**HACKATHON 2021**

**SENECA COLLEGE**

Obtaining insights on ALS and other rare diseases through  
Open data

**\_\_\_**

**By**  **Leonardo Patricelli (Project Manager and data scientist)  
 Nikolas Trivanovic (UX/UI Designer)**



# 

# Introduction

‘With research, possibilities are limitless’. The importance of this truth cannot be overstated, as continual research is necessary in order to unlock the mysteries often presented by rare diseases. In a field where concrete solutions are not immediate, it is essential to mine the different research we already have, to make connections where none existed previously; to complement that information with new findings; and to listen to the voices of rare disease patients, who often provide the foundations of what is known about these medical enigmas.

Till date there are thousands of rare diseases in the world. Six to seven thousand diseases are considered rare and new diseases are regularly described in medical literature. Diseases that are rare are serious, often chronic and progressive. According to various research, usually signs are observed at birth or in childhood.In the United States, a rare disease is defined as a condition that affects fewer than 200,000 people in the US. Just in Canada, a rare condition affects about 1 in 12 canadians.

Rare diseases are hard to study because it’s complicated to gather info about them. If we had more info about them, maybe doctors may be able to find a cure or to improve the lives of people affected by this kind of disease.

Fortunately, Orphadata provides to the scientific community a comprehensive, high-quality and freely-accessible dataset related to rare diseases and orphan drugs, in a reusable format. For this challenge, we analysed the Orphadata dataset to build an application able to visualize aggregated statistics related to rare diseases and to return important information about them, in order to help doctors and researchers in their job while diagnosing a rare disease.

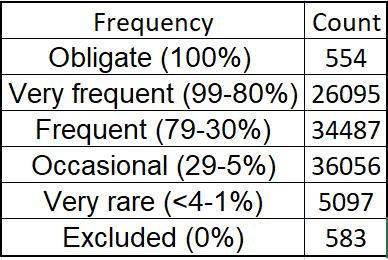
# Data and technologies

For this challenge, we used the data available on the Orphadata website.  
The mission of Orphadata is to provide the scientific community with a comprehensive, high-quality and freely-accessible dataset related to rare diseases and orphan drugs, in a reusable format.

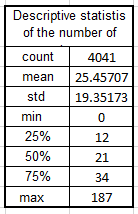
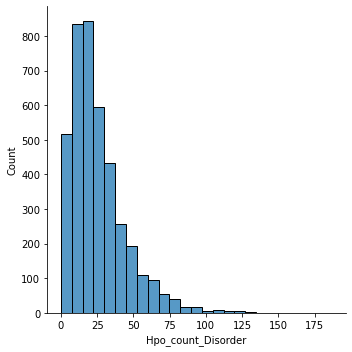
Once downloaded, the data were processed in a Python environment to get insights and to build an automatic way to explore the dataset.

# Human Phenotype Anthology (HPO)

The Orphanet inventory of rare diseases is based on Human Phenotype Ontology (HPO) terms, a standardized and controlled terminology covering phenotypic abnormalities in human diseases.  
The annotation is characterized by frequency; let’s have a look at the frequency categories and their distribution:

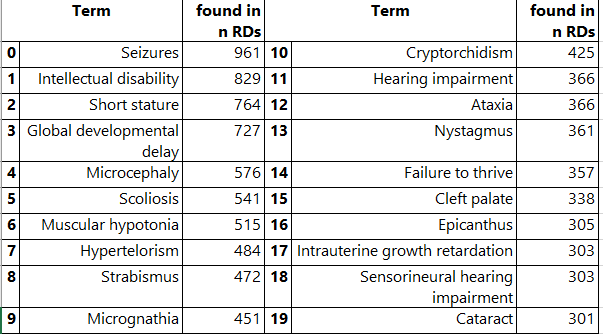


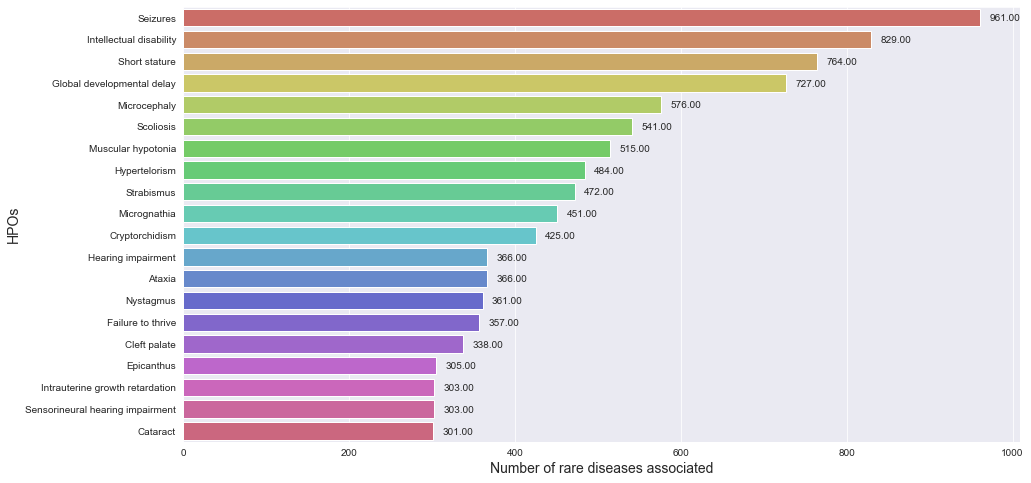
We can also look at the distribution of the number of HPOs associated with each disease.



It seems that each rare disease is associated on average with 25 HPOs, and 50% of them have 21 HPOs or less. Looking at the quantiles and at the histogram bins, we can have an idea of the distribution of the HPOs among the diseases.

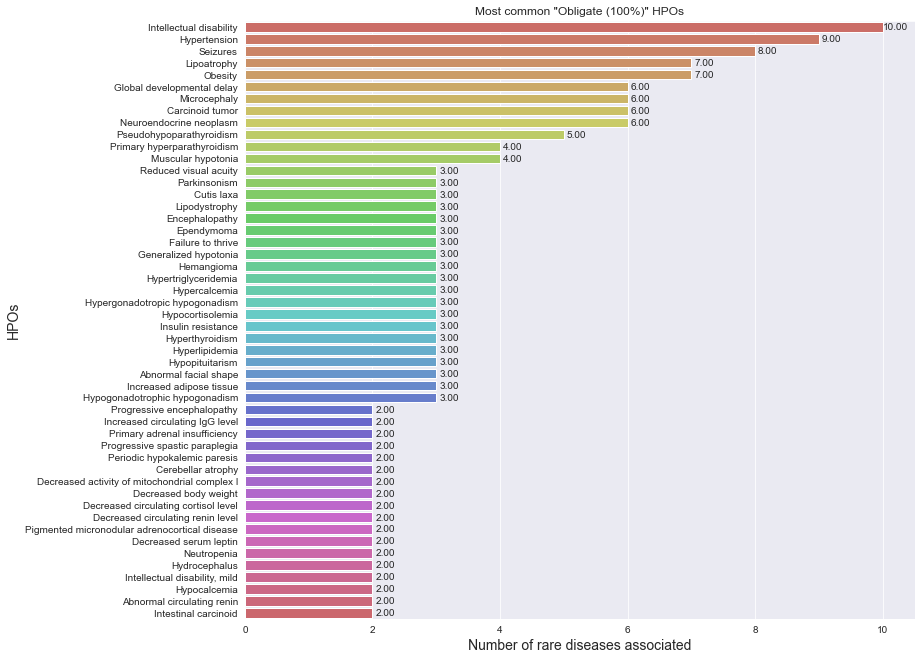
It is also possible to look at the single phenotype to see in how many rare diseases it is present. In the table below we show the 20 most common Phenotypes with the count of the rare diseases involved.

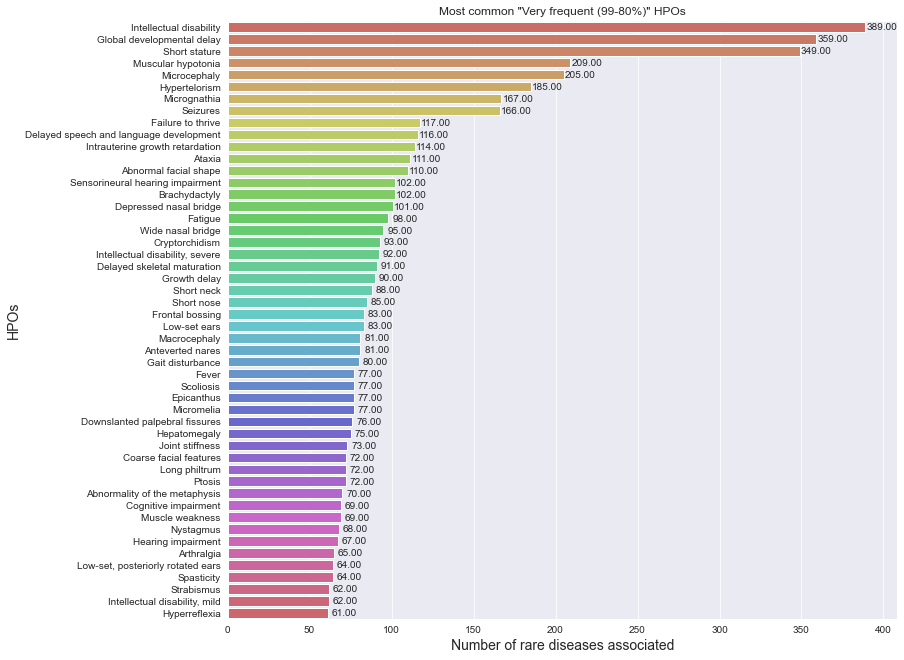


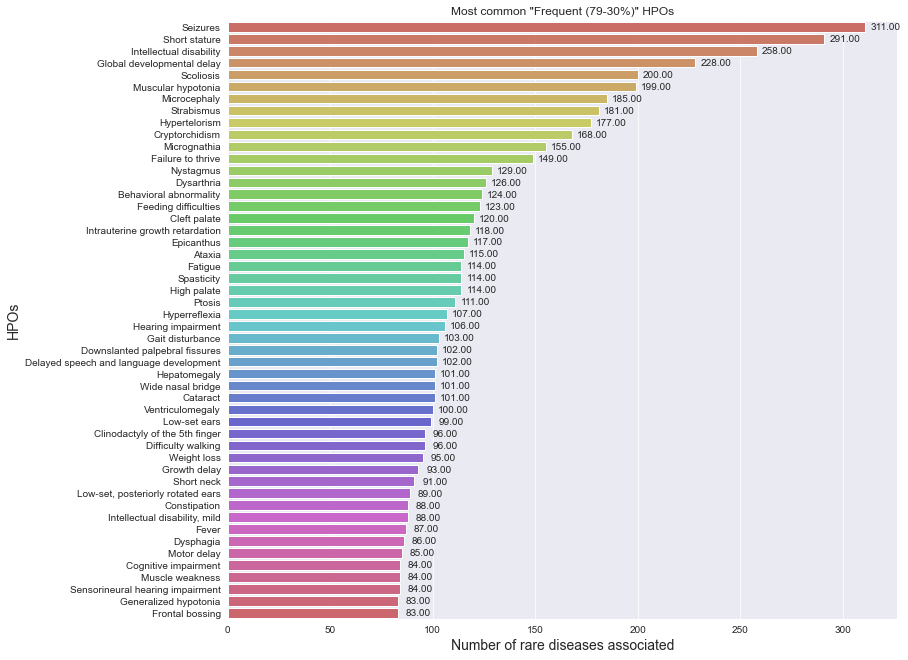


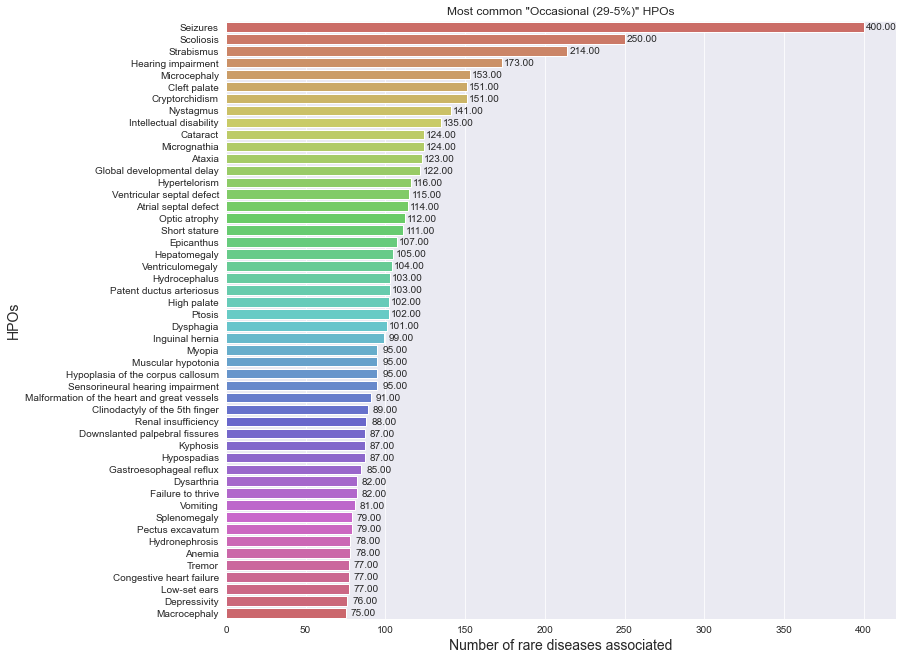
It has been identified that most cases are associated with **seizures** and **intellectual disability** followed by **short stature** and a **global development delay.**

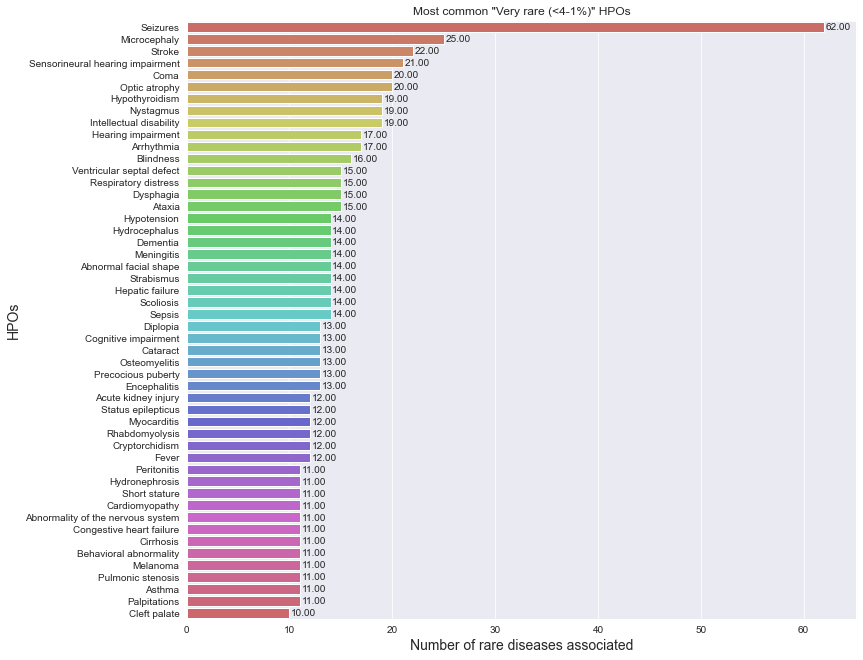
It is also possible to look at the HPOs per frequency class in order to understand what are the most frequent phenotypes in total (and per category).

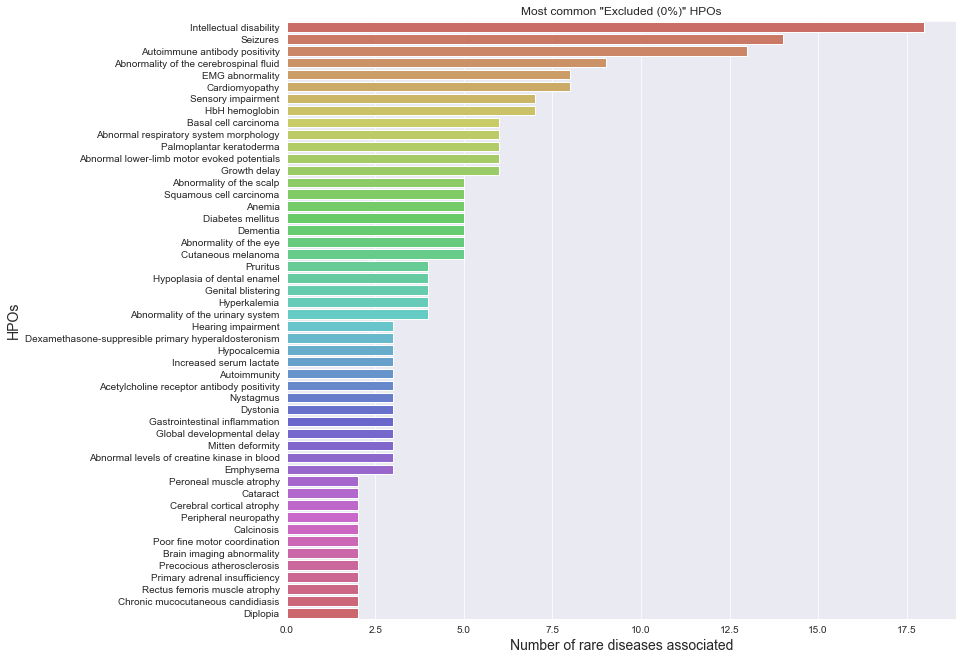


****

****

****

****

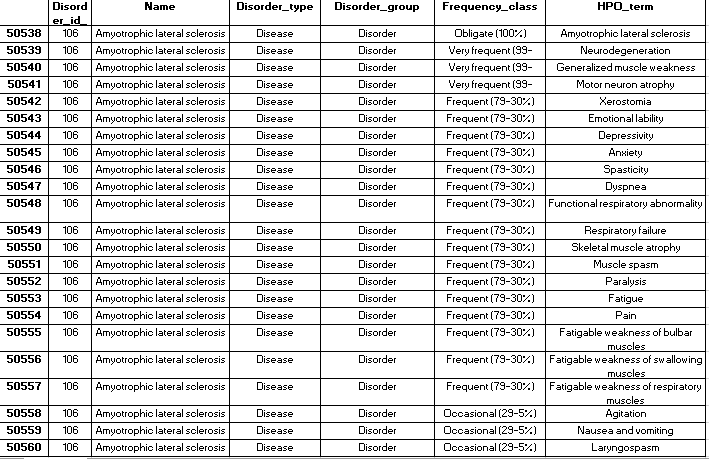
****

## *Looking at the single disease (ALS)*

It is also possible to select a single disease to know what HPOs are related to and the frequency class associated.

In the case of ALS, for example, we see that **neurodegeneration**, **generalized muscle weakness** and **motor neuron atrophy** are very frequent.  
Below is displayed the full table with all the HPO related to the ALS and their frequency class.

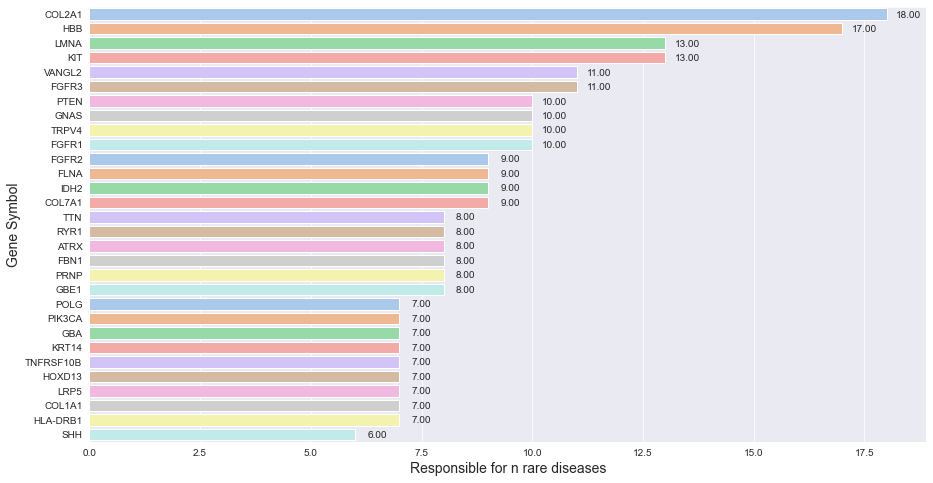
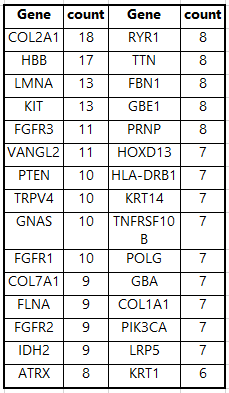
The same output can be obtained for all the diseases.



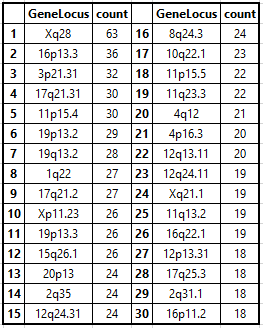
# Genes and their loci

In order to better define rare disorders of genetic origin, Orphanet provides information on every gene related to a rare disorder. This information includes the genetic international nomenclature, the gene typology and the chromosomal location. Orphanet also defines the relationship between genes and their related rare disorders and provides evidence for establishing these gene-disorder relationships.

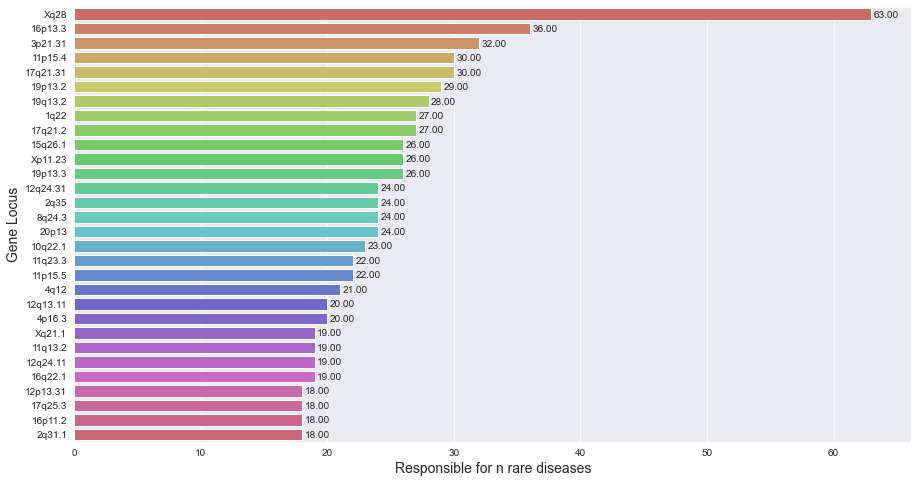
The table below displays the 30 genes (actually their official symbol) most responsible for rare diseases with the count of associated diseases.



In the data, we can find also information related to the location of the Gene.  
We grouped the genes by their location in order to find the locations responsible for most of the rare diseases in the dataset.



These tables can provide insights on what genes should be tested first during the diagnosis of a rare disease, or in what location to see for genetic anomalies. Knowing what are the most “problematic” genes and their location can speed up the process and lead to faster diagnoses.



## *Looking at the single disease (ALS)*

It is also possible to retrieve the genetic information for every disease in the dataset.  
For the ALS, we can see that it is related to the Gene ***cyclin F (***wth related ***symbol)*** located in position ***16p13.3***, with the related association between gene and disorder (in this case ***Disease-causing germline mutations***)



# Functional consequences (Disabilities)

The Orphanet inventory of rare diseases is annotated with activity limitation/participation restriction (functional consequences), using the Orphanet Functioning Thesaurus, derived and adapted from the International Classification of Functioning, Disability and Health – Children and Youth (ICF-CY, WHO 2007). The information provided is assessed taking into account the whole patient population affected by the disease, receiving standard care and management (specific and/or symptomatic management, prevention and prophylaxis, devices and aids, care and support).

Each functional consequence is annotated with the following:

• Frequency in the patient population:

o Very frequent: more than 80%

o Frequent: between 30% and 80%

o Occasional: fewer than 30%

• Temporality:

o Permanent limitation/restriction: the functional consequence is present throughout the life of the patient. It can be congenital, secondary to loss of a skill or participation. It can be a direct or indirect consequence of the disease or of its treatment.

o Transient limitation/restriction: the functional consequence occurs during acute episodes, periodic crises or relapses. It resolves or reduces spontaneously or by the action of treatment or care.

o Delayed acquisition: a skill or participation is performed later than by a healthy person.

• Degree of severity:

o Low: activity or participation can be carried out with little difficulty by the patient alone.

o Moderate: activity or participation can be carried out with some technical and/or human assistance

o Severe: activity or participation cannot be carried out without substantial technical and/or human assistance.

o Complete: activity or participation cannot be carried out, even with technical and/or human assistance.

o Unspecified: limitation/restriction is difficult to quantify or highly variable between patients (ranging from 'Low' to 'Complete').

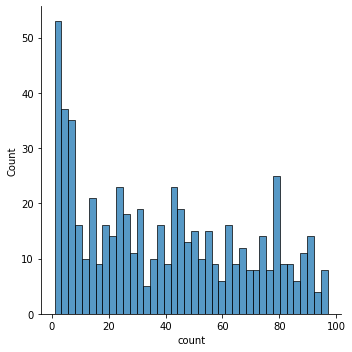
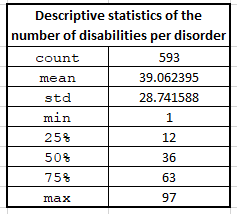
• Loss of ability when relevant, defined by the progressive and definitive loss of a skill or participation over the course of the disease.

A functional limitation is stated to be « undefined » when the current knowledge does not enable information about the extent of the consequences on daily life to be provided.

The unaffected activities and participation are not listed.

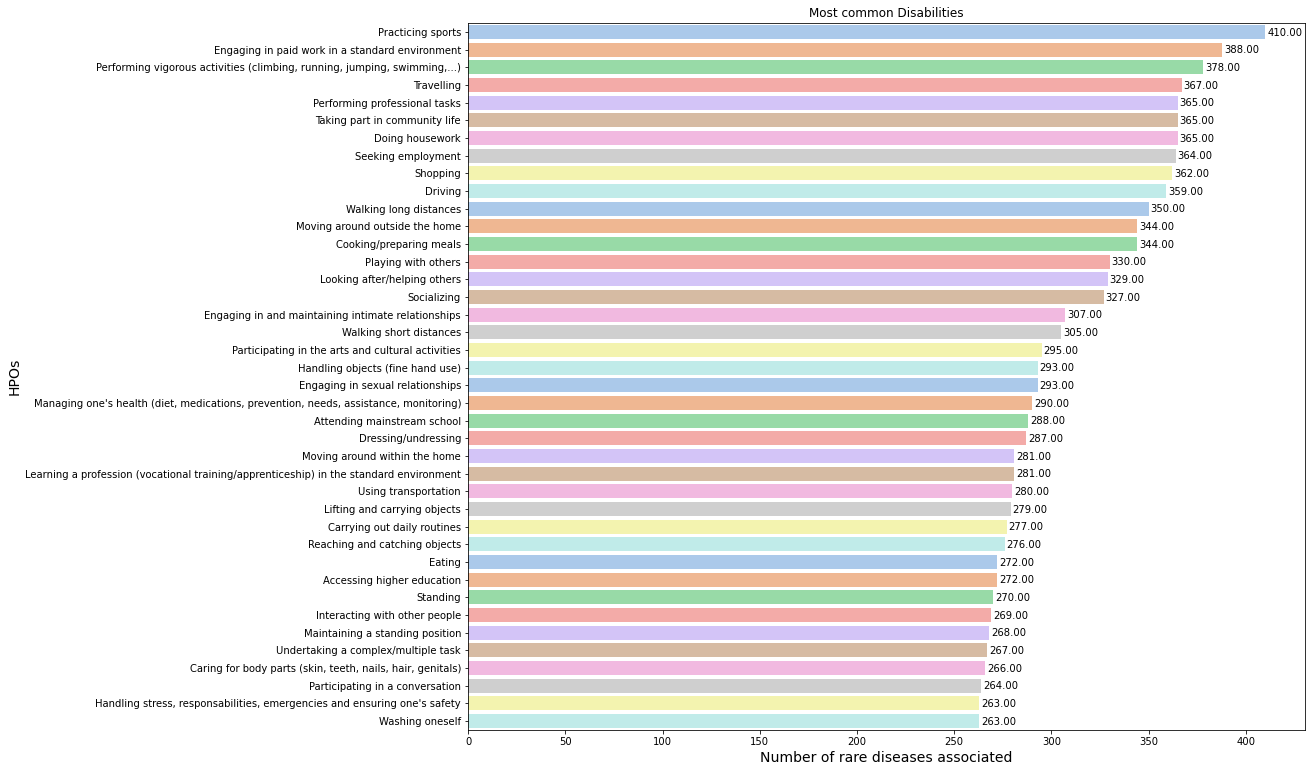
Environmental factors that may have an impact on the daily activities of the patients are also identified and listed when possible.

First of all, let’s have a look at the distribution of the number of disabilities per disorder



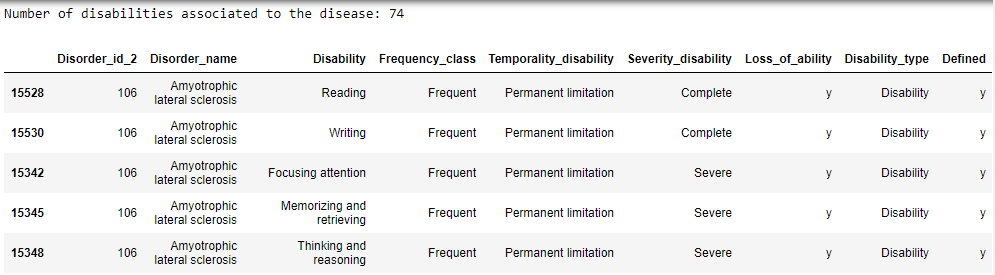
It seems that on average each disease has 39 disabilities related, and half of them has 36 diseases or less (the median). Looking at the barplot we can have a better idea of the distribution. It seems that it varies a lot (in fact this justifies the high standard deviation, with a value of 28).

Here we can have a look at the most common disabilities :

******

## *Looking at the single disease (ALS)*

It’s also possible to analyze what are the disabilities related to a single disease.   
Looking at the ALS we can notice that the disease has 74 associated disabilities.  
Our application sort the values by *Frequency class* and *severity*.  
We are going to pick just the top 5; of these 74 disabilities, **reading** and **writing** result as ***complete*** disabilities, while **Focusing attention**, **memorizing and retrieve** and **thinking and reasoning** are also frequent, but their severity is “***severe***”.



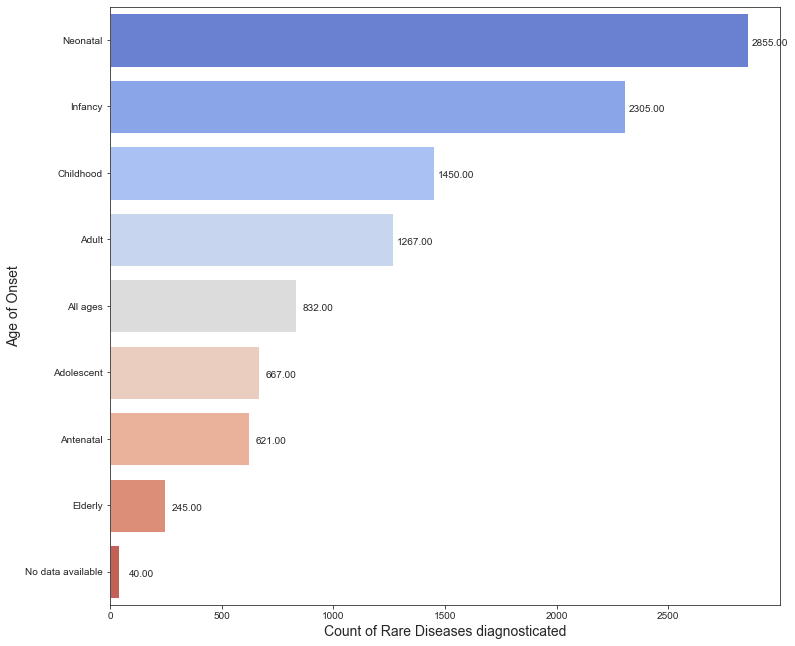
# Average Age of onset and of death

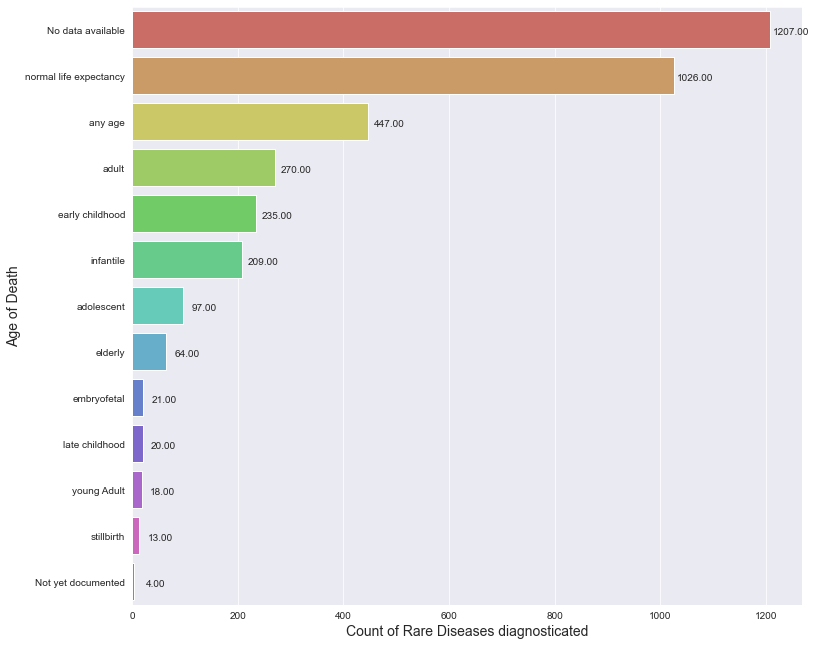
In this section, we are going to present the information related the average age of when the disease is discovered, and the average age of death. To estimate those values, Orphadata referred to published studies and their validity is taken for granted and not re-assessed.

Here an explanation of the variables:

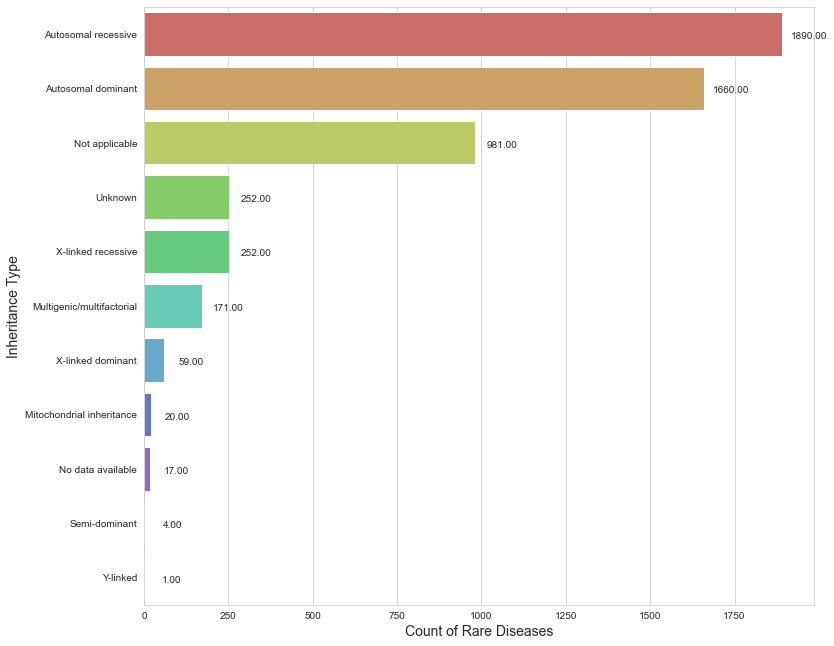
* **AverageAgeOfOnset**: classes based on the estimated average age of clinical entity onset.
* **AverageAgeOfDeath**: classes based on the estimated average age at death for a given clinical entity. There are twelve different population age groups.
* **TypeOfInheritance**: type(s) of inheritance associated with a given clinical entity. There are thirteen different types of inheritance.

Looking at the **AverageAgeOfOnset,** it seems that most of the rare disease is found in early age (infact *Neonatal*, *Infancy*, and *Childhood* are the most recurrent classes).  
For what concerns the **AverageAgeOfDeath**, we have missing data for about a third of the disease, while the second third of diseases allow the patient to have normal life expectancy, and the restant third is distributed mostly between early ages and adult age.

****

****

According to the Orphadataset, most of the rare diseases have an inheritance ***Autosomal recessive*** or ***Autosomal Dominant***. Unfortunately, there is a large part of the dataset in which this information is “*not applicable*”.

****

## *Looking at the single disease (ALS)*

This information can be retrieved for every disease.  
Looking at the ALS disease, we can notice that it has an ***autosomal dominant*** inheritance, and on average it is diagnosed in ***adult age*** and led soon to death, since the average death age is also ***adult***.



# Prevalence

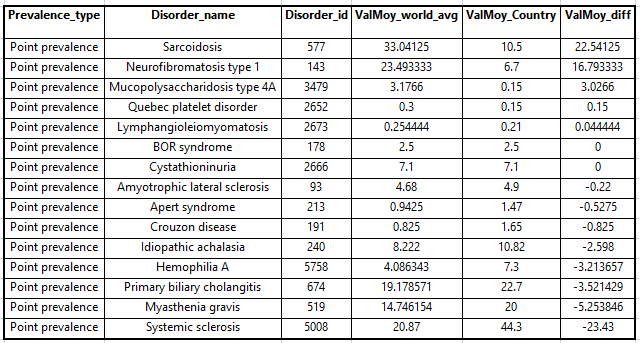
According to Wikipedia, the prevalence is the proportion of a particular population found to be affected by a medical condition (typically a disease or a risk factor such as smoking or seat-belt use) at a specific time. It is derived by comparing the number of people found to have the condition with the total number of people studied, and is usually expressed as a fraction, a percentage, or the number of cases (ex. per 10,000 or 100,000 people).

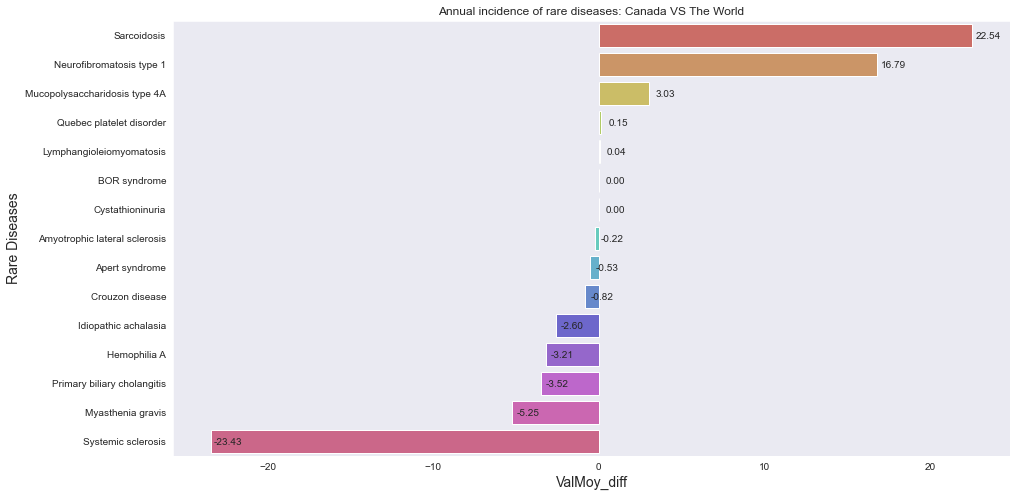
The prevalence is calculated per countries, and can assume 3 values:

1. Point Prevalence: Number of cases scaled up to the general population at a given time;
2. Annual Incidence: is a measure of the [probability](https://en.wikipedia.org/wiki/Probability) of occurrence of a given [medical condition](https://en.wikipedia.org/wiki/Medical_condition) in a population within a year;
3. Prevalence at birth: Number of cases observed at birth relative to the number of children born alive at a given moment;

With our implementation, Selecting a country and the type of prevalence, it is possible to explore the “spread” of rare disease compared to the rest of the world.

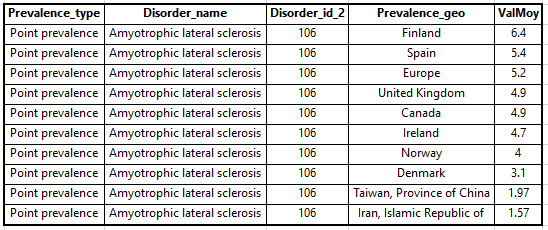
In this case, selecting Canada and “Point Prevalence”, we can compare the Point prevalence of the Canadian rare disease with the rest of the world.   
The Column “ValMoy\_diff” contains the difference between the world average prevalence of a particular disease less the canadian value. Positive values mean that Canada has a higher prevalence, while negative values that Canada has less prevalence. If the difference value is close to 0, than the value of world average and country is very similar.



****

## *Looking at the single disease (ALS)*

Selecting a particular disease, we are able to see the spread among the countries.



In this case, we can notice that Finland and Spain and Europe have the highest value, so we can conclude that they have more cases.  
Thanks to this, researchers can see where certain diseases are more spread and can give a “hint” for future researches.

# Implementations for the Healthcare system

This data can help Healthcare systems in making faster diagnosys and researchers to retrieve useful information during their job.

For example, looking at the information related to the genes, it is possible to implement a procedure that starts from testing the genes and the genes loci associated with the highest number of rare diseases.   
For example, starting with analysing the genes COL2A1, HBB, LMNA and KIT the researcher can check 61 rare diseases, and looking at the loci Xq28, 16p13.3, and 3p21.31 it is possible to check 131 rare diseases.

Or if a doctor suspects a patient may have a particular rare disease, he can search for the HPOs related to the disease and check if the patient has them or not.

Thanks to these tools, it is possible to speed up the diagnosis process.  
Making faster diagnosis has two main advantages:

* Doctors can spend less time on the diagnosis and more on the treatment;
* Testing the genes starting from the ones involved in the higher number of rare diseases can help to save money, since each gene test has a cost of around 300 USD.

In addition, Researchers can use these Apps to retrieve geographical information with the prevalence app and check the spread of a rare disease around the world, and use it as basys to make hypotheses on why certain rare disease are less likely in some countries and more in others.