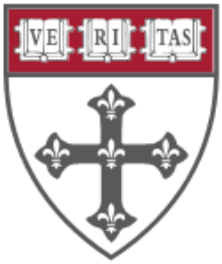
Università degli Studi di Milano-Bicocca

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di Laurea Magistrale in Data Science



BIG DATA IN EPIDEMIOLOGY:

A STUDY OF DIABETES AND GENDER DIFFERENCES IN ALS RISK, IN A POPULATION OF THREE MILLION ISRAELI INDIVIDUALS.

Relatore: Prof. Giorgio Vittadini

Co-relatore: Dr. Andrea Bellavia

Co-relatore: Dr. Ran Rotem

Tesi di Laurea Magistrale di:

Stefano Rola

Matricola: 790383

Anno Accademico 2018-2019

"Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?"
T. S. Eliot

Contents

List of Abbreviations

ALS Amyotrophic Lateral Sclerosis

BMI Body Mass Index

DM Diabetes Mellitus

EEC El Escorial Criteria

EHR Electronic Health Record

EMR Electronic Medical Record

GLM Generalized Linear Model

ICD International Statistical Classification of Diseases

LMN Lower Motor Neuron

MHS Maccabi Healthcare Services

SES Socio-Economic Status

UMN Upper Motor Neuron

1. Introduction

Amyotrophic Lateral Sclerosis (ALS) is a neurodegenerative disease of largely unknown aetiology characterised by rapidly progressive paralysis and respiratory failure.[1] The rarity and severity of ALS make it really hard to inspect; moreover, no definitive diagnostic test or biomarker for ALS exist, and neurologists rely on only clinical criteria for diagnosis. Clinically, ALS is diagnosed according to the El Escorial criteria (EEC), the revised EEC (rEEC), also known as Airlie-House criteria, the Awaji-Shima criteria (ASC), or the updated ASC (uASC). All four sets of diagnostic criteria differentiate between definite, probable, and possible ALS.[2] ALS is clinically characterised by the presence of progressive upper motor neuron (UMN) signs and lower motor neurone (LMN) signs in bulbar, upper limb, axial, and lower limb muscles; the golden standard in clinical diagnosis of ALS is the neuropathological demonstration of the UMN and LMN lesion.[3] Currently there is no effective cure for ALS, and slowing the disease progression is the only benefit provided by existing medications. Therefore, in the absence of a cure, the main focuses are on symptomatic and palliative care of patients with ALS. These treatments may not only alleviate symptoms but also maintain quality of life and improve survival, with greater benefit for patients managed in specialized, multidisciplinary ALS clinics.[3]

While a genetic contribution in the development of the disease is established, multiple studies proved that just the 5-10% of the cases are ascribable to the so-called "familial ALS".[4] This fact, along with other lines of evidence, have led to the suspicion that modifiable behavioral factors may play a crucial role in affecting the remaining 90% of ALS cases. An increase in understanding the disease, and discovering modifiable factors causing it, could then certainly have a decisive influence in improving preventing its development.[5] However, identifying non genetic, modifiable risk factors for ALS and ALS survival has been proven to be difficult.

One class of potential exposures that has provided conflicting results across different studies are medical risk factors. Several studies provide evidence of a connection between metabolic disorders and ALS, but the effects of this relationship are not clear. In particular, while ALS and Diabetes Mellitus seem to share some common pathological mechanisms, several reports produced conflicting results assessing the direction of the connection.[6]

Another challenging element in investigating ALS risk factors is the need of high volume of data: given the rarity and difficulty of inspection of the disease, a large, well characterized collection of medical information is required in order to gain statistically significant and accurate results, as are sophisticated data management and analysis techniques to sort through the very high dimensional data.[7]

This thesis aims to investigate the association between ALS and Diabetes Mellitus in a three millions population of Israeli individuals. The study population's information is stored in a large, yet uncharted database provided by Maccabi Health Services. Data privacy and security considerations will be included as well, because of the several questions and challenges rising within the data revolution, along with a possible solution provided by synthetic data.

2. Background

2.1 Amyotrophic Lateral Sclerosis

What is ALS?

Amyotrophic Lateral Sclerosis (ALS) is a rare and severe neurodegenerative disease of upper and lower motor neurons that results in relentlessly progressive paralysis. The origin of the disease name is to be found in the Greek language. "*A*" means "no", "*Myo*" refers to muscle and "*Trophic*" means "nourishment": *"Amyothrophic"* then means "No muscle nutrition". "*Lateral*" identifies the areas in a person's spinal cord where portions of the nerve cells that signal and control the muscles are located. As this area degenerates, it leads to scarring or hardening ("*Sclerosis*") in the region.

The cause of ALS is largely unknown and today there is no effective cure. Once contracted the disease death usually occurs from respiratory failure. Median survival time is usually 3 years from the first appearance of symptoms.[1] The overall crude worldwide ALS prevalence and incidence in 2018 were 4.42 (95% CI 3.92–4.96) per 100,000 population and 1.59 (95% CI 1.39–1.81) per 100,000 person-years, respectively. Both these rates rise with increasing age until the age

of 70-79. (Fig. 2.1) [5]

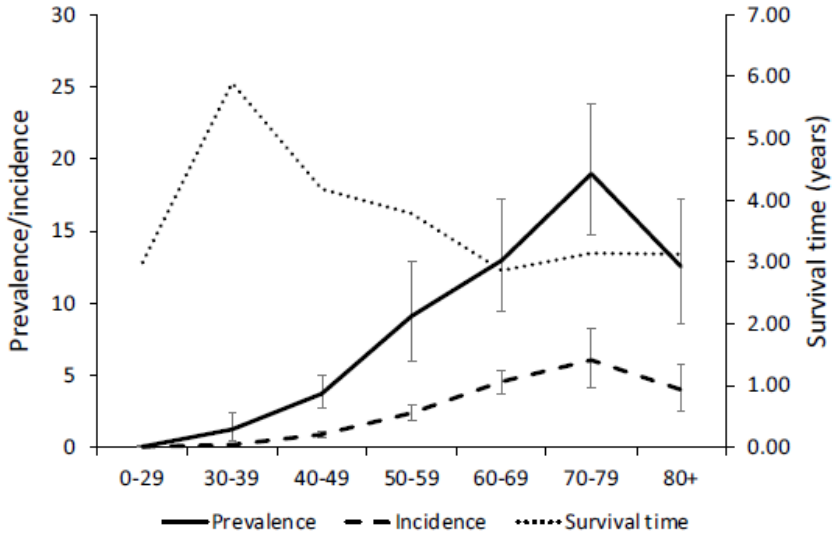


Figure 2.1: Worldwide crude prevalence, crude incidence and survival time of ALS by age. Source: [5]

Nearly all population-based epidemiological studies of ALS have been conducted in Europe, and the findings are relatively consistent.[1] Nevertheless, Marin and colleagues in 2016 showed the heterogeneity of ALS incidence between subcontinents. (Fig. 2.2) This heterogeneity could be related to the ancestral origin of the populations. However, they could not exclude some differences to be caused by environmental factors, given the growing evidence that ALS aetiology relies on the combination of environmental and genetic factors.[8] As a matter of fact only 5%–10% of ALS cases have a hereditary form of the disease, known as "Familial ALS"; if no family history is identified, the diagnosis is assumed to be sporadic.[4] An understanding of the environmental contribution to ALS is essential, as this is the only easily modifiable component of risk.

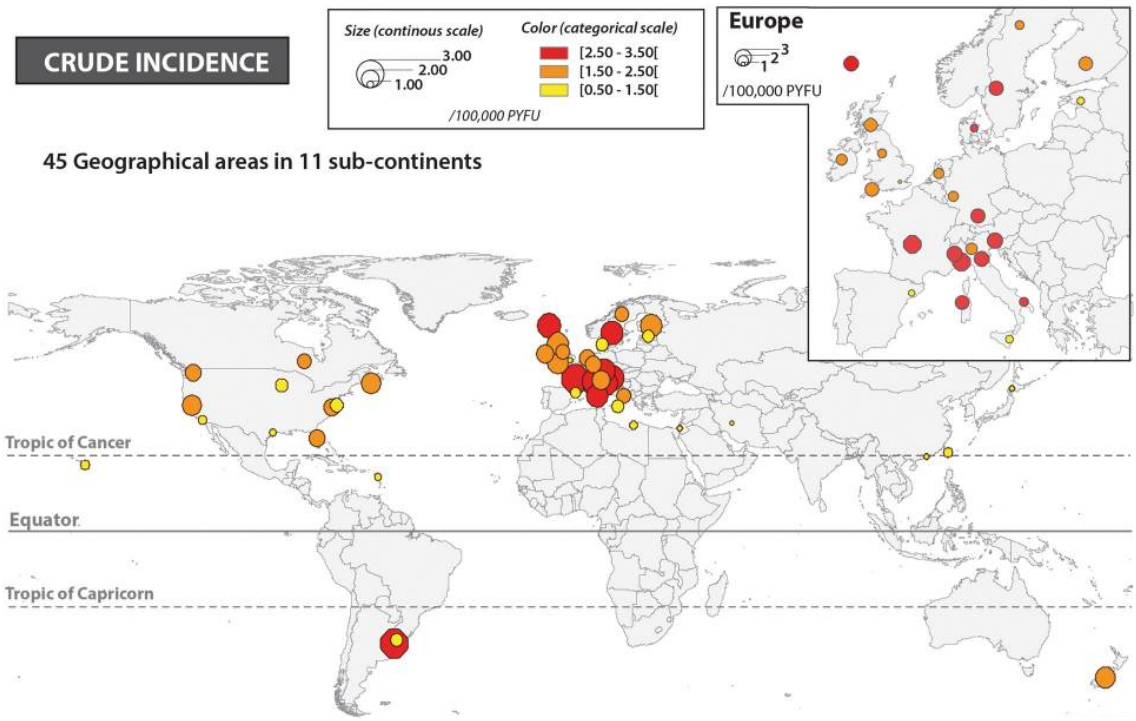


Figure 2.2: Distribution of ALS worldwide (crude incidence). Source: [8]

ALS and Diabetes

Besides motor neuron degeneration, ALS is often associated with - and is potentially caused by - defects in energy metabolism. [9] Patients with ALS frequently experience weight loss, hypermetabolism, hyperlipidemia, insulin resistance, glucose intolerance and increased energy consumption. [9; 10; 11; 12] On the other hand, high BMI and high plasma levels of cholesterol and triglycerides seem to play protective effects towards ALS.[13; 14]

Regarding the risk factors in the development of ALS - which are the main target of this study - several reports provide conflicting evidence concerning a possible role of Diabetes Mellitus (DM). In 2010, for the first time, DM was reported by Jawaid and co-workers as having a possible protective effect and delaying the onset of motor dysfunction in ALS. [6] These findings were along with the stud-

ies mentioned before about metabolic disorders and ALS, and following reports received the same results as Jawaid et al. [15; 16] Moreover, in 2015 Kioumourtzoglou and co-workers used the Danish National Health System's administrative database, covering about 6 million people, to analyse diabetes as a risk factor for ALS in a time period of almost 30 years. In this nested case–control study, a hospital admission for DM was associated with a reduction in the risk of ALS in patients over 40 years of age, but an increased risk in younger patients. When the investigators considered the age at which patients were diagnosed with ALS, diabetes was a risk factor up to 51 years of age, but was protective after age 61 years. [17] Similar results were suggested by another population-based case-control study conducted in the same year on Swedish Patient Register: while diabetes was associated with a decreasing risk of ALS in older patients (70+ years), a higher risk was detected in young patients (under 50 years). [18]

In opposition to these findings, a retrospective cohort study based on Taiwanese's patients with ALS represented that DM increased the risk of ALS in older patients, and the effect of diabetes in increasing the risk of ALS in younger patients was negligible. [19] Different ethnic backgrounds and many environmental factors play an aetiologic role in ALS development and could explain the inconsistent findings reported in European and Asian studies. As Logroscino asserted in 2015, "differences in the risk of ALS between ethnic groups might also be explained by different individual modifiable risk factors, like antecedent diseases, along with the presence of varying genotypes." [20]

In 2018, D'Ovidio et al. conducted a large population cohort study with an accurate assessment of diabetes, also distinguishing between bulbar and spinal ALS onset. This research included individuals with different ethnicities and evaluated the time-dependent association between diabetes and ALS. Evidence deriving from their study showed a strong protective relation between antecedent DM and

development of ALS. [21]

Currently, there are two main suggested hypotheses for the description of the pathological mechanisms shared between DM and ALS (Fig. 2.3). First, hypermetabolism which damages motor neurons in ALS patients could be compensated by hyperlipidemia and high plasma levels of glucose and delayed ALS onset and increased survival. Secondly, mutations in the progranulin gene lead to aberrant processing, aggregation, and mis-localization of TAR DNA-binding protein (TDP-43), causing motor axonopathy which is the pathological hallmark of ALS. An in vivo study demonstrated that overexpression of progranulin reverts the axonopathy induced by TDP-43. Therefore, high concentrations of progranulin in diabetes can rescue the mutant TDP-43-induced axonopathy. [22]

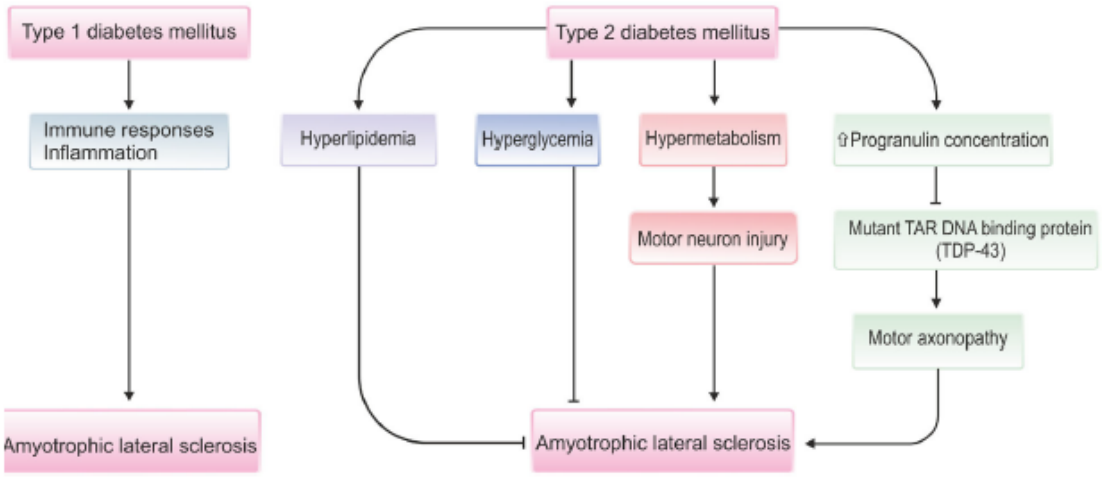


Figure 2.3: Shared pathological mechanisms between Diabetes Mellitus and Amyotrophic Lateral Sclerosis. Source: [22]

Detecting non-genetic, modifiable risk factors for ALS would have great significance improving our knowledge of the disease and thus our contribution in helping preventing its development.

2.2 Methodological considerations

Rare event

The rarity of ALS, given the incidence and prevalence indexes shown in section 2.1, makes it hard to inspect. Investigating a complex and rare disease such as ALS makes the identification of risk factors challenging, and will often require very large dataset to achieve the statistical power required. For this reason, epidemiological studies are scarce, and the few available studies may not be standardized, difficult to combine, may lack firmly established and specific diagnostic criteria, and may be biased depending on the geographical area studied.[23]

Besides, the rapidly progressive degeneration of motor neurons leads to low survival rate: the median survival from onset to death in ALS (in Figure 2.4 is differentiated by subcontinents) is reported to vary from 20 to 48 months.[24] From a methodological perspective the main implications of a low survival rate is that retrieving enough information about patients affected by ALS turns out to be problematic.

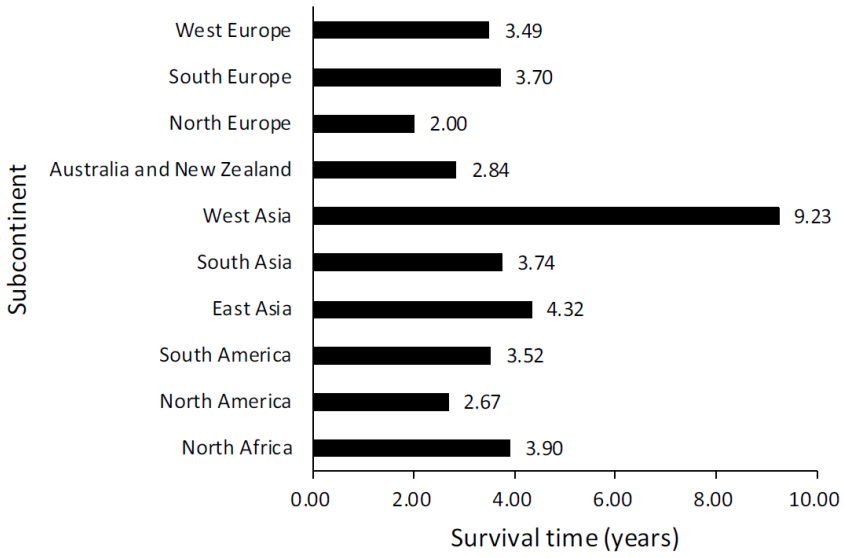


Figure 2.4: ALS survival time in different subcontinents. Source: [5]

While the main focus of this study is not a survival analysis, but the assessment of factors leading to the development of ALS, for a full understanding of the disease it is important to notice that a key element in reducing ALS survival rate - although the main cause still remains the severity of the disease - is the time delay from ALS onset to diagnosis. Primary explanation of this frequent delay is that ALS is hard to diagnose. As a matter of fact no test can provide a definitive diagnosis of ALS. It is primarily diagnosed based on detailed history of the symptoms and signs observed by a physician during physical examination along with a series of tests to rule out other mimicking diseases. ALS symptoms in the early stages of the disease can be similar to those of a wide variety of other, more treatable diseases or disorders. Appropriate tests can exclude the possibility of other conditions; yet ruling out other diseases may take months or years.[25]

In 1994, the El Escorial criteria were published, summarizing a consensus statement of the World Federation of Neurology to facilitate accurate diagnosis of ALS.

The diagnosis by the El Escorial criteria required the presence of upper motor neuron and lower motor neuron degeneration within the same region. In addition, laboratory, electrophysiologic, or neuroimaging studies were supposed to exclude other conditions that might have explained the signs of lower motor neuron and upper motor neuron degeneration. The El Escorial criteria had four levels of diagnostic certainty: definite, probable, possible, and suspected.[26] However, after using them for twenty years as inclusion criteria for clinical trials, concerns have been raised in the ALS community regarding their usage along with the requirements of 21st century clinical research. Moreover, as modern genetics have shed new light on the heterogeneity of ALS and the close relationship between ALS and Frontotemporal Dementia (FTD) has been recognized, the World Federation of Neurology Research Group on ALS/MND has largely discussed to amend and update the criteria.[27]

Nested case-control study

The complexity of ALS diagnosis, as well as the rarity of the disease, have been leading researchers to look for alternative methodologies to traditional prospective studies. For especially difficult to detect diseases a retrospective study may be the only feasible approach, since studying their occurrence with conventional longitudinal research designs is expensive and time-consuming, if possible at all.[28] Moreover, among the positive attributes of the retrospective study is the ability to yield results from presently collectible data, whereas the forward study usually requires future observation of individuals over an extended period.[29] Nonetheless operating on previously collected big amounts of data in retrospective studies could easily lead to issues in analyses' performances. The simplest way to reduce the computational cost of an implementation is to subsample the data before doing anything else. However, uniform subsampling from an imbalanced data set

is inefficient, since it fails to exploit the unequal importance of the data points. Within the investigation of rare diseases, as a typical example of imbalanced data, case-control sampling is then a more effective approach.[30]

In a retrospective nested case-control study - as the one performed in this thesis - the first step is to identify cases (individuals with the disease) that have already occurred within the population of interest. Then, for each case, a pre-determined number of controls is selected from among those in the cohort who have not developed the disease by the time of disease occurrence in the case. The number of controls selected may vary, but the nested case-control literature claims four to five controls per case to be enough to ensure statistically significant results.[31] Several additional points to note are that time-matching is an essential feature of this design, whether controls are matched to cases on age, date of entry into the cohort, length of time in the cohort, or a combination of these measures; that a cohort member who serves as a control at one point in time may later become a case; and that a cohort member may be selected as a control for more than one case.[32]

Complex relationships

Since the aetiology of ALS is believed to be multifactorial, involving both genetic and environmental factors (as shown in 2.1), identifying eligible risk factors is a challenging task. Clinical and epidemiological studies have identified changing in metabolic parameters to be associated with the risk of ALS; however, conflicting results are provided by several reports. Notably, a conventionally 'risky' cardiovascular profile, such as a high Body Mass Index, or Diabetes Mellitus, seems to protect individuals from ALS by delaying the onset of symptoms or slowing clinical progression, whereas, a 'beneficial' cardiovascular profile, with a low Body Mass Index, an athletic lifestyle, and low blood cholesterol levels, may increase the risk

or worsen the prognosis.[12] Some of the detailed results are shown in section 2.1.

Moreover, besides metabolic disorders, other classes of potential exposures have to be considered as a means to further investigate ALS pathogenesis. For instance, medications has not received wide research consideration regarding a possible association with ALS. Yet medications are specifically developed to target biological pathways and in many cases these may coincide with pathways relevant for ALS. While some medications have been included in ALS epidemiological studies, only a small number of medications, one at a time, and often in small datasets have been explored.[33] Assessing single compounds for efficacy in ALS treatments based on suggested biological mechanisms for the disease is a worthwhile and guaranteed method, but it is intrinsically slow. In fact, because of the need to test compounds one at a time, it does not explore effects of combinations of medications in ALS development. Given the large number of medications prescribed, the frequent assumption of multiple medications by a person (particularly in the older age range of usual ALS onset), and the potential interactions between different medications, a more efficient exploration of this huge dataspace is needed, in order to identify the most relevant medications for ALS.

2.3 The need of large dataset and implications

In order to achieve the goal described in the previous section, a very large dataset with detailed disease and medication history, along with accurate identification of ALS cases, is needed. At the same time traditional statistical techniques result inadequate to handling such high-dimensional, potentially inter-correlated data. The need for novel, sophisticated statistical approaches along with large and well characterized population-based studies has been widely recognized.[7] However, the implications of such a procedure are multiple and critical.

Privacy in the age of medical big data

Deciding on the allowable uses of medical data while preserving security and patient's right to privacy is a difficult task. While healthcare organizations store, maintain and transmit huge amounts of data to support the delivery of efficient and proper care, the downsides are the lack of technical support and minimal security. Complicating matters, the healthcare industry continues to be one of the most susceptible to publicly disclosed data breaches. [34]

The first major legal and ethical privacy issue is the use of patient-derived big data leading to discrimination: if employers or insurers learn of sensitive patient information from medical data, such as a debilitating or expensive disease, they may wish not to employ or insure that person, especially since in the United States health insurance is typically tied to employment. US existing laws in health insurance and employment contexts (Genetic Information Nondiscrimination Act, Americans with Disabilities Act and Patient Protection and Affordable Care Act) have favored a more solidaristic view, prohibiting some but not all of this sort of discrimination. These laws represent an attempt to limit privacy harms by limiting consequences of access to data rather than focusing on protecting data themselves: they have important limits. [35]

A second set of privacy harms involves more subjective injuries. Patients whose private health information becomes available can suffer embarrassment, paranoia or mental pain. Even though these injuries may not have measurable external effects — the patients may suffer no financial injury or encounter no stigma from others — they are still injuries. Laws like GINA, the ADA, or the PPACA have little purchase on this type of injury. In some instances, big data permits direct knowledge regarding a person's health by others whom the individual would not want to access the information — whether through inadvertent disclosure or malicious activities such as hacking. A more narrow and more difficult issue raised by

predictive analytics is whether a person's privacy is breached when others make inferences about this individual, inferences enabled more then ever by big data analysis. [35]

It is then crucial that healthcare organizations manage and safeguard personal information and address their legal responsibilities in relation to processing personal data, according to applicable data protection legislation.

A path forward

One reaction to the health privacy violations described above is to significantly limit access to patient data. According to this approach, data should be kept only for limited time, data sharing should be limited to the minimal amount required by the context, or it should be intentionally obfuscated if consequential negative effects are difficult to limit. However, limits on data access can bring their own drawbacks. The basic weakness of privacy overprotection is the limitation it puts on data-driven innovation. Data deidentification, for example, is a common way to comply with the actual data protection legislation; however, deidentified data are much harder to link together when a patient visits different care providers, gets insurance through different payers over time, or moves from one state to another. Fragmented, isolated health data make data-driven innovation hard, imposing both technological and economic difficulties. [35] "Frequently" stated Saha-Chaudhuri and Weinberg in 2017 "data for research undergo anonymization to remove identifying information. However, one may still be able to identify participants using combinations of variables (e.g. if there is a unique person of a certain age, diagnosis, body mass index, town of birth) or auxiliary information available from other sources." [36] "Laws protecting patients' privacy, while beneficial," - add Yale et al. in 2019 - "severely limit access to medical data thus stagnating innovation and limiting educational opportunities. The process of obfuscation of medical data

is costly and time consuming with high penalties for accidental release. Health histories recovered from obfuscated data may result in discrimination." [37]

An alternative solution is proposed by Brad Wible, who asserts that "removal of key information from data can enhance privacy, but this limits data utility and fuels an arms race between deidentification and reidentification. Instead, a generative adversarial network can synthesize data that mimic a protected dataset for analytical purposes but are less likely to reveal any actual private information." [38] Eno and Thompson in 2008 affirmed that "using a mapping to transform a real data set into a synthetic one might (or might not) preserve hidden complex patterns, but at the risk that, if the transformation can be discovered, the original data set can be recovered." Their goal was "to demonstrate that some data mining techniques can discover patterns that can then be used to inverse map into synthetic data sets. These synthetic datasets can be of any size and will faithfully exhibit the same patterns of the original ones, even when they don't conform to simple univariate or multivariate distributions." [39] In 2019 Yale et al. suggested that the application of a Wasserstein GAN model to health real data could lead to the generation of synthetic data while meeting the requirements of privacy preservation. [37] The workflow allowing to export the model they used outside a data-secure environment, in order to create new datasets without undergoing obfuscation, is shown in Figure 2.5.

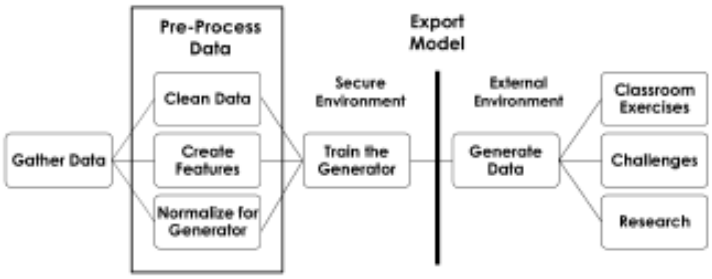


Figure 2.5: Example of workflow used to generate synthetic data

In many contexts a privacy-innovation trade-off seems to still be inevitable. Synthetic dataset could lead to structurally and statistically similar analysis results, while keeping strong privacy guarantees: several studies have adequately demonstrated the feasibility of this method, and the push towards its development is now stronger then ever.

3. Aims of the thesis

The overall aim of this thesis was to manage a large, yet uncharted database of Israeli individuals, in order to evaluate the potential association between Diabetes Mellitus and Amyotrophic Lateral Sclerosis. Secondary goal was to investigate the further contribution of obesity diagnosis and other confounding factors in this association.

Specific aims of this study were:

- To provide an exhaustive picture of benefits and concerns regarding the use of big data in epidemiology.
- To assess the adoption of synthetic data as a possible answer to the big data challenges.
- To mine and manage Maccabi Healthcare Services big data in order to extract useful information.
- To evaluate the shape of the relationship between Diabetes and ALS in MHS database.

4. Materials and methods

4.1 Study population

This thesis used data from a large, national-based, yet uncharted population of Israeli individuals.

Israel has an advanced and centralized health system; 98% of the population is covered by full health insurance and has been for decades using the same linked EMR system. Although operating independently, Israel's four Health Maintenance Organizations (HMOs) and their affiliated hospitals have for the past two decades used the same electronic medical records platform, with access to patient records available to each point of care as needed. As a result, a large body of data about patients, conditions, and treatments has been built up over the years.[40] Maccabi Health Services (MHS) is Israel's second largest integrated health-care organization, serving as both insurer and health-care provider to over 2 million members (25% of the Israeli population). While transferring between health-care providers is possible, it is rare, and attrition is extremely low (1% per year), thus enabling long-term follow-up. MHS members are generally representative of the larger Israeli population, although the average monthly income in MHS is the highest among the four insurers. MHS physicians use an electronic medical record

(EMR) system, which feeds into a central database with administrative and clinically oriented data. The database is linked to records from Israel's Central Bureau of Statistics, permitting linkage to additional information. [41]

In this study the first descriptive analyses were performed on a synthetic dataset built starting from MHS data, provided by MDClone (further details in section 4.2.4). This synthetic population counted 700 ALS cases within 2,876,125 total individuals.

The main analyses, however, were executed on the real MHS data. In order to achieve the goals described in chapter 3 - while keeping the computational costs low - a nested case-control study was then implemented (theoretical details are shown in section 2.2). From the whole MHS database ($\sim$3 million individuals), all ALS cases (662 patients) were extracted; each case was then matched on the year of birth and on being alive at the time of ALS diagnosis to 100 controls. The size of the starting population allowed the creation of such a large matching process, so to maximize statistical power in all analyses despite the outcome rarity. Theoretically, cases could be selected as controls for other cases, with the restriction that their initial date of diagnosis occurred at least 3 years after the date of their matched case. Out of the starting 662 cases, 21 were randomly selected as controls for other cases.

Additional time restrictions were applied to both cases and controls, in order to gain the most accurate and robust results. Further details are provided in sections 4.2.1 and 4.2.2.

Because of the differences in ALS risk and progression between men and women, the analyses were stratified by sex. The final study population consisted in 42 variables on 51,707 individuals, including 24,986 (48%) male and 27,268 (52%) female individuals.

4.2 Data mining

4.2.1 Outcome assessment

The outcome examined was defined as the first occurrence of Amyotrophic Lateral Sclerosis in the MHS database according to International Classification of Diseases (ICD)-9 code 335.20. A binary variable was created in order to represent individuals with (1) and without (0) ALS.

Time restrictions were required in order to gain the most accurate and robust results. As a matter of fact MHS have recorded computerized medical data since 1998. The digitization process of medical information has been performed by physicians, transmitting electronically any medical referrals, results for lab work, consultations and prescriptions. Because of the time needed to record any information preceding 1998, the very first years' information are to be considered unreliable. The first year of ALS diagnosis included in the study was therefore 2001.

Moreover, further limitations were applied to ALS dates of diagnosis. Since MHS is a semi-open cohort (cfr section 4.1) transferring between health-care providers is possible. In order to address for diagnoses previously occurred in other systems, but missing of any documentation, ALS ascertainment was completed including only cases (and the matched controls) being in the system at least one year before ALS diagnosis. For the same reason, all the controls who had been in the system less than one year before ALS diagnosis of the matched case were removed from the analyses.

After exclusion, the final dataset included 547 cases and 51,707 controls (see Table 5.2).

4.2.2 Exposures assessment

Diabetes Mellitus cases were identified in MHS database using ICD-9 code 250.00. A binary variable was created in order to represent individuals with (diabetes = 1) and without (diabetes = 0) diagnosis. In order to minimize the chance that any observed association with diabetes could be due to the effects of underlying, but not yet diagnosed, ALS, the analyses only considered diabetes diagnoses that occurred at least three years prior to the ALS date of diagnosis.

In the following sensitivity analyses, DM cases were detected based on MHS's diabetes registry. Entrance to the registry occurs once a patient meets one of several pre-defined criteria based on lab results, clinical diagnoses and medication dispensing records. Further information about MHS registry criteria can be found in appendix A.

Overweight patients' information were derived as a combination of two variables: obesity and BMI. The obesity diagnoses were obtained based on ICD-9 codes 278.00-278.02. A binary variable was created in order to represent individuals with (obesity = 1) and without (obesity = 0) diagnosis. As for the diabetes, the analyses only considered obesity diagnoses that occurred at least three years prior to the ALS date of diagnosis. Because of the severity of an obesity diagnosis, BMI information was included in the study as well. The last BMI measurement before ALS diagnosis was extracted, with the application of a 3 year lag between the two events. According to the standard World Health Organization categories for BMI, a patient was finally defined to be *overweight* when in presence of an obesity diagnosis OR a BMI greater than 25.0.[42]

4.2.3 Other covariates

This study additionally considered other socio-demographic covariates, in order to address the analyses for potential confounding effects. The included information regarded individuals' birth date (continuous), gender (binary), immigration date (continuous), deceased date (continuous), joining and leaving MHS date (continuous), country of birth (Israel, USSR or other), Socio-Economic Status (SES) and district of residence (Jerusalem&Shfela, center, north, Sharon, south). In particular, SES information was encoded on a discrete scale (1-10), capturing information obtained from many different sources, including on household income, percent recipients of income and unemployment supplements, educational qualifications, crowding, material conditions and car ownership.[43]

The distribution of these variables can be found in Table 5.2.

4.2.4 Data platform

MDClone platform features

MDClone's Healthcare Data Platform combines modern big data technologies to healthcare data organization, management, analysis and sharing. The MDClone engine process can examine medical data, extract their statistical characteristics and dependencies, and extrapolate from these to generate a new data set containing synthetic medical data for fictitious patients, thus overcoming the risk of public disclosure of medical data from actual patients.

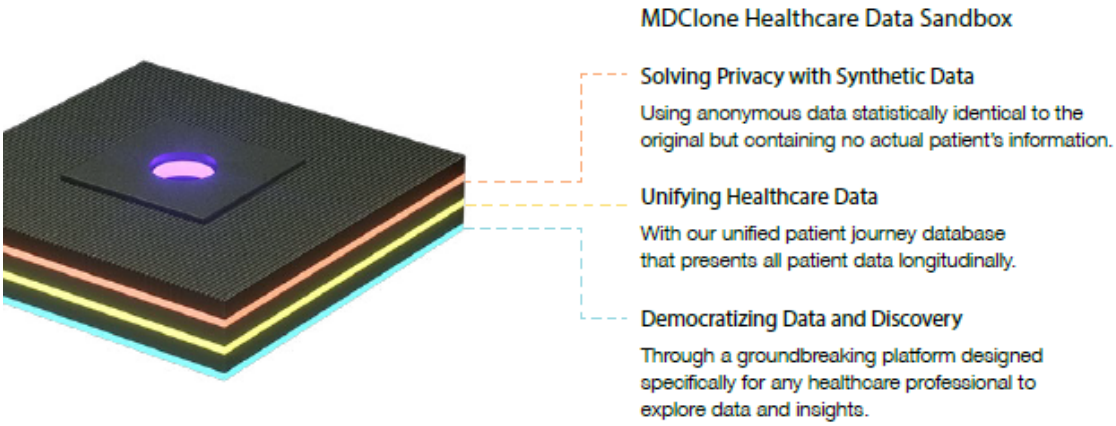


Figure 4.1: MDClone platform benefits

MDClone's synthetic data overcomes one of the major obstacles facing healthcare data analysis: balancing patient privacy and data utilization. Using original data sets, MDClone's Synthetic Data Engine creates non-human subject data statistically comparable to the original but containing no actual patient's information. That way, there's no risk of exposing patient identity and the data can be accessed and analyzed freely. Unlike methods of de-identification, synthetic data is non-reversible, producing non-human subject data without risk of re-identification.

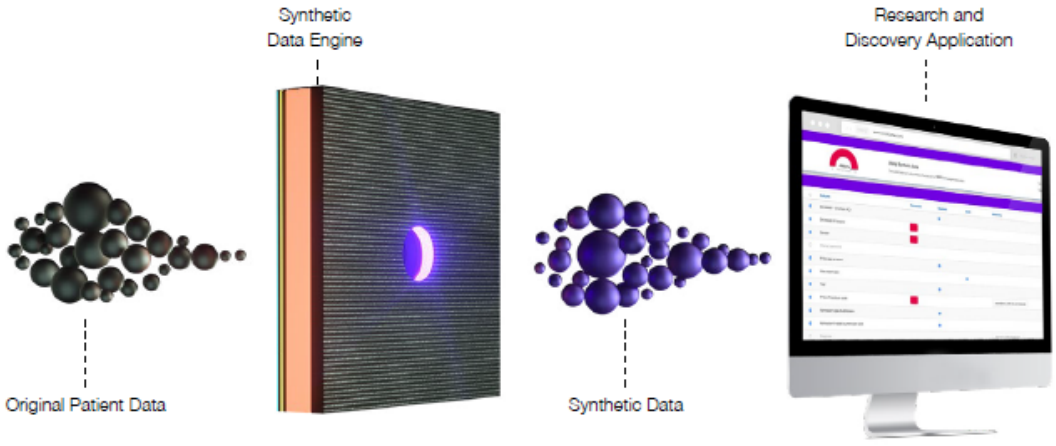


Figure 4.2: MDClone platform workflow

MDClone synthetic data is generated automatically in an interactive process with the end user and without mediators, including without an IRB. It maintains the variables of interest and all correlations between them – including discrete and continuous variables – without advance knowledge of the analysis to be conducted.

MDClone for MHS data

In this study, MDClone platform allowed the download of synthetic data from Maccabi Health System, data required in order to perform preliminary analyses without waiting for an IRB approval.

The first step in the download process was the definition of a "Reference event": as mentioned in section 4.2.1, in this study it was defined as the first diagnosis ever for ALS (ICD-9 code 335.20). Figure 4.3 illustrates the cohort definition based on the reference event in MDClone system. After the query for the diagnosis of interest, other cohort restrictions could be selected. In this study, however, no additional inclusion criteria were applied, based neither on demographics nor on presence/absence of other events.

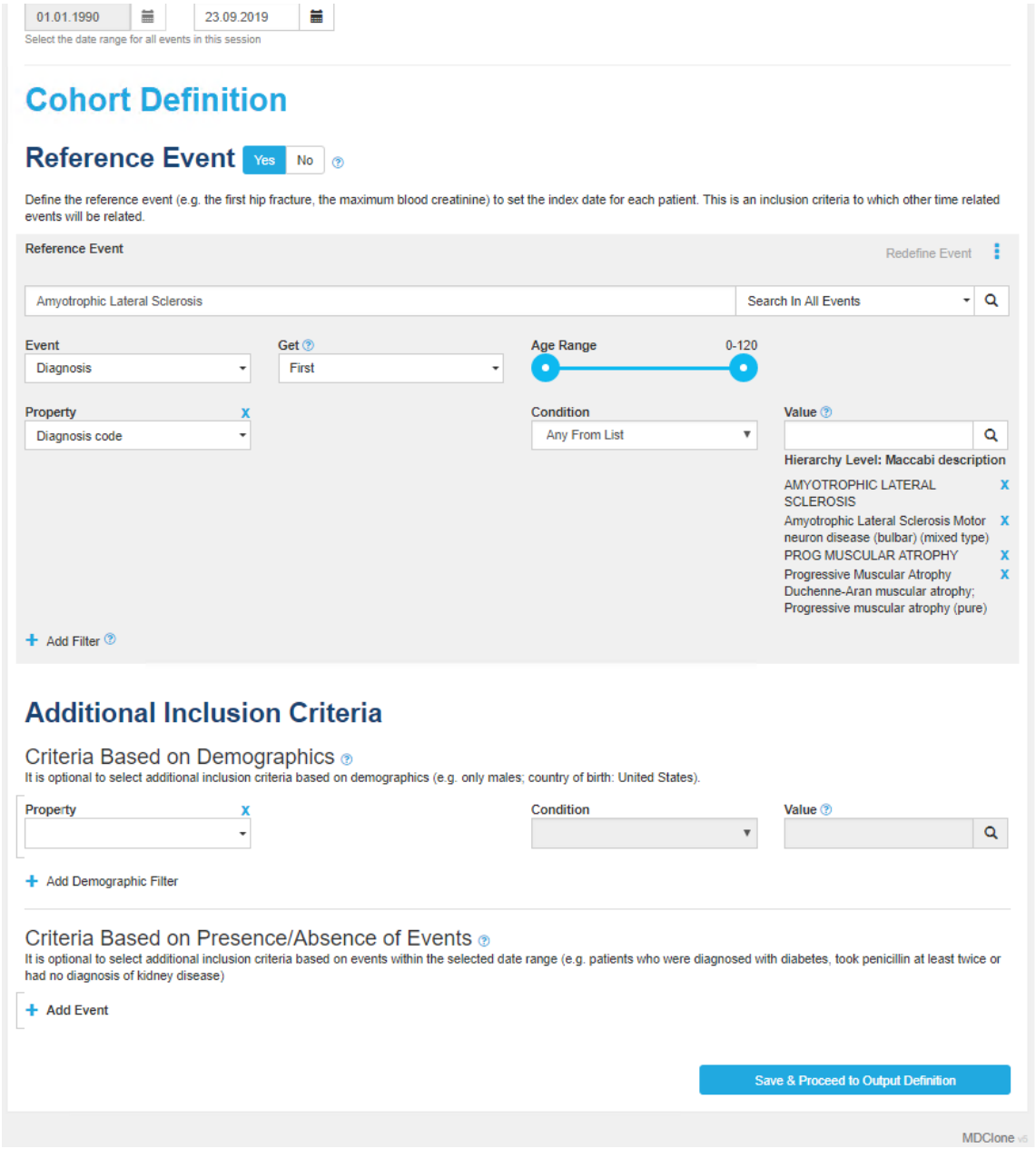


Figure 4.3: Reference event - based cohort definition in MDClone

The second step was the output definition. The data-lake structure of MHS system, enabled by MDClone technologies, allows the user to access a huge amount of data regarding MHS patients. Information about other diseases, as well as so-

cioeconomic status and other demographic documentation, was therefore available for the analyses. In the beginning, events time-related to the reference one were defined: in this study were included diagnoses of Diabetes Mellitus and Obesity. As happened with ALS diagnosis, the system returned the first diagnoses ever, based on their ICD-9 codes (250.0 and 278.0-278.02 respectively). After that, additional information was available about every diagnosis: patient's age at the diagnosis, as well as the diagnosis code, type and source, are just a few examples of the accessible features. Moreover, further information - unrelated to diagnoses yet regarding patients' demographic and socio-economic status - was extracted. The most relevant feature in this study were introduced in section 4.2.3.

The final step in the creation of the synthetic data provides some crucial information about them: in order to guarantee total privacy and protection of patients data, a portion of these data are censored. However, censored data are not suitable for analyses purposes, so that knowing the exact amount of them is extremely important. On top of Figure 4.4 is given the percentage of censored data in the queried population, along with the total number of individuals in the selected cohort. This section allows the definition of further restriction and inclusion criteria as well: it's possible to filter and transform any variable in different formats. The percentage of *Null* values is also shown for every covariate.

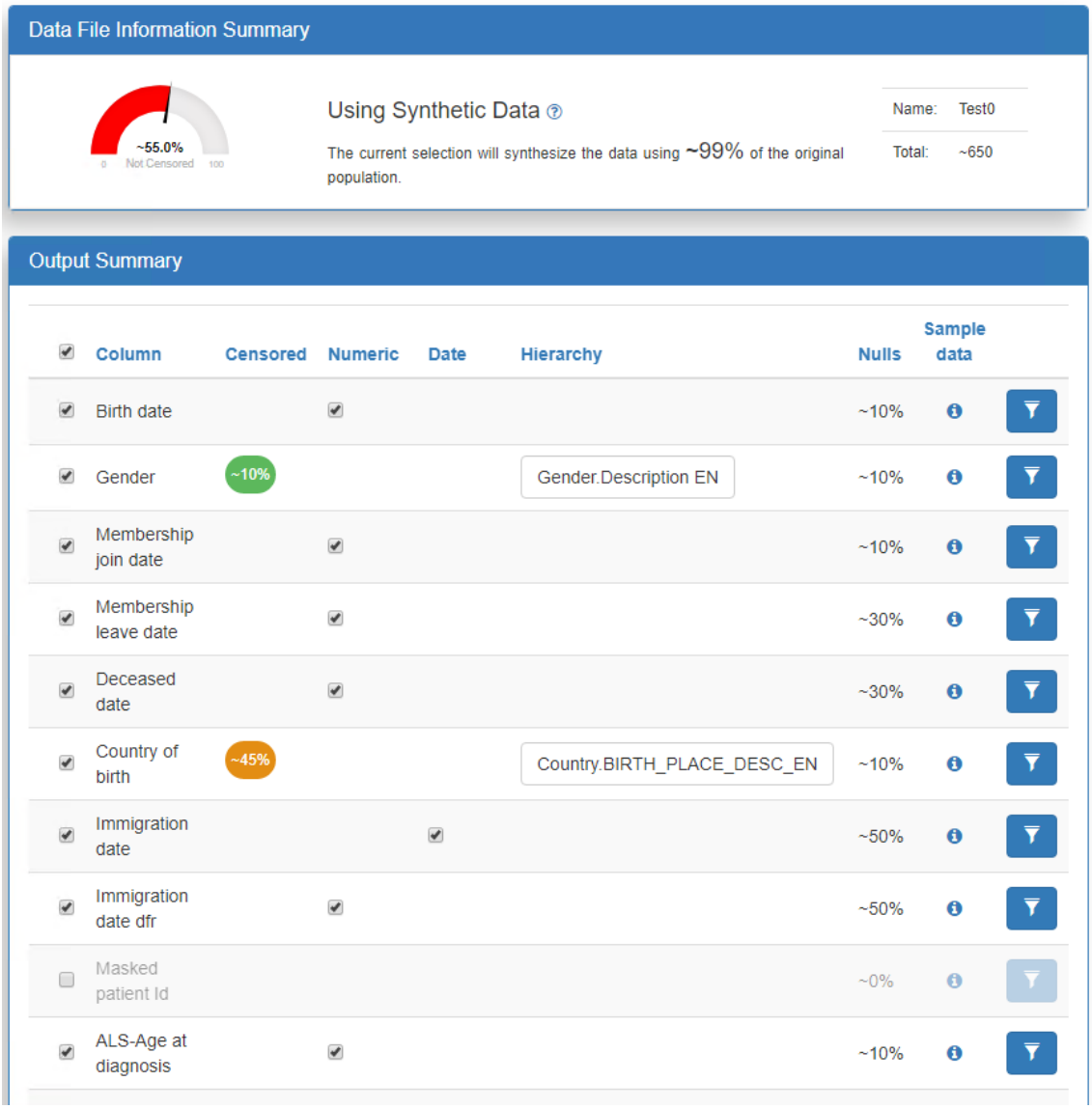


Figure 4.4: Synthetic output summary

Finally, the system allows the download of the synthetic data, along with two extra files: the first containing all the information about the data generated (metadata), the second enabling a comparison between the distribution of synthetic and real data, as a demonstration of the performance power of MDClone system.

4.2.5 Data preprocessing

The variables extracted from MHS database were approximately the same in the synthetic and in the real version of the data. Nonetheless, the download and manipulation techniques applied were different: a distinction in this report is then required.

Synthetic data

The download of the synthetic data was performed in two steps: first, ALS cases were defined and downloaded; then, all the data regarding people without ALS were selected. Since the selected covariates were exactly the same for both people with and without ALS diagnosis, the same operations of preprocessing were executed for both the samples of data.

All the ALS cases were imported as a single data frame of 700 rows and 49 columns. On the other hand - because of the limitation of the downloadable number of rows up to 100,000 - the cohort of people without ALS was downloaded in 30 different data frames. Each data frame contained 38 variables about approximately 100 000 individuals matched by birth date, and was managed as element of a list. A list can be handled with flexibility in an automatic way, avoiding repetitive and potentially leading to error tasks. Moreover, it allows a reduction of the allocated memory of around 30%, improving thus the overall computing performance.

For all the subjects, with and without ALS diagnosis, binary representations of presence or absence of outcome and exposures' diagnoses were created. All the dates included in the analyses were then transformed in four different formats: *date, year, age, days from reference event. Birth dates* and *ages at the diagnosis* were transformed to numeric format. Missing values were uniformed in a common

NA format. Redundant variables - such as the date of diagnosis in *complete* or *age* format - were then removed from the analyses. The final step was the creation of the total dataset, combining the 30 elements of the list in a single big data frame. Descriptive statistics were extracted and results are shown in Table 5.1.

Main analysis

In order to implement high performance analyses while keeping flexibility and dynamism, real MHS data were downloaded via MDClone platform in multiple, smaller datasets. Each dataset contained information about a specific diagnosis or feature regarding only the individuals in the nested case-control previously completed. A list of the extracted datasets, each with its own variables, is provided:

- ALS: [ID | Diagnosis code | Diagnosis date | Diagnosis source | Diagnosis type | Diagnosis start date | Diagnosis end date | Age at diagnosis]
- Diabetes: [ID | Diagnosis code | Diagnosis date | Diagnosis source | Diagnosis type | Diagnosis start date | Diagnosis end date | Age at diagnosis]
- Obesity: [ID | Diagnosis code | Diagnosis date | Diagnosis source | Diagnosis type | Diagnosis start date | Diagnosis end date | Age at diagnosis]
- BMI: [ID | BMI | Date measurement | Age at measurement]
- Riluzole: [ID | Dispensing record | Dispensing date]
- Demographic: [ID | Gender | Birth date | Country of birth | Membership join date | Membership leave date | Deceased date | Immigration date]
- Socio-Economic Status: [ID | SES]
- District: [ID | District]

As shown in the schema above, all the datasets share the ID attribute: this allowed the creation of a merged larger dataset containing, for every individual, all the medical and socio-demographic information needed. The flexibility of the schema gave at the same time the possibility to dynamically merge different subgroups of datasets, assessing just the variables required for every single task.

The first pre-processing operation implemented was the correct codification of the outcome and the exposures. New variables were created, assuming value 1 for the individuals with a diagnosis and value 0 for those without one: this operation was performed for ALS, Diabetes and Obesity diagnosis. To the extent of creating the "overweight" feature, presence or absence of an obesity diagnosis was assessed along with the BMI value: a person was indicated to be overweight if had an obesity record OR a $BMI > 25.0$. (cfr section 4.2.2)

Then, every date required to be managed in a different way, in order to enable operations and comparisons between them. The dates of diagnoses, for example, were stored in a format different from the ones regarding joining or leaving the healthcare system. A common standard was adopted to allow the following operations.

Time relationship between different variables played in fact a crucial role in the analyses. For reasons extensively explained in section 4.2, ALS diagnosis prior to 2001 were removed from the data, as well as every case (along with its matched controls) who had been in the system less than one year before ALS diagnosis. All the controls who had been in the system less than one year before the ALS diagnosis of their matched case were removed too. Finally, a three years lag was imputed between the exposures dates (Diabetes diagnosis, Obesity diagnosis and BMI measurement) and the ALS one. Around 15,000 individuals were removed from the study through this rigorous selection method.

The merging process of different kinds of information caused the creation of a

significant amount of missing values: whether an individual did not have an ALS diagnosis record, for example, all the fields regarding ALS would be filled with *NA*. The absence of information in diagnosis variables, however, is not a real missing value: it states the sanity of the individual. This issue was addressed with the creation of the binary outcome and exposures variables, as previously shown. A solid missing value assessment was anyway required. Further investigation revealed different features revealed different missing values codification systems. The SES missing information, for instance, were codified as "$SES = -1$", while the district missing values were marked as "*missing*"; in order to perform accurate analyses, these and other formats were standardized in a common *Null* format. The amount of missing values in the final dataset is shown in Figure 5.1.

Descriptive statistics were then extracted and results are shown in Table 5.2.

Additional analysis

The sensitivity analysis performed consisted in a different ascertainment of Diabetes cases. In particular, the different definition chosen was based on Maccabi Healthcare System registry. More details about the condition of entrance in the registry can be found in appendix A - the main idea is that this definition combines multiple sources of diagnoses, including clinical ones, medication dispensing records and lab test results. The combination of this multi-factorial diagnoses allows a more accurate and robust Diabetes definition than the one accounting just for the ICD-9 code.

In the registry there are three levels of possible Diabetes diagnosis: low risk pre-diabetes, high risk pre-diabetes and diabetes. The severity of the disease increases on the different levels, so that the number of individuals with a low risk pre-diabetes diagnosis will result much higher than the one of individuals with diabetes (the most severe - and then less common - diagnosis).

In this study, the level of diabetes was codified with a four levels categorical variable. Each individual was assessed for the most severe condition he/she had ever had according to MHS, and so could only have one level of the exposure variable even if in his/her history multiple diagnoses were recorded in different periods. Each level of the exposure variable was imported as a different file and then merged to the main analysis dataset. The same procedures of time management shown in the previous section were required within this analysis as well: the main operations consisted in removing individuals reported to be in the system less than one year in total and applying a three-years lag between the registry diagnosis (no matter which level of severity) and the ALS diagnosis for the cases, as well as a three-years lag between the registry diagnosis and the ALS of the matched cases for all the controls.

The results of this analysis are shown in chapter 5.

4.3 Statistical analysis

The association between ALS and Diabetes Mellitus was modeled in terms of **logistic regression**, a Generalized Linear Model (GLM) widely used in the literature for analyzing the relationship between one (or more) dependent binary variable $\mathbf{Y}$ and one (or more) nominal, ordinal, interval or ratio-level independent variables $\mathbf{X}_{\mathrm{j}}$, also allowing for the assessment of non-linear associations and interactions.

Being i=1,...,n the i-th individual, j=1,...,m the j-th exposure, k=1,...,p the k-th confounding element, a GLM model specifies the relationship between $\mathbf{Y}$ and $\mathbf{X}_{\mathrm{j}}$ as:

$$g(E(Y_{\mathrm{i}})) = \beta_0 + \sum_{j=1}^{m} \beta_{\mathrm{j}} X_{\mathrm{ij}} + \sum_{k=1}^{p} \gamma_{\mathrm{k}} C_{\mathrm{ik}}, \tag{4.1}$$

where E specifies the expected value and g is an invertible and derivable function called *link function* which is determined on the basis of the probability law of $\mathbf{Y}$. The link function allows the linear regression to be generalized to response variables that have error distribution models other than a normal distribution, relating the linear model to a function of the response variable.[44] The GLM allow also to evaluate the interaction between two exposures of interest by simply including a product term of the two exposures in the model.

In a logistic regression model, the link function is the *logit*, the canonical link function for the Bernoulli distribution, defined as follows:

$$logit(p) = log\left(\frac{p}{1-p}\right). \tag{4.2}$$

Being the probability distribution of a logistic function the following:

$$p = Pr(Y_i = 1|X_i) = \frac{exp(\beta_0 + \sum_{j=1}^{m} \beta_{\mathrm{j}} X_{\mathrm{ij}} + \sum_{k=1}^{p} \gamma_{\mathrm{k}} C_{\mathrm{ik}})}{1 + exp(\beta_0 + \sum_{j=1}^{m} \beta_{\mathrm{j}} X_{\mathrm{ij}} + \sum_{k=1}^{p} \gamma_{\mathrm{k}} C_{\mathrm{ik}})}, \tag{4.3}$$

then, from the 4.2 and the 4.3, the logistic regression model can be specified as:

$$logit(p) = \beta_0 + \sum_{j=1}^{m} \beta_{\mathrm{j}} X_{\mathrm{ij}} + \sum_{k=1}^{p} \gamma_{\mathrm{k}} C_{\mathrm{ik}}. \tag{4.4}$$

Since we can define the odds as the probability of an event occurring divided by the probability of the same event not occurring, within the logistic regression model the logit function in 4.2 is the logarithm of the odds of the dependent

variable equaling a case ("p" in 4.4). The Odds Ratio (OR), defined as the odds in presence of an event divided by the odds in absence of the same event, provide an interpretation for the coefficient of the exposures - β_j in 4.4: every one-unit increase in the exposure X_j cause an increase of e^{β_j} in the odds. In other words, the ratio of the odds for a one-unit increase in X_j is e^{β_j}. Given the rarity of the outcome, the odds estimated by the logistic regression can be interpreted as a risk, and having an Odds Ratio interpretation will be crucial in chapter 5: through the ORs and their associated 95% confidence intervals, logistic regression provides a measure of the magnitude of the influence of each predictor on the outcome of interest and of its uncertainty.[45]

Logistic regression also enables adjustments for confounding factors, patient characteristics that are associated with both other predictors and the outcome, so that the measure of the influence of the predictor of interest is not biased by the effect of the confounder. In Figure 4.5 a graphical representation of confounding effects is provided, supposing the presence of just p=2 confounders. To address this situation, both the confounding factors and the predictors of interest are to be included in the model, as shown in models 4.1 and 4.4.

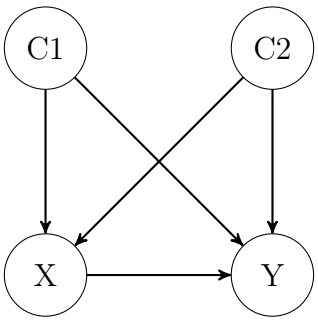


Figure 4.5: Generic X-Y association with p=2 confounders, features affecting both the exposure and the outcome.

The main statistical analysis performed in this study was an implementation of four models of logistic regression. The models were created according to the main

goal of this study - the assessment of an association between ALS and Diabetes Mellitus: while the first model was just an evaluation of the effect of diabetes alone on ALS, the other models were implemented to address for eventual confounding factors. In particular, the second model assessed the confounding effect of being overweight, the third added the interaction between diabetes and overweight, the fourth provided further adjustments for socio-demographic covariates. Finally, each model was implemented both on the overall population and on the population stratified by sex.

Further details about the logistic models are provided:

1. $ALS \sim Diabetes$
2. $ALS \sim Diabetes + Overweight$
3. $ALS \sim Diabetes * Overweight$
4. $ALS \sim Diabetes + Overweight + Country\ of\ birth + District + SES$

The same four models were implemented on the data used for sensitivity analysis purposes as well, assessing the association between ALS and different levels of diabetes - according to MHS registry inclusion criteria. The reference level in the logistic regression model was the non diabetic individuals, and were use to evaluate the influence of low risk pre-diabetes, high risk pre-diabetes and diabetes on ALS. Results are provided in the next chapter.

5. Results

Preliminary analysis

A first preliminary analysis was conducted on the synthetic data provided by MHS via MDClone platform, as shown in section 4.2.4, and results are provided in Table 5.1.

The first thing to notice is how the percentage of individuals with diabetes seems to be really higher within the cases (31%) than in the controls (7%); a similar trend could be detected within people with an obesity diagnosis: 22% of the cases show having one, while only 9.6% of the controls seem to have the same disorder. Moreover, it is interesting to observe how ALS cases are more male patients (58%) than female ones (42%) even if the overall population show a parity in the gender percentages (49.6-50.4%). Finally, synthetic data were partially affected by the presence of *censored* values in all the variables. While theoretically not crucial for statistical purposes, descriptive statistics inevitably show some censored values: an example could be found in the variable *country of birth*.

The analysis of synthetic data reveals partially different results from the main analysis (see Table 5.2); this could certainly be attributed to the different ascertainment method used for exposures variables. Nonetheless, many variables show

a similar distribution, such as country of birth or gender.

Table 5.1: Main covariates distribution, mean (standard deviation) or N(%) in **synthetic** data.

	Overall	**ALS: Yes**	**ALS: No**
N	2,876,125	700	2,875,425
Diabetes	202,575(7%)	220 (31%)	202,355(7%)
Obesity	277,024(9.6%)	114(22%)	276,910(9.6%)
Age		59.6(16.7)	
Birth date	1977(23)	1948(17)	1977(23)
BMI	23(6)	26(5)	23(6)
Gender			
Male	1,428,774(49.6%)	409(58%)	1,428,365(49.6%)
Female	1,447,252(50.4%)	291(42%)	1,446,961(50.4%)
Country of birth			
Israel(78%)	2,243,950	272(39%)	2,243,678(78%)
USSR(12%)	342,474	98(14%)	342,376(12%)
Censored(1%)	27,874	311(44%)	27,563(1%)
Other(9%)	261,827	19(3%)	261,808(9%)

Main analysis

The main analysis was performed on the real MHS data. Descriptive characteristics in the overall population and by levels of ALS are shown in Table 5.2.

Regarding the exposures, in the analyzed dataset 6,602 patients were diagnosed with diabetes, and 78 of them had an ALS diagnosis as well (14% of the total cases); 1,791 individuals were diagnosed with obesity and 13,156 had a BMI greater than 25.0. The overweight patients were then 13,290, and 168 of them had an ALS diagnosis too (31% of the total cases). The difference between the patients with an obesity diagnosis (occurring only at 3% of the population) and the overweight individuals (25% of the population) is a consequence of the inclusion of BMI values to obesity in the overweight definition: while the obesity diagnosis is very severe,

the primary condition for being defined overweight (BMI $>$ 25) is much more common.

All the variables appear to keep a uniform distribution for different values of the outcome: both the socio - demographic information and the medical one seem not to significantly change in presence or absence of ALS. Socio-economic status, for example, occurs with the same mean and standard deviation in cases as well as in controls; the same happens for obesity diagnosis and BMI measurement. It is possible that this generally uniform distribution could be caused by the homogeneity of Maccabi ecosystem, which all people in the study belong to.

An interesting difference has been detected in the amount of male and female patients with ALS diagnosis: while in people without ALS the distribution appears to be mostly uniform, male patients seems more likely to be at risk of ALS. In particular 322 patients out of 547 cases were male (59%), making a crude male : female ratio of 1.4. These findings are consistent with other studies results (cfr [46]) and show the importance of a stratification by gender in the following analysis.

Table 5.2: Main covariates distribution, mean (standard deviation) or N(%) in **real** data.

	Overall	**ALS: Yes**	**ALS: No**
N	52,254	547	51,707
Diabetes	6,602(13%)	78(14%)	6,524(13%)
Overweight	13,290(25%)	168(31%)	13,122(25%)
Age		59.9(15.7)	
Birth date	1949(15.7)	1950(16)	1949(15.7)
Gender			
Male	24,986(48%)	322(59%)	24,664(48%)
Female	27,268(52%)	225(41%)	27,043(52%)
SES (1-10)	6.1(1.8)	6.2(1.8)	6.1(1.8)
Country of birth			
Israel	27,697(53%)	284(52%)	27,692(53%)
USSR	12,341(24%)	124(23%)	12,217(24%)
Other	11,937(23%)	139(25%)	11,798(23%)
District			
Jerusalem&Shfela	12,535(24%)	116(21%)	12,419(24%)
Center	11,257(22%)	124(23%)	11,133(21%)
North	10,159(19%)	88(16%)	10,071(19%)
Sharon	9,934(19%)	122(22%)	9,812(19%)
South	8,065(16%)	72(13%)	7,993(16%)
Obesity	1,791(3%)	16(3%)	1,775(3%)
BMI	27.9(5.3)	27.6(5)	27.9(5.3)

As introduced in section 4.2.5 the amount of missing values in the data was significant for many variables. However, the lack of information for most of the features included in the study was considered as a useful asset itself: an individual without a diagnosis of ALS, for instance, was not to identify as a missing value, but as a person not affected by that disease. In the pre-processing section it has been described how these cases were transformed in information suitable for analysis. In Figure 5.1 is shown the percentage of missing values in each variable. The highest percentages are to be found in diagnosis-related variables, such as diagnosis date,

type or source; on the other hand the lowest percentages of missingness are to be found in socio-demographic information, such as SES, country of birth or gender, as well as in ad-hoc created variables, such as the binary exposure and outcome ones.

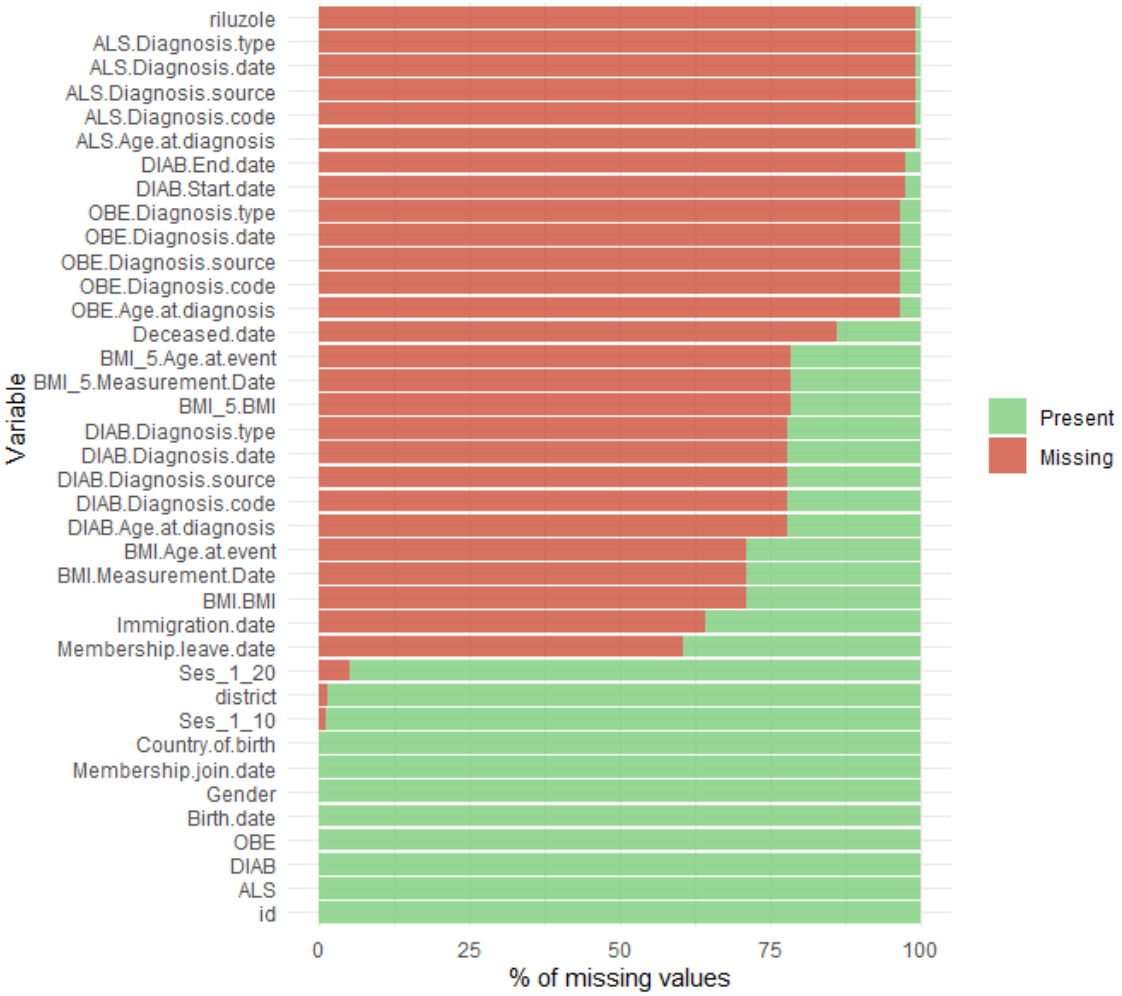


Figure 5.1: Missing values in each variable

Some additional Figures are provided, with the aim of giving a more detailed picture of the analyzed population.

In Figure 5.2 is shown the differences between Socio-Economic Status in multiple districts. There is not such a dramatic contrast between different districts,

even if Center and Sharon ones seem to be the richest: this homogeneity in results probably depends from the region - Maccabi - to which all the individuals in the study belong to.

Figure 5.2: Distribution of Maccabi population across districts and SES levels

Figure 5.3 shows the years distribution of the ALS diagnoses included in this study: as introduced in section 4.2.5 only ALS diagnosis after 2001 were taken into account, in order to satisfy the needs of the nested case control study implemented. The number of cases seems not to significantly change across different years.

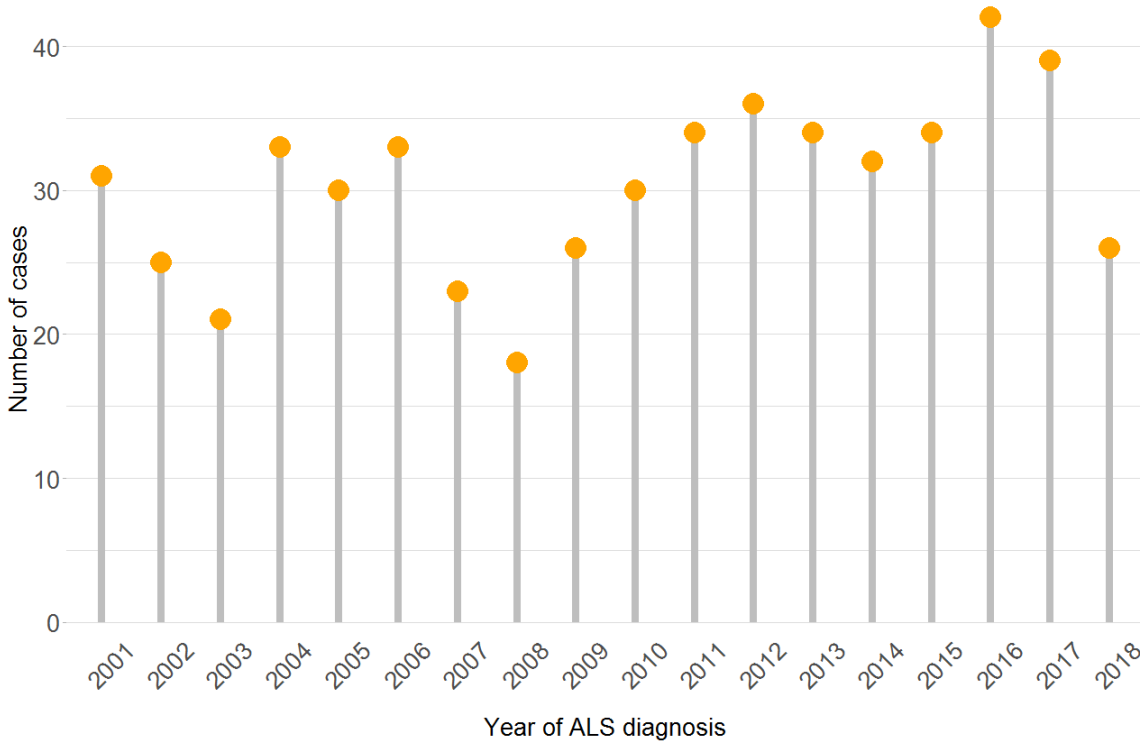


Figure 5.3: Distribution of ALS diagnoses in the considered time lag

Figure 5.4 and Figure 5.5 provide the distribution of the population's age at the time of, respectively, ALS and Diabetes Mellitus diagnosis, divided by gender. For both the diseases the average age at diagnosis is considerably high, as it's been proved in multiple studies (see section 2), and no significant differences between male and female patients were detected.

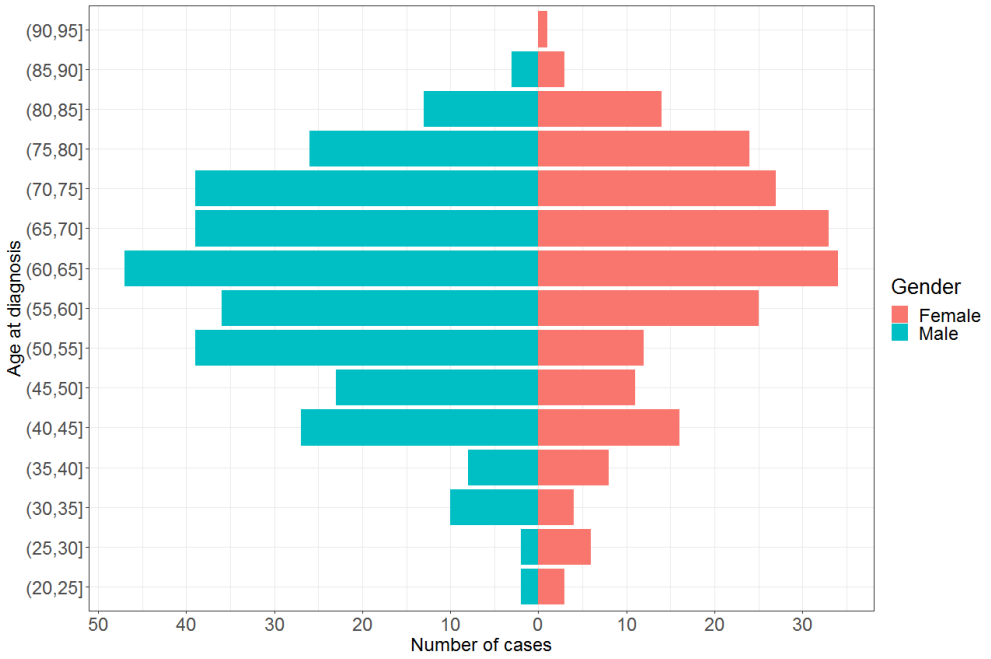


Figure 5.4: Distribution of patients' age at ALS diagnosis, with gender distinction

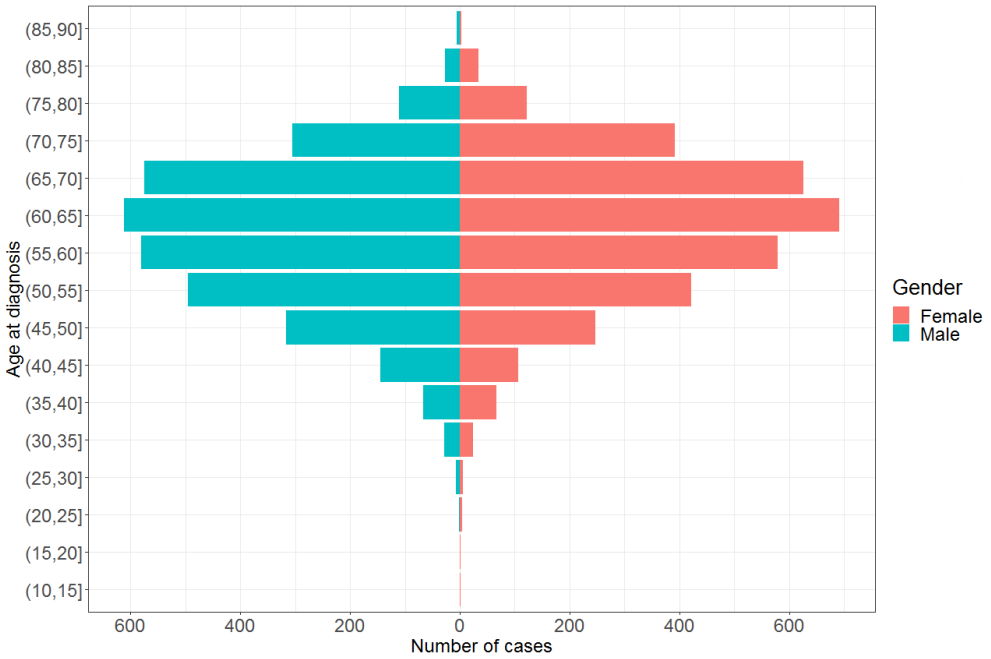


Figure 5.5: Distribution of patients' age at diabetes diagnosis, with gender distinction

As introduced in section 4.3, the main statistical analysis performed in this study consisted of an implementation of four logistic regression models. The results of the models on the overall population are shown in Table 5.3. The Diabetes Odds Ratios seem not to be statistically significant, as each model shows odds ratios close to 1; the results of the single diabetes - ALS assessment appear consistent with the adjustments for possible confounders, as the ones used in models 2, 3 and 4.

An interesting trend could be detected in Tables 5.4 and 5.5, as they show the odds ratios of the same models as before, but estimated on the population stratified by gender. Even if the odds ratio are all close to 1 and no statistically significant results have been found, nonetheless the differences between the two sub-populations seem to be relevant: the odds ratios appear to go in opposite directions, suggesting that diabetes could be more likely to have a positive association with ALS in male individuals than in female ones. This result is consistent across all the different models, including adjustments for confounding factors such as being overweight and socio-demographic features (see section 4.3 for further details). The interaction between gender and diabetes seems however not to be statistically significant (this result is not shown).

Table 5.3: Odds ratio and 95% confidence intervals in different logistic models.

Models	**Diabetes**	**Overweight**	**Interaction**
Model 1	1.17 (0.91 1.49)		
Model 2	1.07 (0.83 1.38)	1.29 (1.06 1.56)	
Model 3	1.24 (0.87 1.78)	1.35 (1.10 1.66)	0.75 (0.46 1.23)
Model 4	1.14 (0.88 1.48)	1.37 (1.13 1.66)	

Table 5.4: Odds ratio and 95% confidence intervals in different logistic models on female patients only.

Models	**Diabetes**	**Overweight**	**Interaction**
Model 1	0.92 (0.60 1.39)		
Model 2	0.83 (0.54 1.27)	1.34 (1.00 1.81)	
Model 3	0.73 (0.37 1.45)	1.31 (0.95 1.80)	1.22 (0.51 2.93)
Model 4	0.89 (0.57 1.40)	1.57 (1.15 2.13)	

Table 5.5: Odds ratio and 95% confidence intervals in different logistic models on male patients only.

Models	**Diabetes**	**Overweight**	**Interaction**
Model 1	1.32 (0.97 1.80)		
Model 2	1.23 (0.89 1.69)	1.24 (0.97 1.59)	
Model 3	1.59 (1.04 2.43)	1.37 (1.05 1.80)	0.59 (0.32 1.10)
Model 4	1.28 (0.93 1.77)	1.25 (0.97 1.60)	

Additional analysis

As introduced in section 4.2.5 an additional analysis was implemented, in order to assess the diabetes diagnoses with different ascertainment methods. In particular, the criteria adopted was the one according to the MHS diabetes registry, which includes a broader set of conditions to be diagnosed as diabetic (check appendix A for more details).

In Table 5.6 the distribution of Diabetes diagnosis according to MHS registry is provided. Since the definition of the diagnosis as the most severe condition appeared in a patient at least three years before ALS onset, some differences between multiple diabetes levels appear to be evident. For instance, the most common disorder in the population of analysis seems to be the low risk pre-diabetes (14% of the total population); this result is reliable according to the definition of

the diagnosis itself, because of the low level of severity of the disease. On the other hand, the number of individuals with a diagnosis of high risk pre-diabetes seems to be the lowest, even if the disease is not the most severe. Digging deeper in the data it has been verified that most of the patients with a high risk pre-diabetes diagnosis showed a diabetes diagnosis as well: because of this, and because of this analysis taking into account just the most severe condition for each individual, many patients with both the conditions were appointed with just the diabetes diagnosis. This consideration gives a reasonable explanation to the high number of diabetic patients (9% of the total population) as well, despite the severity of the diagnosis.

Table 5.6: Distribution - N(%) - of Diabetes diagnoses according to MHS registry.

	Overall	**ALS: Yes**	**ALS: No**
N	52,254	547	51,707
Low risk pre-diabetes	7,418(14%)	109(20%)	7,309(14%)
High risk pre-diabetes	311(0.6%)	4(0.7%)	307(0.6%)
Diabetes	4,843(9%)	53(10%)	4,790(9%)

Comparing Table 5.6 and Table 5.2 it is interesting to notice how different the diabetes ascertainment methods are. As a matter of fact, the diabetes definitions in the registry are more specific, meaning a patient needs to hit a higher threshold of symptoms to join the registry. A single diagnosis, for example, recorded using the ICD-9 code is not enough to join the registry: a patient need to also have at least a confirmatory lab result or a dispensing record for diabetes-specific medication to join. So the ICD-9 definition is very broad, and potentially less accurate: here comes the need for a sensitivity analysis in this direction, since the registry does a much better job weeding out those suspected cases that were never confirmed. As ensured by the Tables, the prevalence based on ICD-9 codes (13% of the total

population) is thus higher than the one based on the registry (9% of the total population).

The following logistic regression analysis was implemented in order to verify the consistency of the previous findings. The Odds Ratios of the most severe diabetes level recorded in MHS registry were extracted from the same four models as before - along with the 95% confidence intervals - and are here provided.

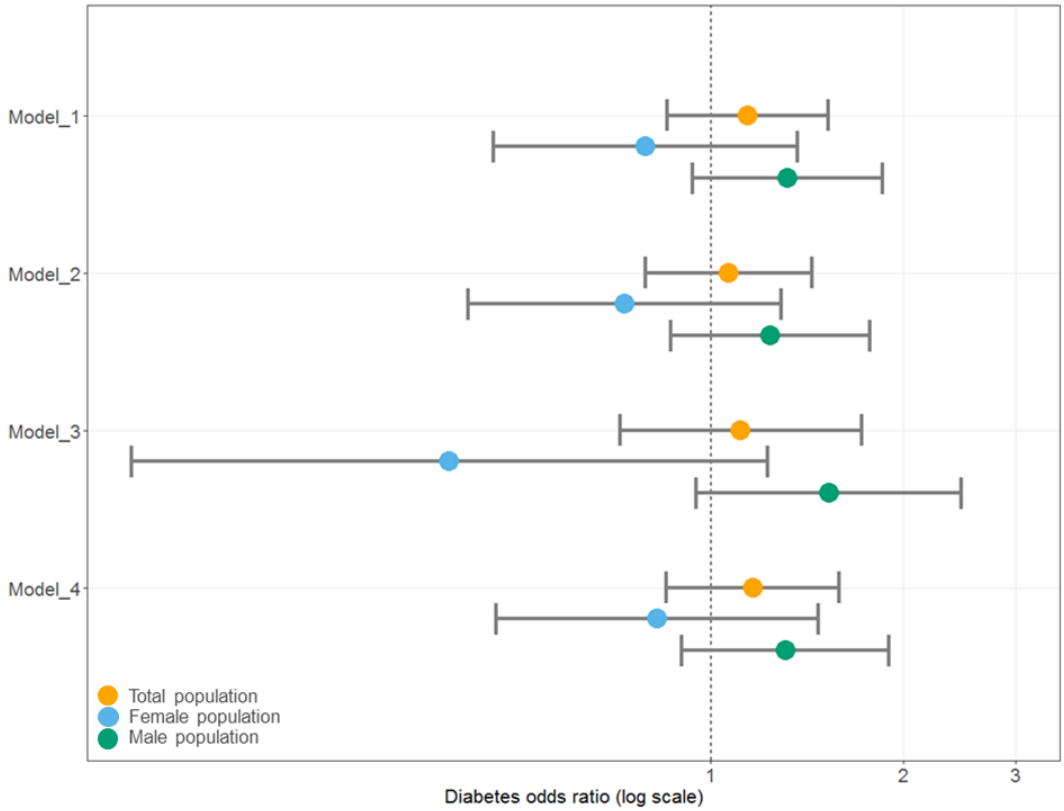


Figure 5.6: Diabetes Odds Ratios according to MHS registry

6. Discussion

In order to evaluate medical risk factors for ALS, the need for a large, well characterized population has been recognized. At the same time, the higher volume of information to be processed requires novel, high-performance technical approaches to be implemented. In this study a nested case control study was performed on a large Israeli population, in order to assess the relationship between ALS and Diabetes Mellitus. After the data mining and pre-processing, a logistic regression model was implemented in order to evaluate the association between the two diseases, adjusting for multiple confounding factors. An additional sensitivity analysis - conducted using the diabetes definition of MHS registry - proved the consistency of the results gained in the previous analysis, allowing two main considerations. First, Diabetes Mellitus does not seem to have a statistically significant association with ALS; this result is suggested to be consistence because of the absence of meaningful variation between the different models. Furthermore, stratifying the population by gender an evident contrast between the two sub-populations appears: despite all the confidence intervals include the 1 (meaning that no evidence of statistical significance was detected), the odds ratios produced by female patients seem to be always smaller than the odds ratios produced by male patients, and most importantly in the opposite directions. This could imply that

the mechanisms relating diabetes and ALS are different between men and women. These findings are consistent with those introduced in the main analysis section. However, they are not intended to give a resolutive and definitive answer to our research question: instead, they call for further assessments in order to deeper investigate the validity of the results of this study.

In this section further considerations about the study conducted, both medical and methodological, will be provided.

6.1 Medical considerations

There is evidence that ALS is more common in men than in women and that gender influences the clinical features of the disease: the male : female ratio has been reported as being between 1 and 3 (cfr [46]) and this thesis' results confirm these numbers. Possible reasons for the differences in ALS between men and women include different exposures to environmental toxins, different biological responses to exogenous toxins, and possibly underlying differences between the male and female nervous systems and different abilities to repair damage.[47] However, no previous record for gender differences in assessing the diabetes-ALS association were detected. Still, as anticipated in section 2.1, multiple studies have investigated the relationship between diabetes and ALS, producing conflicting results: the three main researches conducted on Danish, Swedish and Taiwanese individuals provided opposite results regarding the role of Diabetes Mellitus in the oldest patients with ALS. All of these studies - and this thesis as well - were based on administrative data; the accuracy of the diagnosis, both of the outcome and of the exposure, plays then a crucial role in the reliability of the results. High availability of detailed medical records, such as the one we have in MHS database, certainly helps in this direction: a more robust ascertainment of the outcome and exposure

variables is possible thanks to the combination of medications, lab tests and clinical diagnoses information. In this study the implemented sensitivity analysis showed the consistency of the previously obtained results through the application of a more rigorous definition of diabetes. The results regarding the overweight diagnosis, as well as the interaction between it and diabetes, were provided even though they were not meaningful to the end of this analysis: the variables were introduced only to address their possible confounding effects in assessing the diabetes-ALS relationship. The interpretation of the related Odds Ratio is then not so significant.

6.2 Methodological considerations

The need of a large detailed population to study, because of ALS rarity - see section 2.2, has brought researchers to look closer to the Big Data paradigm.

Every day, people all over the world are generating a massive amount of data. In 2013 the amount of stored information in the world was estimated to be around 1,200 exabytes, of which less than 2 percent non-digital.[48] An exabyte is one billion gigabytes. International Data Corporation (IDC) predicted that the "digital universe" would expand to 40,000 exabytes by the year 2020.[49] Such large amounts of data constitute "Big Data". There is no rigorous definition of big data, but basically we refer on data that is so large in volume, so diverse in variety or moving with such a velocity, that traditional techniques of assessment and analysis are insufficient - characteristics colloquially denoted as the "3 Vs". The declining cost of collection, storage, and processing of data, combined with new sources of data like sensors, cameras, geospatial and other observational technologies, means that we live in a world of near-ubiquitous data collection.[50]

One of the most promising fields where big data can be applied to make a change is healthcare. It is defined as a multi-dimensional system established with

the aim for the prevention, diagnosis, and treatment of health-related issues or impairments in human beings. The health professionals (physicians or nurses) are responsible for different kinds of information such as patient's medical history (diagnosis and prescriptions related data), medical and clinical data (like data from imaging and laboratory examinations), and other private or personal medical data. Previously, the common practice to store such medical records for a patient was in the form of either handwritten notes or typed reports. With the advent of computer systems and its potential, the digitization of all clinical exams and medical records in the healthcare systems has become a standard and widely adopted practice nowadays. Electronic health records (EHR) as defined by Murphy, Hanken and Waters are "computerized medical records for information relating to the past, present or future physical/mental health or condition of an individual which resides in electronic system(s) used to capture, transmit, receive, store, retrieve, link and manipulate multimedia data for the primary purpose of providing healthcare and health-related services".[51] EHR, electronic medical records (EMR), personal health records (PHR), medical practice management software (MPM) and many other healthcare data components collectively have the potential to improve the quality and efficiency of services, predict outbreaks of epidemics, gain valuable insights, avoid preventable diseases, reduce the costs of healthcare delivery as well as the amount of medical errors, while improving the quality of life in general. However, a large proportion of this data is currently unstructured, so that it does not adhere to a pre-defined model or organizational framework. It is difficult to group such varied, yet critical, sources of information into an intuitive or unified data format for traditional data analyses. Nonetheless, the healthcare industry is required to utilize the full potential of these rich streams of information to enhance the patient experience.[49]

Big data can be leveraged to measure hospital quality, to develop scientific

hypotheses, to compare effectiveness of different interventions, to monitor drug and device safety. In order to achieve these goals, we need to manage and analyze these huge amounts of data in a systematic manner; at the same time, the rising of ethical and moral concerns demands the world to think about the so-called data revolution from a responsible perspective. As mentioned in section 2.3 privacy underprotection and overprotection each create cognizable harms to patients both today and tomorrow. To ensure a secure and trustworthy big data environment, it is essential to identify the limitations of existing solutions and envision directions for future research.

7. Final remarks

As the understanding of ALS pathogenesis remains largely unknown, still some steps forward towards its investigation has been made. In this thesis a huge, yet uncharted population has been explored, in order to replicate the results of previous studies regarding ALS and other medical conditions. In particular, following the suggestion of several reports, the association between ALS and diabetes, ascertained according to ICD-9 diagnosis codes, was examined through a logistic regression, with assessment of confounding effects from multiple factors. The findings of this study are mostly consistent with the previous ones, suggesting a stronger positive association between male diabetic patients and ALS rather than the one concerning female diabetic ones. Considerations about rising security and privacy concerns were provided as well, through the analysis of the synthetic data provided by MDClone platform.

7.1 Strengths and limitations

The strengths of this study are considerable. The most innovative element has been represented by the high volume of the analysed data: very few studies considered such a detailed and large population. The size of MHS database allowed the

inclusion of more than 500 ALS cases in the study, granting a high accuracy and robustness of the results. The consistency of the results with previous studies, despite being the first ALS analysis within Maccabi patients, confirms the interest of the findings; at the same time the gender influence in the association between ALS and diabetes has never been recorded, suggesting the need of further analyses.

On the other hand several limitations can be identified. First, the outcome and exposure ascertainment methods just relying on the ICD-9 code could lead to slightly biased results; as we did in the additional analysis for diabetes, a more accurate assessment of ALS - as well of obesity and other confounders - diagnoses could improve the accuracy of the results. This results could be pursued, for example, with the inclusion of medication records in the ascertainment process. Besides, diabetes ascertainment could be improved with the introduction of a further assessment for the patients age, and distinct analyses with type 1 and 2 diabetes. Finally, the absence of data from before the availability of electronic medication data (1998) could lead to outcome and exposure misclassification: the study could have missed a subject diagnosis because of time-related issues. However, this should not be a frequent situation, allowing the gained results to be enough accurate with the application of some time restriction (see section 4.2.5).

7.2 Future research

The data provided by Maccabi Health Services provide tremendous opportunities for the investigation of ALS related disorders.

At the beginning, a further exploration of the association between ALS and Diabetes Mellitus could be performed. The first step could include the distinction between Diabetes type 1 and 2, assessing how age could affect both the diagnoses. Then additional disorders could be taken into account in the relationship,

addressing for more confounding factors such as cardiovascular diseases, hypercholesterolemia, trauma and other conditions.

Then, time factor could be certainly be taken into account. MHS data provide information about every single medical event occurring to each patient, allowing researchers to study trajectories and changing in trends during different years. This element could be used to assess the variation of patients' BMI over time, for example, in relationship with ALS or diabetes diagnoses.

A further and dramatically crucial study could then involve the prescription or dispensing records for medications. While some medications have been explored in epidemiological studies of ALS, researchers have only covered a small number of medications, one at a time, and often in small studies. Given the multitude of medications prescribed, the frequent taking of multiple medications by a single person (especially in the older age range typical of ALS onset), and the possibility of interactions between different medications, there is the need of a more efficient exploration of the very large dataspace MHS provides. This analysis could lead to critical results, in terms both of the identification of the most promising candidates for medications of relevance for ALS and of more accurate results in previously conducted studies. The first element is strongly pursued by researchers, since the only currently effective medication is Riluzole, which just extends the average survival of ALS patients by a few months.[4] But the second point is crucial as well: the inclusion of medication-related information could certainly lead, for instance, to a more robust and accurate ascertainment of each diagnosis, from ALS to Diabetes or Obesity as well. Besides the simple medications prescription or dispensing records, variation in taking time and dosage information could be included in the studies as well.

Finally, in order to adequately evaluate the relationship between complex and multi-factorial elements such as the ones introduced in this section, advanced and

innovative technological solutions are required: from this perspective, more than traditional statistical approaches, novel machine learning techniques could play a central role in the future of epidemiological research.

A. Appendix

Some further details about the criteria for diabetes registry in MHS are provided.

A patient is diagnosed with **Diabetes Mellitus** according to MHS registry in occurrence of at least one of these conditions:

1. Level of Glycated haemoglobin (HbA1c) ≥ 7.25.
2. At least two random Glucose test ≥ 200 with gap of a minimum of 30 days between tests.
3. Oral drugs: purchase of at least two oral medications for diabetes in a three months period, with one of the following: 1) at least one glucose test above 125; 2) $HbA1c \geq 6.5$. Note that:
    - Purchase of oral medications on the same day counts as only one purchase.
    - A woman of childbearing age (18-55) with a GCT or GTT test (no matter the value), whose last purchase of an oral antidiabetic drug was within less than four months from the data of the test does not enter the registry (added to the gestational diabetes registry).

4. Insulin: Purchase of at least two insulin medications in a three-month period. Note that:

    - A patient who entered the registry based on dispensing of insulin alone and who did not have additional dispensing records for insulin in the 6 consecutive months following the initial dispensing date will be removed from the registry.
    - Purchase of multiple insulin drugs on the same day counts as only one purchase.
    - A woman of childbearing age (18-55) with a GCT or GTT test (no matter the value), whose last purchase date of insulin was within less than four months from the date of the test does not enter the registry (added to the gestational diabetes registry).
    - Diagnoses (based on ICD-9 codes given by a family physician/ pediatrician, diabetes consultant, endocrinologist, or ophthalmologist) with one of the following
        (a) $HbA1c \geq 6.5$ within six months before or after diagnosis.
        (b) Two glucose tests ≥ 126, one six months before or after diagnosis and one at any point in time.

A patient is diagnosed with **High risk of Pre-Diabetes** according to MHS registry in occurrence of at least one of these conditions:

1. Level of Glycated haemoglobin (HbA1c) between 6.5 and 7.24 (ever)
2. Fasting glucose between 126 and 199 (at least two abnormal values in a 3 year period)
3. Two hour OGTT 75 gram ≥ 200 (ever)

A patient is diagnosed with **Low risk of Pre-Diabetes** according to MHS registry in occurrence of at least one of these conditions:

1. Level of Glycated haemoglobin (HbA1c) between 5.7 and 6.4 (ever)
2. Fasting glucose between 100 and 125 (at least two abnormal values in a 3 year period)
3. Two hour OGTT 75 gram between 140 and 199 (ever)
4. Diagnosis with ICD-9 code 790.2

Acknowledgements

The work of these months has been possible thanks to the support and contribution of many people.

First, I would like to thank **Giorgio Vittadini, Ran Rotem** and **Andrea Bellavia**, my thesis supervisors. Thanks for giving me the opportunity of working with you, receiving continuous support and mentoring during the last six months (and even before). Besides the knowledge and teachings you've shared with me, I really want to thank you for having me introduced to the world of academia, making me freely exploring benefits and downsides of research. The attention you have given me is the most sincere proof of how wonderful teachers you are: I have learnt a lot.

A special thanks to **Marc Weisskopf** and all his wonderful **team**: you guys have welcome me like an old friend, and I have always felt like being home with all of you. Thanks in particular to **Vy** for the precious tips on my terrible English.

A particular thanks to **Fondazione C.E.U.R.** for making this thesis possible with your generous scholarship.

Thank also to all the Environmental Health Department of the T. H. Chan Harvard School of Public Health for the hospitality of these months. Thank particularly to **Allison** for being such a precious guide in all the (almost infinite) bureaucracy I had to correctly follow in order to join Harvard.

Last, but not least, thanks to all my family and friends both in Italy and USA. I will give you all special credits further on up the road.

References

[1] P. Couratier, P. Corcia, G. Lautrette, M. Nicol, P.-M. Preux, and B. Marin, "Epidemiology of amyotrophic lateral sclerosis: A review of literature," *Revue Neurologique*, vol. 172, pp. 37–45, Jan. 2016. 1, 3, 4

[2] J. Finsterer, "Accuracy of clinical and electrophysiological criteria for diagnosing amyotrophic lateral sclerosis," *Clinical Neurophysiology*, vol. 127, pp. 2682–2683, July 2016. 1

[3] N. Riva, F. Agosta, C. Lunetta, M. Filippi, and A. Quattrini, "Recent advances in amyotrophic lateral sclerosis," *Journal of Neurology*, vol. 263, pp. 1241–1254, June 2016. 1

[4] C. Ingre, P. M. Roos, F. Piehl, F. Kamel, and F. Fang, "Risk factors for amyotrophic lateral sclerosis," *Clinical Epidemiology*, vol. 7, pp. 181–193, Feb. 2015. 2, 4, 55

[5] L. Xu, T. Liu, L. Liu, X. Yao, L. Chen, D. Fan, S. Zhan, and S. Wang, "Global variation in prevalence and incidence of amyotrophic lateral sclerosis: a systematic review and meta-analysis," *Journal of Neurology*, Dec. 2019. 2, 4, 9

[6] A. Jawaid, A. R. Salamone, A. M. Strutt, S. B. Murthy, M. Wheaton, E. J. McDowell, E. Simpson, S. H. Appel, M. K. York, and P. E. Schulz, "ALS disease onset may occur later in patients with pre-morbid diabetes mellitus: Diabetes Mellitus and ALS," *European Journal of Neurology*, vol. 17, pp. 733–739, May 2010. 2, 5

[7] G. Logroscino, B. J. Traynor, O. Hardiman, A. Chio', P. Couratier, J. D. Mitchell, R. J. Swingler, E. Beghi, and for EURALS, "Descriptive epidemiology of amyotrophic lateral sclerosis: new evidence and unsolved issues," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 79, pp. 6–11, Jan. 2008. 2, 12

[8] B. Marin, F. Boumédiene, G. Logroscino, P. Couratier, M.-C. Babron, A. L. Leutenegger, M. Copetti, P.-M. Preux, and E. Beghi, "Variation in worldwide incidence of amyotrophic lateral sclerosis: a meta-analysis," *International Journal of Epidemiology*, vol. 46, pp. 57–74, Feb. 2017. 4, 5

[9] L. Dupuis, P.-F. Pradat, A. C. Ludolph, and J.-P. Loeffler, "Energy metabolism in amyotrophic lateral sclerosis," *The Lancet Neurology*, vol. 10, pp. 75–82, Jan. 2011. 5

[10] E. T. Reyes, O. H. Perurena, B. W. Festoff, R. Jorgensen, and W. V. Moore, "Insulin resistance in amyotrophic lateral sclerosis," *Journal of the Neurological Sciences*, vol. 63, pp. 317–324, Mar. 1984. 5

[11] J. C. Desport, P. M. Preux, L. Magy, Y. Boirie, J. M. Vallat, B. Beaufrère, and P. Couratier, "Factors correlated with hypermetabolism in patients with amyotrophic lateral sclerosis," *The American Journal of Clinical Nutrition*, vol. 74, pp. 328–334, Sept. 2001. 5

[12] A. Jawaid, R. Khan, M. Polymenidou, and P. E. Schulz, "Disease-modifying effects of metabolic perturbations in ALS/FTLD," *Molecular Neurodegeneration*, vol. 13, p. 63, Dec. 2018. 5, 12

[13] J. O'Reilly, H. Wang, M. G. Weisskopf, K. C. Fitzgerald, G. Falcone, M. L. McCullough, M. Thun, Y. Park, L. N. Kolonel, and A. Ascherio, "Premorbid body mass index and risk of amyotrophic lateral sclerosis," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 14, pp. 205–211, Apr. 2013. 5

[14] P. Zeng, X. Yu, and H. Xu, "Association Between Premorbid Body Mass Index and Amyotrophic Lateral Sclerosis: Causal Inference Through Genetic Approaches," *Frontiers in Neurology*, vol. 10, p. 543, May 2019. 5

[15] S. Paganoni, T. Hyman, A. Shui, P. Allred, M. Harms, J. Liu, N. Maragakis, D. Schoenfeld, H. Yu, N. Atassi, M. Cudkowicz, and T. M. Miller, "Pre-morbid Type 2 Diabetes Mellitus is not a prognostic factor in ALS," *Muscle & nerve*, vol. 52, pp. 339–343, Sept. 2015. 6

[16] C. S. Mitchell, S. K. Hollinger, S. D. Goswami, M. A. Polak, R. H. Lee, and J. D. Glass, "Antecedent Disease is Less Prevalent in Amyotrophic Lateral Sclerosis," *Neuro-Degenerative Diseases*, vol. 15, no. 2, pp. 109–113, 2015. 6

[17] M.-A. Kioumourtzoglou, R. S. Rotem, R. M. Seals, O. Gredal, J. Hansen, and M. G. Weisskopf, "Diabetes Mellitus, Obesity, and Diagnosis of Amyotrophic Lateral Sclerosis: A Population-Based Study," *JAMA Neurology*, vol. 72, p. 905, Aug. 2015. 6

[18] D. Mariosa, F. Kamel, R. Bellocco, W. Ye, and F. Fang, "Association between diabetes and amyotrophic lateral sclerosis in Sweden," *European Journal of Neurology*, vol. 22, pp. 1436–1442, Nov. 2015. 6

[19] Y. Sun, C.-J. Lu, R.-C. Chen, W.-H. Hou, and C.-Y. Li, "Risk of Amyotrophic Lateral Sclerosis in Patients With Diabetes: A Nationwide Population-Based Cohort Study," *Journal of Epidemiology*, vol. 25, no. 6, pp. 445–451, 2015. 6

[20] G. Logroscino, "Are diabetes and amyotrophic lateral sclerosis related?," *Nature Reviews Neurology*, vol. 11, pp. 488–490, Sept. 2015. 6

[21] F. D'Ovidio, A. d'Errico, P. Carnà, A. Calvo, G. Costa, and A. Chiò, "The role of pre-morbid diabetes on developing amyotrophic lateral sclerosis," *European Journal of Neurology*, vol. 25, pp. 164–170, Jan. 2018. 7

[22] A. Nasrolahi, J. Mahmoudi, A. Noori-Zadeh, K. Haghani, S. Bakhtiyari, and S. Darabi, "Shared Pathological Mechanisms Between Diabetes Mellitus and Neurodegenerative Diseases," *Current Pharmacology Reports*, vol. 5, pp. 219–231, Aug. 2019. 7

[23] S. Auvin, J. Irwin, P. Abi-Aad, and A. Battersby, "The Problem of Rarity: Estimation of Prevalence in Rare Disease," *Value in Health*, vol. 21, pp. 501–507, May 2018. 8

[24] A. Chiò, G. Logroscino, O. Hardiman, R. Swingler, D. Mitchell, E. Beghi, B. G. Traynor, and O. B. o. t. E. Consortium, "Prognostic factors in ALS: A critical review," *Amyotrophic Lateral Sclerosis*, vol. 10, pp. 310–323, Jan. 2009. 8

[25] "Amyotrophic Lateral Sclerosis (ALS) Fact Sheet | National Institute of Neurological Disorders and Stroke." 9

[26] E. Tiryaki and H. A. Horak, "ALS and Other Motor Neuron Diseases:," *CONTINUUM: Lifelong Learning in Neurology*, vol. 20, pp. 1185–1207, Oct. 2014. 10

[27] F. Agosta, A. Al-Chalabi, M. Filippi, O. Hardiman, R. Kaji, V. Meininger, I. Nakano, P. Shaw, J. Shefner, L. H. van den Berg, and A. Ludolph, "The El Escorial criteria: Strengths and weaknesses," *Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration*, vol. 16, pp. 1–7, Mar. 2015. 10

[28] M. G. Lacy, "Efficiently Studying Rare Events: Case-Control Methods for Sociologists," *Sociological Perspectives*, vol. 40, no. 1, pp. 129–154, 1997. 10

[29] N. Mantel and W. Haenszel, "Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease," *JNCI: Journal of the National Cancer Institute*, vol. 22, pp. 719–748, Apr. 1959. 10

[30] W. Fithian and T. Hastie, "Local case-control sampling: Efficient subsampling in imbalanced data sets," *The Annals of Statistics*, vol. 42, pp. 1693–1724, Oct. 2014. arXiv: 1306.3706. 11

[31] V. L. Ernster, "Nested Case-Control Studies," *Preventive Medicine*, vol. 23, pp. 587–590, Sept. 1994. 11

[32] S. Wacholder, J. K. McLaughlin, D. T. Silverman, and J. S. Mandel, "Selection of Controls in Case-Control StudiesI. Principles," *American Journal of Epidemiology*, vol. 135, pp. 1019–1028, May 1992. 11

[33] Z. Zheng, L. Sheng, and H. Shang, "Statins and amyotrophic lateral sclerosis: a systematic review and meta-analysis," *Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration*, vol. 14, pp. 241–245, Aug. 2013. 12

[34] K. Abouelmehdi, A. Beni-Hessane, and H. Khaloufi, "Big healthcare data: preserving security and privacy," *Journal of Big Data*, vol. 5, no. 1, pp. 1–18, 2018. 13

[35] W. N. Price and I. G. Cohen, "Privacy in the age of medical big data," *Nature medicine*, vol. 25, no. 1, pp. 37–43, 2019. 13, 14

[36] P. Saha-Chaudhuri and C. R. Weinberg, "Addressing data privacy in matched studies via virtual pooling," *BMC medical research methodology*, vol. 17, p. 136, Sept. 2017. 14

[37] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. Bennett, "Privacy Preserving Synthetic Health Data," Apr. 2019. 15

[38] B. Wible, "Synthetic data, privacy, and the law," *Science*, vol. 364, pp. 348–349, Apr. 2019. 15

[39] J. Eno and C. W. Thompson, "Generating Synthetic Data to Match Data Mining Patterns," *IEEE Internet Computing*, vol. 12, no. 3, 2008. 15

[40] A. Mizroch, "How Israel Turned Decades Of Medical Data Into Digital Health Gold." 18

[41] R. S. Rotem, G. Chodick, M. Davidovitch, R. Hauser, B. A. Coull, and M. G. Weisskopf, "Congenital Abnormalities of the Male Reproductive System and Risk of Autism Spectrum Disorders," *American Journal of Epidemiology*, vol. 187, pp. 656–663, Apr. 2018. 19

[42] "Body mass index - BMI," Feb. 2020. 21

[43] "Characterization and Classification of Geographical Units by the Socio-Economic Level of the Population, 2013." 22

[44] D. A. Freedman, "Statistical Models," p. 458. 33

[45] A. Grimes and F. Schulz, "Making Sense of Odds and Odds Ratios," *Obstetrics & Gynecology*, vol. 111, no. 2, Part 1, pp. 423–426, 2008. 34

[46] Z. R. Manjaly, K. M. Scott, K. Abhinav, L. Wijesekera, J. Ganesalingam, L. H. Goldstein, A. Janssen, A. Dougherty, E. Willey, B. R. Stanton, M. R. Turner, M.-A. Ampong, M. Sakel, R. W. Orrell, R. Howard, C. E. Shaw, P. N. Leigh, and A. Al-Chalabi, "The sex ratio in amyotrophic lateral sclerosis: A population based study," *Amyotrophic lateral sclerosis : official publication of the World Federation of Neurology Research Group on Motor Neuron Diseases*, vol. 11, pp. 439–442, Oct. 2010. 38, 49

[47] P. A. Mccombe and R. D. Henderson, "Effects of gender in amyotrophic lateral sclerosis," *Gender Medicine*, vol. 7, no. 6, pp. 557–570, 2010. Publisher: EM Inc USA. 49

[48] V. Mayer-Schönberger and K. Cukier, *Big Data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013. 50

[49] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: management, analysis and future prospects," *Journal of Big Data*, vol. 6, p. 54, Dec. 2019. 50, 51

[50] E. O. of the President, "Big Data: seizing opportunities, preserving values," 2014. 50

[51] M. Reisman, "EHRs: The Challenge of Making Electronic Data Usable and Interoperable," *Pharmacy and Therapeutics*, vol. 42, pp. 572–575, Sept. 2017. 51