\ years = \frac{int}{20} \times dur \tag{2}$$

int = cigarettes per day, dur = duration of smoking calculated for current smokers as [current age – age started smoking] or for former smokers as [age stopped smoking – age started smoking].

The Comprehensive or Lifetime Smoking Index (I will refer to it as the lifetime smoking index) is a non-linear model of smoking exposure that takes all measures into account. To calculate a lifetime smoking value for an individual, values of half-life (τ) and lag time (δ) are calculated (Leffondré, Abrahamowicz, Xiao, & Siemiatycki, 2006). Half-life captures the exponentially decreasing effect of smoking on an outcome and lag-time accounts for the reverse causality seen with other diseases when quitting smoking (Leffondré, Abrahamowicz, Siemiatycki, & Rachet, 2002). Lifetime smoking index is calculated using equations 3-5:

$$tsc^* = \max(tsc - \delta, 0) \tag{3}$$

$$dur^* = \max(dur + tsc - \delta, 0) - tsc^* \tag{4}$$

$$lifetime\ smoking\ index\ value = \left(1 - 0.5^{dur^*/\tau}\right)\left(0.5^{tsc^*/\tau}\right)\ln(int + 1) \tag{5}$$

where τ = half-life, δ = lag time, tss = time started smoking (age of initiation), dur = duration of smoking (calculated as above), tsc = time since cessation, int = cigarettes per day.

Values of tau and delta can be set *a priori,* in this case values of δ and τ were simulated and those that fit the dataset best, determined using the Aikike Information Criteria, were used to calculate values of smoking exposure that can be included in a regression model (Akaike, 1998). Calculation of tau and delta was undertaken in R.

3.4 Logistic regression

Research investigating the relationship between an exposure (such as smoking) and a binary outcome (such as disease state) involves quantifying the chance of each outcome in the case of exposure to the factor under consideration. Calculating chance might involve working out 'risk', the chance of an outcome as a proportion of all outcomes, or 'odds' which refers to the chance of an

outcome occurring compared to the chance of the outcome not occurring. To calculate risk, the total number of people at risk is required however, to calculate an odds ratio two groups can be compared without knowing the overall prevalence to exposure, so it is suitable for case-control studies (Bland and Altman, 2000). Odds can be calculated using the equation 6 where *p(x)* refers to probability of an outcome occurring.

$$Odds = \frac{p(x)}{1 - p(x)} \tag{6}$$

The ratio of the odds is a univariate analysis, which compares the odds of an outcome in a group of people exposed to the factor under consideration compared to the event occurring in a group not exposed to the factor. Table 1 is a contingency table of a hypothetical scenario being investigated in a case-control study. Using the example in table 1, a hypothetical odds ratio can be calculated using equation 7.

Group	Exposed	Not exposed
Cases	a	b
Controls	c	d

Table 3-1: Example contingency table for calculating odds ratio

$$Odds\ ratio = \frac{(a/c)}{(b/d)} \tag{7}$$

Log transforming the odds of an outcome allows it to take any value between negative and positive infinity, so it can be used as an outcome in multiple regression. This means that two groups can be compared while controlling for confounding variables. Regression using the transformed odds, or 'logit' function as the outcome is called logistic regression (Sperandei, 2014).

$$\log(\frac{p(x)}{1 - p(x)}) = \beta_0 + \beta_1 X \tag{8}$$

Taking the inverse of the logit function gives the sigmoid function, which can be used to predict the probability of being in either group 1 or 2 given the values of the regression coefficients.

$$p(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \tag{9}$$

Maximum likelihood estimation is a machine learning method which finds the set of regression coefficients for which the likelihood of observing the data is maximum and it is typically used to

estimate regression coefficients. Power calculation for logistic regression can be performed using the power calculation method for a chi squared test.

3.5 Mendelian randomisation

A method of analysis that is designed to address confounder bias is instrumental variable analysis, a technique originating from the field of econometrics. An instrumental variable is one that is correlated with the variable of interest but not confounded with other variables that may bias analysis. For example, it was first used to investigate lifetime earnings in people who were eligible for military deployment, rather than those that were deployed, to mitigate the effect of confounding from socioeconomic status (Angrist, 1990). During the Vietnam War, people were drafted based on their date of birth, with dates of birth chosen randomly by lottery. This set up a natural experiment, 'randomising' people to military service by birth date, so reducing the effect of confounding. Using draft eligibility of 19-year olds in the period of 1950 to 1953 as an instrumental variable for deployment an unbiased effect of military deployment on lifetime earnings was calculated.

The concept of using genotype as an instrumental variable for epidemiological exposures such as smoking, has been developed into a method called Mendelian Randomisation (Davey Smith and Ebrahim, 2003; Davey Smith and Hemani, 2014). Two-sample Mendelian Randomisation compares the effect size of genetic variants from summary statistics of two genome-wide association studies, one looking for effect of genotype on the exposure, the other at the effect of genotype on the outcome. If the assumptions are met, an unbiased effect of an exposure on an outcome can be calculated by calculating the ratio of the effect size on likelihood of exposure and the effect size on likelihood of outcome. The instrument can be made up of a single variant, in which case the calculation is called the Wald ratio, in the case of multiple variants the ratios are meta-analysed using inverse variance weighted analysis.

More explanation about the assumptions of MR and the assumptions of different methods of analysis can be found in section 6.11.5 on page 68.

Mendelian randomisation was performed in R using the Two sample MR package (Hemani et al., 2018b).

3.6 Polygenic risk score analysis

If individual level genome-wide genetic data are available, it is possible to calculate the genetic liability to a phenotype using summary statistics from a genome-wide association study (Choi, Mak, & O'Reilly, 2020). The genetic liability to the phenotype takes is a single, composite score of the risk alleles for a phenotype, weighted by the effect size of the allele and is called a polygenic risk score

(PRS). Variants that meet a lower threshold than genome-wide significance are included in the composite score as they increase predictive power. We used data from the UK Biobank to calculate polygenic risk scores for smoking and used the score as a covariate in logistic regression to analyse whether it increased the risk of ALS in the UK Biobank sample.

3.7 Time-to-event analysis

Time-to-event data (also termed survival data) is the name for data that describes the time from a starting point, for example study entry or disease onset to an end point, for example disease recurrence or death (Altman, 1990). As the follow-up time will not cover everyone having experienced the event, other methods for analysis of continuous data are not suitable because they do not fully account for the uncertainty produced by censored events. Instead, data are usually visualised using Kaplan-Meier plots that make it easy to interpret differences in time-to-event between groups and obtain a rough estimate of median survival (Bland and Altman, 1998). Kaplan-Meier plots show a step function as the proportion changes only when there is an event of interest. The Kaplan-Meier curve makes the following assumptions, that losing people to follow-up is unrelated to their prognosis, that the event of interest, for example death or disease recurrence, happened at the time recorded, and that survival probabilities are unrelated to study recruitment time (Bland and Altman, 1998).

The log rank test is used to test the null hypothesis that there is no difference in the probability of an event happening between groups, at any given time (Altman, 1990). Proportions of people experiencing the event change over time (and will start and end as equal proportions) so the log rank test works by calculating the expected number of events every time an event happens. From the calculations at each event, the total observed and expected deaths are calculated by adding up the values at each time point and a chi squared test is used to compare them and generate a p-value. Median survival, or survival comparisons at a pre-specified time can be used to quantify the difference between groups.

Cox proportional hazards regression is used to quantify the overall difference in likelihood of an event between groups, and to control for confounder variables (Cox, 1972). The hazard rate is the likelihood that an event will happen at any time – and is the same as the incidence. The hazard ratio is the ratio of the hazard rate in one group compared to another group and is analogous to the odds ratio. As is the case with the odds ratio, taking a logarithm of the hazard ratio allows it to be modelled as a regression. This is demonstrated in equation 10 where h(t) represents the hazard rate, $h_0(t)$ represents the hazard rate of the baseline or comparator group, β_1 is the estimate of covariate effects derived from the Cox regression model, and *X* a value of a given covariate.

$$\frac{h(t)}{h_0(t)} = e^{\beta_1 X} \quad (10)$$

Cox proportional hazards regression models are referred to as semi-parametric (Cox, 1972). The baseline hazard function can take any shape and there are no restrictions put on it, however the covariates are assumed to act linearly on the outcome – known as the proportional hazards assumption. This can be tested visually using a log(-log(S(t))) v t plot (where S(t) is the survival function and represents the probability that an individual is still alive at time t) and checking the lines are parallel, as well as testing whether there is a significant relationship between the Schoenfield residuals and time, if there is a relationship then the proportional hazards assumption is violated.

Violated assumptions may be corrected by creating a variable with a time interaction term and re-running the Cox proportional hazards model. Alternatively, other models such as accelerated failure time models may be more appropriate for the dataset. The variable that is time dependent, may still be associated with survival, but the hazard ratio may not be a reliable indicator of increased risk at any time point.

Kaplan-Meier analysis, log rank tests and Cox proportional hazards regression were performed in R using 'ggplot2' and 'Survival' (Therneau, 2020; Wickham, 2016).

Chapter 4 Motor Neuron Disease Register for England, Wales and Northern Ireland – an analysis of incidence in England

Sarah Opie-Martin [a], Lynn Ossher [b], Andrea Bredin [a], Anna Kulka [a], Neil Pearce [c], Kevin Talbot [b], Ammar Al-Chalabi* [a]

[a] Maurice Wohl Clinical Neuroscience Institute, Department of Basic and Clinical Neuroscience, King's College London; [b] Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK; [c] Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London

**corresponding author: ammar.al-chalabi@kcl.ac.uk*

Joint senior authors: Professor Kevin Talbot and Professor Ammar Al-Chalabi

Word count: 3000

Tables: 3

Figures: 2

References: 45

Statement of contribution: AA-C, KT conceived the MND Register for England, Wales and Northern Ireland. SO-M created data collection storage tool and conducted all statistical analysis. AA-C, KT, NP, AB, AK, LO provided intellectual input for data interpretation of this manuscript. SO-M, LO, AB, AK organised data collection, ethical approval and set up of sites across England, Wales and Northern Ireland. AA-C and SO-M wrote the first draft of the manuscript. All authors reviewed and approved the final manuscript.

4.1 Abstract

Introduction

Amyotrophic lateral sclerosis (ALS) has a reported incidence of 1-2/100,000 person-years. It is estimated that there are 5000 people with ALS in the UK at any one time; however the true figure, and geographical distribution, is unknown. In this study we describe the establishment of a population register for England, Wales and Northern Ireland and report estimated incidence.

Methods

People with a diagnosis of ALS given by a consultant neurologist and whose postcode of residence is within England, Wales or Northern Ireland were eligible. The catchment area was based on six data contributors that had been participating since 2016. All centres included in this analysis were in England, and therefore Wales and Northern Ireland are not included in this report. Crude age- and sex-specific incidence rates were estimated using population census records for the relevant postcodes from Office of National Statistics census data. These rates were standardised to the UK population structure using direct standardisation.

Results

There were 232 people in the database with a date of diagnosis between 2017 and 2018, when missing data were imputed there were an estimated 287-301 people. The denominator population of the catchment area is 7,251,845 according to 2011 UK census data. Age- and sex- adjusted incidence for complete cases was 1.61/100,000 person-years (95% confidence interval 1.58, 1.63), and for imputed datasets was 2.072 /100,000 person-years (95% CI 2.072, 2.073)

Discussion

We found incidence in this previously unreported area of the UK to be similar to other published estimates. As the MND Register for England, Wales and Northern Ireland grows we will update incidence estimates and report on further analyses.

Keywords: epidemiology, incidence, population register

4.2 Introduction

Motor neuron disease, also known as amyotrophic lateral sclerosis (ALS), is an adult-onset neurodegenerative disease affecting upper and lower motor neurons. Estimated global incidence of ALS has been reported as 1-2/100,000 person years (Logroscino et al., 2018; Marin et al., 2017; Xu et al., 2020) . Due to geographical variation the range of estimated incidence by subcontinent is 1-3/100,000 person years, with the highest incidence rates reported in Western Europe (Marin et al., 2017; Xu et al., 2020). ALS causes progressive weakness and paralysis, with death from respiratory failure usually between 2 and 3 years after diagnosis, but clinical presentation and disease progression are highly variable (Westeneng et al., 2018). There is currently no cure for ALS although riluzole and, more recently in some countries, edaravone are approved drugs that modestly extend survival for some people (Abe et al., 2017; Bensimon et al., 1994). Since the initial discovery that mutations in the *SOD1* gene can cause ALS, there has been considerable progress in the identification of genetic risk factors (Bowling, Schulz, Brown, & Beal, 1993). Despite these advances, disease aetiology in the majority of cases is not understood. Heritability estimates are compatible with the possibility that non-genetic factors such as stochastic biological events in ageing, environmental exposures or lifestyle choices contribute to disease risk, but there is no consensus on what these factors are.

Population registers collect information about every person diagnosed with a given condition in a defined geographical area, providing a source of representative data that can be used by researchers and authorities responsible for healthcare funding and organisation (Chiò et al., 2013).

ALS is highly variable in its presentation and clinical course. Collection of population level data about the clinical features in ALS, including cognitive impairment, has led to a greater understanding of prognostic significance of phenotypic subgroups of ALS (Ganesalingam et al., 2009; Phukan et al., 2012). Data from several European population registers has been used to create an accurate disease progression model which has helped inform care planning and communication with patients about prognosis (Westeneng et al., 2018). Population register data was used to show that a multistep model of disease may be relevant to ALS aetiology, and to estimate the number of 'steps' likely to be involved (Balendra et al., 2014; Chiò et al., 2018; Vucic et al., 2019). Information from population registers has also been used to estimate the projected number of people with ALS in the future if current population demographic trends persist, as well as for modelling the potential effects of future disease modifying treatments (Arthur et al., 2016; Gowland et al., 2019).

Population-based datasets eliminate the inherent ascertainment bias of intervention and case control studies based on referral cohorts (Hardiman et al., 2017), providing unbiased estimates of

the effects of exposure to risk factors associated with ALS (de Jong et al., 2012). Comparisons of clinical characteristics of patients enrolled in drug trials and population-based data from the same recruiting area show large differences that might help explain lack of generalisability of results from intervention trials.

There are some trends in ALS that are consistently reported between countries, for example, most people present with symptoms in the limbs (Marin et al., 2016). However, the incidence, pattern of disease progression and phenotypic spectrum of ALS differ between countries as shown by studies quantifying peak age of onset of symptoms, proportions of people with different sites of presentation, and survival time (Marin et al., 2018; Marin et al., 2016). These differences are probably due to complex demographic and healthcare factors; therefore, it is important to collect information at a local level to inform patients and healthcare professionals. In the UK there are five regional population registers for ALS: the South East ALS (SEALS) Register, Peninsula Network, South Wales Register, Northern Ireland register, and MND Care in Scotland, as well as many long-standing databases that document the attendees of specialist ALS clinics (Abhinav et al., 2007b; Donaghy et al., 2010; Forbes, Colville, Parratt, & Swingler, 2007; Imam, Ball, Wright, Hanemann, & Zajicek, 2010; J. D. Mitchell, Gatrell, Al-Hamad, Davies, & Batterby, 1998). All have provided insight into the overall picture of ALS in the UK and have contributed to estimates of UK-wide incidence, prevalence, and lifetime risk. There are an estimated 5,000 people living with ALS in the UK at any one time, but whether this is the true figure and how people with ALS are geographically distributed is not known.

In this paper we describe the creation of the MND Register for England, Wales and Northern Ireland through the incorporation of local population registers, use of data collected routinely to organise ALS clinics, and involvement of people with ALS directly through a self-registration website. We also report initial findings on incidence for areas with complete case ascertainment.

4.3 Methods

4.3.1. Patient eligibility

Eligible individuals were defined as having been diagnosed with ALS, Primary Lateral Sclerosis (PLS), or Progressive Muscular Atrophy (PMA) by a consultant neurologist. Where motor neuron involvement appeared to be restricted to the upper or lower motor neurons (including flail limb variants), but time since diagnosis was less than 4 years, the diagnostic category was recorded as ‘upper motor neuron predominant ALS’ or ‘lower motor neuron predominant ALS’, with a free text box available for provision of more detail if needed. Site of onset of first focal weakness, El Escorial category, and co-existing dementia were considered as phenotypic modifiers and recorded as

separate variables (Al-Chalabi et al., 2016). People with cognitive impairment, including those with fronto-temporal dementia were eligible. People with Kennedy's disease were not included.

People with ALS provided informed consent for the inclusion of identifiable data in the register to their healthcare professional or via our website (details below). As an informed consent discussion may not always be appropriate during a clinic appointment, or progression may be so rapid as to preclude an approach for informed consent, an anonymised data capture protocol was devised.

4.3.2. Identifying data sources

It is estimated that 90% of people with ALS will visit a specialist ALS service as part of their pathway of care in the UK; many of these services are funded in part by the ALS charity the Motor Neurone Disease Association and are referred to as MND Care and Research Centres or Networks (*90% figure from internal report from Motor Neurone Disease Association*). Therefore, we specifically invited all of these services to contribute data. To ensure complete case ascertainment, we identified other services, including general neurology clinics, community services, clinical nurse specialists and hospices where people with ALS also receive care.

4.3.3. Catchment areas

Specialist centres generally oversee a defined geographic area of the country. We asked each site to identify the areas in which every incident case of ALS would be referred to them to map areas of complete ascertainment. This information was generally provided in the form of UK postcode districts (for example, SE22 or SE5), unitary authorities, or counties. Many areas were overlapping between centres, so cases were sometimes reported more than once.

4.3.4. Data collection and transfer

The project has been designed to avoid duplication of data collection efforts for health professionals and researchers. Where there was a local population register or long-standing clinical database already in use, the local dataset was aligned with the agreed Register dataset. Where there was no database in use we provided a Microsoft Access template with data export functionality. There are many pieces of information that are collected to facilitate routine care organisation and some of these, such as postcode, name, hospital identifier, and sex are also part of our dataset. The template database was designed to be compatible with use as part of routine care to avoid duplication of the data collection effort. A minimum dataset of name, date of birth, unique national health service identifier, date of diagnosis, diagnosis (subtype of ALS), date of first weakness, site of first weakness, sex, and postcode of residence was requested to ensure the ability to estimate incidence and identify duplicate records.

People with ALS in the UK will encounter a variety of services, including tertiary referral centres, general neurology clinics, general practitioners, clinical nurse specialists, and local therapy teams. ALS is a clinical diagnosis which needs to be monitored over time, and people often see more than one consultant neurologist to confirm the diagnosis. As a consequence, duplicate records could be generated for the same patient by different data contributors. We used pseudonymisation to differentiate duplicate records while maintaining confidentiality of participants.

4.3.5. Website for patient self-registration

A website was developed to allow people living with MND to self-register for inclusion in the database and to provide consent for access to their clinical information (https://mndregister.ac.uk) with the aim of increasing direct patient participation and case ascertainment. At registration, participants were asked to indicate the neurologist who provided their diagnosis or ongoing care, to facilitate confirmation of clinical details from the medical record.

4.3.6. Statistical analysis

For this analysis data were extracted from local databases and sent to the central database during September 2019, our final cut-off for data transfer was October 2019. The data included complete records from people who had provided consent, as well as de-identified records from individuals who could not be approached for informed consent.

We used disease diagnosis date to estimate incidence, focusing on the years 2017 and 2018 to include the most complete dataset based on available records.

Patients were grouped by age at diagnosis and sex in five-year age bands. We had an open-ended cohort for individuals over 85 years at the time of disease diagnosis, so everyone with an age of diagnosis of more than 85 years were analysed together. Crude age-and sex-specific incidence rates were estimated using age- and sex-specific 2011 population census records for the relevant postcodes from Office of National Statistics (ONS) census data, the estimates are reported in person-years, taking into account that data were extracted over two years (*LC1117EW – Sex by age*, 2011). These rates were standardised to the UK population structure as measured by the 2011 UK census, the US population structure as measured by the US 2010 census and the European standard population using the direct standardisation method. We received residential data from people who had not provided consent for transfer of identifiable data at postcode area level (e.g., SE) to ensure anonymity. Our denominator population was made up of postcode areas where we had 100% capture, which was a subset of our total catchment area (darker grey areas in Figure 1).

Confidence intervals for crude rates were estimated using the exact method for Poisson intervals (Ulm, 1990). Confidence intervals for overall age- and sex-adjusted incidence rates were estimated at the 95% level using an approximation of the standard error for a binomial proportion (Keyfitz, 1966).

4.3.7. Multiple imputation

To estimate date and age of diagnosis for those records with date of onset only we used predictive mean matching to generate 22 datasets (as 22% of cases were missing diagnostic delay) and calculated standardised incidence for all datasets (Bodner, 2008; White et al., 2011). The model for predictive mean matching included data collection centre, age of onset, gender, diagnosis subtype, site of onset and family history. We calculated pooled estimates of incidence and 95% confidence intervals using Rubin's rules (Rubin, 1987). Imputation datasets were generated using the R package 'mice'. (Groothuis-Oudshoorn, 2011; Team, 2018).

4.4 Results

We defined an area of complete data capture by combining the catchment areas of six data collection centres that had been participating continuously since 2016. The complete postcode area for the catchment zone, the denominator population, represents a population of over seven million people and is indicated by the darker grey areas in Figure 1.

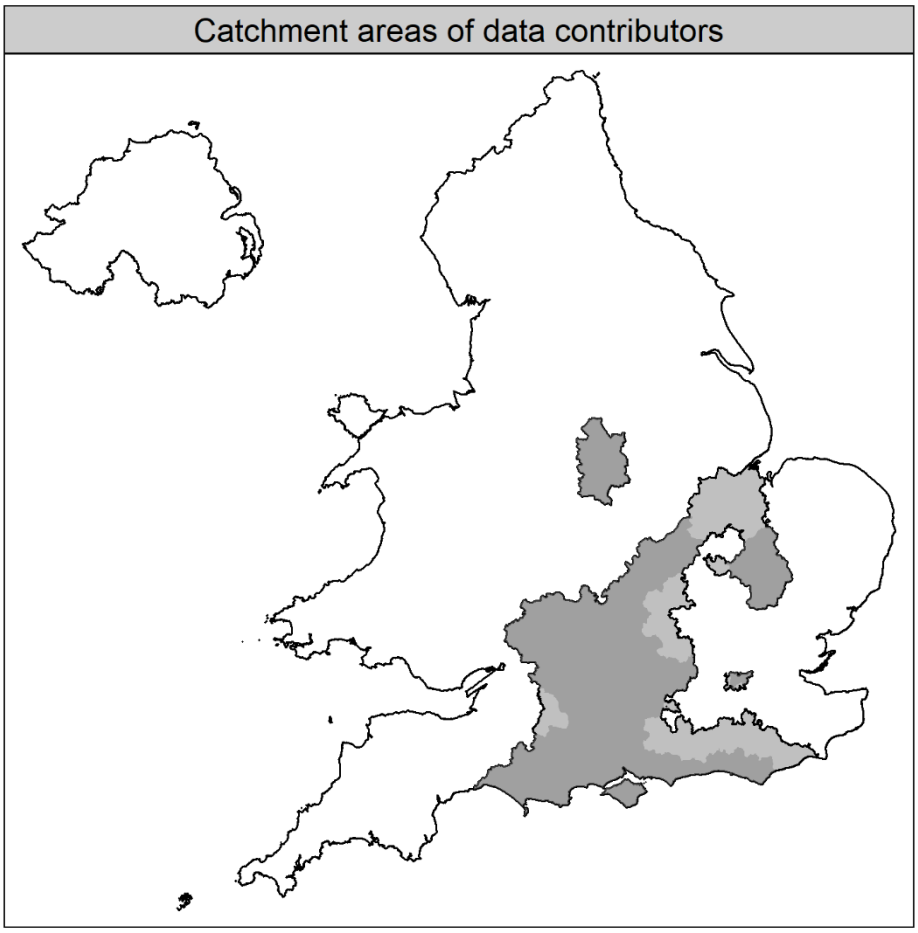


Figure 4-1 Map of catchment areas of clinics included in the incidence estimate

The map shows catchment areas of each clinic in light grey, with the whole postcode areas shaded in darker grey.

As of October 2019, there were 5066 records in the MND Register, through data transfers from 17 centres, including 426 people who had signed up online. We extracted data based on postcode area of residence at diagnosis. After data cleaning there were 1748 records, of these, 232 recorded a date of diagnosis between 2017 and 2018, referred to as the complete case dataset. 312 people had no date of diagnosis recorded, but all had a date of onset. We used imputed values of diagnostic delay to estimate date of diagnosis (figure 2).

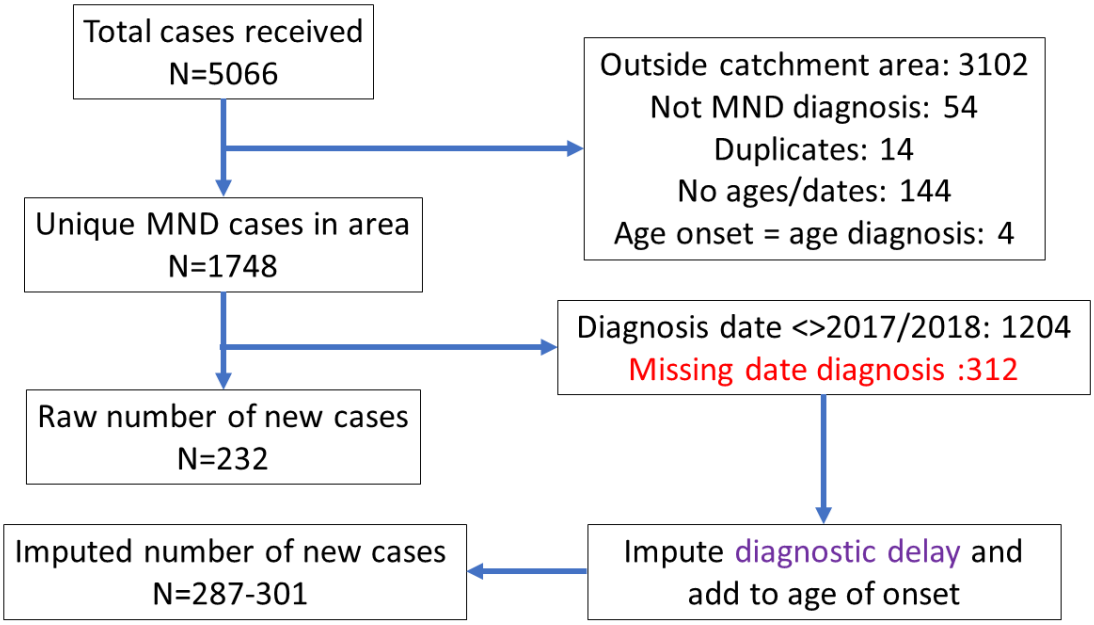


Figure 4-2 Modified consort diagram showing details of included cases

The case mix of the complete case dataset is shown in table 1.

Variable		Total = 232	Female	Male
Sex ratio F:M (n not recorded)		1:1.3 (2)	100	130
Diagnosis N (%)	Amyotrophic lateral sclerosis	106 (45.7)	91 (91)	107 (82)
	Lower motor neuron predominant ALS	8 (3.4)	2 (2)	6 (4.6)
	Upper motor neuron predominant ALS	8 (3.4)	2 (2)	6 (4.6)
	Primary lateral sclerosis	7 (3.01)	2 (2)	5 (3.9)
	Progressive muscular atrophy	3 (1.3)	3 (3)	0 (0)
Site of onset N (%)	Bulbar	67 (28.9)	42 (42)	25 (19.2)
	Spinal	118 (50.9)	41 (41)	76 (58.5)
	Respiratory	6 (2.6)	2 (2)	4 (3.01)
	Generalised	13 (5.6)	5 (5)	8 (6.2)
	Not recorded	28 (12.1)	10 (10)	17 (13.1)
Mean age of onset (SD)		64 (12) 27 records missing data	67 (12) 13 records missing data	62 (12) 13 records missing data

Table 4-1 Basic demographics and clinical features of people with ALS

Counts of people in the catchment area by age and sex (from the complete case dataset) are reported in table 2.

	Overall		Female		Male	
	n	Local population	n	Local population	n	Local population
16-44	16	3507543	5	1746724	21	1760819
45-49	8	664362	3	334775	5	329587
50-54	17	572399	7	286996	10	285403
55-59	32	498531	8	252548	23	245983
60-64	31	530802	13	271113	18	259689
65-69	31	423035	14	217876	17	205159
70-74	41	339173	16	177671	24	161502
75-79	25	281413	13	153323	12	128090
80+	31	434587	21	272241	10	162346

Table 4-2 Population counts for men and women in each age group

Ages 16-44 are shown as one category but were analysed in 5- year age bands (except 16-19 which was 4 years). Ages 80-84 85+ were also analysed separately but are displayed in aggregate. The local population numbers were multiplied by 2 to calculate person-years.

After multiple imputation, estimated numbers of people diagnosed between 2017 and 2018 ranged from 287-301. The pooled and complete case incidence estimates are presented in table 3. The estimated age and sex adjusted incidence for the UK is 1.61/100,000 (95% confidence interval 1.58, 1.63) based on complete case analysis and 2.07/100,000 (95% CI 2.072, 2.073) people based on the imputed dataset.

		Overall	Female	Male
Complete case analysis	England, Wales and Northern Ireland	1.61 (1.58, 1.63)	1.35 (1.31, 1.38)	1.85 (1.8, 1.9)
	European standard	1.76 (1.73, 1.78)	1.41 (1.37, 1.45)	2.07 (2, 2.13)
	US 2010 census	1.45 (1.43, 1.47)	1.21 (1.18, 1.24)	1.67 (1.63, 1.71)

Imputed	England, Wales and Northern Ireland	2.072 (2.072, 2.073)	1.775 (1.774, 1.777)	2.356 (2.353, 2.359)
	European standard	2.267 (2.266, 2.268)	1.874 (1.872, 1.876)	2.63 (2.626, 2.635)
	US 2010 census	1.874 (1.873, 1.874)	1.59 (1.589, 1.591)	2.133 (2.13, 2.135)

Table 4-3 Standardised incidence estimates

Incidence calculations are presented for men and women separately, standardised to different reference populations. Imputed rates are shown to three decimal places to reflect the accuracy needed to display the pooled 95% confidence intervals.

4.5 Discussion

We have established a population register that identifies records from multiple sources and uses data that are often available as part of routine data collection for care. The base population for our incidence calculation is over seven million people and therefore represents a large register compared to others globally. Once the MND Register includes data from all areas of the UK, which is the aim of the project, it will represent a database of a scale not yet reported. There are significant challenges in organising and maintaining a database for a population this size, including the co-ordination of data collection across independent hospital systems.

Organising a population register as a federated database can result in selection bias because of boundary effects. Through annexation of areas of complete ascertainment over time, it will provide a more complete, unbiased picture of the disease. Although the catchment area is constructed from the catchment areas of 6 centres, extracting data by postcode of residence instead of by attendance at a centre meant including data from 8 centres rather than 6, so some databases may be missing cases from their databases. 25 cases were transferred by the two extra centres.

UK geography is organised into many partially overlapping administrative units. Postcode is ubiquitously recorded but hospital catchment areas and population estimates are made up of county or unitary authority boundaries that are not always congruent with postcodes. This and the transfer of anonymised data that includes high-level postcode rather than the full postcode data has made estimating incidence challenging while there are few centres participating. This is expected to improve as the MND Register includes more data contributors over time. Our study is part of the UK Clinical Research Network, so other services not included in this mapping effort will potentially be notified of the project and be incentivised to participate. The MND Register team regularly attend symposia and local conferences and use social media in order to raise awareness about the project, including the self-registration website. There are regular campaign efforts from the MND Association including a spread about the Register in their quarterly magazine and videos to help people self-register.

The advantages of collecting clinical data from already existing databases is that it reduces burden on healthcare professionals who may have to collect similar data for a range of different reporting processes and care tasks and is relatively inexpensive. It is a system that is successfully used by other population registers in the UK. The disadvantages are that there is less control over the format of data collected and it cannot be easily modified to incorporate other data collection. Although centralised databases have worked successfully in smaller areas such as Scotland and Northern Ireland, the scale of NHS services in England and Wales mean this is unlikely to be possible at

present. Through establishment of contributing centres in many different locations throughout England, Wales and Northern Ireland we have encountered variation in care processes locally.

The use of patient reported data as an additional source of information is used by the National Amyotrophic Lateral Sclerosis Registry in the US and by the TREAT-NMD neuromuscular network (Antao and Horton, 2012; Bladen et al., 2014). As well as using a website, the US Registry collects clinical data from administrative datasets, a method we would like to use to supplement our data collection in the future. Detailed analysis of case ascertainment of the US Registry shows variation by race and insurance use (Kaye, Wagner, Wu, & Mehta, 2018). Although everyone is eligible to use NHS services in the UK we may not be counting privately treated patients who prefer not to register online, although this is expected to be a small number of people as NHS services provide high-quality multidisciplinary care.

Using this new register, we have estimated the incidence of ALS for previously unreported areas of England. We estimate that age- and sex-adjusted UK incidence is 1.61/100,000 person-years and 2.07/100,000 person-years using imputed data. The comparison of rates will focus on the imputed incidence because it is less likely to be an under-estimate. . The imputed estimated incidence is slightly lower than what has been reported in some smaller population registers in the UK, for example the rate of 2.52/100,000 in Devon and Cornwall, 2.1/100,000 in the South East ALS Register, and higher than 1.76/100,000 previously reported in Lancashire (Gowland et al., 2019; Imam et al., 2010; J. D. Mitchell et al., 1998).

Incidence was reported as 1.4/100,000 person-years in Northern Ireland, standardised to the European standard population (Donaghy et al., 2010). Our imputed estimate standardised to the European standard population is higher than this at 2.26/100,000 person years. In the most recent report from the Scottish register the incidence rate standardised to the US 2010 census population was reported as 3.83/100,000 person years, our imputed estimated standardised to the same population is lower at 1.87/100,000 person-years, but is comparable to the 1.89/100,000 person years reported for Northern Europe in 2017, also standardised to the US 2010 census population (Leighton et al., 2019; Marin et al., 2017).

The EURALS consortium reported average crude incidence rate of 2.16/100,000 person-years, the imputed crude incidence rate is similar to this, being 2.06/100,000 person-years (Logroscino et al., 2010). Our crude imputed incidence rate is between the 2.40 and 1.49/100,000 person-years reported for Northern and Southern Europe in a recent global incidence study (Xu et al., 2020).

Our estimates are based on data from areas that have not been sampled before, so the results may reflect true lower incidence in these parts of the UK. It is also possible that there are areas of low case ascertainment in our sample. The MND Register as a federated database is relatively new, and the collection of data was initiated at different times by individual participating sites. Detailed reporting by population register in the Republic of Ireland and Scotland have shown that data quality and ascertainment improves over time (Leighton et al., 2019; Rooney et al., 2017; Rooney et al., 2013). As more centres contribute data, we will be able to perform capture-recapture analysis of overlapping areas allowing more accurate incidence estimates.

In the future we will estimate prevalence and lifetime risk, as well as mapping incidence compared to healthcare provision. Collecting large, national datasets has helped improve care and understanding of other diseases and we have laid the groundwork and generated the momentum to do this for ALS as well.

4.6 Acknowledgements

We are grateful to the Motor Neurone Disease Association, Betty Messenger Foundation and a family trust that wishes to remain anonymous for funding this study. This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. This is an EU Joint Programme - Neurodegenerative Disease Research (JPND) project. The project is supported through the following funding organisations under the aegis of JPND - www.jpnd.eu *(United Kingdom, Medical Research Council* (MR/L501529/1 (STRENGTH); MR/R024804/1 (BRAIN-MEND)) *and Economic and Social Research Council* (ES/L008238/1; ALS-CarE)*).*

Chapter 5 UK Case control study of smoking and risk of Amyotrophic Lateral Sclerosis

S. Opie-Martin[a]*, A. Jones[a], A. Iacoangeli[a], A. Al-Khleifat[a], M. Oumar[a], P. J. Shaw[b], C. E. Shaw[a], K. E. Morrison[c], R. E. Wootton[d,f,g], G. Davey-Smith[d], N. Pearce[e], A. Al-Chalabi[f]

[a] Maurice Wohl Clinical Neuroscience Institute, King's College London, Department of Basic and Clinical Neuroscience, London, UK; [b] Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK; [c] Faculty of Medicine, University of Southampton, UK; [d] MRC Integrative Epidemiology Unit at the University of Bristol, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, UK; [e] London School of Hygiene and Tropical Medicine, London,

UK; [f] School of Psychological Science, University of Bristol, 12a Priory Road, Bristol, BS8 1TU, UK; [g] NIHR Bristol Biomedical Research Centre, University Hospitals Bristol NHS Foundation Trust and University of Bristol, Bristol, UK

* Corresponding author Sarah Opie-Martin, Maurice Wohl Clinical Neuroscience Institute, 5 Cutcombe Road, Camberwell, London SE5 9RX email sarah.martin@kcl.ac.uk telephone: 0207 848 5258

Word count: 2984

Keywords: amyotrophic lateral sclerosis, smoking, case control, comprehensive smoking index, motor neuron disease

Statement of contribution: AA-C and SO-M conceived and planned the study. SO-M conducted all statistical analysis. AA-C, AJ, AI, AA-K, MO, GDS, NP provided intellectual input for data interpretation. PS, CS, KM and AA-C organised and led collection of the epidemiology dataset. AA-C and SO-M wrote the first draft of the manuscript. All authors reviewed and approved the final manuscript.

5.1 Abstract

Introduction

Susceptibility to amyotrophic lateral sclerosis (ALS) is associated with smoking in some studies, but it is not clear which aspect of smoking behaviour is related. Using detailed records of lifetime smoking we investigated the relationship between smoking and ALS in a UK population.

Methods

In this retrospective case-control study, smoking status was collected using environmental questionnaires from people diagnosed with ALS between 2008 and 2013 and from age, sex and geographically matched controls. Categorical measures of smoking behaviour were: smoking at time of survey and smoking initiation; continuous measures were intensity (cigarettes per day), duration (years from starting to stopping or time of survey), cigarette pack years, and comprehensive smoking index (CSI), a measure of lifetime smoking. We used logistic regression to assess risk of ALS with different combinations of smoking variables adjusted for age at survey, gender, level of education, smoking status and alcohol initiation, selecting the best model using the Akaike Information Criterion.

Results

There were 388 records with full smoking history. The best fitting model used CSI and smoking status at time of survey. We found a weak association between current smoking and risk of ALS, OR 3.63 (95% CI 1.02-13.9) p-value 0.05. Increase in CSI score did not increase risk of ALS: OR 0.81 (95% CI 0.58-1.11) p-value 0.2.

Conclusion

There is weak evidence of a positive effect of current smoking on risk of ALS which does not show dose-dependence with higher levels of lifetime smoking and may be a false positive result.

5.2 Introduction

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease characterised by progressive death of motor neurons leading to relentlessly worsening weakness and death, usually from respiratory failure due to involvement of the diaphragm, 2-3 years after diagnosis (Brown and Al-Chalabi, 2017; Westeneng et al., 2018). Although there is an evident genetic component, heritability studies indicate that environmental (and probably stochastic) factors also contribute (Al-Chalabi et al., 2010; Longinetti and Fang, 2019; McLaughlin et al., 2015; Smith, 2011).

There is evidence from multiple studies that smoking is associated with ALS, but no agreement over which aspect of smoking behaviour is related to ALS (Alonso et al., 2010a; Alonso et al., 2010b; de Jong et al., 2012; F. Fang, Bellocco, Hernán, & Ye, 2006; Gallo et al., 2009; Kamel, Umbach, Munsat, Shefner, & Sandler, 1999; H. Wang et al., 2011; Weisskopf et al., 2004). Despite an evidence-based literature review that concluded that smoking can be considered a risk factor for ALS, it remains unclear if there is a dose-response effect, or what the biological mechanism might be (Armon, 2009). In addition, confounding cannot be discounted, since ALS is also associated with military service, education and socioeconomic status, which are also associated with smoking status (Beard and Kamel, 2015; Sutedja et al., 2009). It is biologically plausible that smoking could be a risk factor through oxidative stress or exposure to potentially neurotoxic chemicals, and so it remains an attractive candidate for studies of environmental aetiology (D'Amico, Factor-Litvak, Santella, & Mitsumoto, 2013; Roberts, Johnson, Cudkowicz, Eum, & Weisskopf, 2015).

The comprehensive smoking index (CSI) estimates lifetime smoking by combining duration, intensity and time since cessation into a score allowing all factors to be considered while avoiding issues of multicollinearity between smoking exposure variables (Leffondré et al., 2006). CSI has not previously been used to investigate the role of smoking in ALS risk.

We therefore analysed retrospective case-control data to determine whether smoking is related to ALS in a UK population, investigating the relationship between different smoking variables including CSI and other regularly used measures, and risk of ALS.

5.3 Methods

5.3.1. Case-control study design

The data were obtained from the Motor Neurone Disease Association of England, Wales and Northern Ireland (MNDA) Collections collected as part of the MNDA Epidemiology Study, REC reference 07/MRE01/57. People diagnosed with definite, probable or possible ALS according to the El Escorial criteria between 2008 and 2013 were included (Brooks, 1994). Three tertiary centres in London, Sheffield and Birmingham acted as data collection hubs but people with ALS were recruited

at secondary centres such as district general hospitals, therefore these are incident cases representative of the ALS population. General practitioners from the general practice of the person with ALS were asked to invite 10 healthy controls to participate in the study via post. The research team matched people on age (within 5 years of the person with ALS) and gender in a 1:1 ratio. 413 participants provided informed consent, 405 undertook a telephone interview about their lifestyle including smoking undertaken by a trained nurse. 3 participants gave no information on smoking behaviour.

5.3.2. Definition of smoking status

Categorical measures were: smoking at time of survey (current, former, never), smoking initiation (ever, never).

To define former smokers we used logistic regression modelling to compare ALS risk between current smokers and ex-smokers, using never smokers as a reference. Few people had recently quit (n=3 within one year of survey) so we grouped ex-smokers into 5-year time since cessation intervals up to 20 years which was aggregated to 20+. ALS risk reduced from an odds ratio of 2.02 to 0.79 for current smokers compared to people who had quit within 5 years so former smokers were defined as having given up at least a day before the survey.

Continuous measures included: intensity (cigarettes per day), duration of smoking (years from starting to stopping or time of survey), pack years (intensity x duration), and CSI. The CSI is a non-linear model of smoking exposure that combines duration of smoking, time since cessation and smoking intensity into a continuous score which can be used in a regression model representing lifetime smoking (Leffondré et al., 2006). The model involves simulation of tau and delta from the dataset. Delta, or half-life, reflects the exponential decay in the effect of smoking on health outcomes during a lifetime. Tau, or lag-time, reflects that smokers may be at a higher risk of disease immediately after quitting due to reverse causality. The equations for CSI are as follows:

$tsc^* = \max(tsc - \delta, 0)$

$dur^* = \max(dur + tsc - \delta, 0) - tsc^*$

$\text{comprehensive smoking index} = (1 - 0.5^{dur^*/\tau})\,(0.5^{tsc^*/\tau})\,\ln(int+1)$

tsc = time since cessation, δ = lag time, tss = time started smoking, dur = duration of smoking (calculated as age-tss for people currently smoking or [age-tsc]-tss for former smokers), τ = half-life, int = cigarettes per day.

5.3.3. Logistic Regression

Data were analysed using R. Continuous demographic characteristics were compared by Student's *t*-Test or Mann-Whitney U test. Categorical variables were compared by chi-squared or Fisher's exact test. The primary outcome, whether smoking increases risk of ALS, was analysed using logistic regression with maximum likelihood estimation. We generated 8 models with combinations of one categorical and one continuous measure of smoking, comparing the Akaike Information Criterion (AIC) of the models to assess fit (Akaike, 1998). Odds ratios were adjusted for age, educational attainment, gender and alcohol consumption.

Assuming an odds ratio of 1.8, a 20% smoking rate in the control population and alpha of 0.05, we had 71% power with a sample size of 400 cases and controls in a 1:1 ratio.

5.4 Results

There were 202 cases and 200 control records available for analysis. The two groups were similar except for educational attainment and alcohol status. The details are shown in table 1.

Demographic/behavioural measure		**Case (n=202)**	**Control (n=200)**	**p-value (test)**
Gender ratio, Female:Male % (n)		41:59 (85:117)	44:56 (88:112)	0.77 (Chi squared test)
Educational attainment % (n)	Primary school	1.5 (3)	1 (2)	0.0041 (Fisher's exact test)
	Secondary school	38.1 (77)	30.5 (61)	
	College	31.2 (63)	23.5 (47)	
	Technical school	8.4 (17)	12 (24)	
	University	14.4 (29)	29 (58)	
	Other	5.5 (12)	3.5 (7)	
	Missing	0.5 (1)	0.5 (1)	Not analysed
Mean age at survey (standard deviation)		63.1 (10.53)	64.5 (10.52)	0.12 (*t*-test)
Alcohol use % (n)	Alcohol status Never : Ever	8:62 (17:184)	12:88 (24:176)	0.32 (Fisher's exact test)
Site of onset % (n)	Bulbar	21.7 (44)	n/a	
	Spinal	73.3 (148)	n/a	
	Not known/recorded	5 (10)	n/a	
Mean age at onset (SD)		60.7 (10.6)	n/a	
Median months onset – diagnosis (IQR)		12 (13)	n/a	
Median months onset – survey (IQR)		28.1 (21.5)	n/a	

Table 5-1 Unadjusted comparisons of demographics and behaviour in ALS cases and controls.

The three centres are tertiary referral centres with about a third of the patients diagnosed at the centre, and the remainder diagnosed elsewhere first. SD = standard deviation, IQR = interquartile range n/a =Not applicable.

The optimal CSI variables were tau = 2 and delta = 3.6. There were no differences between groups in unadjusted smoking behaviours, as shown in table 2.

Smoking measure		**Case**	**Control**	**p-value**
Smoking behaviour % (n)	Smoking initiation (ever smokers)	47 (94)	53 (105)	0.27
	Smoking status Never : Former : Current	47:44:9 (94:90:18)	53:44:3 (105:88:7)	0.065
Median age smoking initiation (n)		16 +- 3 (108)	16 +- 3.5(94)	0.94
Median cigarettes per day (n)		17 +- 10	15 +- 10	0.88
Median duration smoking		23.5 +- 25.6	23 +- 24	0.58
Median cigarette pack years		20+-27.27	16+-28.4	0.7
Median comprehensive smoking index values		0.031 +- 1.85	0.0053 +- 1.36	0.33

Table 5-2 Smoking variables and crude comparisons.

Chi squared tests were used for categorical variables and Mann-Whitney U for continuous variables as all were non-normally distributed. 6 people were missing duration information, 4 missing smoking intensity. Records with missing data were excluded from analysis.

388 records had full smoking history available for logistic regression analysis. Table 3 gives the results of the best fitting logistic regression model which included the CSI and smoking status at time of survey with AIC 543.77. The highest AIC, representing the worst fitting model, was for smoking initiation and number of cigarettes per day at 553.23. An increase in the value of CSI did not increase the risk of ALS: OR 0.81 (95% CI 0.57-1.11) p-value 0.2. Current smoking increased the risk of ALS, OR 3.62 (95% CI 1.02-13.9) p-value=0.05, a Bonferroni correction shows that this is likely a false positive result because of multiple testing.

Variable		**Odds ratio**	**Lower CI**	**Upper CI**	**P value**
Smoking status	Current smoker	3.62	1.02	13.8	0.05
	Former smoker	1.08	0.67	1.74	0.74
Comprehensive smoking index		0.81	0.58	1.11	0.2
Age		0.98	0.96	1	0.07

Ever drinker		1.33	0.65	2.75	0.43
Male		1.05	0.68	1.63	0.83
Education level	Primary School	2	0.05	78.4	0.68
	Secondary School	1.27	0.05	33.2	0.87
	Technical School	0.73	0.03	19.4	0.83
	College	1.32	0.05	34.4	0.85
	University	0.44	0.02	11.7	0.58
	Other	1.6	0.06	45.7	0.75

Table 5-3 Best fitting logistic regression model for smoking and risk of ALS.

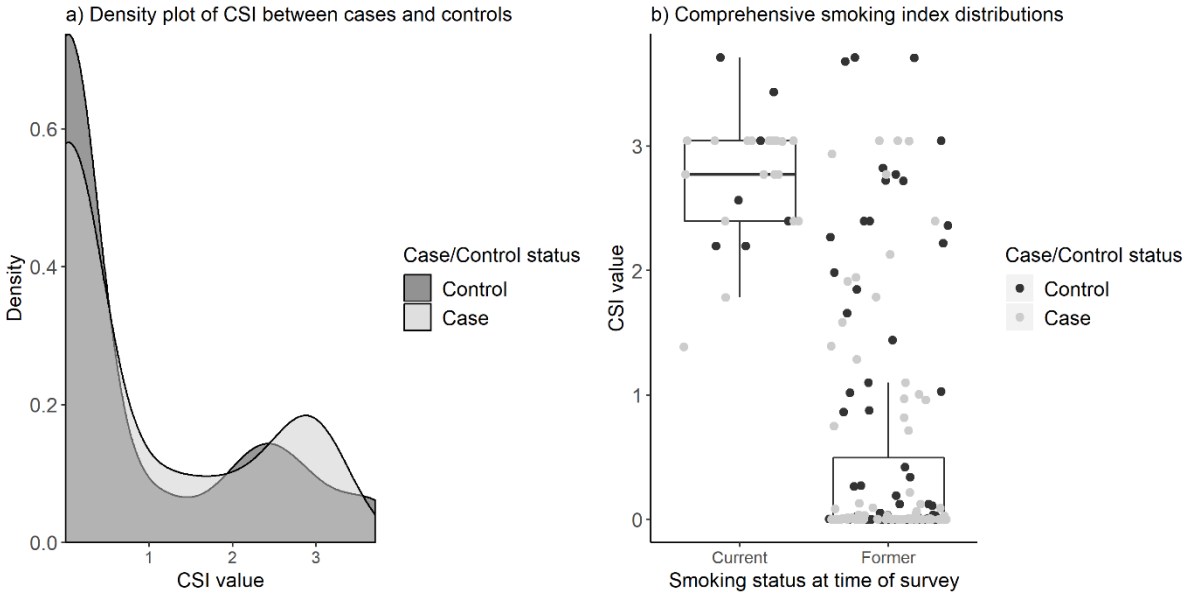


Figure 5-1 Comprehensive smoking index distributions by case control status

a) density plot of CSI value by case control status b) box plot of CSI value by smoking status at time of survey, points coloured by case control status. Both graphs are in ever smokers only.

5.5 Discussion

We found a weak association between current smoking and risk of ALS using traditional epidemiology methods to explore association. We report an uncorrected p-value of 0.05, and several models tested for fit, suggesting that this is in fact a false positive result. We also found that using CSI to measure lifetime smoking exposure resulted in a better fitting model for our data than using cigarette pack years, but we found no evidence of a dose-dependent response of ALS risk to smoking.

Our results are similar to those from a study conducted in the Netherlands which found current smoking to be associated with ALS in an incident cohort but no strong dose-dependent relationship (de Jong et al., 2012). The strength of association between smoking and ALS was reported as weak in a meta-analysis of case-control and cohort studies, with a higher effect in women (Alonso et al., 2010a). This weakness may be due to the reliance on prevalent and clinic cohorts which would under-represent smokers because their survival is shorter (de Jong et al., 2012).

A pooled analysis of prospective studies found that there was an increased risk of ALS in former and current smokers (H. Wang et al., 2011). Two large prospective cohort studies included in the pooled analysis were originally set up as prospective studies into environmental exposures and cancer risk (Gallo et al., 2009; Weisskopf et al., 2004). People with ALS were identified from death certificates, which may over-represent people who smoke as their survival is shorter.

The CSI is more useful than cigarette pack years to investigate dose-dependency, as it formally considers the decreased risk of disease after smoking cessation. The CSI had a bimodal distribution of smoking exposure in both cases and controls, corresponding to smoking at time of survey. The mean CSI of current smokers is slightly higher in cases than controls and so dose-dependency in current smokers should be investigated further.

Median age of smoking initiation was around the late teens in both groups, and it has been reported that frontotemporal dementia, a behavioural change that occurs in some people with ALS is not associated with smoking behaviours, so association is unlikely to reflect reverse causality (Tremolizzo et al., 2017).

The strengths of this study are that we have detailed environmental data on incident cases of ALS and controls. A limitation is the sample size which means it is only powered to detect relatively large effect sizes with odds ratios of the order of 1.8 or higher. Retrospective case-control studies generally suffer from recall bias. This study may suffer the effect of two opposing sample biases: people in an environmental study of lifestyle may be more likely to smoke heavily, and some people in this ALS study attended specialist clinics so may be less likely to smoke. Additionally, we do not know how many controls who were contacted declined to participate, so the control population may be biased. There were no current smokers in the controls recruited in London, although a subgroup analysis in the other two areas show that odds ratios for current smoking are consistent between the remaining areas.

We found that people with ALS were less likely to drink alcohol, but our survey responses do not support a protective relationship as ALS was cited as the reason for not drinking in most cases.

Despite controlling for drinking and educational status, it is not possible to completely rule out the effects of confounding.

In this study of smoking and ALS, we do not find strong evidence to support smoking as a risk factor, even using lifetime smoking exposure as measured by the CSI.

5.6 Acknowledgements

This project was funded through the Motor Neurone Disease Association. Data used in this research were entirely obtained from the UK MND Collections – epidemiology data for MND Research, funded by the MND Association and the Wellcome Trust. We would like to thank people with MND and their families for their participation in this project. This is in part an EU Joint Programme - Neurodegenerative Disease Research (JPND) project. The project is supported through the following funding organisations under the aegis of JPND - www.jpnd.eu (United Kingdom, Medical Research Council (MR/L501529/1; MR/R024804/1) and Economic and Social Research Council (ES/L008238/1)). This study represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The work leading up to this publication was funded by the European Community's Health Seventh Framework Programme (FP7/2007–2013; grant agreement number 259867) and Horizon 2020 Programme (H2020-PHC-2014-two-stage; grant agreement number 633413).

5.7 Disclosure of interest

None of the authors had competing interests to declare.

Appendix for Chapter 5 – Data collection questions and options for lifestyle and socioeconomic covariates

Data were collected using paper forms and entered into a Microsoft Access database.

Tobacco history questions	**Tobacco History options**
Q34. Frequent cigarette smoker?	Yes/No
Q35. Age started smoking	Integer
Q36. Still smoke	Yes/No
Q36a. Age stopped smoking	Integer
Q37. Other periods stopped	Yes/No/Don't know
Q37a. Years not smoking	Integer
Q38. Cigarettes per day	Integer
Alcohol consumption questions	**Lifetime Alcohol Consumption options**
Q27. Drunk alcohol in 6 month period	Yes/No/Don't know
Q28. Age started drinking	Integer
Q29. Still drink alcohol	Yes/No
Q30. Age stopped drinking	Integer
Q31. Reason for stopping drinking	Free text
Q32. Frequency of drinking (amount)	Integer
Q32. Frequency of drinking (per?)	Day/Month/Year
Q33. Number of drinks per session	Integer
sr_sociobackground	**Socio-economic Background options**
Q19. Level of Education	Primary/Secondary/Technical School/College/University/Other

Table 5-4 Questions as worded on questionnaires for variables analysed

Chapter 6 Relationship between smoking and ALS: Mendelian randomization interrogation of causality

Sarah Opie-Martin[1], Robyn E Wootton[2,3,4], Ashley Budu-Aggrey[2], Aleksey Shatunov[1], Ashley Jones[1], Alfredo Iacoangeli[1], Ahmad Al Khleifat[1], George Davey-Smith[4], Ammar Al-Chalabi*[1]

1) Maurice Wohl Clinical Neuroscience Institute, King's College London, Department of Basic and Clinical Neuroscience, London, UK 2) MRC Integrative Epidemiology Unit at the University of Bristol, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, UK 3) School of Psychological Science, University of Bristol, 12a Priory Road, Bristol, BS8 1TU, UK 4) NIHR Bristol Biomedical Research Centre, University Hospitals Bristol NHS Foundation Trust and University of Bristol, Bristol, UK

*Corresponding author email ammar.al-chalabi@kcl.ac.uk Maurice Wohl Clinical Neuroscience Institute, 5 Cutcombe Road, London SE5 9RX

Word count: 2255

References: 36

Acceptance date: 11/05/2020 – Journal of Neurology, Neurosurgery and Psychiatry

Statement of contribution: AA-C and SO-M conceived and planned the study. ABA conducted GRS analysis, S-OM conducted all other statistical analysis supported by RW. AA-C, GDS, ABA and RW provided intellectual input for data interpretation. AA-C and SO-M wrote the first draft of the manuscript. All authors reviewed and approved the final manuscript.

6.1 Abstract

Objective: Smoking has been widely studied as a susceptibility factor for ALS, but results are conflicting and at risk of confounding bias. We used the results of recently published large genome-wide association studies and Mendelian randomisation methods to reduce confounding, to assess the relationship between smoking and ALS.

Methods: Two genome-wide association studies investigating lifetime smoking (n=463,003) and ever smoking (n=1,232,091) were identified and used to define instrumental variables for smoking. A genome-wide association study of ALS (20,806 cases; 59,804 controls) was used as the outcome for inverse variance weighted Mendelian randomisation, and four other Mendelian randomisation methods, to test whether smoking is causal for ALS. Analyses were bi-directional to assess reverse causality.

Results: There was no strong evidence for a causal or reverse causal relationship between smoking and ALS. The results of Mendelian randomisation using the inverse variance weighted method were: lifetime smoking odds ratio 0.94 (95% confidence intervals 0.74,1.19), p-value 0.59; ever smoking odds ratio 1.10 (95% CI 1,1.23), p-value 0.05.

Conclusions: Using multiple methods, large sample sizes, and sensitivity analyses, we find no evidence with Mendelian randomisation techniques that smoking causes ALS. Other smoking phenotypes, such as current smoking, may be suitable for future Mendelian randomisation studies

6.2 Introduction

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease of motor neurons, resulting in progressive paralysis of skeletal and bulbar muscles, with death from neuromuscular respiratory failure typically occurring within two to three years of symptom onset (Brown and Al-Chalabi, 2017). ALS has an incidence of 1-2 per 100,000 person-years and a lifetime risk of about 1 in 300 (Chio et al., 2013; C. A. Johnston et al., 2006). It has a peak age of onset of 58 and affects men slightly more frequently than women (Chio et al., 2013). In 5% there is a family history of ALS in a first degree relative, but twin and other studies have shown that apparently sporadic ALS has a heritability of 60%, leaving the possibility that up to 40% of the contribution could be environmental (Al-Chalabi et al., 2010). There are currently no agreed environmental risk factors for ALS, although smoking has been widely studied with mixed results (Alonso et al., 2010a; Alonso et al., 2010b; de Jong et al., 2012; Pamphlett and Ward, 2012; H. Wang et al., 2011; Yu et al., 2014).

Summary statistics from genome-wide association studies allow us to use genetic predisposition for environmental risk factors to investigate causality, using Mendelian randomisation (Davey Smith and Ebrahim, 2003; Richardson, Harrison, Hemani, & Davey Smith, 2019). Mendelian randomisation is based on Mendel's laws of inheritance, allowing genotype to be used as an instrumental variable when studying the effect of an environmental exposure on an outcome (Lawlor, Harbord, Sterne, Timpson, & Davey Smith, 2008). Genotype is considerably less likely to be confounded with other exposures that may bias the results of observational studies (Davey Smith et al., 2007). Mendelian randomisation can also help to reduce bias from reverse causation, because the genetic variants one is born with are unchanged through a lifetime. This means the outcome, for example ALS, cannot change an individual's genetic predisposition for the exposure, for example smoking. In population-based genetic association studies, as opposed to parent-offspring or between-sibling studies, the randomisation is only approximate, and horizontal pleiotropy (where the genetic variants being tested increase the risk of both the environmental exposure and of ALS) can in any setting distort findings. A series of sensitivity analyses are available that can uncover such biases (Davey Smith and Ebrahim, 2003; Davey Smith and Hemani, 2014; Davies et al., 2019).

Mendelian randomisation analysis has previously been used to assess the causal relationship between smoking and ALS with conflicting results (Bandres‐Ciga et al., 2019; Zhan and Fang, 2019). Since then, updated genome-wide association studies for smoking and ALS have been published, with much larger numbers, allowing a new analysis with the advantages of sufficient power and updated methods. We performed two-sample Mendelian randomisation analyses to assess whether

there was evidence for smoking being causal for ALS and, as a sensitivity analysis, in the other direction to test if ALS liability might be causal for smoking.

6.3 Methods

Two-sample Mendelian randomisation analysis enables the summary statistics of genome-wide association studies to be used to estimate the causal effect of an exposure on an outcome based on the effect sizes of genetic variations on the exposure and on the outcome in the separate samples (Pierce and Burgess, 2013). The effect estimate from a Mendelian randomisation study is an estimation of the true causal effect of an exposure on the outcome of interest. In the case of ALS diagnosis, this will be expressed as an odds ratio.

We defined an instrument for lifetime smoking index, a continuous measure of smoking exposure from a genome-wide association study of 463,033 people, with 126 independently associated single nucleotide polymorphisms (SNPs) of genome-wide significance that explained 0.31% of the variance in lifetime smoking (Wootton et al., 2018). We also defined an instrument for 'Ever smoking' a binary measure of smoking exposure from a genome-wide association study of 1,232,091 individuals, with 378 genome-wide significant SNPs accounting for 4% of variance (M. Liu et al., 2019). The variance explained in both studies is in line with genome-wide association studies that have been used to assess the causality of smoking for other conditions.

More details of how the phenotypes were defined can be found in the supplementary file.

We used summary data from the most recently published genome-wide association study for ALS (Nicolas et al., 2018). The study reported 10 SNPs to be independently associated with risk of ALS, in a population of 80,610 (20,806 cases and 59,804 controls).

6.3.1. Statistical Analyses

To perform the Mendelian randomisation analysis, we used the 'TwoSampleMR' package, an R package and genome-wide association study summary data library, developed as a platform for performing Mendelian randomisation tests and sensitivity analyses (Hemani et al., 2018b).

We applied five different Mendelian randomisation methods: inverse-variance weighted, MR Egger, weighted median, weighted mode, and MR using robust associated profile score (MR RAPS). Each of these methods make different assumptions about pleiotropy and instrument strength so a consistent effect across the multiple methods gives us the strongest evidence for causality (Hemani, Bowden, & Davey Smith, 2018a). For details of these and other sensitivity analyses, please see supplementary materials.

As an additional, supportive analysis, we defined genetic risk scores for smoking and used the UK Biobank data to test whether genetic risk score for smoking predicts ALS case control status. Details of the methods can be found in the supplementary file under the heading 'Genetic risk score analysis'.

6.4 Results

Details of sensitivity analyses performed can be found in supplementary information (tables S1-S5). All instruments passed sensitivity analyses except for: heterogeneity tested using Cochran's Q (table S3); and a minority of SNPs did not pass Steiger filtering analysis (table S4), however re-running the Mendelian randomisation analyses without these SNPs did not change the results (table S5). After quality control, the number of SNPs making up each instrument were n=119 for lifetime smoking exposure and n=353 for ever smoking. Genome-wide association study summary statistics for each SNP making up the instruments are found in supplementary file 2. Graphs showing scatter plots of effect sizes of SNP on exposure and outcome variable, leave-one-out analyses and single SNP analyses are shown in figures S1-S12.

The results of the Mendelian randomisation analyses for each instrumental variable are shown in figure 1. We did not find strong evidence that lifetime smoking or ever smoking were causal for ALS. The result of the inverse variance weighted method for lifetime smoking index was odds ratio 0.94 (95% confidence intervals 0.74, 1.19) p-value = 0.59, and for ever smoking, OR 1.10 (95%CI 1.00,1.23) p-value=0.05. We did not find that ALS liability was causal of smoking status (table S6). Odds ratios tended to be >1 for the instruments testing if having ever smoked was associated with ALS, and <1 for the lifetime smoking instrument. A forest plot of results is shown in figure 1.

Exposure	Method	OR (95% CI)	P value
Lifetime smoking index	Inverse variance weighted	0.94 (0.74, 1.19)	0.59
	Weighted median	0.81 (0.60, 1.08)	0.15
	Weighted mode	0.72 (0.41, 1.26)	0.25
	Robust adjusted profile score (RAPS)	0.89 (0.70, 1.14)	0.36
	MR-Egger SIMEX - unweighted	0.47 (0.19, 1.12)	0.09
Ever smoking	Inverse variance weighted	1.10 (1.00, 1.23)	0.05
	Weighted median	1.06 (0.93, 1.21)	0.39
	Weighted mode	0.94 (0.60, 1.49)	0.81
	Robust adjusted profile score (RAPS)	1.11 (1.00, 1.23)	0.04
	MR-Egger SIMEX - unweighted	1.47 (1.04, 2.09)	0.18

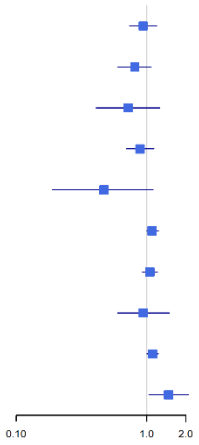


Figure 6-1 Forest plot of Mendelian randomisation analyses

Using genetic risk score analysis we found no association between smoking and ALS case control status (table S7).

6.5 Discussion

Using instruments defined from recently published, large-scale genome-wide association studies, and numerous Mendelian randomisation methodologies and sensitivity analyses, we found no evidence that smoking causes ALS. Our result is supported by a lack of association found when performing genetic risk score analysis. We also found no relationship between genetic liability to ALS and likelihood of smoking, suggesting that reverse causality is not driving the association between smoking and risk of ALS reported in some epidemiological studies.

Two previous studies have used 2-sample Mendelian randomisation analysis to assess the causal relationship between smoking and ALS. One reported a positive association between ever smoking and ALS using inverse variance weighted Mendelian randomisation analysis, but the result was not replicated with other Mendelian randomisation analyses and no other sensitivity analyses were reported. The ever smoking instrument was defined from (the Social Science Genetic Association Consortium (SSGAC)) genome-wide association study, which has a smaller sample size than GSCAN. The outcome SNPs were from an ALS genome-wide association study published previously to the one used here (van Rheenen et al., 2016). With a larger genome-wide association study we replicate the inverse variance weighted result with borderline significance, p-value 0.05, (figure 1) but do not consider this evidence of causality in the context of the results of the other Mendelian randomisation analyses presented. The other Mendelian randomisation study found no association using smoking instruments from a smaller smoking genome-wide association study and the same ALS genome-wide association study used here (Bandres‐Ciga et al., 2019; Zhan and Fang, 2019).

Our study is necessary to analyse the larger genome-wide association studies available, and to report all sensitivity analyses needed to interpret Mendelian randomisation results.

Large numbers of SNPs and high F statistic values mean the genetic instruments had enough strength to detect associations using the inverse variance weighted method (Burgess and Thompson, 2011). The I^2 statistic quantifies regression dilution, which can be caused by measurement error (Bowden et al., 2016b). When using linear regression analysis (as is the case with MR Egger), measurement error in the exposure variable will cause the effect size to tend to the null, and in the outcome will reduce statistical power (K. Liu, 1988). All smoking instruments had I^2 values of <0.9, indicating some regression dilution. In cases where I^2 >0.6 SIMEX modelling was undertaken to

estimate regression values and the results support the findings from other Mendelian randomisation models used.

For Mendelian randomisation to be valid, the genetic variants used as instrumental variables must mediate an effect only through the exposure of interest (the risk of ALS is only increased due to smoking, not through another effect of the variants), i.e. there should be no horizontal pleiotropy. A limitation of this study is that we are unable to fully discount this pleiotropy using statistical techniques. All genetic instruments tested positive for heterogeneity using Cochran's Q statistic, which may be caused by pleiotropy (Bowden, Hemani, & Davey Smith, 2018). MR Egger intercept analysis did not find evidence to support the presence of directional pleiotropy. However, low I^2 values may invalidate the intercept estimation from MR Egger regression. A previous study used linkage disequilibrium regression score analysis and found that exposure to tobacco smoke in the home and being a light smoker (<100 cigarettes in a lifetime) are genetically related to ALS, which may support pleiotropy (Bandres‐Ciga et al., 2019). Horizontal pleiotropy can cause false positives and false negatives. We used 5 Mendelian randomisation models that vary in their assumptions of pleiotropy to try account for these potential errors, although future models of Mendelian randomisation and how they account for pleiotropy may be better suited to identifying association between ALS and smoking. Use of a binary outcome measure (which is the case in this study) or otherwise invalidated modelling assumptions may cause heterogeneity tests to produce positive results(Hemani et al., 2018a).

Inverse variance weighted method will estimate the true causal effect of an exposure if all Mendelian randomisation assumptions hold; if not, other methods have been developed (Bowden, Davey Smith, & Burgess, 2015; Bowden, Davey Smith, Haycock, & Burgess, 2016a; Bowden et al., 2016b; Hartwig, Davey Smith, & Bowden, 2017; Kang, Zhang, Cai, & Small, 2016). It is suggested best practice to use multiple Mendelian randomisation methods to check for consistency of estimated effect (Hemani et al., 2018a). Following this approach, we find a consistent lack of relationship between smoking and ALS. However, UK Biobank data contribute to the study cohorts of the lifetime smoking index and ever smoking instrument so the exposure phenotype populations are not completely independent.

The advantage of the instrumental variables we used is that they can be used in an unstratified population (we do not need to know the smoking status of people in the outcome genome-wide association study)(Wootton et al., 2018). However, ever smoking is not consistently associated with ALS in epidemiology studies. A meta-analysis of case-control and cohort studies did not find strong supportive evidence of risk of ALS in people who had ever smoked (OR 1.12, 95%CI 0.98, 1.27)

(Alonso et al., 2010a). Since then an association has been reported in some studies but not others (Alonso et al., 2010b; H. Wang et al., 2011). Lifetime smoking index can be used to assess dose-dependency, important contributory evidence to showing causality. Evidence of a dose-dependent effect of smoking on ALS risk is rarely shown, although a dose-dependent effect of reduced risk of ALS with increased time since smoking cessation when comparing former to current smokers has recently been reported (Peters et al., 2019). The only ALS risk study to use the lifetime smoking index did not find an association (unpublished data).

The most powerful Mendelian randomisation evidence on a potential effect of heaviness of smoking on ALS risk would require individual level data on a large sample, in which the CHRNA5 variant – related to heaviness of smoking amongst smokers – can be related to ALS risk by strata of smoking behaviour (Millard, Munafo, Tilling, Wootton, & Davey Smith, 2019). There is currently no adequately powered study allowing such analyses.

Triangulation of multiple strands of epidemiological evidence makes findings more robust (Lawlor et al., 2017). Since the 1960's, smoking rates in many countries globally have been falling. It may be possible in the future to detect decreased rates of ALS in the population if smoking is a causal risk factor.

Using robust methods to detect association and estimate causal effects with summary statistics from genome-wide association studies we do not find strong evidence to support a relationship between smoking and ALS.

6.6 Declaration of interests

There are no conflicts of interest to declare.

6.7 Acknowledgements

The project is supported through the following funding organizations under the egis of JPND—www.jpnd.eu (United Kingdom, Medical Research Council [MR/L501529/1] and Economic and Social Research Council [ES/ L008238/1]). The work leading up to this publication was funded by the European Community's Health Seventh Framework Program [FP7/2007–2013; grant agreement number 259867], Horizon 2020 Framework Programme [H2020-PHC-2014-two-stage; grant agreement number 633413] and Programme Grants for Applied Research. This project is also supported by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London Maudsley Foundation Trust and King's College London. This research has been conducted using data from the UK Biobank Resource (application number 19278). We have also received funding from the Motor Neurone Disease Association, ALS Association, Patients Like

Me and the Psychiatry Research Trust. This research was also supported by the NIHR Bristol Biomedical Research Centre at University Hospitals Bristol NHS Foundation Trust and the University of Bristol. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care.

6.8 Funding statement

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors

6.9 Contributorship

AA-C and SO-M conceived and planned the study. ABA conducted GRS analysis, S-OM conducted all other statistical analysis supported by RW. AA-C, GDS, ABA and RW provided intellectual input for data interpretation. AA-C and SO-M wrote the first draft of the manuscript. All authors reviewed and approved the final manuscript.

6.10 Ethical approval

This research involves analysing publicly available data, with no collection of new data. Ethical approval had been obtained by the original study authors.

6.11 Supplementary file

6.11.1. Smoking phenotypes

We investigated whether smoking is causal of ALS using two smoking phenotypes: the lifetime smoking index, a continuous measure of smoking exposure; and ever smoking, a binary measure of smoking exposure.

The lifetime smoking index is a model that combines multiple aspects of smoking behaviour including smoking initiation, heaviness, duration and cessation with two constants: tau, the levelling off of risk with increased smoking exposure; and delta, the lag in drop of disease risk after cessation due to reverse causality (Leffondré et al., 2006). The lifetime smoking index model can be used to provide a lifetime smoking exposure score per individual that can be used in statistical models. The UK Biobank has been used to develop an instrumental variable for lifetime smoking index which has been validated with positive control MR studies in diseases in which smoking is established as a risk factor such as coronary heart disease (Wootton et al., 2018).

Ever smoking, was defined by Genome-wide association study and Sequencing Consortium of Alcohol and Nicotine use (GSCAN) as those people who had ever reported; 1) being a regular smoker

in their life, 2) having smoked over 100 cigarettes over the course of your life or 3) have you ever smoked every day for at least a month? (M. Liu et al., 2019).

6.11.2. Instrument strength

Instrument strength (a measure of how related the SNPs are to the exposure) can be quantified using the F statistic, and regression dilution as a result of random error in the SNP-exposure effects can be summarised using the I^2_{GX} statistic(Bowden et al., 2016b). The results of each test are detailed in table S1.

Smoking causal for ALS			
Instrument	F statistic	I^2_{GX} unweighted	I^2_{GX} weighted
Lifetime smoking index	44.0	0.64	0.42
Ever smoking	44.9	0.60	0.47
ALS liability causal for smoking			
Instrument	F statistic	I^2_{GX} unweighted	I^2_{GX} weighted
Lifetime smoking index	50.4	0.60	0.47
Ever smoking	50.4	0.91	0.71

Supplementary table 6-1 mean F statistic of each SNP, unweighted and weighted I^2_{GX} statistics for MR analyses in both directions.

6.11.3. Pleiotropy tests

Cochran's *Q* statistic is used to detect heterogeneity which can be caused by horizontal pleiotropy, the results of this test can be found in table S2 (Bowden et al., 2018). The MR Egger intercept test to estimates the bias from directional pleiotropy and results are shown in table S3 (Bowden et al., 2015).

Smoking causal for ALS				
Instrument (exposure)	Method	Q	Q_df	Q_pval
Lifetime smoking index	MR Egger	185.62	117	5.55E-05
	Inverse variance weighted	187.77	118	4.64E-05
	Rucker's Q	2.16	1	1.42E-01
Ever smoking	MR Egger	467.64	351	2.96E-05
	Inverse variance weighted	469.53	352	2.68E-05
	Rucker's Q	1.89	1	1.70E-01
ALS liability causal for smoking				
Instrument (outcome)	Method	Q	Q_df	Q_pval

Lifetime smoking index	MR Egger	7.89	7	0.34
	Inverse variance weighted	11.40	8	0.18
	Rucker's Q	3.51	1	0.06
Ever smoking	MR Egger	9.73	7	0.20
	Inverse variance weighted	10.77	8	0.21
	Rucker's Q	1.04	1	0.31

Supplementary table 6-2 heterogeneity analyses using Cochran Q statistic for MR analyses in both directions

Smoking causal for ALS			
Instrument	MR Egger intercept	SE	p-value
Lifetime smoking index	0.008	0.007	0.25
Ever smoking	0.005	0.004	0.24
ALS liability causal for smoking			
Instrument	MR Egger intercept	SE	p-value
Lifetime smoking index	-0.003	0.002	0.12
Ever smoking	-0.002	0.003	0.42

Supplementary table 6-3 pleiotropy tests using MR Egger intercept

6.11.4. SNP filtering

Steiger filtering is used to check for reverse causation by testing whether the association between instrumental variable SNPs and the MR exposure measure is greater than the association between those same SNPs and the MR outcome measure (Hemani, Tilling, & Davey Smith, 2017). The results of this test are shown in table S4.

Smoking causal for ALS				
Instrument	Total SNPs	Unable to find LD proxy	Palindromic SNPs with intermediate frequencies	Failed Steiger filtering
Lifetime smoking index	126	0	7	4
Ever smoking	378	5	20	52
ALS liability causal of smoking				
ALS and lifetime smoking	10	1	0	0
ALS and ever smoking	10	1	0	0

Supplementary table 6-4 Numbers of SNPs that were removed from instruments for either not finding a match in the outcome dataset, being ambiguous matches, as well as those not passing Steiger filtering.

Exposure Instrument	Method	N SNPs	P-value	Odds ratio	95% CI
Smoking causal for ALS					
Lifetime smoking index	Inverse variance weighted	115	0.51	0.92	0.75, 1.16
	Weighted median	115	0.14	0.81	0.61, 1.07
	Weighted mode	115	0.28	0.71	0.39, 1.32
	Robust adjusted profile score (RAPS)	115	0.33	0.89	0.71, 1.12
Ever smoking	Inverse variance weighted	301	0.17	1.07	0.97,1.17
	Weighted median	301	0.58	1.04	0.91,1.19
	Weighted mode	301	0.67	0.90	0.56,1.44
	Robust adjusted profile score (RAPS)	301	0.05	1.07	0.97,1.18

Supplementary table 6-5 Results of MR analyses with only variants passing Steiger filtering.

6.11.5. Mendelian Randomisation analyses

For results of a Mendelian randomisation analysis to be valid, three assumptions must be satisfied (Lawlor et al., 2008). One assumption is that the instrument, in this case SNPs, used to assess the relationship between the exposure and outcome must be strongly associated with the exposure, in this case smoking. Another assumption is that SNPs are not confounded with factors that also confound the exposure, for example, that SNPs are not confounded with a factor such as socioeconomic status, also associated with smoking. The final assumption is that the effect of the SNPs on the outcome is only through their effect on the exposure. If the SNPs affect the outcome through another effect of the gene (in this case if the SNPs cause ALS in some other way not just through increased smoking) this is referred to as horizontal pleiotropy.

Inverse variance weighted Mendelian randomisation is a meta-analysis of the ratio of SNP-exposure effects on SNP-outcome effects weighted by the inverse variance of the outcome effects (Johnson, 2012). Random effects Inverse variance weighted Mendelian randomisation will return an unbiased estimate of the effect of the exposure on the outcome if model assumptions are met and the direction of effect of horizonal pleiotropy is balanced; it is the default inverse variance weighted method in the TwoSampleMR package (Bowden et al., 2017). We used this as our main analysis with each of the other methods providing a sensitivity analysis.

MR Egger analysis regresses SNP-exposure effects on SNP-outcome effects but does not constrain the intercept to pass through the origin as is the case in inverse variance weighted Mendelian randomisation (Bowden et al., 2016b). Not constraining the intercept to pass through the origin means the intercept from MR Egger can be used to estimate bias from directional pleiotropy (other effects of the SNPs than just those on the exposure of interest) and that the slope estimate can still

be valid in the presence of unbalanced pleiotropy. We used simulation extrapolation (SIMEX) corrections to MR Egger estimates where there was evidence of regression dilution (low I^2_{GX} value, see table S1) (Bowden et al., 2016b).

The weighted median method assumes half of all SNPs are valid instruments and more heavily weights the SNPs that are more strongly associated with the exposure (Bowden et al., 2016a).

The weighted mode method clusters SNPs into groups and estimates causal effect from the largest group, weighting each SNPs contribution to the clustering by inverse variance of its outcome (Hartwig et al., 2017).

MR RAPS penalises SNPs based on their individual estimated pleiotropic effect calculated using a robust adjusted profile score (Zhao, Wang, Hemani, Bowden, & Small, 2018).

The results of the Mendelian randomisation analyses in both directions are found in table S6.

Exposure Instrument	**Method**	**N SNPs**	**P-value**	**Odds ratio**	**95% CI**
Smoking causal for ALS					
Lifetime smoking index	Inverse variance weighted	119	0.59	0.94	0.74, 1.19
	Weighted median	119	0.15	0.81	0.60, 1.08
	Weighted mode	119	0.25	0.72	0.41, 1.26
	Robust adjusted profile score (RAPS)	119	0.36	0.89	0.70, 1.14
	MR-Egger, SIMEX - unweighted	119	0.09	0.47	0.19,1.12
Ever smoking	Inverse variance weighted	353	0.05	1.10	1.00,1.23
	Weighted median	353	0.39	1.06	0.93,1.21
	Weighted mode	353	0.81	0.94	0.60,1.49
	Robust adjusted profile score (RAPS)	353	0.04	1.11	1.00,1.23
	MR-Egger, SIMEX - unweighted	353	0.18	1.47	1.04,2.09
ALS liability causal for smoking					
Outcome Instrument	method	N SNPs	p-value	Odds ratio	95% CI
Lifetime smoking index	Inverse variance weighted	9	0.99	1.00	0.99, 1.01
	Weighted median	9	0.88	1.00	0.99, 1.01
	Weighted mode	9	0.79	1.00	0.99, 1.02
	Robust adjusted profile score (RAPS)	9	0.63	1.00	0.99, 1.02
	MR-Egger, SIMEX - unweighted	9	0.98	1.00	0.97,1.03
Ever smoking	Inverse variance weighted	9	0.71	1.00	0.99,1.02
	Weighted median	9	0.83	1.00	0.98,1.01
	Weighted mode	9	0.59	0.99	0.97,1.04
	Robust adjusted profile score (RAPS)	9	0.93	1.00	0.99,1.01
	MR-Egger	9	0.38	1.01	0.98,1.05

Supplementary table 6-6 Table S6 results of MR analysis in both directions

6.11.6. Genetic risk score analysis

Genetic risk score analysis uses genome-wide association summary statistics to create individual genetic risk scores for a trait that can be used to predict development of the same trait in different people or can be correlated with development of other traits. *P*-value thresholds for included SNPs can be relaxed compared to those used to define genetic instruments in Mendelian randomisation, and typically genetic risk score analysis is used as a first step to generate hypotheses for causality that can be more rigorously tested using Mendelian randomisation (Richardson et al., 2019). Smoking has already been associated with ALS in some observational literature so in this case we have used genetic risk score to check for concordance with Mendelian randomisation analysis.

We generated genetic risk scores for lifetime smoking index and ever smoking using the SNPs reported as reaching genome-wide significance from the GWASs defined above and used them to predict ALS in data from UK Biobank (194 ALS cases; 384,970 controls). Lifetime smoking index score was derived using 126 SNPs and GSCAN ever smoking score using 378 SNPs. To test bi-directionality, the association between a genetic risk score for ALS and various smoking variables in UK Biobank was assessed (385,164 individuals with available genotype and phenotype data). For the 135 variants associated with ALS (threshold for association with ALS $p < 5x10^{-5}$), 128 SNPs were used to generate a genetic risk score for ALS (Nicolas et al., 2018). Smoking variables included participant responses to having "ever smoked" (UK Biobank number 20160), "number of cigarettes currently smoked daily" (UK Biobank number 3456), "number of cigarettes previously smoked daily" (UK Biobank 2887), "current tobacco smoking" (UK Biobank number 1239) and "past tobacco smoking" (UK Biobank number 1249).

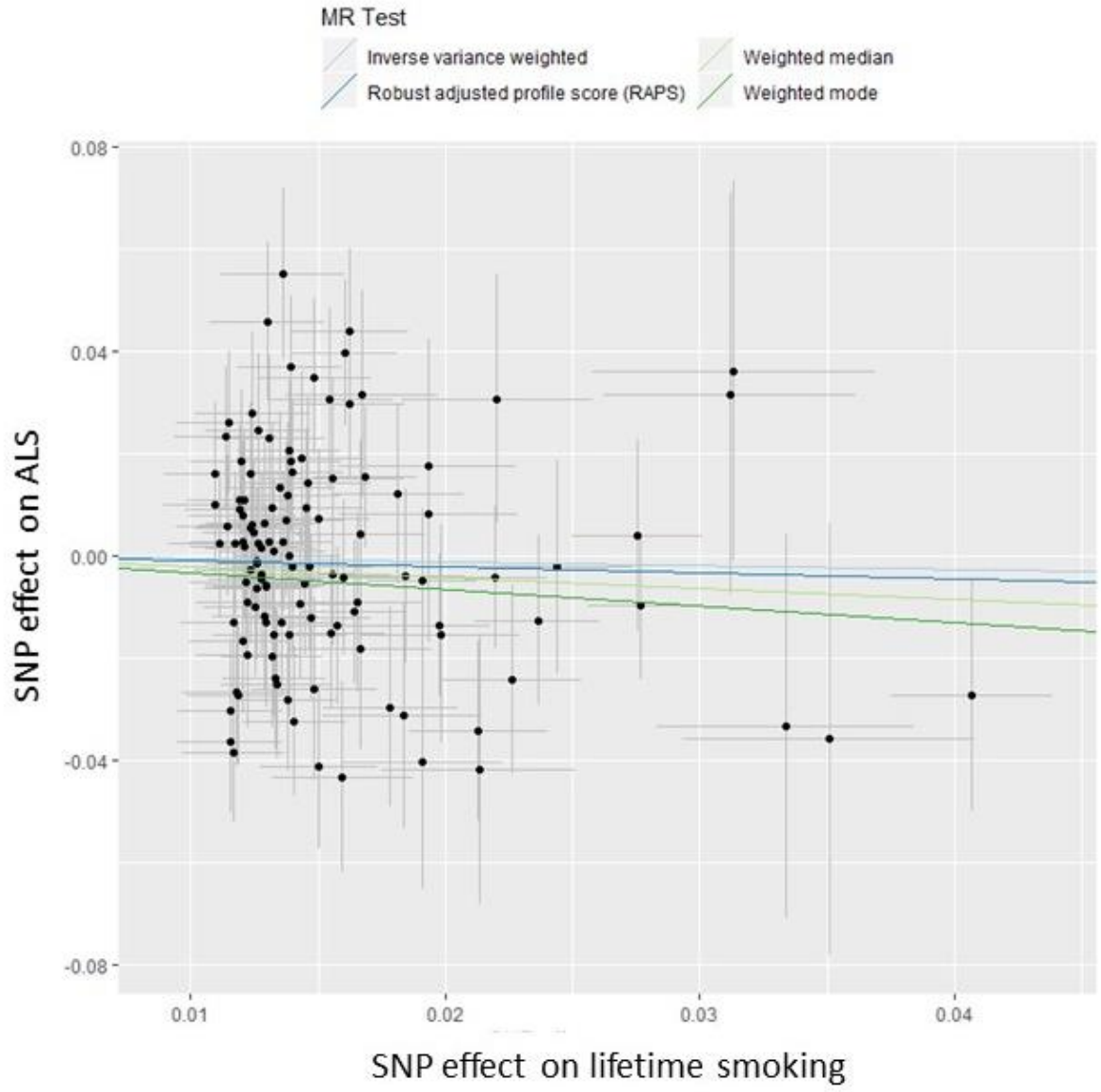
All SNPs were ensured to be independent of each other ($r^2 < 0.001$) using genotype data from European individuals (CEU) from phase 3 of the 1000 Genomes Project as a reference. For SNPs that were not present in the UK Biobank genotype data, a suitable proxy was selected ($r^2 > 0.8$). The risk score was weighted by the effect size (beta) of the reported effect allele and normalised to have a mean of zero and a standard deviation of one. Individuals with evidence of genetic relatedness or who were not of European ancestry were excluded, as well as those who had withdrawn consent. Further QC measures that have been applied to the UK Biobank genetic data set have been described previously (R. Mitchell, Hemani, G., Dudding, T., Corbin, L., Harrison, S., Paternoster, L. , 2018).

Regression analyses were completed in R, where linear regression was applied to analyse continuous traits, logistic regression for binary traits and ordinal logistic regression for ordered categorical traits. All analyses were adjusted for age, sex, 10 principal components of ancestry and the genotyping chip used to generate the genetic data for participants. Results of GRS analyses are found in table S7.

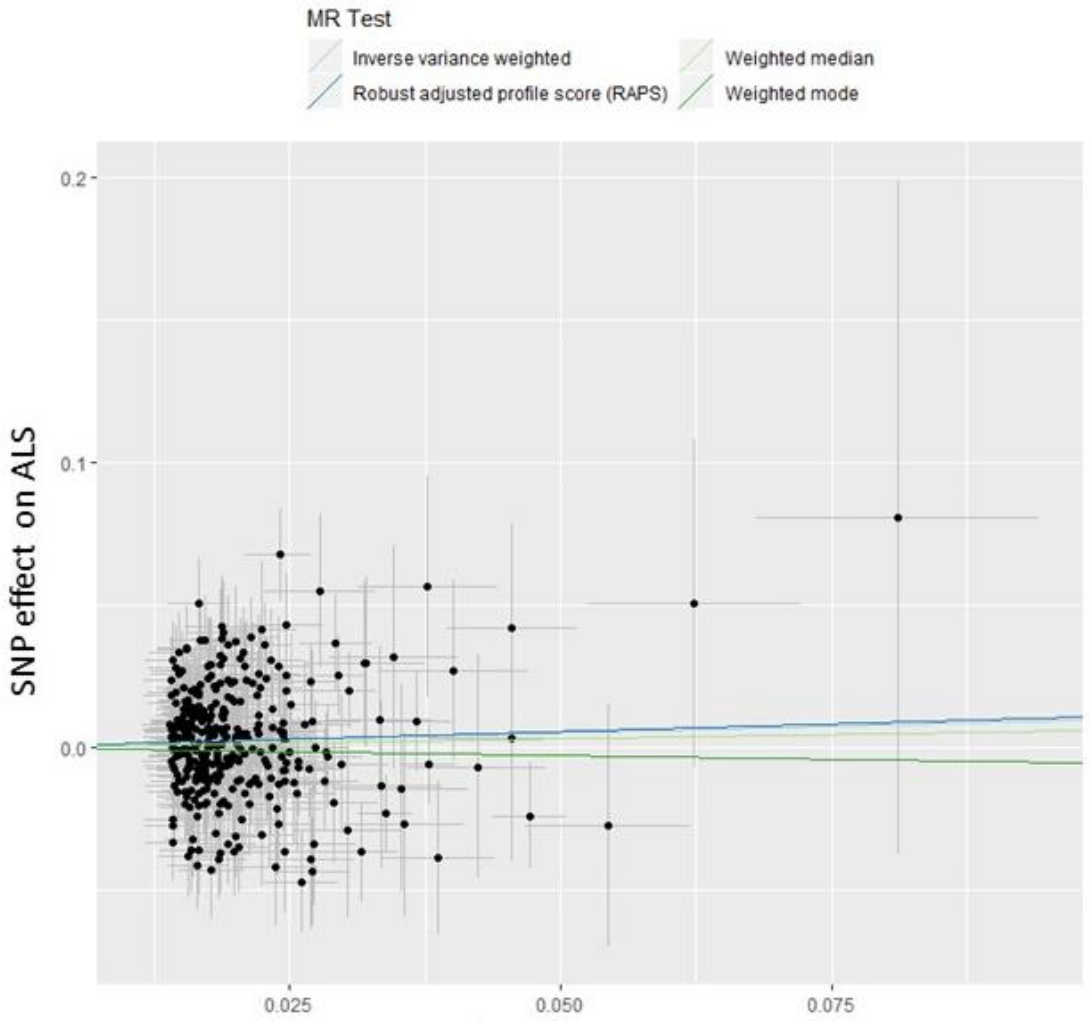
Smoking GRS	OR	95% CI	Beta	SE	P	N	Model	AUC
Directionality: Smoking causal for ALS								
Lifetime smoking index	1.07	0.93, 1.23	0.07	0.07	0.34	385,164	OLR	0.66
Ever smoking	0.96	0.84, 1.11	-0.04	0.07	0.59	385,164	OLR	0.66
Directionality: ALS liability causal for smoking								
Ever smoked	1.00	0.99, 1.01	0.00	0.003	0.91	383,828	GLM	0.58
Current tobacco smoking	1.01	1.00, 1.02	0.01	0.01	0.12	384,917	OLR	n/a
Past tobacco smoking	1.00	0.99, 1.01	0.00	0.003	0.82	356,231	OLR	n/a
Number of cigarettes currently smoked daily (current cigarette smokers)	n/a	n/a	0.03	0.05	0.60	27,226	LM	n/a
Number of cigarettes previously smoked daily	n/a	n/a	-0.03	0.03	0.37	91,492	LM	n/a

Supplementary table 6-7 Bi-directional genetic risk score analysis. OLR: ordinal logistic regression, GLM: generalized linear model, LM: linear model. AUC: area under the curve

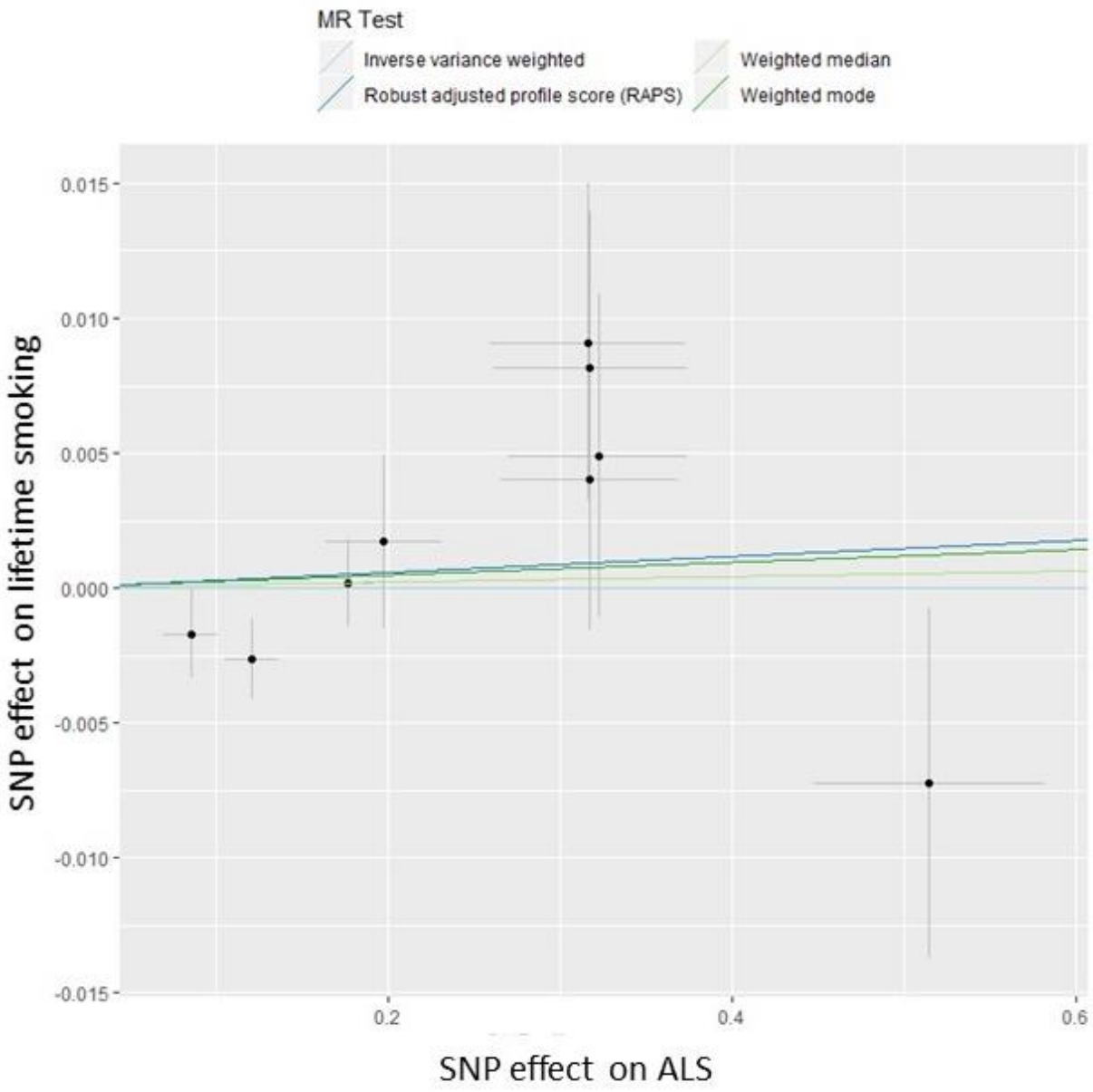
6.11.7. Scatter plots



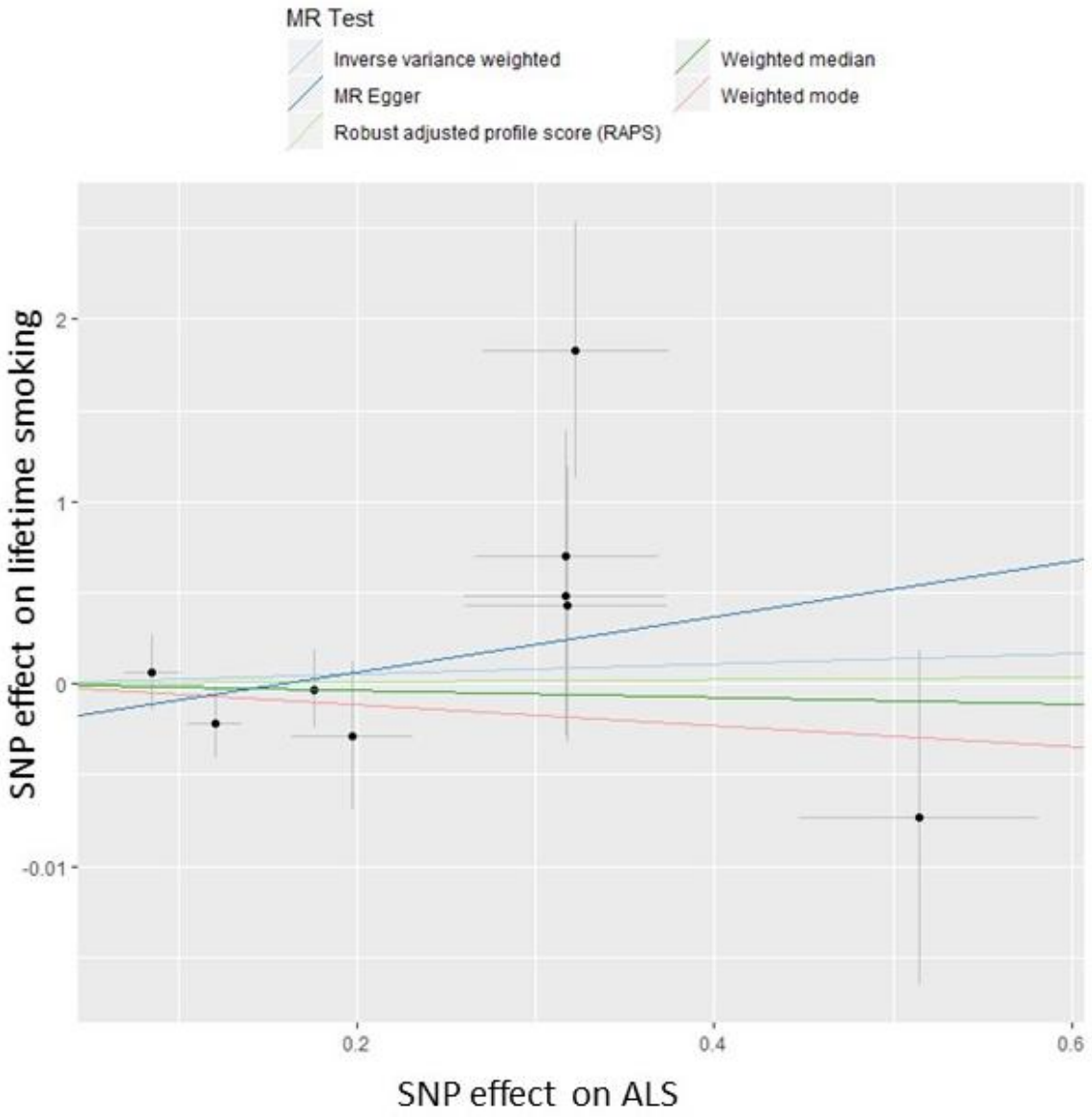
Supplementary figure 6-1 Scatter plot of SNP effect on Lifetime smoking index and ALS



Supplementary figure 6-2 Scatterplot of SNP effect on ever smoking and ALS



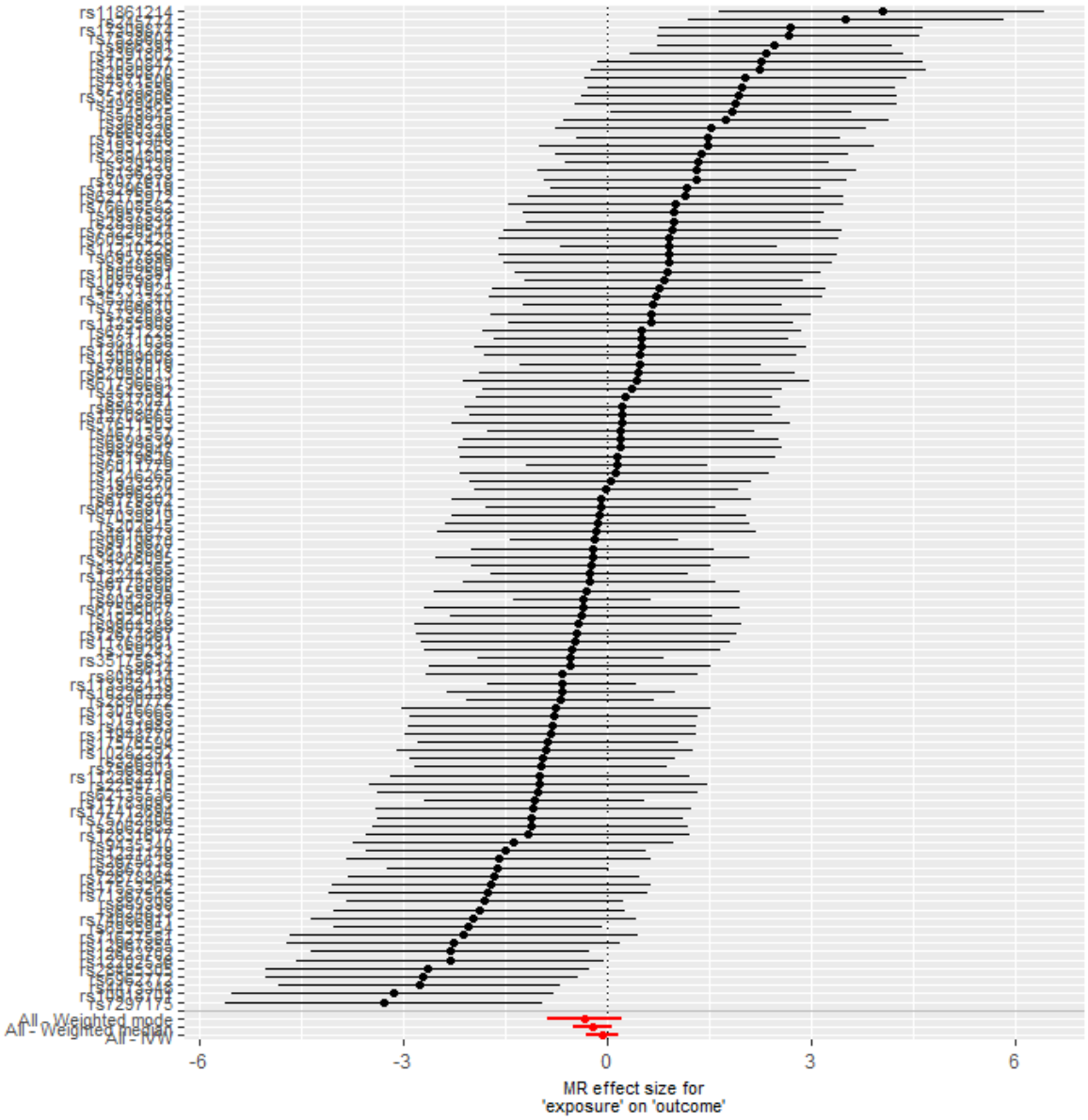
Supplementary figure 6-3 Scatter plot of SNP effect on ALS and lifetime smoking



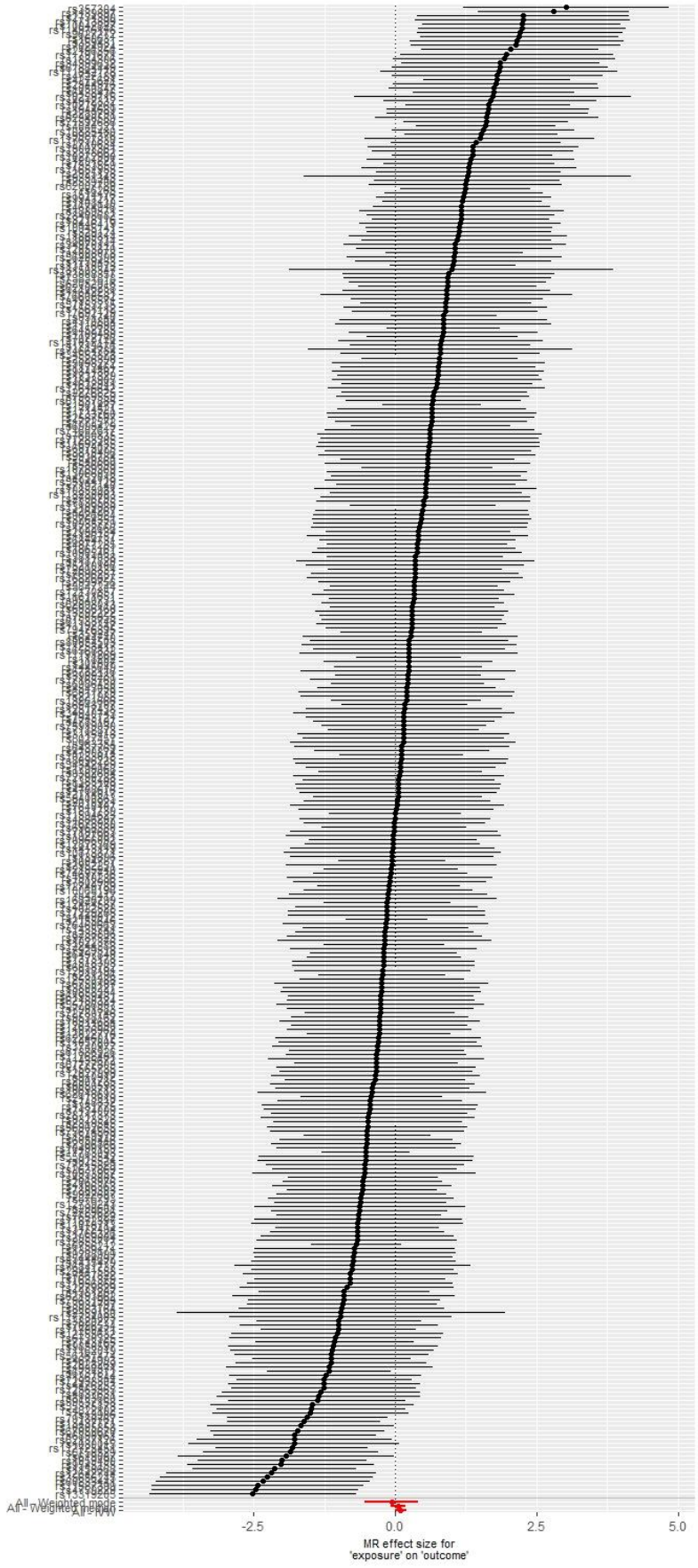
Supplementary figure 6-4 Scatter plot of SNP effects in analysis of ALS liability being causal of ever smoking

6.11.8. *Individual SNP plots*

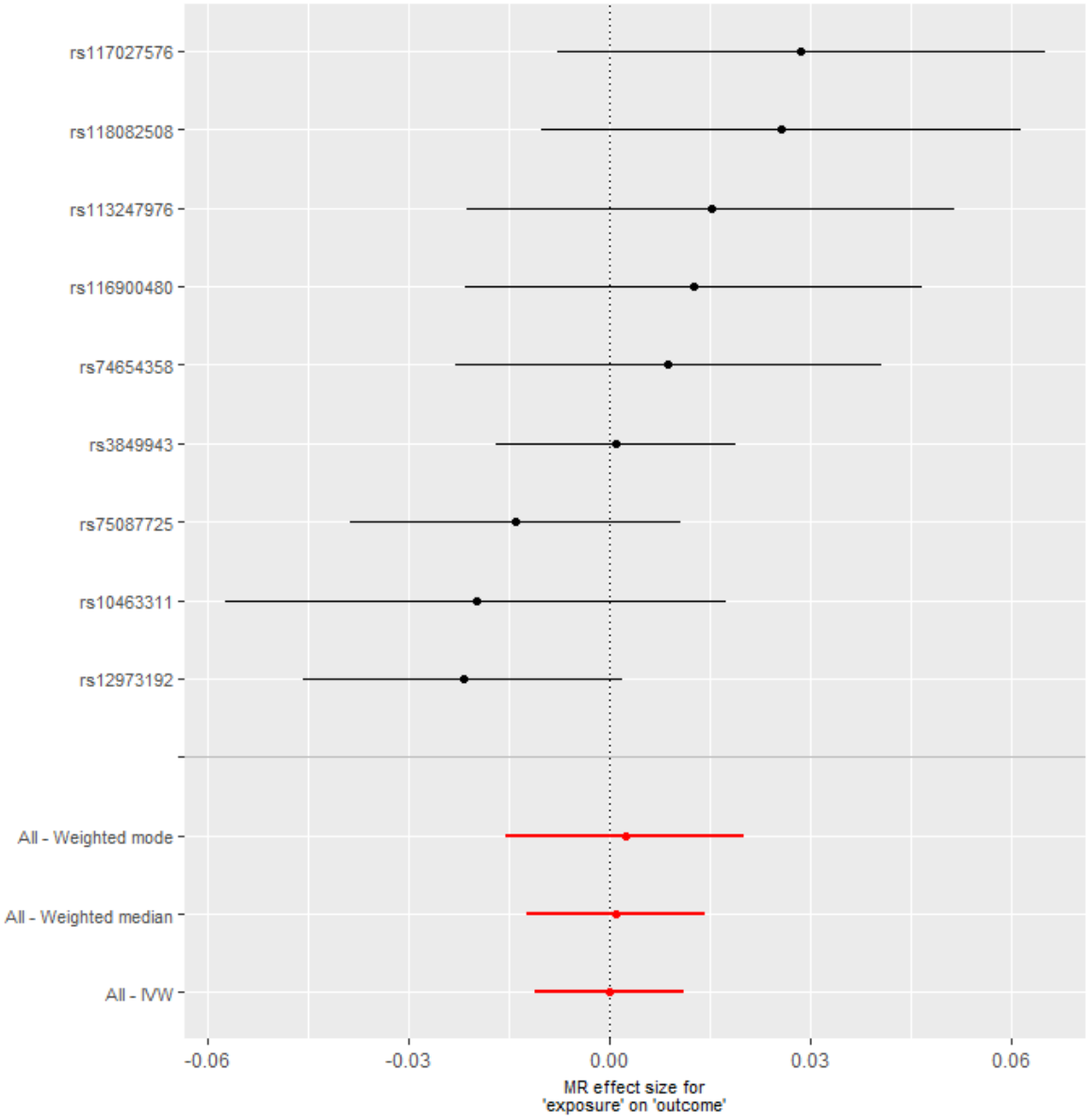
Individual SNP plots show the results of single SNP analyses compared to the methods using multiple SNPs.



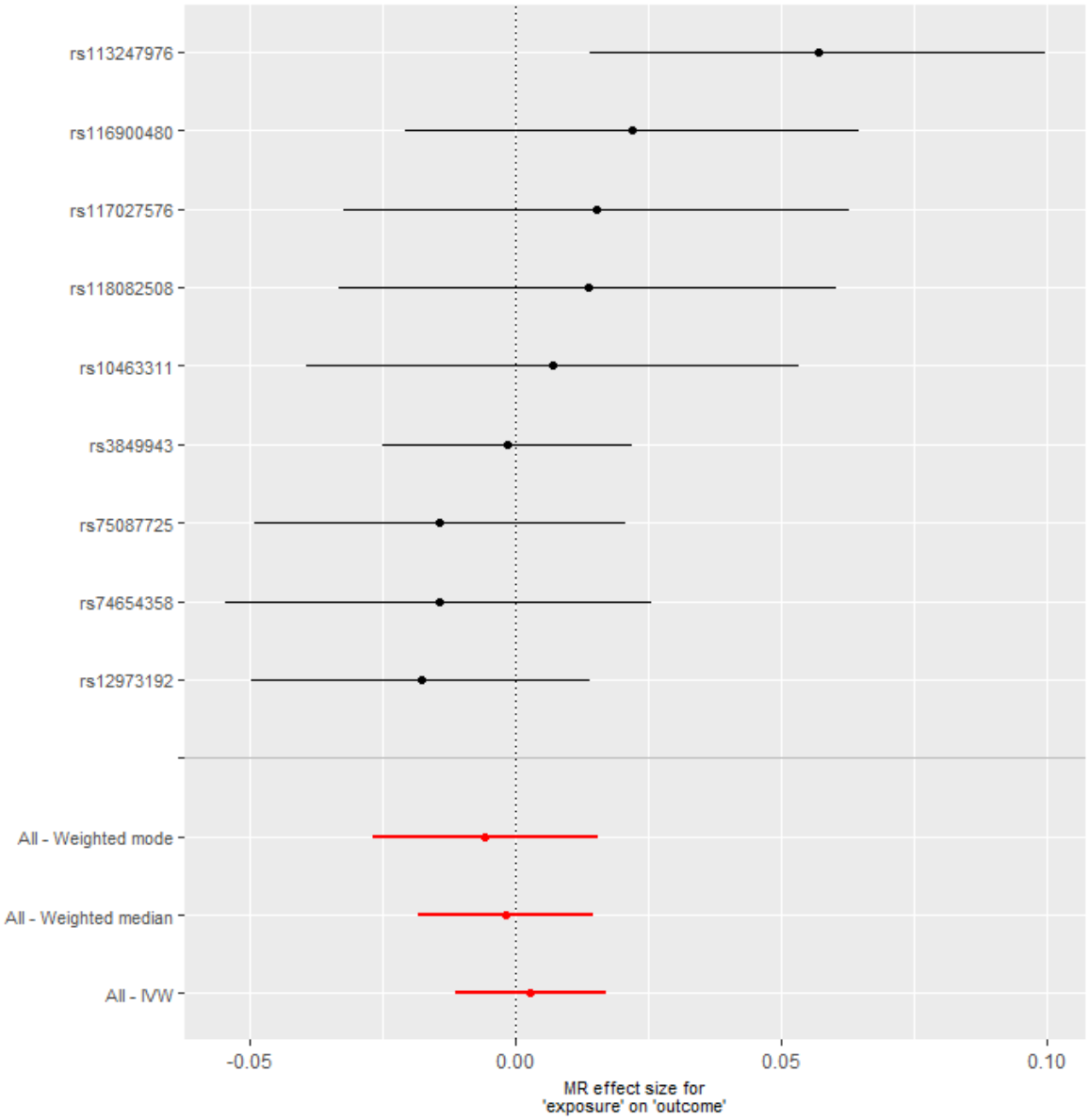
Supplementary figure 6-5 Single SNP analysis of lifetime smoking index and ALS



Supplementary figure 6-6 Single SNP analysis ever smoking and ALS



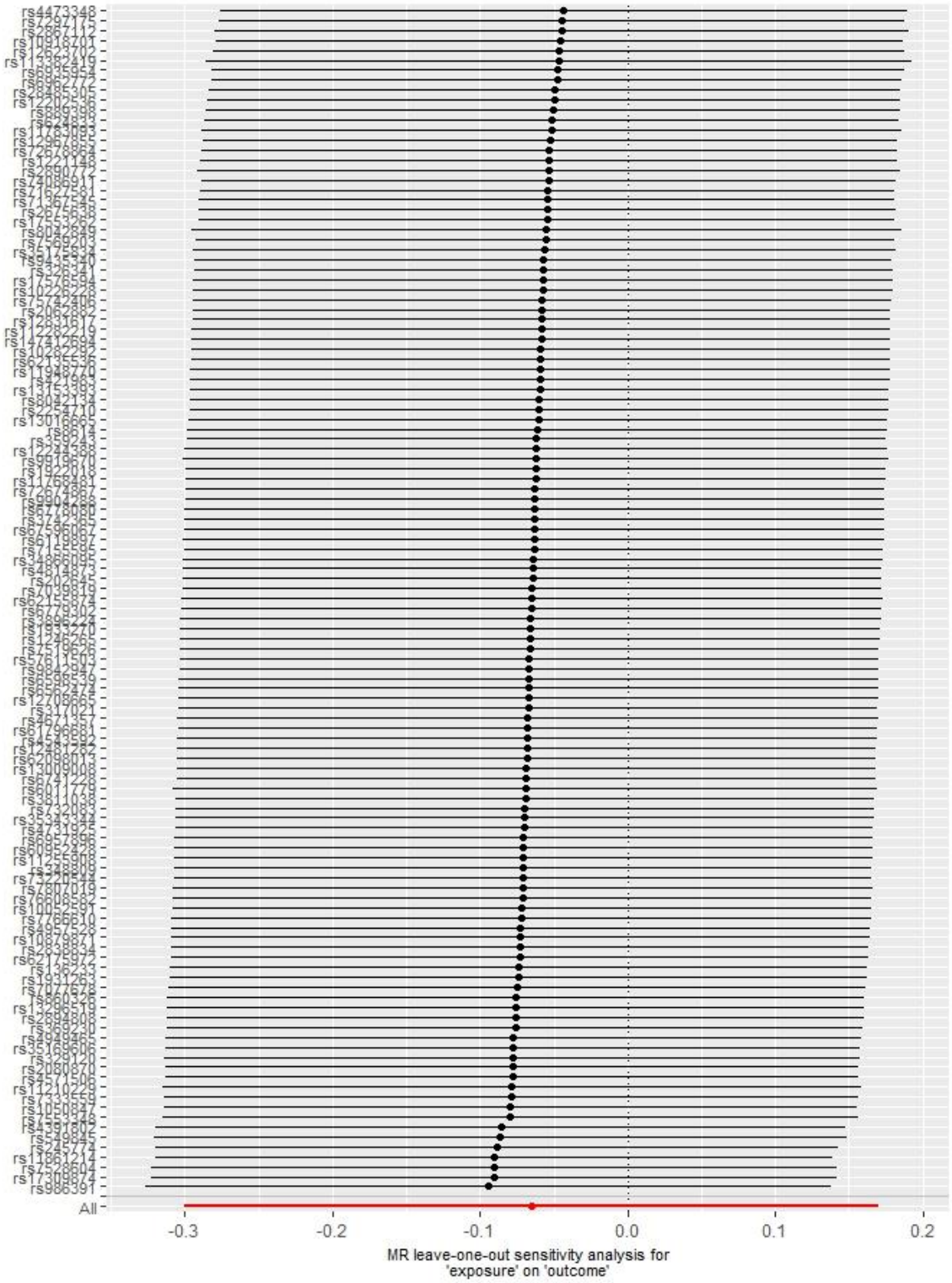
Supplementary figure 6-7 Single SNP analysis of ALS liability and lifetime smoking index



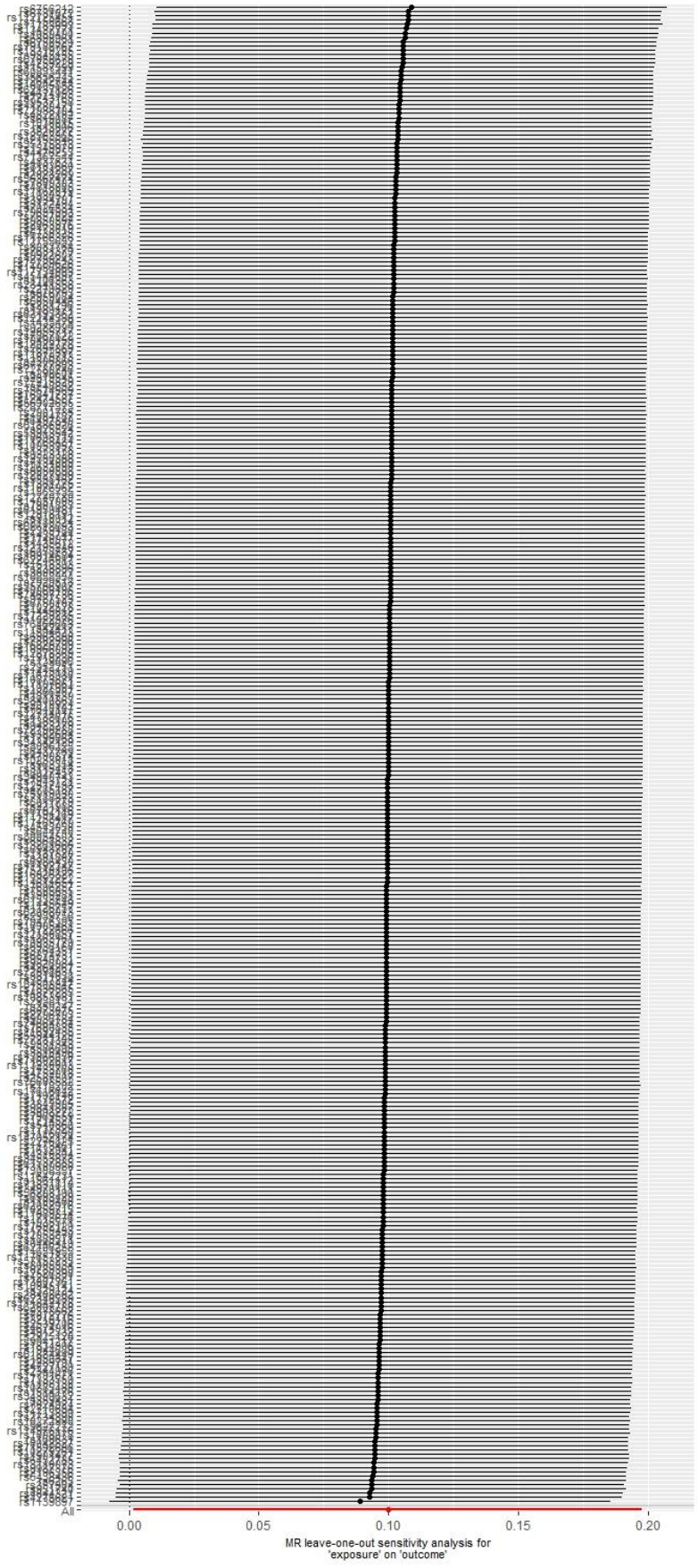
Supplementary figure 6-8 Single SNP analysis of ALS libaility ever smoking

6.11.9. Leave one out analyses

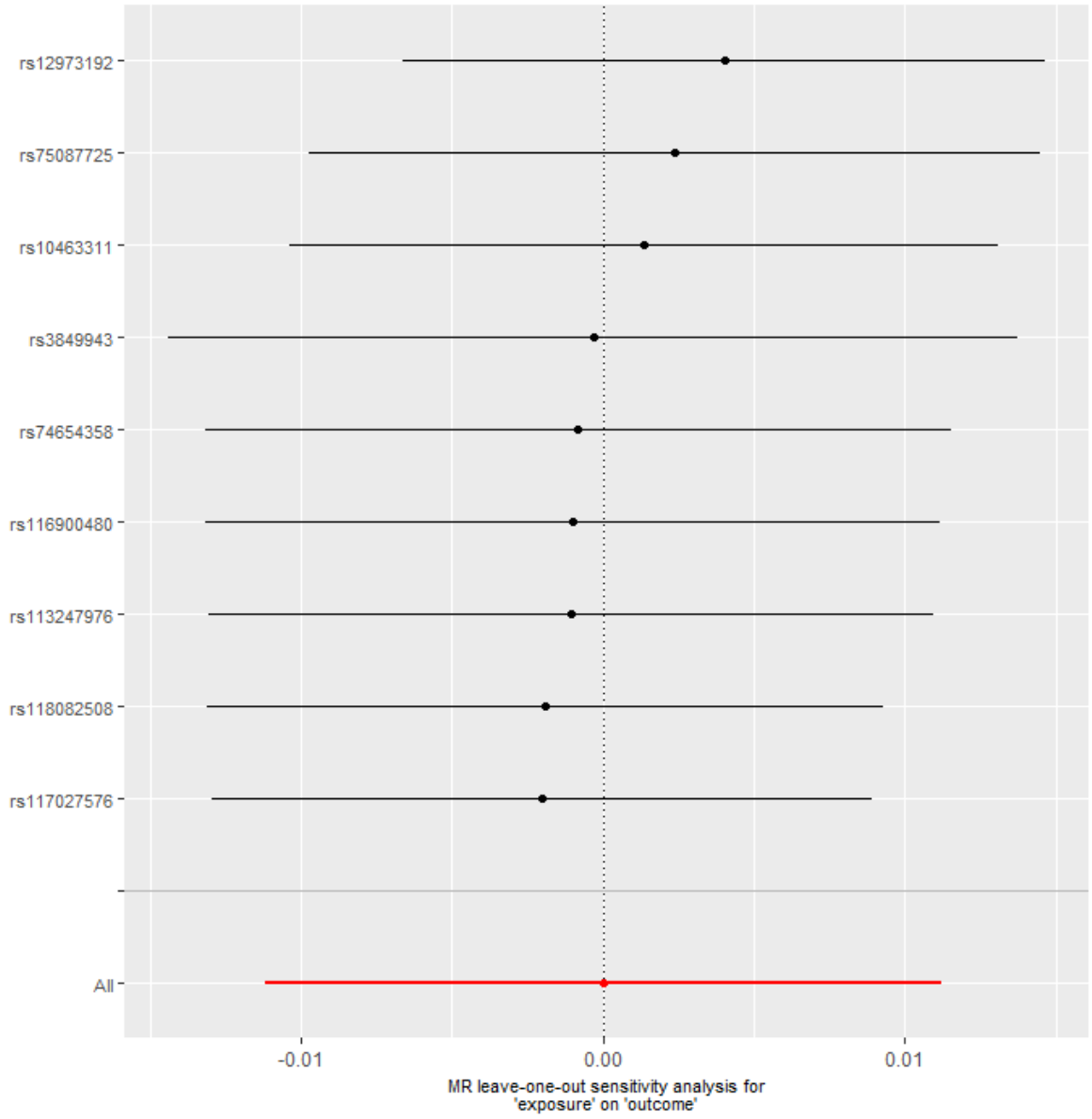
In this analysis the rows represent MR analysis of smoking on ALS using all of the SNPs available in each instrument except for the SNP listed on the y-axis. The point shows the effect size of the analysis with that SNP removed and the line represents the standard error.



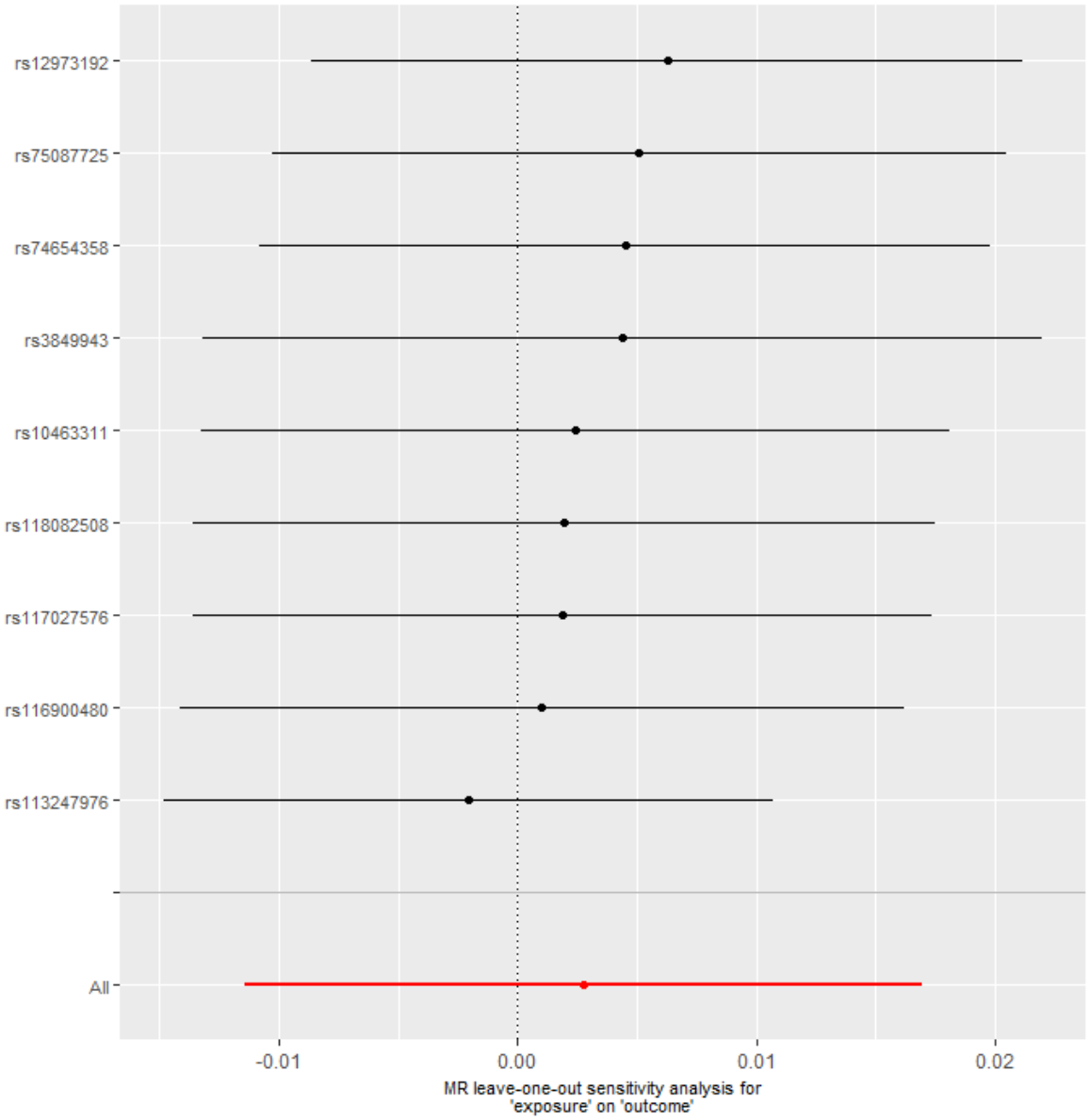
Supplementary figure 6-9 Leave one out analyis of lifetime smoking index on ALS



Supplementary figure 6-10 Leave one out analyis of ever smoking on ALS



Supplementary figure 6-11 Leave one out analysis of ALS liability on CSI



Supplementary figure 6-12 Leave one out analysis of ALS liabilty on ever smoking

Chapter 7 Analysing phenotype by variant in people with *SOD1* ALS

7.1 Introduction

In 1993, variants in the gene *superoxide dismutase 1 (SOD1)* were identified as a causal factor in people with ALS through linkage analysis of 13 different families with 11 different *SOD1* missense mutations (Rosen et al., 1993). *SOD1* variants are reported in 15% of people with familial ALS in European populations and 30% of people with familial ALS in Asian populations, and 1-2% of people with sporadic ALS in both populations (Zou et al., 2017). Since the discovery that variants in *SOD1* can cause an ALS phenotype, over 160 variants throughout the gene have been reported (https://alsod.ac.uk). It is generally assumed all variants reported in *SOD1* are causal variants, with a few exceptions such as the variant N20S (Vela et al., 2012).

Several clinical and demographic factors have been associated with differences in disease progression in ALS. In the ENCALS model of survival prediction the site of onset, age of onset, presence of a *C9orf72* expansion variant, diagnostic category, lung capacity at diagnosis and ALSFRS-R score at diagnosis were all shown to be predictive of ALS survival – although this was a study to optimise a predictive model of ALS rather than determine which variables cause shorter survival (Westeneng et al., 2018). Within the *SOD1* ALS population, certain variants are associated with atypical disease progression. For example, the A5V variant is associated with shorter survival and the D91A variant is associated with longer survival (Bali et al., 2017; Parton et al., 2002). Demographic factors also correlate with survival, for example men with ALS and *SOD1* variants have shorter survival than women (Tang, Ma, Liu, Chen, & Fan, 2019).

SOD1 is commonly expressed in cells of the central nervous system, making up 1-2% of total soluble protein. SOD1 is a cytosolic and mitochondrial antioxidant enzyme, which catalyses the dismutation of superoxide radicals. It may have other roles such as an endoplasmic reticulum activating zinc sensor, a transcription factor and an autophagy regulator (Bunton-Stasyshyn, Saccon, Fratta, & Fisher, 2014). The wild-type SOD1 protein goes through a complex process to maturation, the final step of which is homodimerization and there are many steps along the way that could be disrupted and lead to misfolding. SOD1 aggregates are found in post-mortem CNS tissue of people with *SOD1* ALS and the aggregates are thought to protect against the toxic effects of soluble misfolded SOD1 protein (Gill et al., 2019). Other than the dimer interface, the main functional domains of the enzyme are the electrostatic loop and the zinc binding domain (Galaleldeen et al., 2009). The effect of the change in amino acid by codon location on *SOD1* ALS phenotype in a large dataset has not been previously investigated.

In this study we aimed to collect a large, international dataset of people with ALS that have a recorded *SOD1* variant to inspect the demographic and phenotypic characteristics and analyse survival by variant.

7.2 Methods

7.2.1. Data sources

Data were collected from a variety of sources including: case reports of individuals or families with *SOD1* ALS; anonymised records from specialist ALS centres that perform genetic testing; and entries from the ALS Online Database (https://alsod.ac.uk), a resource that compiles data from studies and submitted by people who treat people with ALS. Additionally, we included data from Project MinE, a global whole genome sequencing project. In the instance of missing data from case reports, corresponding authors were contacted to ask for further information on cases.

7.2.2. Clinical and demographic variables

We collected or requested amino acid change (codon number was assigned taking into account the initial methionine for example variants were labelled as A5V/D91A instead of A4V/D90A). People were eligible if they had a recorded diagnosis of ALS made by a neurologist, or their diagnosis was published as ALS in the literature. Two people had ALS-flail limb, and these were coded as ALS. We collected sex at birth and age of onset in years of first motor symptoms of ALS. Site of onset was coded as bulbar, spinal, respiratory and mixed. We asked whether people had a family history of ALS as reported by their clinician with no specific definition. To record disease progression, we collected or requested the time in months from onset of motor symptoms to diagnosis as well as the months onset to death, or their most recent appointment date. Finally, we asked whether the person had been diagnosed with dementia; this was not specified as being a formal diagnosis of frontotemporal dementia.

7.2.3. Annotation of amino acid changes

Amino acids that are within 6Å of the dimer interface were classed as being within the dimer interface. The codons making up the electrostatic loop and dimer interface were defined according to those amino acids identified as being in those areas according to the literature (Galaleldeen et al., 2009). If the codon was in the dimer interface and the electrostatic loop or the zinc loop, they were classified in those locations rather than in the dimer interface. The codon numbers and their corresponding location are shown in Table 7-1.

Location	Codon number
Dimer interface	4-10, 18-20, 50-55, 60-62, 112-116, 148-154

Electrostatic loop	122-143
Zinc loop	51-84

Table 7-1 Codon numbers by functional protein region

7.2.4. Statistical analysis

Data were analysed and graphs made in R using the packages 'ggplot2', 'rworldmap' and 'surv'. The methods used to analyse the survival distributions and whether variant location had an effect on survival were Kaplan-Meier analysis, the log-rank test and Cox proportional hazards regression (detailed in section 3.7).

7.3 Results

Once data were cleaned, there were 1,383 cases, each with a non-synonymous variant we were able to analyse. Of these, 1086 had information on disease duration available. For more details please see consort diagram (Figure 7-1).

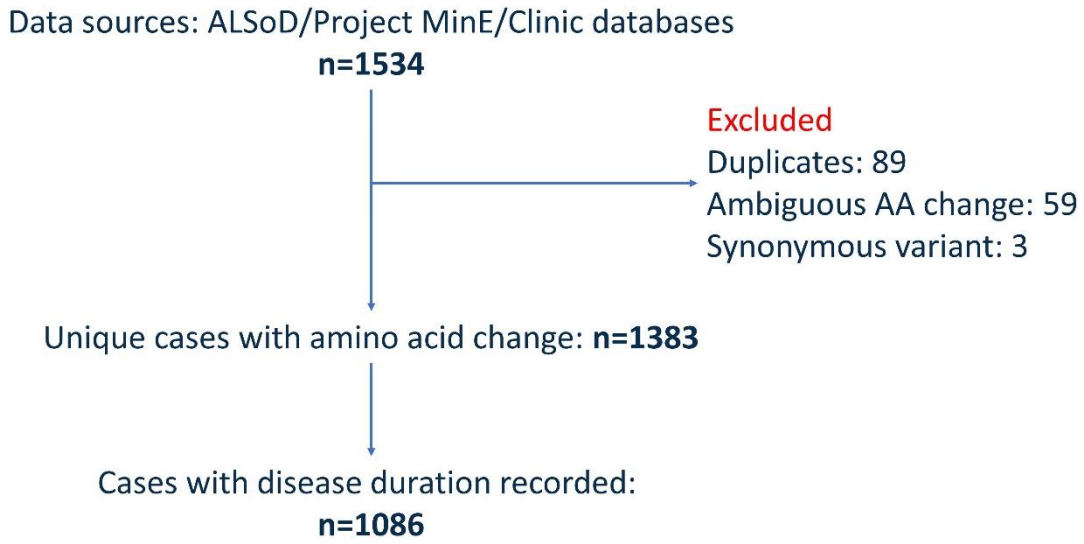


Figure 7-1 Consort diagram of people included in the study

Modified consort diagram showing number of records in the dataset and numbers of records excluded.

Variable		Total n = 1383	Percent	Total n = 1086	Percent
Diagnosis	ALS (incl flail limb)	1366	98.7	1071	98.6
	PLS	1	0.1	1	0.1
	PMA	16	1.2	14	1.3
Site of onset	Spinal	1026	74.2	842	77.5
	Bulbar	108	7.8	93	8.6
	Mixed	8	0.58	6	0.55
	Respiratory	8	0.58	6	0.55
	Not recorded	233	16.8	139	12.8
Mean age of onset years		48.9 (12.8)	NA	49.2 (12.6)	NA
Gender	Female : Male : Not recorded	655 : 726 : 2	47.4 : 52.5 : 0.1	524 : 564 : 1	48.2 : 51.7 : 0.1
Family history	Yes : No : Not recorded	969 : 185 : 229	70.1 : 13.4 : 16.5	798 : 144 : 144	73.4 : 13.3 : 13.3
Median diagnostic delay months (723 missing)		10 (19.3)	NA	10 (24.9)	NA
Median disease duration months		27.7 (61.0)	NA	27.7 (61.0)	NA
Dead	Yes : No	861: 522	62.3 :37.7	828:258	76.2:23.8

Table 7-2 Demographic features of people with SOD1 ALS.

The table includes the full dataset, and the characteristics of those people in the dataset where disease duration was available for all records.

Table 7-2 shows the clinical and demographic variables collected as part of the dataset. The percentages of cases in each category are similar between then the larger dataset of complete records and the smaller dataset of people with disease duration recorded.

Country	Number of records
Australia	27
Austria	5
Belgium	6
Bulgaria	4
Canada	1
China	89
Finland	9
France	50
Germany	2
Hungary	4
Iran (Islamic Republic of)	3
Ireland	2
Italy	68
Japan	41
Korea (the Republic of)	11
Netherlands (the)	7
Pakistan	3
Poland	11
Portugal	3
Russia	9
Slovenia	8
Spain	18
Sweden	8

Taiwan (Province of China)	3
Turkey	66
United Kingdom of Great Britain and Northern Ireland (the)	49
United States of America (the)	542
Not recorded	37

Table 7-3 Number of records by country

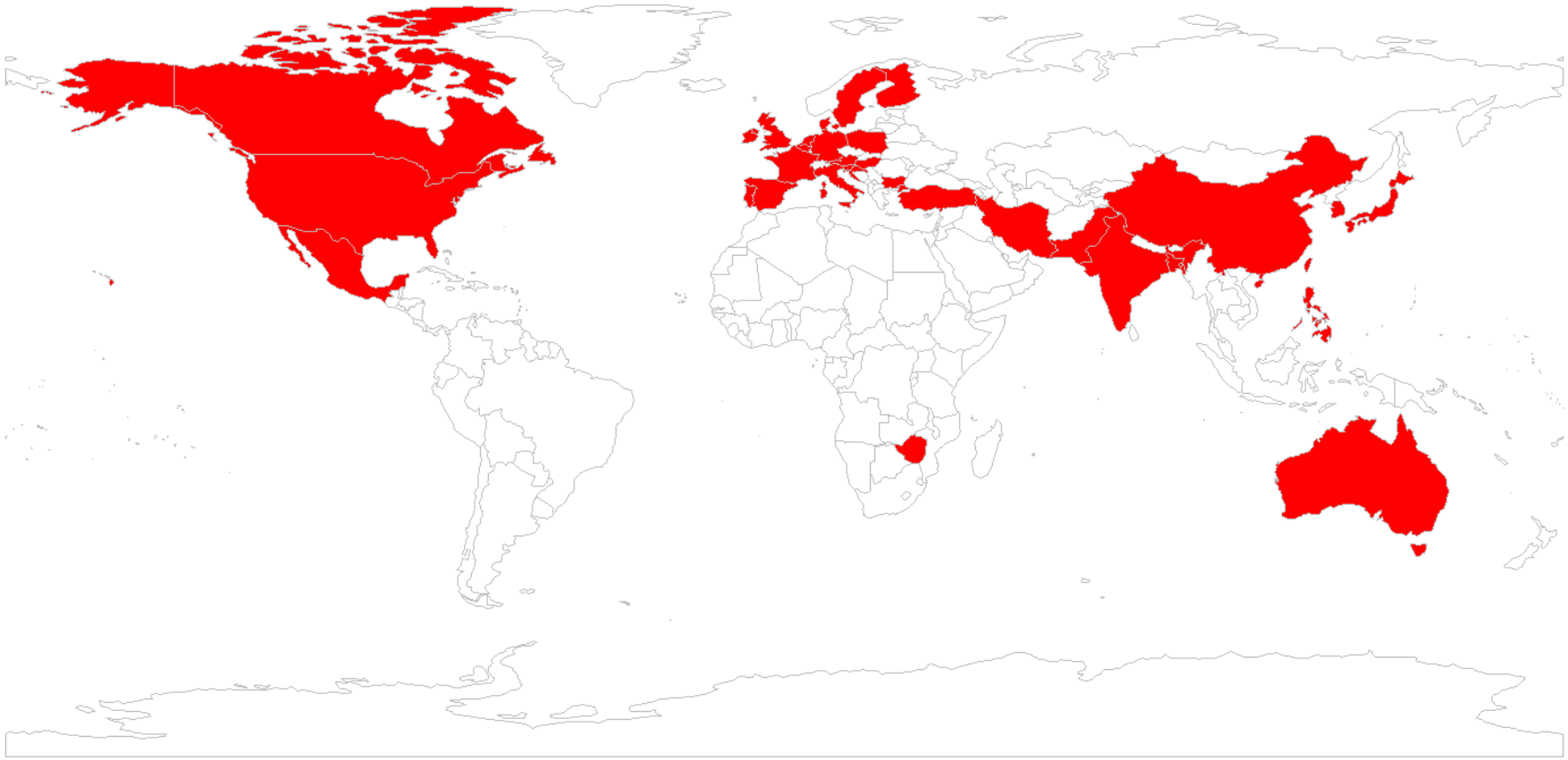


Figure 7-2 World map showing the 34 countries data were obtained from.

Data from 34 countries were included in the dataset. The majority of cases came from US medical datasets, making up almost half of the cases in the dataset with survival data. Accordingly, the A5V variant is highly represented in the dataset.

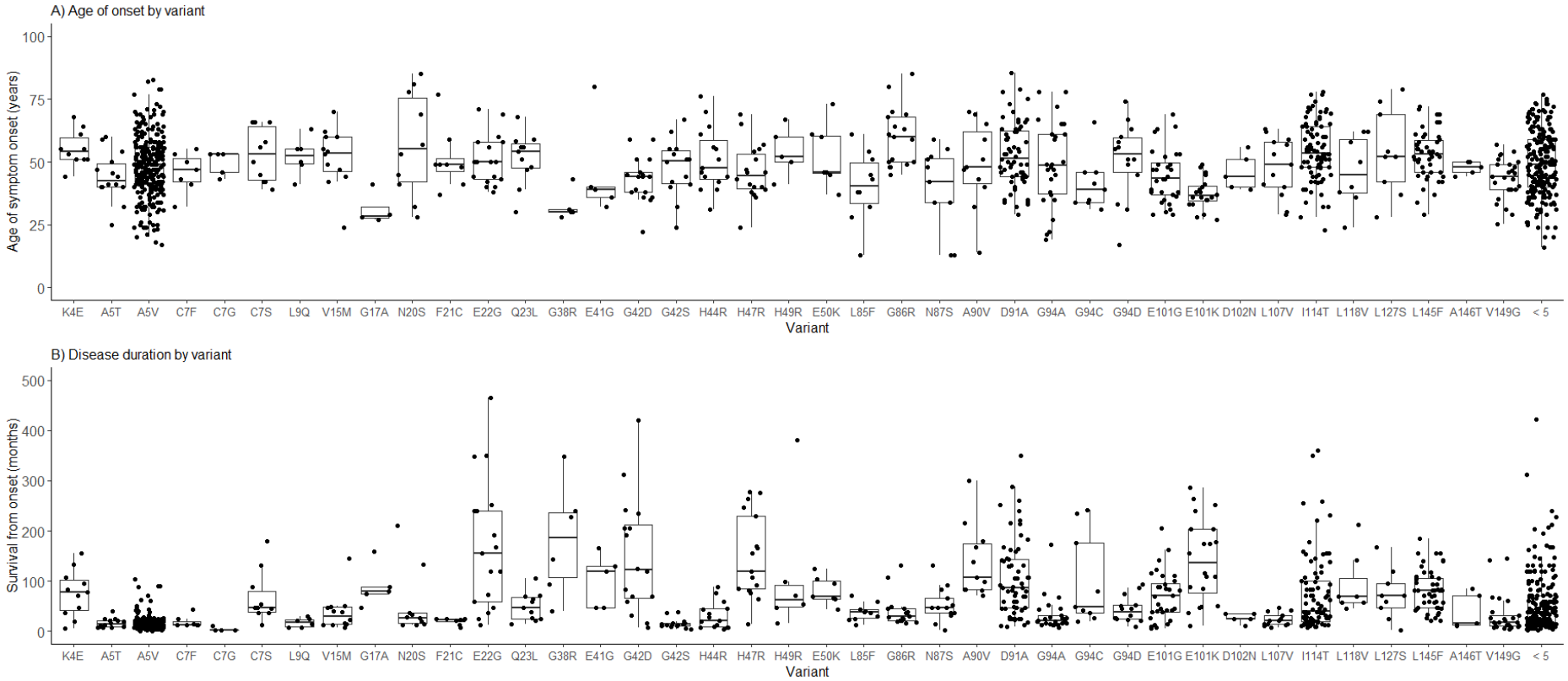


Figure 7-3 Box plots of age of onset and disease duration by variant

a) box plot showing age of onset by variant b) box plot of survival by variant for those variants where there were >5 cases. Both graphs were created with the dataset of people with complete duration data only (n=1086).

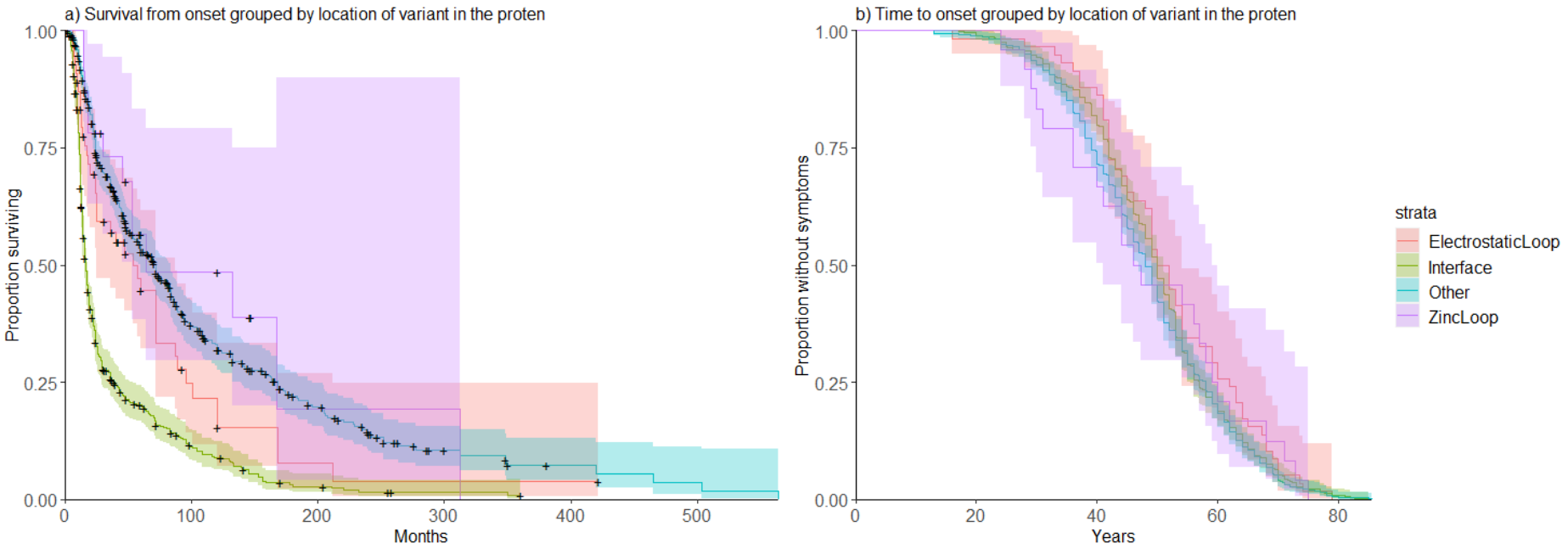


Figure 7-4 Kaplan-Meier curves of survival and time to onset of symptoms compared by location of variant in the SOD1 protein

a) survival in months from onset of symptoms with cases grouped by location in the variant, b) time to onset in years group by location in the variant.

Functional location	Number of records	Median survival (months)	Median age of onset (years)
Electrostatic loop	59	37	51
Dimer interface	478	15	50
Other	525	51	48
Zinc loop	24	50	46

Table 7-4 Median survival and age of onset by functional location of codon

Log rank tests to determine whether there was a difference between groups showed an effect for survival ($p < 0.001$) but not for age of onset ($p = 0.4$). Running a Cox proportional hazards model to quantify the contribution of location to the hazard ratio and control for age, sex, site of onset and whether someone has an A5V variant gave the following results as shown in table 7-5

Variable	Hazard ratio (se) p-value		
	Duration dataset	Duration dataset (time-varying covariate adjusted)	Duration dataset no A5V
Electrostatic loop	1.30 (0.18) 0.17	0.95 (0.75) 0.30	1.30 (0.19) 0.17
Interface	1.29 (0.1) 0.01	1.44 (0.35) 0.60	1.28 (0.1) 0.02
Zinc loop	0.87 (0.28) 0.6	0.56 (1.1) 0.66	0.89 (0.68)
Other	1	1	1

Table 7-5 Cox PH results of functional location by survival

On testing the proportional hazards assumption, this is violated for functional location, with a log(-log) plot showing the two lines crossing (zinc loop and other). Testing the Schoenfeld residuals with time found a significant relationship. Re-running the model with location of variant as a time-dependent covariate changed both their estimates and the p-values.

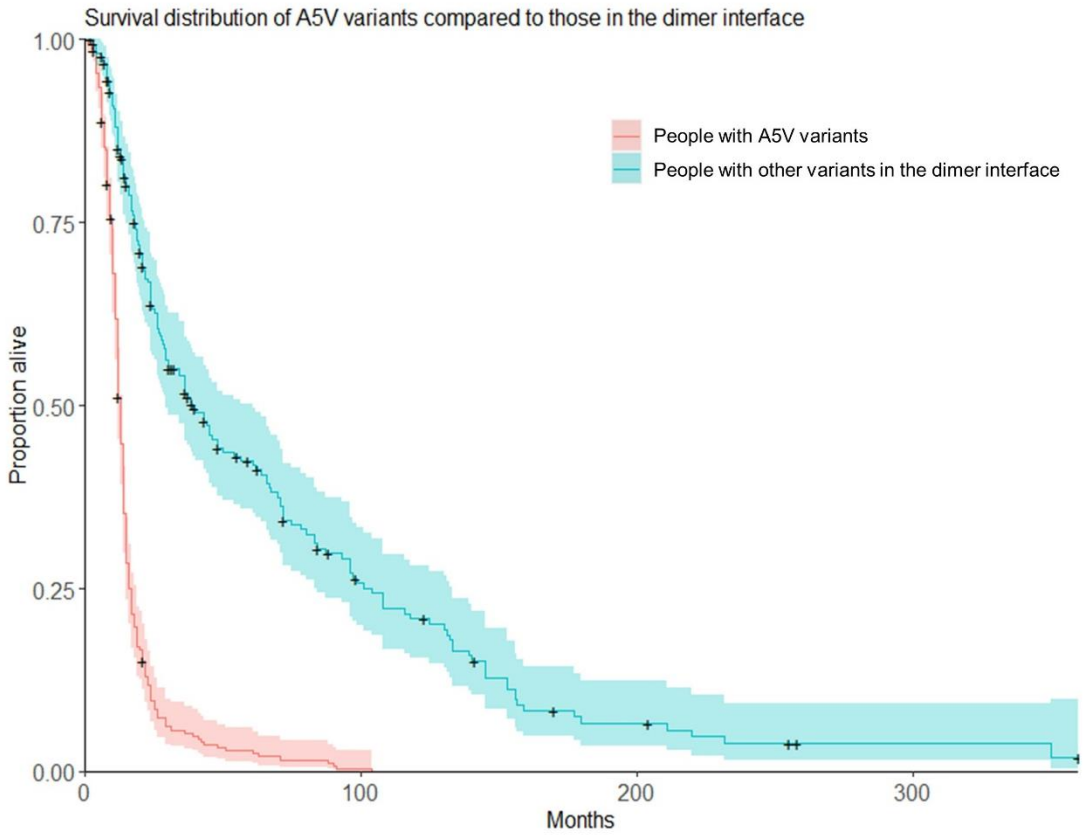


Figure 7-5 Kaplan-Meier curve comparing survival distribution of A5V variants with other variants in the dimer interface

Median survival for people with A5V variants was 12 months, compared to 29 months for people with variants in all other parts of the dimer interface, log rank test p-value < 0.001.

7.4 Discussion

In this study we report on the most comprehensive dataset of ALS cases with an *SOD1* variant ever recorded. *SOD1* ALS does appear to have a distinct phenotypic profile compared to sporadic ALS: there is a higher proportion of people with classic ALS, almost 99% of cases were given this diagnosis, although relative motor neuron contribution was not requested.

Additionally, there was a higher proportion of limb onset, 75% compared to 60%, which is consistent with evidence from Italy whereby *SOD1* variants increased the likelihood of limb onset (Chiò et al., 2020). There were more men than women with *SOD1* ALS, which would not be expected from a disease with high genetic load that is not X-linked. *SOD1* is not 100% penetrant, and in multistep model analysis there are two remaining steps, so *SOD1* does not fully account for risk. It is possible that whatever risk is associated with being male still has some effect. Average age of onset is lower than in sporadic ALS, by about 15 years, but not much lower than those people with ALS of a genetic cause (Mehta et al., 2019). As would be expected from a dataset of people with a genetic risk factor, there is a high proportion of people with familial disease, although there are still some people with no recorded family history. Differences from the general ALS population may be reflective of different cellular pathological changes in *SOD1* ALS – for example TDP-43 pathology is not observed (Mackenzie et al., 2007).

I have investigated whether grouping variants by the site of the protein they are in is an informative way of classifying variants. There is an association with survival time, where variants in the dimer interface and electrostatic loop are associated with with shorter survival time, as shown on a Kaplan-Meier plot. The Cox proportional hazards model estimates should not be considered an indication of the true hazard, as there is violation of the proportional hazards assumption and a different, parametric model may be more appropriate to quantify the effect of location on survival time. However, the violation of proportional hazards is an interesting finding, as it is possible that the variant has an effect on survival differently at different times during the disease. If it has a higher effect later on for example, then starting treatment as soon as symptoms start may have more of an effect than later on. This should be tested by further modelling.

In addition to the violation of proportional hazards, it is not clear whether the dimer interface as a location overall is important, or whether that area happens to be the site of a few very pathological

variants. Since A5V has poor prognosis and is located in the dimer interface we might expect the survival distribution of other variants in the dimer interface to have the same survival distribution as A5V if codon location is important, this is not the case as shown in figure 7-5. It is possible there is an interaction between location and other properties of the changed amino acid. In Figure 7-3, the shorter survival groups seem to be at the N terminal of the protein, and particularly, changes at residues 5 and 7 look severely pathological. There are four different variants at codon position 7 and variants of Cys-7 have been shown to cause protein misfolding (Leinartaitė and Johansson, 2013; Toichi, Yamanaka, & Furukawa, 2013).

There does not appear to be a correlation with location of the variant and age of onset, adding to evidence that onset and disease process are two different mechanisms. As shown in figure 7-3 the age of onset looks much more uniform between variants than the survival time.

Those variants that change survival are likely to show the highest benefit in clinical trials of antisense oligonucleotides where progression time is an outcome. The results of trials will therefore likely vary by variant population and this should be considered in interpretation, as well as trying to balance variants on randomisation. If anti-sense oligonucleotides are used to prevent disease onset the particular variant change may affect outcomes differently. As people born with no functioning SOD1 protein tend to have severely adverse outcomes, it may not be safe to administer anti-sense oligonucleotides from a young age for disease protection (Andersen et al., 2019). In addition, the tolerability of spinal delivery multiple times per year for gene therapy strategies that require this, and the costs involved, would need to be assessed.

There are several disadvantages to this study. Of the 150 variants in the dataset of cases with disease duration data, 80 only have one or two records, so there are too few in each category to analyse the effect of variants at each codon. The majority of the data are from the US dataset and the A5V variant is therefore over-represented. The A5V variant has such a strong effect on survival that this effect will skew the survival data. The dataset is not population-based so results are likely not generalisable to the whole *SOD1* ALS population. There are many cases from specialist centres or that were published because they represented the identification of a new variant and subsequent cases may not have been considered interesting enough to publish. The data definitions were not tightly specified and we only analysed missense mutations rather than structural genetic variants. There is substantial missing data, and work should be done trying to impute the data.

Despite these limitations, this study remains the largest and most comprehensive analysis of the relationship between genotypic variation and phenotype for any gene in ALS, and one of the largest in any neurodegenerative disease. Although *C9orf72* phenotype has been extensively studied, that

represents a single variant in a gene rather than the many variants studied here. The insights and approach used show that similar methods could be applied to other Mendelian causes of ALS and are likely to yield insights about mechanism and genetic clinical trial interpretation.

Chapter 8 Conclusions and further studies

The estimated incidence of MND in the UK from MND Register data is similar to previous estimates from the UK and incidence reported in other European population registers. The MND Register is organised to collect data from a variety of sources including specialist centres, general neurology clinics and hospices, as well as allowing self-registration by people with ALS. Although the data represents a subset of the total area the register aims to cover, and there are some issues with missing data, the design seems to facilitate collection of accurate population level data. Collecting data from a variety of sources in such a densely populated area, combined with data from other registers in Scotland, Ireland and in other areas of Europe means the MND Register will become a hugely important resource. In the future we will investigate prevalence, update estimates of lifetime risk of ALS and geographical patterns of spread.

I have performed an observational study on the risk of smoking and ALS, followed by instrumental variable analysis and polygenic risk score analysis to triangulate evidence and investigate causality. The results strongly indicate that smoking is not a risk factor for the general ALS population. This does not exclude the possibility that there is a gene-environment interaction in a sub-group. Smoking has been shown to affect disease progression, which may make smoking an important explanatory variable in studies of prognosis. It is possible that environmental factors in ALS have a similar effect on risk as is thought to be the case in ALS genetics where a small number of exposures have medium effect on risk. It may not be possible to detect risk factors using individual study and environmental-wide association studies may be more appropriate (Patel, Bhattacharya, & Butte, 2010). Unlike in the case of genome-wide association studies on single nucleotide polymorphisms, the coding of environmental exposures would be very challenging. To use two-sample MR to test causality in newly identified subgroups of ALS, GWAS studies would need to be repeated on these subgroups, and this may be possible with Project MinE data. It may also be possible to run a much larger scale polygenic risk score analysis to re-run the logistic regression with polygenic risk score as a covariate presented in this thesis. A further study on smoking and risk of ALS would be to use methylation information to proxy smoking exposure and compare risk between smokers and non-smokers by presence of *CHRNA5* variation that predicts smoking intensity, a study that would also be possible using Project MinE data.

Using a dataset with phenotype and genotype information about people with *SOD1* ALS, I have found that survival distribution in people with *SOD1* ALS, may be related to the location of the variant in the protein. The location of the variant within the protein does not appear to affect age of onset, implying that disease onset and risk are not equally affected by variant in *SOD1* ALS.

Subgrouping in ALS is probably one of the factors that would accelerate the understanding and treatment of the disease and it appears that even within an already small subgroup of people with ALS there are further subgroups. It is also likely the case that there should be separate subgroups for risk and progression.

The location of the variant in the protein may affect the survival benefit elicited from a trial so the likely makeup of people entering trials should be considered in power calculations. In *SOD1* trials, randomisation to match genetic variant may also be useful, although it is possible that this would slow recruitment.

References

Abe, K., Aoki, M., Tsuji, S., Itoyama, Y., Sobue, G., Togo, M., . . . Yoshino, H. (2017). Safety and efficacy of edaravone in well defined patients with amyotrophic lateral sclerosis: a randomised, double-blind, placebo-controlled trial. *The Lancet Neurology, 16*(7), 505-512. doi:10.1016/S1474-4422(17)30115-1

Abhinav, K., Al-Chalabi, A., Hortobagyi, T., & Leigh, P. N. (2007a). Electrical injury and amyotrophic lateral sclerosis: a systematic review of the literature. *J Neurol Neurosurg Psychiatry, 78*(5), 450-453. doi:10.1136/jnnp.2006.104414

Abhinav, K., Stanton, B., Johnston, C., Hardstaff, J., Orrell, R. W., Howard, R., . . . Al-Chalabi, A. (2007b). Amyotrophic lateral sclerosis in South-East England: a population-based study. The South-East England register for amyotrophic lateral sclerosis (SEALS Registry). *Neuroepidemiology, 29*(1-2), 44-48. doi:10.1159/000108917

Abramzon, Y. A., Fratta, P., Traynor, B. J., & Chia, R. (2020). The Overlapping Genetics of Amyotrophic Lateral Sclerosis and Frontotemporal Dementia. *14*(42). doi:10.3389/fnins.2020.00042

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (pp. 199-213). New York, NY: Springer New York.

Al-Chalabi, A. (2017). Perspective: Don't keep it in the family. *Nature, 550*(7676), S112-S112. doi:10.1038/550S112a

Al-Chalabi, A., Calvo, A., Chio, A., Colville, S., Ellis, C. M., Hardiman, O., . . . Pearce, N. (2014). Analysis of amyotrophic lateral sclerosis as a multistep process: a population-based modelling study. *Lancet Neurol, 13*(11), 1108-1113. doi:10.1016/S1474-4422(14)70219-4

Al-Chalabi, A., Fang, F., Hanby, M. F., Leigh, P. N., Shaw, C. E., Ye, W., & Rijsdijk, F. (2010). An estimate of amyotrophic lateral sclerosis heritability using twin data. *Journal of Neurology, Neurosurgery & Psychiatry, 81*(12), 1324. doi:10.1136/jnnp.2010.207464

Al-Chalabi, A., & Hardiman, O. (2013). The epidemiology of ALS: a conspiracy of genes, environment and time. *Nat Rev Neurol, 9*(11), 617-628. doi:10.1038/nrneurol.2013.203

Al-Chalabi, A., Hardiman, O., Kiernan, M. C., Chiò, A., Rix-Brooks, B., & van den Berg, L. H. (2016). Amyotrophic lateral sclerosis: moving towards a new classification system. *Lancet Neurol, 15*(11), 1182-1194. doi:10.1016/s1474-4422(16)30199-5

Alonso, A., Logroscino, G., & Hernán, M. A. (2010a). Smoking and the risk of amyotrophic lateral sclerosis: a systematic review and meta-analysis. *Journal of Neurology, Neurosurgery & Psychiatry*, 1249-1252.

Alonso, A., Logroscino, G., Jick, S. S., & Hernán, M. A. (2010b). Association of smoking with amyotrophic lateral sclerosis risk and survival in men and women: a prospective study. *BMC neurology, 10*(1), 6.

Althubaiti, A. (2016). Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of multidisciplinary healthcare, 9*, 211-217. doi:10.2147/JMDH.S104807

Altman, D. G. (1990). *Practical statistics for medical research*: CRC press.

Amado, D. A., & Davidson, B. L. (2021). Gene therapy for ALS: A review. *Molecular Therapy*. doi:https://doi.org/10.1016/j.ymthe.2021.04.008

Andersen, P. M., Nordström, U., Tsiakas, K., Johannsen, J., Volk, A. E., Bierhals, T., . . . Santer, R. (2019). Phenotype in an Infant with SOD1 Homozygous Truncating Mutation. *N Engl J Med, 381*(5), 486-488. doi:10.1056/NEJMc1905039

Andrews, J. A., Jackson, C. E., Heiman-Patterson, T. D., Bettica, P., Brooks, B. R., & Pioro, E. P. (2020). Real-world evidence of riluzole effectiveness in treating amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 21*(7-8), 509-518. doi:10.1080/21678421.2020.1771734

Angrist, J. D. (1990). Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *The American Economic Review, 80*(3), 313-336.

Antao, V. C., & Horton, D. K. (2012). The National Amyotrophic Lateral Sclerosis (ALS) Registry. *Journal of environmental health, 75*(1), 28-30.

Armon, C. (2009). Smoking may be considered an established risk factor for sporadic ALS. *Neurology, 73*(20), 1693-1698. doi:10.1212/WNL.0b013e3181c1df48

Arthur, K. C., Calvo, A., Price, T. R., Geiger, J. T., Chiò, A., & Traynor, B. J. (2016). Projected increase in amyotrophic lateral sclerosis from 2015 to 2040. *Nature Communications, 7*, 12408-12408. doi:10.1038/ncomms12408

Balendra, R., Al Khleifat, A., Fang, T., & Al-Chalabi, A. (2019). A standard operating procedure for King's ALS clinical staging. *Amyotroph Lateral Scler Frontotemporal Degener, 20*(3-4), 159-164. doi:10.1080/21678421.2018.1556696

Balendra, R., Jones, A., Jivraj, N., Knights, C., Ellis, C. M., Burman, R., . . . Al-Chalabi, A. (2014). Estimating clinical stage of amyotrophic lateral sclerosis from the ALS Functional Rating Scale. *Amyotroph Lateral Scler Frontotemporal Degener, 15*(3-4), 279-284. doi:10.3109/21678421.2014.897357

Bali, T., Self, W., Liu, J., Siddique, T., Wang, L. H., Bird, T. D., . . . Miller, T. M. (2017). Defining SOD1 ALS natural history to guide therapeutic clinical trial design. *J Neurol Neurosurg Psychiatry, 88*(2), 99-105. doi:10.1136/jnnp-2016-313521

Bandres‐Ciga, S., Noyce, A. J., Hemani, G., Nicolas, A., Calvo, A., Mora, G., . . . Traynor, B. J. (2019). Shared polygenic risk and causal inferences in amyotrophic lateral sclerosis. *Annals of Neurology, 85*(4), 470-481. doi:10.1002/ana.25431

Beard, J. D., & Kamel, F. (2015). Military service, deployments, and exposures in relation to amyotrophic lateral sclerosis etiology and survival. *Epidemiol Rev, 37*(1), 55-70. doi:10.1093/epirev/mxu001

Bedlack, R., & Hardiman, O. (2009). ALSUntangled (ALSU): A Scientific Approach to Off-Label Treatment Options for People with ALS Using Tweets and Twitters. *Amyotrophic Lateral Sclerosis, 10*(3), 129-130. doi:10.1080/17482960903015986

Bedlack, R., Vaughan, T., Wicks, P., Heywood, J., Sinani, E., Selsov, R., . . . Sherman, A. (2016). How common are ALS plateaus and reversals? *Neurology, 86*(9), 808-812. doi:10.1212/WNL.0000000000002251

Bensimon, G., Lacomblez, L., & Meininger, V. (1994). A Controlled Trial of Riluzole in Amyotrophic Lateral Sclerosis. *330*(9), 585-591. doi:10.1056/nejm199403033300901

Bladen, C. L., Thompson, R., Jackson, J. M., Garland, C., Wegel, C., Ambrosini, A., . . . Lochmüller, H. (2014). Mapping the differences in care for 5,000 Spinal Muscular Atrophy patients, a survey of 24 national registries in North America, Australasia and Europe. *Journal of Neurology, 261*(1), 152-163. doi:10.1007/s00415-013-7154-1

Bland, J. M., & Altman, D. G. (1998). Survival probabilities (the Kaplan-Meier method). *Bmj, 317*(7172), 1572. doi:10.1136/bmj.317.7172.1572

Bland, J. M., & Altman, D. G. (2000). The odds ratio. *320*(7247), 1468. doi:10.1136/bmj.320.7247.1468 %J BMJ

Blecher, R., Elliott, M. A., Yilmaz, E., Dettori, J. R., Oskouian, R. J., Patel, A., . . . Chapman, J. R. (2019). Contact Sports as a Risk Factor for Amyotrophic Lateral Sclerosis: A Systematic Review. *Global spine journal, 9*(1), 104-118. doi:10.1177/2192568218813916

Bodner, T. E. (2008). What Improves with Increased Missing Data Imputations? *Structural Equation Modeling: A Multidisciplinary Journal, 15*(4), 651-675. doi:10.1080/10705510802339072

Boekestein, W. A., Kleine, B. U., Hageman, G., Schelhaas, H. J., & Zwarts, M. J. (2010). Sensitivity and specificity of the 'Awaji' electrodiagnostic criteria for amyotrophic lateral sclerosis: retrospective comparison of the Awaji and revised El Escorial criteria for ALS. *Amyotroph Lateral Scler, 11*(6), 497-501. doi:10.3109/17482961003777462

Bourke, S. C., Tomlinson, M., Williams, T. L., Bullock, R. E., Shaw, P. J., & Gibson, G. J. (2006). Effects of non-invasive ventilation on survival and quality of life in patients with amyotrophic lateral sclerosis: a randomised controlled trial. *Lancet Neurol, 5*(2), 140-147. doi:10.1016/s1474-4422(05)70326-4

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol, 44*(2), 512-525. doi:10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., & Burgess, S. (2016a). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol, 40*(4), 304-314. doi:10.1002/gepi.21965

Bowden, J., Del Greco, M., Minelli, C., Davey Smith, G., Sheehan, N. A., & Thompson, J. R. (2016b). Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int J Epidemiol, 45*(6), 1961-1974. doi:10.1093/ije/dyw220

Bowden, J., Del Greco, M., Minelli, C., Davey Smith, G., Sheehan, N., & Thompson, J. (2017). A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in Medicine, 36*(11), 1783-1802. doi:10.1002/sim.7221

Bowden, J., Hemani, G., & Davey Smith, G. (2018). Invited Commentary: Detecting Individual and Global Horizontal Pleiotropy in Mendelian Randomization—A Job for the Humble Heterogeneity Statistic? *Am J Epidemiol, 187*(12), 2681-2685. doi:10.1093/aje/kwy185 %J American Journal of Epidemiology

Bowling, A. C., Schulz, J. B., Brown, R. H., Jr., & Beal, M. F. (1993). Superoxide dismutase activity, oxidative damage, and mitochondrial energy metabolism in familial and sporadic amyotrophic lateral sclerosis. *J Neurochem, 61*(6), 2322-2325. doi:10.1111/j.1471-4159.1993.tb07478.x

Bozzoni, V., Pansarasa, O., Diamanti, L., Nosari, G., Cereda, C., & Ceroni, M. (2016). Amyotrophic lateral sclerosis and environmental factors. *Funct Neurol, 31*(1), 7-19. doi:10.11138/fneur/2016.31.1.007

Brooks, B. R. (1994). El escorial World Federation of Neurology criteria for the diagnosis of amyotrophic lateral sclerosis. *Journal of the Neurological Sciences, 124*, 96-107. doi:https://doi.org/10.1016/0022-510X(94)90191-0

Brooks, B. R., Crumpacker, D., Fellus, J., Kantor, D., & Kaye, R. E. (2013). PRISM: a novel research tool to assess the prevalence of pseudobulbar affect symptoms across neurological conditions. *PLoS One, 8*(8), e72232. doi:10.1371/journal.pone.0072232

Brooks, B. R., Miller, R. G., Swash, M., & Munsat, T. L. (2000). El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Other Motor Neuron Disord, 1*(5), 293-299. doi:10.1080/146608200300079536

Brown, R. H., & Al-Chalabi, A. (2017). Amyotrophic Lateral Sclerosis. *New England Journal of Medicine, 377*(2), 162-172. doi:10.1056/NEJMra1603471

Bunton-Stasyshyn, R. K. A., Saccon, R. A., Fratta, P., & Fisher, E. M. C. (2014). SOD1 Function and Its Implications for Amyotrophic Lateral Sclerosis Pathology: New and Renascent Themes. *The Neuroscientist, 21*(5), 519-529. doi:10.1177/1073858414561795

Burgess, S., & Thompson, S. G. (2011). Avoiding bias from weak instruments in Mendelian randomization studies. *Int J Epidemiol, 40*(3), 755-764. doi:10.1093/ije/dyr036

Byrne, S., Elamin, M., Bede, P., & Hardiman, O. (2012). Absence of consensus in diagnostic criteria for familial neurodegenerative diseases. *J Neurol Neurosurg Psychiatry, 83*(4), 365-367. doi:10.1136/jnnp-2011-301530

Byrne, S., Walsh, C., Lynch, C., Bede, P., Elamin, M., Kenna, K., . . . Hardiman, O. (2011). Rate of familial amyotrophic lateral sclerosis: a systematic review and meta-analysis. *Journal of Neurology, Neurosurgery & Psychiatry, 82*(6), 623. doi:10.1136/jnnp.2010.224501

Calvo, A., Canosa, A., Bertuzzo, D., Cugnasco, P., Solero, L., Clerico, M., . . . Chiò, A. (2016). Influence of cigarette smoking on ALS outcome: a population-based study. *J Neurol Neurosurg Psychiatry, 87*(11), 1229-1233. doi:10.1136/jnnp-2016-313793

Cedarbaum, J. M., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B., & Nakanishi, A. (1999). The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. *Journal of the Neurological Sciences, 169*(1), 13-21. doi:10.1016/S0022-510X(99)00210-5

Charcot, J. (1874). Amyotrophies spinales deuteropathiques sclérose latérale amyotrophique & Sclérose latérale amyotrophique. *2*(Oeuvres Complétes)), 234-266.

Chiò, A., Calvo, A., Dossena, M., Ghiglione, P., Mutani, R., & Mora, G. (2009a). ALS in Italian professional soccer players: The risk is still present and could be soccer-specific. *Amyotrophic Lateral Sclerosis, 10*(4), 205-209. doi:10.1080/17482960902721634

Chiò, A., Calvo, A., Ghiglione, P., Mazzini, L., Mutani, R., & Mora, G. (2010). Tracheostomy in amyotrophic lateral sclerosis: a 10-year population-based study in Italy. *Journal of Neurology, Neurosurgery & Psychiatry, 81*(10), 1141. doi:10.1136/jnnp.2009.175984

Chiò, A., Hammond, E., Mora, G., Bonito, V., & Filippini, G. (2015). Development and evaluation of a clinical staging system for amyotrophic lateral sclerosis. *86*(1), 38-44. doi:10.1136/jnnp-2013-306589 Journal of Neurology, Neurosurgery & Psychiatry

Chiò, A., Logroscino, G., Hardiman, O., Swingler, R., Mitchell, D., Beghi, E., & Traynor, B. J. (2009b). Prognostic factors in ALS: A critical review. *Amyotroph Lateral Scler, 10*(5-6), 310-323. doi:10.3109/17482960802566824

Chio, A., Logroscino, G., Traynor, B. J., Collins, J., Simeone, J. C., Goldstein, L. A., & White, L. A. (2013). Global epidemiology of amyotrophic lateral sclerosis: a systematic review of the published literature. *Neuroepidemiology, 41*(2), 118-130. doi:10.1159/000351153

Chiò, A., Logroscino, G., Traynor, B. J., Collins, J., Simeone, J. C., Goldstein, L. A., & White, L. A. (2013). Global epidemiology of amyotrophic lateral sclerosis: a systematic review of the published literature. *Neuroepidemiology, 41*(2), 118-130. doi:10.1159/000351153

Chiò, A., Mazzini, L., D'Alfonso, S., Corrado, L., Canosa, A., Moglia, C., . . . Al-Chalabi, A. (2018). The multistep hypothesis of ALS revisited: The role of genetic mutations. *Neurology, 91*(7), e635-e642. doi:10.1212/WNL.0000000000005996

Chiò, A., Moglia, C., Canosa, A., Manera, U., Ovidio, F., Vasta, R., . . . Calvo, A. (2020). ALS phenotype is influenced by age, sex, and genetics. *Neurology, 94*(8), e802. doi:10.1212/WNL.0000000000008869

Choi, S. W., Mak, T. S.-H., & O'Reilly, P. F. (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols, 15*(9), 2759-2772. doi:10.1038/s41596-020-0353-1

Coggon, D., Rose, G. A., & Barker, D. J. P. (2003). *Epidemiology for the uninitiated* (5th ed ed.). London: BMJ.

Cole, S. R., Platt, R. W., Schisterman, E. F., Chu, H., Westreich, D., Richardson, D., & Poole, C. (2010). Illustrating bias due to conditioning on a collider. *Int J Epidemiol, 39*(2), 417-420. doi:10.1093/ije/dyp334

Cox, D. R. (1972). Regression models and life‐tables. *Journal of the Royal Statistical Society: Series B, 34*(2), 187-202.

Critchley, J. (2004). Epidemiology for the uninitiated, 5th ed. *Journal of Epidemiology and Community Health, 58*(11), 963.

Cruz, M. P. (2013). Nuedexta for the treatment of pseudobulbar affect: a condition of involuntary crying or laughing. *P & T : a peer-reviewed journal for formulary management, 38*(6), 325-328.

D'Amico, E., Factor-Litvak, P., Santella, R. M., & Mitsumoto, H. (2013). Clinical Perspective of Oxidative Stress in Sporadic ALS. *Free radical biology & medicine, 65*, 509-527. doi:10.1016/j.freeradbiomed.2013.06.029

D'Gama, A. M., & Walsh, C. A. (2018). Somatic mosaicism and neurodevelopmental disease. *Nature Neuroscience, 21*(11), 1504-1514. doi:10.1038/s41593-018-0257-3
Davey Smith, G., & Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol, 32*(1), 1-22. doi:10.1093/ije/dyg070
Davey Smith, G., & Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet, 23*(R1), R89-98. doi:10.1093/hmg/ddu328
Davey Smith, G., Lawlor, D. A., Harbord, R., Timpson, N., Day, I., & Ebrahim, S. (2007). Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Med, 4*(12), e352. doi:10.1371/journal.pmed.0040352
Davies, N. M., Howe, L. J., Brumpton, B., Havdahl, A., Evans, D. M., & Davey Smith, G. (2019). Within family Mendelian randomization studies. *Hum Mol Genet, 28*(R2), R170-r179. doi:10.1093/hmg/ddz204
de Carvalho, M., Dengler, R., Eisen, A., England, J. D., Kaji, R., Kimura, J., . . . Swash, M. (2008). Electrodiagnostic criteria for diagnosis of ALS. *Clin Neurophysiol, 119*(3), 497-503. doi:10.1016/j.clinph.2007.09.143
de Carvalho, M., Kiernan, M. C., & Swash, M. (2017). Fasciculation in amyotrophic lateral sclerosis: origin and pathophysiological relevance. *J Neurol Neurosurg Psychiatry, 88*(9), 773-779. doi:10.1136/jnnp-2017-315574
de Jong, S. W., Huisman, M. H. B., Sutedja, N. A., van der Kooi, A. J., de Visser, M., Schelhaas, H. J., . . . van den Berg, L. H. (2012). Smoking, Alcohol Consumption, and the Risk of Amyotrophic Lateral Sclerosis: A Population-based Study. *Am J Epidemiol, 176*(3), 233-239. doi:10.1093/aje/kws015
de Jongh, A. D., van Eijk, R. P. A., & van den Berg, L. H. (2019). Evidence for a multimodal effect of riluzole in patients with ALS? *Journal of Neurology, Neurosurgery & Psychiatry, 90*(10), 1183. doi:10.1136/jnnp-2018-320211
Delzor, A., Couratier, P., Boumédiène, F., Nicol, M., Druet-Cabanac, M., Paraf, F., . . . Marin, B. (2014). Searching for a link between the L-BMAA neurotoxin and amyotrophic lateral sclerosis: a study protocol of the French BMAALS programme. *BMJ Open, 4*(8), e005528. doi:10.1136/bmjopen-2014-005528
Desport, J. C., Preux, P. M., Truong, T. C., Vallat, J. M., Sautereau, D., & Couratier, P. (1999). Nutritional status is a prognostic factor for survival in ALS patients. *Neurology, 53*(5), 1059-1063. doi:10.1212/wnl.53.5.1059
Dharmadasa, T., Matamala, J. M., Howells, J., Vucic, S., & Kiernan, M. C. (2020). Early focality and spread of cortical dysfunction in amyotrophic lateral sclerosis: A regional study across the motor cortices. *Clin Neurophysiol, 131*(4), 958-966. doi:10.1016/j.clinph.2019.11.057
Dickerson, A. S., Hansen, J., Kioumourtzoglou, M.-A., Specht, A. J., Gredal, O., & Weisskopf, M. G. (2018). Study of occupation and amyotrophic lateral sclerosis in a Danish cohort. *Occupational and Environmental Medicine, 75*(9), 630. doi:10.1136/oemed-2018-105110
Donaghy, C., Clarke, J., Patterson, C., Kee, F., Hardiman, O., & Patterson, V. (2010). The epidemiology of motor neuron disease in Northern Ireland using capture-recapture methodology. *Amyotroph Lateral Scler, 11*(4), 374-378. doi:10.3109/17482960903329569
Enders, C. K. (2010). *Applied missing data analysis*: Guilford press.
Falconer, D. S. (1996). *Introduction to quantitative genetics*: Pearson Education India.
Fang, F., Bellocco, R., Hernán, M. A., & Ye, W. (2006). Smoking, snuff dipping and the risk of amyotrophic lateral sclerosis–a prospective cohort study. *Neuroepidemiology, 27*(4), 217-221.
Fang, T., Al Khleifat, A., Meurgey, J. H., Jones, A., Leigh, N., Bensimon, G., & Al-Chalabi, A. (2018). Stage at which riluzole treatment prolongs survival in patients with amyotrophic lateral

sclerosis: a retrospective analysis of data from a dose-ranging study. *The Lancet Neurology, 17*(5), 416-422. doi:10.1016/S1474-4422(18)30054-1
Fang, T., Jozsa, F., & Al-Chalabi, A. (2017). Nonmotor Symptoms in Amyotrophic Lateral Sclerosis: A Systematic Review. *International review of neurobiology, 134*, 1409.
Ferrari, R., Kapogiannis, D., Huey, E. D., & Momeni, P. (2011). FTD and ALS: a tale of two diseases. *Current Alzheimer research, 8*(3), 273-294. doi:10.2174/156720511795563700
Fischer, H., Kheifets, L., Huss, A., Peters, T. L., Vermeulen, R., Ye, W., . . . Feychting, M. (2015). Occupational Exposure to Electric Shocks and Magnetic Fields and Amyotrophic Lateral Sclerosis in Sweden. *Epidemiology, 26*(6), 824-830. doi:10.1097/ede.0000000000000365
Forbes, R. B., Colville, S., Parratt, J., & Swingler, R. J. (2007). The incidence of motor neuron disease in Scotland. *J Neurol, 254*(7), 866-869. doi:10.1007/s00415-006-0454-y
Galaleldeen, A., Strange, R. W., Whitson, L. J., Antonyuk, S. V., Narayana, N., Taylor, A. B., . . . Hart, P. J. (2009). Structural and biophysical properties of metal-free pathogenic SOD1 mutants A4V and G93A. *Archives of Biochemistry and Biophysics, 492*(1), 40-47. doi:https://doi.org/10.1016/j.abb.2009.09.020
Gallo, V., Bueno-De-Mesquita, H. B., Vermeulen, R., Andersen, P. M., Kyrozis, A., Linseisen, J., . . . Riboli, E. (2009). Smoking and risk for amyotrophic lateral sclerosis: analysis of the EPIC cohort. *Ann Neurol, 65*(4), 378-385. doi:10.1002/ana.21653
Ganesalingam, J., Stahl, D., Wijesekera, L., Galtrey, C., Shaw, C. E., Leigh, P. N., & Al-Chalabi, A. (2009). Latent Cluster Analysis of ALS Phenotypes Identifies Prognostically Differing Groups. *PLoS One, 4*(9), e7107. doi:10.1371/journal.pone.0007107
Garton, F. C., Trabjerg, B. B., Wray, N. R., & Agerbo, E. (2020). Cardiovascular disease, psychiatric diagnosis and sex differences in the multistep hypothesis of amyotrophic lateral sclerosis. *European Journal of Neurology*. doi:https://doi.org/10.1111/ene.14554
Gill, C., Phelan, J. P., Hatzipetros, T., Kidd, J. D., Tassinari, V. R., Levine, B., . . . Vieira, F. G. (2019). SOD1-positive aggregate accumulation in the CNS predicts slower disease progression and increased longevity in a mutant SOD1 mouse model of ALS. *Scientific Reports, 9*(1), 6724. doi:10.1038/s41598-019-43164-z
Gowland, A., Opie-Martin, S., Scott, K. M., Jones, A. R., Mehta, P. R., Batts, C. J., . . . Al-Chalabi, A. (2019). Predicting the future of ALS: the impact of demographic change and potential new treatments on the prevalence of ALS in the United Kingdom, 2020–2116. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 20*(3-4), 264-274. doi:10.1080/21678421.2019.1587629
Greenland, S., & Neutra, R. (1980). Control of confounding in the assessment of medical technology. *Int J Epidemiol, 9*(4), 361-367. doi:10.1093/ije/9.4.361
Groothuis-Oudshoorn, S. v. B. K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*(3), 1-67.
Group, P. S. (2015). Gastrostomy in patients with amyotrophic lateral sclerosis (ProGas): a prospective cohort study. *The Lancet Neurology, 14*(7), 702-709.
Hardiman, O., Al-Chalabi, A., Brayne, C., Beghi, E., van den Berg, L. H., Chio, A., . . . Rooney, J. (2017). The changing picture of amyotrophic lateral sclerosis: lessons from European registers. *Journal of Neurology, Neurosurgery & Psychiatry, 88*(7), 557-563.
Hardiman, O., & van den Berg, L. H. (2017). Edaravone: a new treatment for ALS on the horizon? *The Lancet Neurology, 16*(7), 490-491. doi:10.1016/S1474-4422(17)30163-1
Harrison, D., Mehta, P., van Es, M. A., Stommel, E., Drory, V. E., Nefussy, B., . . . Bedlack, R. (2018). "ALS reversals": demographics, disease characteristics, treatments, and co-morbidities. *Amyotroph Lateral Scler Frontotemporal Degener, 19*(7-8), 495-499. doi:10.1080/21678421.2018.1457059
Hartwig, F. P., Davey Smith, G., & Bowden, J. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int J Epidemiol, 46*(6), 1985-1998. doi:10.1093/ije/dyx102

Hemani, G., Bowden, J., & Davey Smith, G. (2018a). Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human Molecular Genetics, 27*(R2), R195-R208. doi:10.1093/hmg/ddy163 %J Human Molecular Genetics

Hemani, G., Tilling, K., & Davey Smith, G. (2017). Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLOS Genetics, 13*(11), e1007081. doi:10.1371/journal.pgen.1007081

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., . . . Haycock, P. C. (2018b). The MR-Base platform supports systematic causal inference across the human phenome. *eLife, 7*, e34408. doi:10.7554/eLife.34408

Hinchcliffe, M., & Smith, A. (2017). Riluzole: real-world evidence supports significant extension of median survival times in patients with amyotrophic lateral sclerosis. *Degenerative neurological and neuromuscular disease, 7*, 61-70. doi:10.2147/DNND.S135748

Hirakawa, A., Asano, J., Sato, H., & Teramukai, S. (2018). Master protocol trials in oncology: Review and new trial designs. *Contemp Clin Trials Commun, 12*, 1-8. doi:10.1016/j.conctc.2018.08.009

Huynh, W., Simon, N. G., Grosskreutz, J., Turner, M. R., Vucic, S., & Kiernan, M. C. (2016). Assessment of the upper motor neuron in amyotrophic lateral sclerosis. *Clin Neurophysiol, 127*(7), 2643-2660. doi:10.1016/j.clinph.2016.04.025

Imam, I., Ball, S., Wright, D., Hanemann, C. O., & Zajicek, J. (2010). The epidemiology of motor neurone disease in two counties in the southwest of England. *J Neurol, 257*(6), 977-981. doi:10.1007/s00415-009-5448-0

Johnsen, B., Pugdahl, K., Fuglsang-Frederiksen, A., Kollewe, K., Paracka, L., Dengler, R., . . . de Carvalho, M. (2019). Diagnostic criteria for amyotrophic lateral sclerosis: A multicentre study of inter-rater variation and sensitivity. *Clin Neurophysiol, 130*(2), 307-314. doi:10.1016/j.clinph.2018.11.021

Johnson, T. (2012). Efficient calculation for Multi-SNP genetic risk scores. *American Society of Human Genetics Annual Meeting*.

Johnston, C. A., Stanton, B. R., Turner, M. R., Gray, R., Blunt, A. H.-M., Butt, D., . . . Al-Chalabi, A. (2006). Amyotrophic lateral sclerosis in an urban setting. *Journal of Neurology, 253*(12), 1642-1643.

Johnston, C. A., Stanton, B. R., Turner, M. R., Gray, R., Blunt, A. H., Butt, D., . . . Al-Chalabi, A. (2006). Amyotrophic lateral sclerosis in an urban setting: a population based study of inner city London. *J Neurol, 253*(12), 1642-1643. doi:10.1007/s00415-006-0195-y

Kamel, F., Umbach, D. M., Munsat, T. L., Shefner, J. M., & Sandler, D. P. (1999). Association of cigarette smoking with amyotrophic lateral sclerosis. *Neuroepidemiology, 18*(4), 194-202.

Kang, H., Zhang, A., Cai, T. T., & Small, D. S. (2016). Instrumental Variables Estimation With Some Invalid Instruments and its Application to Mendelian Randomization. *Journal of the American Statistical Association, 111*(513), 132-144. doi:10.1080/01621459.2014.994705

Kaye, W. E., Wagner, L., Wu, R., & Mehta, P. (2018). Evaluating the completeness of the national ALS registry, United States. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 19*(1-2), 112-117. doi:10.1080/21678421.2017.1384021

Kenna, K. P., van Doormaal, P. T., Dekker, A. M., Ticozzi, N., Kenna, B. J., Diekstra, F. P., . . . Landers, J. E. (2016). NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nat Genet, 48*(9), 1037-1042. doi:10.1038/ng.3626

Keyfitz, N. (1966). 3. Sampling variance of standardized mortality rates. *Hum Biol, 38*(3), 309-317.

Khalaf, R., Martin, S., Ellis, C., Burman, R., Sreedharan, J., Shaw, C., . . . Al-Chalabi, A. (2019). Relative preservation of triceps over biceps strength in upper limb-onset ALS: the 'split elbow'. *Journal of Neurology, Neurosurgery & Psychiatry, 90*(7), 730-733. doi:10.1136/jnnp-2018-319894

Kiernan, M. C., Vucic, S., Talbot, K., McDermott, C. J., Hardiman, O., Shefner, J. M., . . . Turner, M. R. (2020). Improving clinical trial outcomes in amyotrophic lateral sclerosis. *Nature Reviews Neurology*. doi:10.1038/s41582-020-00434-z

Klim, J. R., Vance, C., & Scotter, E. L. (2019). Antisense oligonucleotide therapies for Amyotrophic Lateral Sclerosis: Existing and emerging targets. *The International Journal of Biochemistry & Cell Biology, 110*, 149-153. doi:https://doi.org/10.1016/j.biocel.2019.03.009

Koeman, T., Slottje, P., Schouten, L. J., Peters, S., Huss, A., Veldink, J. H., . . . Vermeulen, R. (2017). Occupational exposure and amyotrophic lateral sclerosis in a prospective cohort. *Occupational and Environmental Medicine, 74*(8), 578. doi:10.1136/oemed-2016-103780

Lacorte, E., Ferrigno, L., Leoncini, E., Corbo, M., Boccia, S., & Vanacore, N. (2016). Physical activity, and physical activity related to sports, leisure and occupational activity as risk factors for ALS: A systematic review. *Neurosci Biobehav Rev, 66*, 61-79. doi:10.1016/j.neubiorev.2016.04.007

Lawlor, D. A., Ebrahim, S., Kundu, D., Bruckdorfer, K. R., Whincup, P. H., & Smith, G. D. (2005). Vitamin C is not associated with coronary heart disease risk once life course socioeconomic position is taken into account: prospective findings from the British Women's Heart and Health Study. *Heart (British Cardiac Society), 91*(8), 1086-1087. doi:10.1136/hrt.2004.048934

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., & Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *27*(8), 1133-1163. doi:10.1002/sim.3034

Lawlor, D. A., Tilling, K., & Davey Smith, G. (2017). Triangulation in aetiological epidemiology. *International Journal of Epidemiology, 45*(6), 1866-1886. doi:10.1093/ije/dyw314

LC1117EW – Sex by age. (2011). Retrieved from: https://www.nomisweb.co.uk/

Leffondré, K., Abrahamowicz, M., Siemiatycki, J., & Rachet, B. (2002). Modeling smoking history: a comparison of different approaches. *Am J Epidemiol, 156*(9), 813-823. doi:10.1093/aje/kwf122

Leffondré, K., Abrahamowicz, M., Xiao, Y., & Siemiatycki, J. (2006). Modelling smoking history using a comprehensive smoking index: application to lung cancer. *Stat Med, 25*(24), 4132-4146. doi:10.1002/sim.2680

Leighton, D. J., Newton, J., Stephenson, L. J., Colville, S., Davenport, R., Gorrie, G., . . . Pal, S. (2019). Changing epidemiology of motor neurone disease in Scotland. *J Neurol, 266*(4), 817-825. doi:10.1007/s00415-019-09190-7

Leinartaitė, L., & Johansson, A. S. (2013). Disulfide scrambling in superoxide dismutase 1 reduces its cytotoxic effect in cultured cells and promotes protein aggregation. *PLoS One, 8*(10), e78060. doi:10.1371/journal.pone.0078060

Liu, K. (1988). Measurement error and its impact on partial correlation and multiple linear regression analyses. *Am J Epidemiol, 127*(4), 864-874. doi:10.1093/oxfordjournals.aje.a114870

Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D. M., Chen, F., . . . Vrieze, S. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature Genetics, 51*(2), 237-244. doi:10.1038/s41588-018-0307-5

Logroscino, G., Piccininni, M., Marin, B., Nichols, E., A, F., Abdelalim, A., . . . M, C. J. L. (2018). Global, regional, and national burden of motor neuron diseases 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol, 17*(12), 1083-1097. doi:10.1016/S1474-4422(18)30404-6

Logroscino, G., Traynor, B. J., Hardiman, O., Chio, A., Couratier, P., Mitchell, J. D., . . . Beghi, E. (2008). Descriptive epidemiology of amyotrophic lateral sclerosis: new evidence and unsolved issues. *J Neurol Neurosurg Psychiatry, 79*(1), 6-11. doi:10.1136/jnnp.2006.104828

Logroscino, G., Traynor, B. J., Hardiman, O., Chiò, A., Mitchell, D., Swingler, R. J., . . . Beghi, E. (2010). Incidence of amyotrophic lateral sclerosis in Europe. *J Neurol Neurosurg Psychiatry, 81*(4), 385-390. doi:10.1136/jnnp.2009.183525

Longinetti, E., & Fang, F. (2019). Epidemiology of amyotrophic lateral sclerosis: an update of recent literature. *Current opinion in neurology, 32*(5), 771-776. doi:10.1097/WCO.0000000000000730

Lunetta, C., Moglia, C., Lizio, A., Caponnetto, C., Dubbioso, R., Giannini, F., . . . Calvo, A. (2020). The Italian multicenter experience with edaravone in amyotrophic lateral sclerosis. *J Neurol, 267*(11), 3258-3267. doi:10.1007/s00415-020-09993-z

Mackenzie, I. R., Bigio, E. H., Ince, P. G., Geser, F., Neumann, M., Cairns, N. J., . . . Trojanowski, J. Q. (2007). Pathological TDP-43 distinguishes sporadic amyotrophic lateral sclerosis from amyotrophic lateral sclerosis with SOD1 mutations. *Ann Neurol, 61*(5), 427-434. doi:10.1002/ana.21147

Maier, A., Holm, T., Wicks, P., Steinfurth, L., Linke, P., Münch, C., . . . Meyer, T. (2012). Online assessment of ALS functional rating scale compares well to in-clinic evaluation: a prospective trial. *Amyotroph Lateral Scler, 13*(2), 210-216. doi:10.3109/17482968.2011.633268

Malek, A. M., Barchowsky, A., Bowser, R., Heiman-Patterson, T., Lacomis, D., Rana, S., . . . Talbott, E. O. (2015). Exposure to hazardous air pollutants and the risk of amyotrophic lateral sclerosis. *Environ Pollut, 197*, 181-186. doi:10.1016/j.envpol.2014.12.010

Marin, B., Boumédiene, F., Logroscino, G., Couratier, P., Babron, M., Leutenegger, A., . . . Beghi, E. (2017). Variation in worldwide incidence of amyotrophic lateral sclerosis: a meta-analysis. *International Journal of Epidemiology, 46*(1), 57-74. doi:10.1093/ije/dyw061

Marin, B., Fontana, A., Arcuti, S., Copetti, M., Boumédiene, F., Couratier, P., . . . Logroscino, G. (2018). Age-specific ALS incidence: a dose-response meta-analysis. *Eur J Epidemiol, 33*(7), 621-634. doi:10.1007/s10654-018-0392-x

Marin, B., Logroscino, G., Boumédiene, F., Labrunie, A., Couratier, P., Babron, M., . . . Beghi, E. (2016). Clinical and demographic factors and outcome of amyotrophic lateral sclerosis in relation to population ancestral origin. *European Journal of Epidemiology, 31*(3), 229-245. doi:10.1007/s10654-015-0090-x

Martin, S., Trevor-Jones, E., Khan, S., Shaw, K., Marchment, D., Kulka, A., . . . Al-Chalabi, A. (2017). The benefit of evolving multidisciplinary care in ALS: a diagnostic cohort survival comparison. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 18*(7-8), 569-575. doi:10.1080/21678421.2017.1349151

McCrate, M. E., & Kaspar, B. K. (2008). Physical activity and neuroprotection in amyotrophic lateral sclerosis. *Neuromolecular Med, 10*(2), 108-117. doi:10.1007/s12017-008-8030-5

McLaughlin, R. L., Schijven, D., van Rheenen, W., van Eijk, K. R., O'Brien, M., Kahn, R. S., . . . Schizophrenia Working Group of the Psychiatric Genomics, C. (2017). Genetic correlation between amyotrophic lateral sclerosis and schizophrenia. *Nature Communications, 8*(1), 14774. doi:10.1038/ncomms14774

McLaughlin, R. L., Vajda, A., & Hardiman, O. (2015). Heritability of Amyotrophic Lateral Sclerosis: Insights From Disparate Numbers. *JAMA Neurol, 72*(8), 857-858. doi:10.1001/jamaneurol.2014.4049

Mehta, P. R., Jones, A. R., Opie-Martin, S., Shatunov, A., Iacoangeli, A., Al Khleifat, A., . . . Al-Chalabi, A. (2019). Younger age of onset in familial amyotrophic lateral sclerosis is a result of pathogenic gene variants, rather than ascertainment bias. *Journal of Neurology, Neurosurgery & Psychiatry, 90*(3), 268. doi:10.1136/jnnp-2018-319089

Menon, P., Kiernan, M. C., Yiannikas, C., Stroud, J., & Vucic, S. (2013). Split-hand index for the diagnosis of amyotrophic lateral sclerosis. *Clinical Neurophysiology, 124*(2), 410-416. doi:https://doi.org/10.1016/j.clinph.2012.07.025

Millard, L. A. C., Munafo, M. R., Tilling, K., Wootton, R. E., & Davey Smith, G. (2019). MR-pheWAS with stratification and interaction: Searching for the causal effects of smoking heaviness identified an effect on facial aging. *PLoS Genet, 15*(10), e1008353. doi:10.1371/journal.pgen.1008353

Miller, R. G., Mitchell, J. D., & Moore, D. H. (2012). Riluzole for amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND). *Cochrane database of systematic reviews*(3).
Miller, T., Cudkowicz, M., Shaw, P. J., Andersen, P. M., Atassi, N., Bucelli, R. C., . . . Ferguson, T. A. (2020). Phase 1–2 Trial of Antisense Oligonucleotide Tofersen for SOD1 ALS. *New England Journal of Medicine, 383*(2), 109-119. doi:10.1056/NEJMoa2003715
Mitchell, J. D., Gatrell, A. C., Al-Hamad, A., Davies, R. B., & Batterby, G. (1998). Geographical epidemiology of residence of patients with motor neuron disease in Lancashire and south Cumbria. *J Neurol Neurosurg Psychiatry, 65*(6), 842-847. doi:10.1136/jnnp.65.6.842
Mitchell, R., Hemani, G., Dudding, T., Corbin, L., Harrison, S., Paternoster, L. . (2018). UK Biobank Genetic Data: MRC-IEU Quality Control, version 2 - Datasets. *data.bris*. doi:doi:10.5523/bris.1ovaau5sxunp2cv8rcy88688v
Moglia, C., Calvo, A., Grassano, M., Canosa, A., Manera, U., Ovidio, F., . . . Chiò, A. (2019). Early weight loss in amyotrophic lateral sclerosis: outcome relevance and clinical correlates in a population-based cohort. *Journal of Neurology, Neurosurgery Psychiatry, 90*(6), 666. doi:10.1136/jnnp-2018-319611
Neugebauer, R., & Ng, S. (1990). Differential recall as a source of bias in epidemiologic research. *J Clin Epidemiol, 43*(12), 1337-1341. doi:10.1016/0895-4356(90)90100-4
Neumann, M., Sampathu, D. M., Kwong, L. K., Truax, A. C., Micsenyi, M. C., Chou, T. T., . . . Lee, V. M. (2006). Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science, 314*(5796), 130-133. doi:10.1126/science.1134108
Nicolas, A., Kenna, K. P., Renton, A. E., Ticozzi, N., Faghri, F., Chia, R., . . . Landers, J. E. (2018). Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron, 97*(6), 1268-1283.e1266. doi:10.1016/j.neuron.2018.02.027
O'Reilly, D. F., Brazis, P. W., & Rubino, F. A. (1982). The misdiagnosis of unilateral presentations of amyotrophic lateral sclerosis. *Muscle Nerve Official Journal of the American Association of Electrodiagnostic Medicine, 5*(9), 724-726.
Pamphlett, R., & Rikard-Bell, A. (2013). Different Occupations Associated with Amyotrophic Lateral Sclerosis: Is Diesel Exhaust the Link? *PLoS One, 8*(11), e80993. doi:10.1371/journal.pone.0080993
Pamphlett, R., & Ward, E. C. (2012). Smoking is not a risk factor for sporadic amyotrophic lateral sclerosis in an Australian population. *Neuroepidemiology, 38*(2), 106-113. doi:10.1159/000336013
Parton, M. J., Broom, W., Andersen, P. M., Al-Chalabi, A., Nigel Leigh, P., Powell, J. F., & Shaw, C. E. (2002). D90A-SOD1 mediated amyotrophic lateral sclerosis: a single founder for all cases with evidence for a Cis-acting disease modifier in the recessive haplotype. *Hum Mutat, 20*(6), 473. doi:10.1002/humu.9081
Patel, C. J., Bhattacharya, J., & Butte, A. J. (2010). An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus. *PLoS One, 5*(5), e10746. doi:10.1371/journal.pone.0010746
Peters, S., Visser, A. E., D'Ovidio, F., Vlaanderen, J., Portengen, L., Beghi, E., . . . van den Berg, L. H. (2019). Effect modification of the association between total cigarette smoking and ALS risk by intensity, duration and time-since-quitting: Euro-MOTOR. *Journal of Neurology, Neurosurgery & Psychiatry*, jnnp-2019-320986. doi:10.1136/jnnp-2019-320986
Phukan, J., Elamin, M., Bede, P., Jordan, N., Gallagher, L., Byrne, S., . . . Hardiman, O. (2012). The syndrome of cognitive impairment in amyotrophic lateral sclerosis: a population-based study. *J Neurol Neurosurg Psychiatry, 83*(1), 102-108. doi:10.1136/jnnp-2011-300188
Pierce, B. L., & Burgess, S. (2013). Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am J Epidemiol, 178*(7), 1177-1184. doi:10.1093/aje/kwt084
Prignot, J. (1987). Quantification and chemical markers of tobacco-exposure. *Eur J Respir Dis, 70*(1), 1-7.

Pupillo, E., Bianchi, E., Vanacore, N., Montalto, C., Ricca, G., Robustelli Della Cuna, F. S., . . . Beghi, E. (2020). Increased risk and early onset of ALS in professional players from Italian Soccer Teams. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 21*(5-6), 403-409. doi:10.1080/21678421.2020.1752250

Raaphorst, J., de Visser, M., Linssen, W. H. J. P., de Haan, R. J., & Schmand, B. (2010). The cognitive profile of amyotrophic lateral sclerosis: A meta-analysis. *Amyotrophic Lateral Sclerosis, 11*(1-2), 27-37. doi:10.3109/17482960802645008

Ravits, J. M., & La Spada, A. R. (2009). ALS motor phenotype heterogeneity, focality, and spread: deconstructing motor neuron degeneration. *Neurology, 73*(10), 805-811. doi:10.1212/WNL.0b013e3181b6bbbd

Read, A. P. (2018). *Human molecular genetics*: Garland Science.

Richards, D., Morren, J. A., & Pioro, E. P. (2020). Time to diagnosis and factors affecting diagnostic delay in amyotrophic lateral sclerosis. *Journal of the Neurological Sciences, 417*. doi:10.1016/j.jns.2020.117054

Richardson, T. G., Harrison, S., Hemani, G., & Davey Smith, G. (2019). An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *eLife, 8*, e43657. doi:10.7554/eLife.43657

Ringholz, G. M., Appel, S. H., Bradshaw, M., Cooke, N. A., Mosnik, D. M., & Schulz, P. E. (2005). Prevalence and patterns of cognitive impairment in sporadic ALS. *Neurology, 65*(4), 586-590. doi:10.1212/01.wnl.0000172911.39167.b6

Roberts, A. L., Johnson, N. J., Cudkowicz, M. E., Eum, K.-D., & Weisskopf, M. G. (2015). Job-related formaldehyde exposure and ALS mortality in the USA. *Journal of Neurology, Neurosurgery and Psychiatry, 87*(7), 786-788. doi:10.1136/jnnp-2015-310750

Rooney, J., Brayne, C., Tobin, K., Logroscino, G., Glymour, M. M., & Hardiman, O. (2017). Benefits, pitfalls, and future design of population-based registers in neurodegenerative disease. *Neurology, 88*(24), 2321. doi:10.1212/WNL.0000000000004038

Rooney, J., Byrne, S., Heverin, M., Corr, B., Elamin, M., Staines, A., . . . Hardiman, O. (2013). Survival Analysis of Irish Amyotrophic Lateral Sclerosis Patients Diagnosed from 1995–2010. *PLoS One, 8*(9), e74733. doi:10.1371/journal.pone.0074733

Rooney, J., Byrne, S., Heverin, M., Tobin, K., Dick, A., Donaghy, C., & Hardiman, O. (2015). A multidisciplinary clinic approach improves survival in ALS: a comparative study of ALS in Ireland and Northern Ireland. *Journal of Neurology, Neurosurgery & Psychiatry, 86*(5), 496-501.

Rooney, J., Vajda, A., Heverin, M., Crampsie, A., Tobin, K., McLaughlin, R., . . . Hardiman, O. (2016). No association between soil constituents and amyotrophic lateral sclerosis relative risk in Ireland. *Environ Res, 147*, 102-107. doi:10.1016/j.envres.2016.01.038

Rosen, D. R., Siddique, T., Patterson, D., Figlewicz, D. A., Sapp, P., Hentati, A., . . . et al. (1993). Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature, 362*(6415), 59-62. doi:10.1038/362059a0

Ross, J. P., Leblond, C. S., Catoire, H., Volkening, K., Strong, M., Zinman, L., . . . Rouleau, G. A. (2019). Somatic expansion of the C9orf72 hexanucleotide repeat does not occur in ALS spinal cord tissues. *J Neurology Genetics, 5*(2), e317. doi:10.1212/NXG.0000000000000317

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys* (Vol. 31): Wiley.

Ryan, M., Heverin, M., Doherty, M. A., Davis, N., Corr, E. M., Vajda, A., . . . Hardiman, O. (2018). Determining the incidence of familiality in ALS: A study of temporal trends in Ireland from 1994 to 2016. *Neurology. Genetics, 4*(3), e239-e239. doi:10.1212/NXG.0000000000000239

Ryan, M., Heverin, M., McLaughlin, R. L., & Hardiman, O. (2019). Lifetime Risk and Heritability of Amyotrophic Lateral Sclerosis. *JAMA Neurol, 76*(11), 1367-1374. doi:10.1001/jamaneurol.2019.2044

Saville, B. R., & Berry, S. M. (2016). Efficiencies of platform clinical trials: A vision of the future. *Clin Trials, 13*(3), 358-366. doi:10.1177/1740774515626362

Scarmeas, N., Shih, T., Stern, Y., Ottman, R., & Rowland, L. P. (2002). Premorbid weight, body mass, and varsity athletics in ALS. *Neurology, 59*(5), 773-775. doi:10.1212/wnl.59.5.773

Scotter, E. L., Chen, H. J., & Shaw, C. E. (2015). TDP-43 Proteinopathy and ALS: Insights into Disease Mechanisms and Therapeutic Targets. *Neurotherapeutics, 12*(2), 352-363. doi:10.1007/s13311-015-0338-x

Shatunov, A., & Al-Chalabi, A. (2020). The genetic architecture of ALS. *Neurobiol Dis, 147*, 105156. doi:10.1016/j.nbd.2020.105156

Shefner, J. M., Al-Chalabi, A., Baker, M. R., Cui, L.-Y., de Carvalho, M., Eisen, A., . . . Kiernan, M. C. (2020). A proposal for new diagnostic criteria for ALS. *Clinical Neurophysiology, 131*(8), 1975-1978. doi:https://doi.org/10.1016/j.clinph.2020.04.005

Simon, N. G., Lee, M., Bae, J. S., Mioshi, E., Lin, C. S. Y., Pfluger, C. M., . . . Kiernan, M. C. (2015). Dissociated lower limb muscle involvement in amyotrophic lateral sclerosis. *Journal of Neurology, 262*(6), 1424-1432. doi:10.1007/s00415-015-7721-8

Skelly, A. C., Dettori, J. R., & Brodt, E. D. (2012). Assessing bias: the importance of considering confounding. *Evidence-based spine-care journal, 3*(1), 9-12. doi:10.1055/s-0031-1298595

Smith, G. D. (2011). Epidemiology, epigenetics and the 'Gloomy Prospect': embracing randomness in population health research and practice. *International Journal of Epidemiology, 40*(3), 537-562. doi:10.1093/ije/dyr117

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia medica, 24*(1), 12-18. doi:10.11613/BM.2014.003

Steinbach, R., Batyrbekova, M., Gaur, N., Voss, A., Stubendorff, B., Mayer, T. E., . . . Grosskreutz, J. (2020). Applying the D50 disease progression model to gray and white matter pathology in amyotrophic lateral sclerosis. *NeuroImage: Clinical, 25*, 102094. doi:https://doi.org/10.1016/j.nicl.2019.102094

Sutedja, N. A., Veldink, J. H., Fischer, K., Kromhout, H., Heederik, D., Huisman, M. H., . . . van den Berg, L. H. (2009). Exposure to chemicals and metals and risk of amyotrophic lateral sclerosis: a systematic review. *Amyotroph Lateral Scler, 10*(5-6), 302-309. doi:10.3109/17482960802455416

Swinnen, B., & Robberecht, W. (2014). The phenotypic variability of amyotrophic lateral sclerosis. *Nat Rev Neurol, 10*(11), 661-670. doi:10.1038/nrneurol.2014.184

Tai, H., Cui, L., Shen, D., Li, D., Cui, B., & Fang, J. (2017). Military service and the risk of amyotrophic lateral sclerosis: A meta-analysis. *J Clin Neurosci, 45*, 337-342. doi:10.1016/j.jocn.2017.08.035

Tang, L., Ma, Y., Liu, X.-l., Chen, L., & Fan, D.-s. (2019). Correction to: Better survival in female SOD1-mutant patients with ALS: a study of SOD1-related natural history. *Translational Neurodegeneration, 8*(1), 10. doi:10.1186/s40035-019-0150-3

Team, R. C. (2018). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*.

Therneau, T. (2020). A Package for Survival Analysis in R.

Toichi, K., Yamanaka, K., & Furukawa, Y. (2013). Disulfide scrambling describes the oligomer formation of superoxide dismutase (SOD1) proteins in the familial form of amyotrophic lateral sclerosis. *J Biol Chem, 288*(7), 4970-4980. doi:10.1074/jbc.M112.414235

Traynor, B. J., Codd, M. B., Corr, B., Forde, C., Frost, E., & Hardiman, O. (2000). Amyotrophic lateral sclerosis mimic syndromes: a population-based study. *Arch Neurol, 57*(1), 109-113. doi:10.1001/archneur.57.1.109

Tremolizzo, L., Bianchi, E., Susani, E., Pupillo, E., Messina, P., Aliprandi, A., . . . Ferrarese, C. (2017). Voluptuary Habits and Risk of Frontotemporal Dementia: A Case Control Retrospective Study. *J Alzheimers Dis, 60*(2), 335-340. doi:10.3233/jad-170260

Turner, M. R., & Al-Chalabi, A. (2020). REM sleep physiology and selective neuronal vulnerability in amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry, 91*(7), 789-790. doi:10.1136/jnnp-2020-323100

Ulm, K. (1990). A simple method to calculate the confidence interval of a standardized mortality ratio (SMR). *Am J Epidemiol, 131*(2), 373-375. doi:10.1093/oxfordjournals.aje.a115507
van Eijk, R. P. A., Jones, A. R., Sproviero, W., Shatunov, A., Shaw, P. J., Leigh, P. N., . . . van Es, M. A. (2017). Meta-analysis of pharmacogenetic interactions in amyotrophic lateral sclerosis clinical trials. *Neurology, 89*(18), 1915-1922. doi:10.1212/wnl.0000000000004606
van Eijk, R. P. A., & van Den Berg, L. (2020). In pursuit of the normal progressor: the holy grail for ALS clinical trial design? *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 21*(1-2), 1-2. doi:10.1080/21678421.2019.1675710
van Eijk, R. P. A., Westeneng, H. J., Nikolakopoulos, S., Verhagen, I. E., van Es, M. A., Eijkemans, M. J. C., & van den Berg, L. H. (2019). Refining eligibility criteria for amyotrophic lateral sclerosis clinical trials. *Neurology, 92*(5), e451-460. doi:10.1212/wnl.0000000000006855
van Es, M. A., Goedee, H. S., Westeneng, H. J., Nijboer, T. C. W., & van den Berg, L. H. (2020). Is it accurate to classify ALS as a neuromuscular disorder? *Expert Rev Neurother, 20*(9), 895-906. doi:10.1080/14737175.2020.1806061
van Rheenen, W., Shatunov, A., Dekker, A. M., McLaughlin, R. L., Diekstra, F. P., Pulit, S. L., . . . Group, N. S. (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature Genetics, 48*(9), 1043-1048. doi:10.1038/ng.3622
Vela, A., Galán, L., Valencia, C., de la Torre, P., Cuadrado, L., Esteban, J., . . . Matías-Guiu, J. (2012). SOD1-N196 mutation in a family with amyotrophic lateral sclerosis. *Neurología (English Edition), 27*(1), 11-15. doi:https://doi.org/10.1016/j.nrleng.2011.02.004
Veldink, J. H. (2017). ALS genetic epidemiology 'How simplex is the genetic epidemiology of ALS?'. *Journal of Neurology, Neurosurgery & Psychiatry, 88*(7), 537. doi:10.1136/jnnp-2016-315469
Verber, N. S., Shepheard, S. R., Sassani, M., McDonough, H. E., Moore, S. A., Alix, J. J. P., . . . Shaw, P. J. (2019). Biomarkers in Motor Neuron Disease: A State of the Art Review. *Frontiers in neurology, 10*, 291-291. doi:10.3389/fneur.2019.00291
Visser, A. E., Rooney, J. P. K., D'Ovidio, F., Westeneng, H. J., Vermeulen, R. C. H., Beghi, E., . . . van den Berg, L. H. (2018). Multicentre, cross-cultural, population-based, case-control study of physical activity as risk factor for amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry, 89*(8), 797-803. doi:10.1136/jnnp-2017-317724
Vucic, S. (2019). Split elbow sign: more evidence for the importance of cortical dysfunction in ALS. *Journal of Neurology, Neurosurgery & Psychiatry, 90*(7), 729-729. doi:10.1136/jnnp-2019-320534
Vucic, S., Higashihara, M., Sobue, G., Atsuta, N., Doi, Y., Kuwabara, S., . . . Consortium, P. (2020). ALS is a multistep process in South Korean, Japanese, and Australian patients. *Neurology, 94*(15), e1657-e1663. doi:10.1212/WNL.0000000000009015
Vucic, S., Westeneng, H.-J., Al-Chalabi, A., Van Den Berg, L. H., Talman, P., & Kiernan, M. C. (2019). Amyotrophic lateral sclerosis as a multi-step process: an Australia population study. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 20*(7-8), 532-537. doi:10.1080/21678421.2018.1556697
Wang, H., O'Reilly, É. J., Weisskopf, M. G., Logroscino, G., McCullough, M. L., Thun, M., . . . Ascherio, A. (2011). Smoking and risk of amyotrophic lateral sclerosis: a pooled analysis of five prospective cohorts. *Arch Neurol, 68*(2), 207-213. doi:10.1001/archneurol.2010.367
Wang, M. D., Little, J., Gomes, J., Cashman, N. R., & Krewski, D. (2017). Identification of risk factors associated with onset and progression of amyotrophic lateral sclerosis using systematic review and meta-analysis. *Neurotoxicology, 61*, 101-130. doi:10.1016/j.neuro.2016.06.015
Weisskopf, M. G., McCullough, M. L., Calle, E. E., Thun, M. J., Cudkowicz, M., & Ascherio, A. (2004). Prospective Study of Cigarette Smoking and Amyotrophic Lateral Sclerosis. *Am J Epidemiol, 160*(1), 26-33. doi:10.1093/aje/kwh179
Westeneng, H. J., Debray, T. P. A., Visser, A. E., van Eijk, R. P. A., Rooney, J. P. K., Calvo, A., . . . van den Berg, L. H. (2018). Prognosis for patients with amyotrophic lateral sclerosis:

development and validation of a personalised prediction model. *Lancet Neurol, 17*(5), 423-433. doi:10.1016/s1474-4422(18)30089-9

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*(4), 377-399. doi:10.1002/sim.4067

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*: Springer-Verlag New York.

Wilbourn, A. J. (2000). The "split hand syndrome". *Muscle Nerve, 23*(1), 138. doi:10.1002/1097-459820000123:13.0.2-7

Wingo, T. S., Cutler, D. J., Yarab, N., Kelly, C. M., & Glass, J. D. (2011). The Heritability of Amyotrophic Lateral Sclerosis in a Clinically Ascertained United States Research Registry. *PLoS One, 6*(11), e27985. doi:10.1371/journal.pone.0027985

Wootton, R. E., Richmond, R. C., Stuijfzand, B. G., Lawn, R. B., Sallis, H. M., Taylor, G. M. J., . . . Munafò, M. R. (2018). Evidence for causal effects of lifetime smoking on risk for depression and schizophrenia: a Mendelian randomisation study. *Psychological Medicine*, 1-9. doi:10.1017/S0033291719002678

Xu, L., Liu, T., Liu, L., Yao, X., Chen, L., Fan, D., . . . Wang, S. (2020). Global variation in prevalence and incidence of amyotrophic lateral sclerosis: a systematic review and meta-analysis. *J Neurol, 267*(4), 944-953. doi:10.1007/s00415-019-09652-y

Yang, J., Visscher, P. M., & Wray, N. R. (2010). Sporadic cases are the norm for complex disease. *Eur J Hum Genet, 18*(9), 1039-1043. doi:10.1038/ejhg.2009.177

Yoshida, H., Yanai, H., Namiki, Y., Fukatsu-Sasaki, K., Furutani, N., & Tada, N. (2006). Neuroprotective effects of edaravone: a novel free radical scavenger in cerebrovascular injury. *CNS Drug Rev, 12*(1), 9-20. doi:10.1111/j.1527-3458.2006.00009.x

Yu, Y., Su, F. C., Callaghan, B. C., Goutman, S. A., Batterman, S. A., & Feldman, E. L. (2014). Environmental risk factors and amyotrophic lateral sclerosis (ALS): a case-control study of ALS in Michigan. *PLoS One, 9*(6), e101186. doi:10.1371/journal.pone.0101186

Zhan, Y., & Fang, F. (2019). Smoking and amyotrophic lateral sclerosis: A mendelian randomization study. *85*(4), 482-484. doi:10.1002/ana.25443

Zhao, Q., Wang, J., Hemani, G., Bowden, J., & Small, D. S. (2018). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. doi: arXiv:1801.09652.

Zou, Z.-Y., Zhou, Z.-R., Che, C.-H., Liu, C.-Y., He, R.-L., & Huang, H.-P. (2017). Genetic epidemiology of amyotrophic lateral sclerosis: a systematic review and meta-analysis. *Journal of Neurology, Neurosurgery & Psychiatry, 88*(7), 540. doi:10.1136/jnnp-2016-315018

Acknowledgements

I would like to acknowledge the contribution of my supervisors: Professor Ammar Al-Chalabi and Professor Kevin Talbot, both of whom are supportive and inspirational researchers.

MND is a tragic disease, despite this people living with MND and their families and carers approach their lives with dignity and bravery. They participate in research studies and in the research community enthusiastically and without them none of this work would be possible.

Members of the lab group have provided feedback, ideas and support when I have needed it throughout my PhD. In particular, Andrea Bredin and Lynn Ossher have been efficient and enthusiastic project managers; without their input the number of sites participating and data collection for the paper presented in this thesis would not be at the level it is. Alfredo Iacoangeli, Ashley Jones, Aleksey Shatunov, Isabella Fogh, Tom Spargo, Harry Bowles, Ahmad Al Khleifat and Simon Topp have all helped me in various ways throughout my degree, proofreading documents, discussing ideas, improving my interpretation of results, and providing code.

At King's College Hospital the clinic coordinators, Catherine Knights, Caty Bailey and Angeline Brooks have all helped the project enormously by testing the template database, explaining what data need to be collected for care purposes and collecting data for the project at King's College London. Theresa Chiwera has contributed to the MND Register project by organising consenting for research studies while also organising the many other research studies that take place at King's College Hospital. Rachel Tuck, the clinic administrator has provided so much help in updating records and organising patient data. Adrian Broughton is a caring and dedicated clinical nurse specialist and is always an excellent conversationalist. There are too many people to name individually, but thanks to all the people in the centres in England, Wales and Northern Ireland for consenting people and collecting data for the MND Register.

I would like to thank the MND Association for funding me throughout my time at the MND Register and thank the NIHR CLAHRC for funding the first year of my Ph.D.