# The Needle in a Haystack

## *How to find relevant information in Genomic Information Systems*

by
Yuri M. van Geffen

This thesis is written as part of the master thesis process at Utrecht University (UU), *Master Business Informatics*. The first supervisor is *Dr. Verónica Burriel Coll*, with *Prof. Yannis Velegrakis* as the second supervisor. The main study is performed at the Universitat Politècnica de València (UPV) under the supervision of *Prof. Óscar Pastor López*, from February to August 2019, with the thesis finished in Utrecht, in November 2019. The exchange was partly funded through the *Erasmus+ Traineeship grant*.

## Acknowledgements

No academic work can be performed without the support, as well as the critical eye of others. Many people have been involved in the creation of this work. I would like to thank Dr. Verónica Burriel Coll for handing me the opportunity and connections to work at her alma mater as a visiting student, as well as her guidance through the master thesis process. I would also like to thank Prof. Óscar Pastor López, who was vital in steering me towards our common goal when I strayed, and in provoking creative thinking when my ideas were too linear. My appreciation also goes out to the rest of the PROS research team for the fruitful and animated discussions, during both formal and informal meetings. Valuable input has also been given by Prof. Yannis Velegrakis, especially in the later stages of the project. With his balanced and outside perspective he was able to push the thesis to a higher level, for which I would like to thank him.

Last but not least, I would like to thank my family and friends for their support when I needed it, and allowing me to ramble on about statistics, phenotypes, APIs, sequencing, $F_{score}$, regression, prediction and all that other good stuff!

## Abstract

This thesis is about data quality and automation of retrieval, within the domain of genomic information systems. In recent years, large scale genomic studies have become common due to lower cost and improved tools and software for analysis. With the relative ease of performing these studies, the pool of genomic research data has grown massively, to the point that information systems such as the GWAS Catalog and Ensembl are used to collect, manage, and distribute study results. Researchers and practitioners have to make sense of the data contained in these systems manually. This boils down to choosing which data is relevant to them, and which data is not, with the end goal of generating new knowledge. Apart from taking a lot of time, manual evaluation introduces errors. Automation is necessary to reduce errors and save valuable time.

We explored the genetic information system domain using a bottom-up approach. The SILE method was used as a framework. The study focusses on the Identification step within this framework. An exploratory analysis was performed on the data contained in both the GWAS Catalog and the Ensembl genome browser. With the knowledge gained from this analysis, a solution is proposed to automate the selection process within these information systems. This solution involves a combined classification and regression model, ranking entries within the information system on relevance. We built these models by identifying relevant entries by hand and training the models on this manually created data set. The models then provided the ability to identify relevant entries with a high certainty in a, previously unseen, validation set.

It is shown that an understanding of the domain with regards to data quality, is key to developing automated solutions. Important factors here are the difference in entries between phenotypes, and over time. Another important factor to consider is the difference between theoretical ideal measures, and the availability of these measures in practice. This study provides a basis for automation of relevant entry retrieval within the genomic information system domain.

# Contents

# Chapter 1

# Introduction

Over the past decade, genome-wide studies have become a hallmark of the genetic research field. When looking at the Google search interest, Genome-Wide Association Studies (GWAS) and Next Generation Sequencing seem to have been at a plateau since 2010 (Figure 1.1). A similar trend can be seen in the amount of published molecular databases in the Nucleic Acids Research database issue [23]. As an example, about a hundred databases are added every year. Because the field is still relatively new there is a large dispersion of knowledge. To remedy this problem several public general variant databases are being developed, such as the GWAS Catalog [9] and the Ensembl database [66]. They both have different thresholds for admittance in the database, different query models, and different structures. The process of finding relevant information in these databases is currently largely a manual one. Because of this, extra errors are introduced and valuable research time is lost. A method based on Search, Identification, Load, and Exploitation (SILE)[1], developed by León at the PROS research center, aims to formalize this process to reduce mistakes and quantify data and information quality [34]. In another paper by León, it is shown that when following a set method for finding important entries the results vary over time, which means that knowledge of clinicians and researchers is quickly outdated [45]. The retrieval of entries is a largely manual process that is in dire need of automation. Variant databases often provide programmatical access to the data contained in them. However, the filtering and analysis, after pulling the data, are still done manually. Automation is complicated because the demands are highly specific to the researched phenotype and the ultimate goal of the extracted information. At the moment no automated solution exists to determine relevant criteria and extract entries with these criteria. The fact that often used information systems source their information from all publicly available research induces more issues. The organizations that publish these information system use criteria for determining wether research should be included, but the criteria are often very minimal. The goal of these criteria is to only exclude research that is lacking proper statistical basis, and it is up to the user to further filter the results.

This filtering is an important step in the SILE method (Section 2.3), developed

[1] https://anleopa.github.io/SILEWebPage/

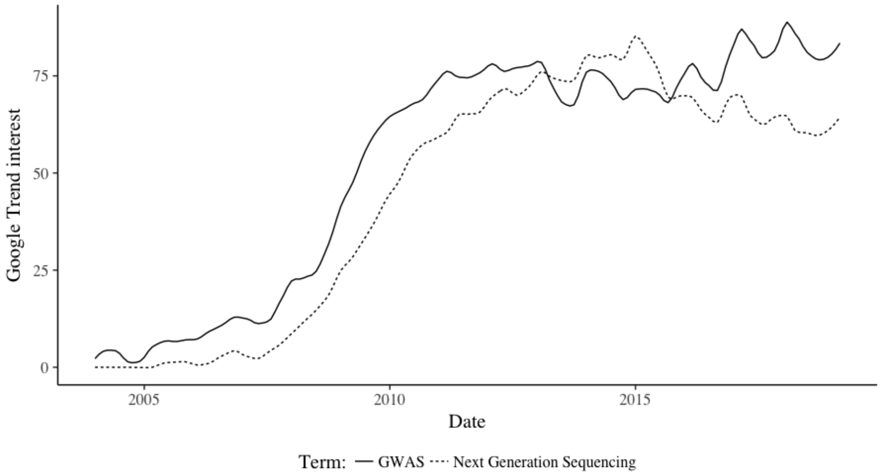


Figure 1.1: Google search trend for *GWAS* and *Next Generation Sequencing.*

to formalize the process of information extraction from genomic information systems. Formalizing this process not only reduces errors, it also opens up the possibility for automation. The automated selection of measures and criteria when retrieving entries is the next step in usability and efficient resource use in the field of genomic information systems. To develop a useful automation solution the domain needs to be properly understood, not only from a top-down perspective (SILE), but also from a bottom-up perspective. We need to understand what kind of data we are dealing with, what the distribution is of the data over different factors, and what data is missing. The overarching question that we aim to answer is the following: *How can data quality, and by extension information quality, be improved when identifying relevant entries for different phenotypes from established and public genomic information systems, in an automated way?*

This thesis is an account of the research towards increased data quality and automation publicly available genomic information systems. We perform an exploratory analysis into the fundamentals of the data contained in publicly available genomic information systems and propose a solution to automatically retrieve relevant data from these systems. It is a model-based solution that learns from manually extracted base data and extends to unseen data. It can be used as a tool in the Identification step of the SILE method.

The next chapter will introduce the basic concepts in genetics research and gives a relevant statistical background. Chapter 3 will state the problem along with the relevant research questions. It also goes into the chosen research methodology. Chapter 4 will account of the exploratory data analysis performed, as well as propose a treatment based on this analysis. Chapter 5 will state the results obtained using the treatment, which will be discussed in chapter 6.

# Chapter 2

# Background

## 2.1 Genetics

Genetics is a field of study concerned with hereditary traits and variation within species [21], and specifically tries to answer two questions:

- what makes a species?
- what causes variation within a species?

A large part of the answer can be found within one of the building blocks of life; deoxyribonucleic acid (DNA). This large molecule, shaped like a double helix, is contained within the cells of all known organisms. DNA contains the instructions for the functioning of all parts of the organism, encoded in *nucleotides* [24]. A nucleotide is a smaller molecule that forms the building block of DNA and can be represented using a single letter (A, C, T, and G). During the reproduction of cells, the strains of DNA are copied to the new cells, essentially creating a copy of the original cell. This process also ensures that DNA is copied (and recombined) from the parent(s) to offspring organisms. This process is not perfect. Therefore, every time a copy is made local changes are introduced. DNA can also slightly change under the influence of external factors (e.g. radiation). These changes, called alleles, mutations, variants, or variations, depending on the context and specific criteria, introduce variance between individual organisms, which can lead to different species [17]. Studying them within one organism or a population[1] can lead to new knowledge about phenotypes[2], proteins and DNA itself. The following sections will explain the relevant parts of the genetics field to this research. Section 2.1.1 explains how and why alleles within a population are studied. Section 2.1.2 and 2.1.3 will go into how these alleles can be found. Section 2.1.4 will explain how the alleles can influence the way you look and to which diseases you are susceptible. Sections 2.1.5, 2.1.6, 2.1.7, and 2.1.8 will go into some of the relevant nomenclature.

[1] Population: the group of organisms of one species living and interbreeding in one area.

[2] Phenotype: the external and internal trait of an organism (including physical appearance, biochemistry, and development).

### 2.1.1 More about variation

A variant is a change in the DNA present in the population. Usually, variants are based on a reference genome or relative to the most common nucleotide. The most common type of variation is the notion of single-nucleotide polymorphism (SNP). A SNP is a change of the DNA on one specific *base pair* (a combination of A and T, or C and G) of the string of molecules. A SNP has the constraint that it has to be present in a substantial amount of the population, however, there is no consensus on the percentage. An often used threshold is 5%, such as in the HapMap project [15]. More than 84 million SNPs have been identified as of 2015 [58]. Every human has about 4 to 5 million of these variations.

### 2.1.2 Reading the genome

Reading of the genome can be done in multiple ways. One way is to sequence the whole genome, scanning all of the more than 3 billion base pairs. This was first done in the year 2000. Until recently, this method (called Whole Genome Sequencing, WGS for short), was not cost-effective for most researchers. They relied, and continue to rely, heavily on genetic micro-array technology. This technology allows for checking the existence of specific variations within the DNA. Usualy, micro-array chips bind to about one million positions in the genome. Statistical measures can be used to infer the presence of other variations within the genome. In short; variations that lie close together are more likely to be transferred from parent to child together. Using this knowledge, and the million data points retrieved from the micro-array, one can infer the chance that a specific variation is present. The following section goes into why this is useful when studying large populations. Nowadays, the price sequencing the whole genome is going down through the development of new techniques (Next Generation Sequencing [6]), allowing researchers to upscale the amount of people that have their DNA sequences for academic research, and in a practical setting.

### 2.1.3 How to study populations

As SNPs have the constraint to be present in a substantial part of the population they are a prime target to be studied with statistical methods. A common way of finding interesting SNPs is by performing a Genome-Wide Association Study (GWAS). A GWAS usually has a high number of participants with, and without the researched phenotype. Participants can run in the thousands, with studies sometimes having over a 100.000 participants [57, 65]. For each participant the genome is sequenced, and statistical techniques are applied to the presence of variations and the expression of a certain phenotype. This way a relation between the variation and the phenotype can be found. The book chapter *Genome-Wide Association Studies* of Bush and Moore [12] goes into more detail of how GWAS works.

### 2.1.4 Gene expression

A large part of the DNA encodes for proteins used in the functioning of the organism. Simplified, a chunk of the DNA is transcribed to RNA, which in turn is translated to a protein (Figure 2.1). A change in this part of the DNA can lead to a differently functioning or broken protein. The variation of a part of the DNA does not necessarily mean that the gene at that location will not function. Some variants may result in a synonymous protein or a protein that functions similarly. Variants may also result in a protein with reduced function. Broken proteins, non-existing proteins, or proteins with reduced function can seriously affect the organism and lead to disease. Apart from protein-encoding DNA, there are many other known and unknown ways DNA can influence the functioning of an organism.

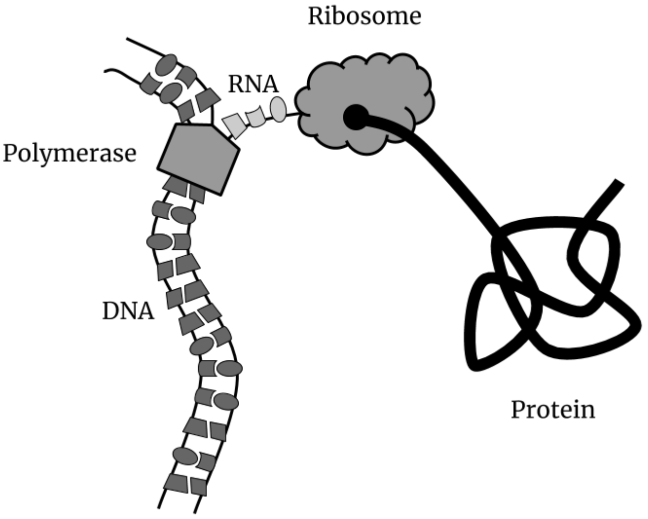


Figure 2.1: Simplified process of protein synthesis from DNA.

### 2.1.5 Pathogenicity

Pathogenicity refers to how disease-causing a variant is. An often-used standard for assigning pathogenicity to a variant is created by Richards et al. [51], where classifications such as *pathogenic*, *likely pathogenic*, *benign*, *likely benign*, and *uncertain* are used. For each category, different evidence criteria are supporting either a benign or a pathogenic variant. An example of pathogenic evidence from this standard:

> Null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multi-exon deletion) in a gene where the loss of function (LOF) is a known mechanism of disease ( . . . ).

Strong evidence from the same standard for a variant being benign is:

> Allele frequency is above 5% in Exome Sequencing Project, 1000 Genomes, or ExAC.

It is important to note that although the above evidence criteria are objective, there are other criteria which are not as objective and demand the judgment of professionals, e.g the benign-supporting evidence:

> Well-established in vitro or in vivo functional studies show no damaging effect on protein function or splicing.

In general, the more different sources specify the same pathogenicity the more trustworthy the result is. In Van Gijn et al. [59] it is shown that finding consensus can be quite complex but possible in most cases.

### 2.1.6 Frequencies

In genetics, frequencies refer to the amount of the population that has a specific variant [19]. Generally, it is reported in a factor or percentage and is relative to a specific population, e.g. *variant X has a frequency of 5% in the population of Northern Europeans* means that within the researched population of northern European descent, 1 in 20 carried the variant X in their DNA.

An often-used statistic is the Minor Allele Frequency (MAF) which refers to the frequency of the second most common allele in a population[3]. It is interesting because it is a measure of commonality. A high MAF (near 50%) means that a lot of people carry the allele and it is thus very unlikely to be very impactful in a disease (given that the disease kills or disables the organism), as highly pathogenic alleles naturally become rare through natural selection. A low MAF variant, less than a few percents, is often not of interest to researchers as most subjects will not carry the variant, and in a clinical environment, a targeted remedy will only help a small number of patients. Statistically, it is also important to find enough cases of the researched allele, which is more practical with a higher MAF. The Risk (Allele) Frequency is another often reported statistic and refers to the allele that induces an increased risk of developing a phenotype. It is often reported on a 0%-100% range.

Usually, it is assumed that common variants result in less extreme or lethal phenotypes (or phenotypes that are expressed at a later age) as those will survive in the population. Lethal and extreme phenotypes will often be weeded out by natural/sexual selection and will thus be less common in the population. This makes that research interest often lies in less common variants. However, as Gibson describes in *Rare and Common Variants: Twenty arguments*, there are arguments against this trend as many common variants can combine into an extreme or lethal phenotype [18].

[3] *The frequency of the minor variant is the minor allele frequencies for heterozygous and homozygous SNPs with reference to the frequency of all alleles at a particular SNP location (https://www.ncbi.nlm.nih.gov/books/NBK44476/).*

### 2.1.7 The chance of disease

Another important statistic is the chance of developing disease given the presence of a variant. The odds ratio (OR, or risk ratio), is the relative probability of an organism developing a specific phenotype when an allele is present. An $OR < 1$ means that the presence of the allele lowers the chance of developing the phenotype, and an $OR > 1$ means the chance is increased relative to the dominant allele. $OR = 1$ means there is no effect of the allele on the phenotype. When an odds ratio of 2 is reported, it indicates that the researched minor allele doubles the chances of developing the phenotype, relative to the dominant allele. It is important to note that this is only relevant within a specific population.

**The consequence of human error**

Human mistakes in genetic screening can have dire consequences for the patient. Particularly bitter is the case of Elisha Cooke-Moore who, after a yearly scan which indicated she had a strong chance of developing breast cancer and other types of cancer, decided to have her breast and uterus removed [8]. Reportedly, a nurse reviewed the results of a genetic test and concluded Cook-Moore suffered from a mutation in the MLH1 gene, leading to Lynch syndrome. This syndrome increases the chances of developing specific kinds of cancers, among others breast cancer and cancer of the uterus. After independent surgeons, who did not confirm the results of the test, performed the two invasive operations, Cook-Moore discovered a mistake made by the nurse. The results of the test were negative, not positive as the nurse had determined.

Many things went wrong in this specific case; the results should have been interpreted by a trained professional, a second opinion should have been given before any invasive operation, and the two surgeons should have independently confirmed the results of the test and not blindly have operated. However, the overarching lesson to be learned is that human judgment is fallible and that proper procedures and guidelines are needed for a safe environment to practice genetic screening.

### 2.1.8 Genomic Information Systems

A genomic (or genetic) information system (GIS) is a repository for genomic data to be used for clinical research or treatment. These repositories can contain a combination of raw or annotated sequence data[4], variant information[5], protein information[6], genetic studies[7],[8], and more. Many serve specific research purposes, while others are more general. The analyses done on the data contained in these information systems are often specific to a problem domain. Some genome browsers allow for generic analysis tools to be run on their platform. All in all, they make for very complicated pieces of software that are hard to compare to software used in other contexts. For this reason, genomic information science is a research field within many academic institutions. Another fac-

[4] https://www.ncbi.nlm.nih.gov/sra

[5] https://www.ncbi.nlm.nih.gov/clinvar

[6] https://www.ncbi.nlm.nih.gov/protein

[7] https://www.ebi.ac.uk/gwas/

[8] https://www.ensembl.org/index.html

tor is that GISs often have to handle vast amounts of data. A data file coming off of a sequencing machine can contain about 200GB of data [52]. To analyze this data (or multiples of it), big data analysis and statistical techniques are needed.

**A complicated landscape**

A painfull case in genetic testing is reported by the Mayo clinic [1]. In this case, a 13-year-old received a surgically implanted defibrillator which turned out to be unneeded. Many of his family members were told they possibly had fatal heart disease, this also turned out to be false:

> "To interpret a sequence where DNA letters differ from the *reference human genome* (...) researchers scour public or proprietary databases to see if the misspelling[a] is disease-causing or benign. Unfortunately, databases often disagree. And many misspellings once thought to be dangerous - and still listed that way in databases - have since been determined not to be. A recent study estimated that people have, on average, 54 mutations listed as pathogenic, of which 41 are almost certainly not."

This case provides a clear example of the precarious landscape clinicians and researchers walk in. Standardization and automation can help prevent similar mistakes from being made.

[a] variant, allele, etc.

## 2.2 Relevant statistical background

To interpret and understand the results of our data analysis and modeling it is important to have an understanding of some statistical concepts and definitions. An important concept for this research is statistical modeling. When talking about modeling in general, the goal is to create a representation of a part of the real world. Examples of things that can be modeled in the real world are physical processes like gravity, heat exchange, and fluid dynamics. These physical processes are studied to better understand the world around us. But also non-physical processes can be studied. A company might engage in business process modeling (BPM) to better understand what is going on and where efficiencies and risks lie. Statistical modeling is a form of mathematical modeling that usually aims to find the relation between dependent and independent variables. A dependent variable, as the name suggests, is assumed to be dependent on the independent variable(s). One of the benefits of statistical modeling is that the influence of the independent variables can be shown. Many statistical models can also be used to predict the dependent variable based on the independent variables. Both of these characteristics will be used throughout this research. In this research, we will use two different kinds of models. The first kind is a regression model and the second one is a classification model.

### Regression

When building a regression model we want to determine the influence of each of the independent variables on the dependent variable. Often the influence of the independent variables is assumed to be linear; the size of the independent variable corresponds directly to the size of the dependent variable. This so-called linear regression model generally takes the following shape:

$$dependent = intercept + (influence * independent) \tag{2.1}$$

Here the *intercept* is a base value off-setting the resulting value. It is equal to the mean value of the dependent variable when the independent variable is 0. The *influence* variable is the constant that we wish to determine. This is called the *beta-coefficient* ($\beta$). In a linear model, the independent variable is often used as-is. When the correlation between the dependent and independent variable is not linear, a function can be applied to the independent variable shaping it into a linear relationship, essentially performing an exponential regression using lineair regression algorithms. In this research we assume relationships to be linear or near-linear and will not consider exponential regression for the sake of simplicity in the resulting model.

In many models, there are more independent variables. Each of these variables will also have its beta-coefficient. The model is usually represented in the following shape:

$$y = (\sum_i \beta_i * X_i) + \epsilon \tag{2.2}$$

Note that the intercept is not explicitly written down. It hides in the sum, with $X_1 = 1$ and the corresponding coefficient having the intercept value. Training algorithms aim to find the best beta-coefficients to represent a specific data set, called a training set. Usually, the best fitting coefficients do not fit the training data perfectly. The epsilon ($\epsilon$) is a way to represent the error; how far is our model away from the truth. This error can be explained by missing variables in the model, or by other factors such as noise in the data or the relationship of the dependent and independent variable not being linear. The error is generally unknown but can be approached given some extra information. It is also key to measuring the performance of our model. A model with an error of 0 is a perfect model, it fits the given data perfectly. Any dependent variable that is different between our training data and the predictions by our model incurs a value to this epsilon. After training the model, several measures can be calculated. For each coefficient (including the intercept) a *z-value* and a *P-value* can be calculated. They are related to the trustworthiness of the found coefficients. The z-value is related to the influence this particular independent variable had over the dependent variable. The higher this value, the more influence the variable has on the outcome of predictions. The P-value says something about the trustworthiness of this z-value. The lower the P-value the more trustworthy the z-value is.

### Classification

Another type of statistical model is a classification model. Such a model aims to classify entries with predefined labels. The algorithm that creates this kind model out of training data also tries to reduce the error between the data and the prediction that the model would perform on this data. A specific type of classification model that is used in this research is a *classification tree*. A classification tree algorithm tries to find a split in the data that reduces the total error the most, iteratively applying this to the created splits. The resulting tree is then usually pruned to create a model that is not too complicated and does not *overfit* on the data. Overfitting occurs when a model is too tuned to the training data and thus fails on other data because it is also fitted to the noise contained in the training data. A classification model can be represented in many ways. An example representation is shown in figure 2.2. It shows the root node at the top, with 100% of the data (as no splits have occurred yet) and with a 50% male-female ratio. After an initial split on the *height* variable being larger than 175cm, two groups are made. One containing 35% of the data and consisting largely of males (represented in black). The other side naturally contains 65% of the data and consists of mostly females. This right node is split another time. This time the split occurs on the *weight* variable. If the weight is over 75kg (and $height \leq 175cm$) the data is categorized as male, otherwise as female. Of course, these two variables are by no means enough to determine the sex of a person (one could argue that two groups are not even enough to represent al sexes). However, using these variables we can determine the sex *with some confidence*. The quality of our model and our confidence of the model are directly related.

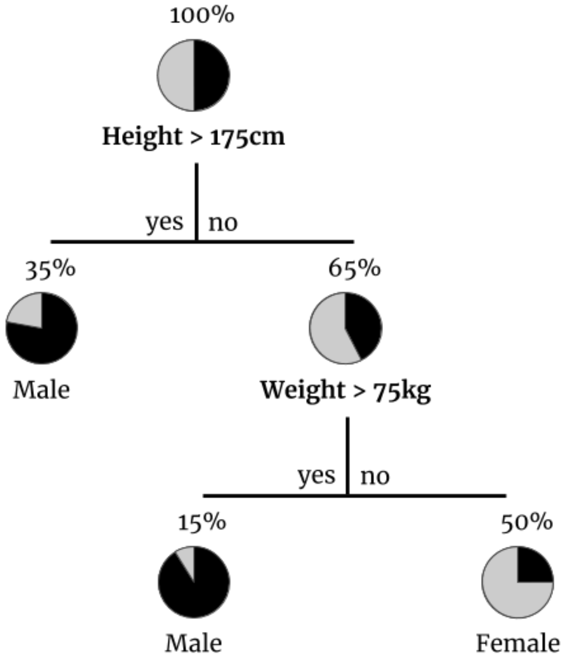


Figure 2.2: An example of a classification tree that splits a group of people into two different groups, male and female, using height and weight as variables.

| | | **True condition** | |
|---|---|---|---|
| | **Total population** | **Positive** | **Negative** |
| **Predicted condition** | **Positive** | True positive (TP) | False positive (FP) |
| | **Negative** | False negative (FN) | True negative (TN) |

Table 2.1: Structure of a confusion matrix. A confusion matrix compares the predicted values with the real values in a data set.

### Evaluating the models

To choose which models perform well and which do not a test set is often used. This test set is usually taken from the same source as the test set and the right outcomes for the dependent variable is known. When this is the case a prediction can be made using the models on this previously unseen data which, can be compared to the actual outcome. In order to determine how good a model perfoms on the test set, statisticians can use many different methods of evaluation. An often used method is called the *Root Mean Square(d) Error*, or *RMSE*. For a model $M$, and $n$ observations in training set $O$, the RMSE is defined as followed:

$$RMSE(M,O) = \sqrt{\frac{\sum\limits_{i}^{n} predict(M,O_i)^2}{n}} \tag{2.3}$$

In short, the RMSE takes the average of the difference between observations and the value that the model would predict for that observation. The prediction is squared, and subsequently square-rooted to make all errors positive; predictions higher or lower than the observation are treated the same. The RMSE is often used to include or exclude independent variables. The RMSE is mainly aimed at being a performance measure in regression, not in classification. As we use both a regression and a classification model, but want to compare them, we will use different measures in this research; true positive and negative rate, sensitivity, positive and negative predictive value, accuracy, and $F_{score}$. They can be calculated with the help of a confusion matrix. Table 2.1 shows the structure of a binary confusion matrix. A confusion matrix compares the prediction of a model with true values and categorizes all entries in one of four categories. *True positiveness* is the amount of entries that are predicted as true, which are also true in reality. *False positiveness* are entries predicted as false, but that are true in reality. It is also referred to as Type I error. *False negatives* are values predicted as false, but that are true in reality. It is also referred to as Type II error. The final category is *true negative.* These are values predicted as false that are also false in reality. Naturally, a perfect model maximizes true positiveness and true negativeness.

The other measures can be extracted from these basic values. Sensitivity (or recall, equation 2.4) is related to the number of positive values being identified as positive. It gives a quick view of how well positive values are predicted. Positive predictive value (PPV, precision) and negative predictive value (NPV) can be used to determine how big the distribution of the predictions are. A high PPV (equation 2.5) essentially means that a relatively high amount of the results are relevant results, with the opposite for NPV and irrelevant results (equation 2.6). Accuracy is related to the closeness of predicted values to the actual values. High accuracy means that your predicted values will, *on average*, be close to real values. The accuracy is the rate of properly predicted values, as can be seen in equation 2.7. Another important measure is the $F_{score}$ (equation 2.8). It takes both the sensitivity and the positive predictive value into account is a good way to determine the overall performance on the classification of positive examples.

$$sensitivity = \frac{TP}{positives} \tag{2.4}$$

$$PPV = \frac{TP}{TP + FP} \tag{2.5}$$

$$NPV = \frac{TN}{TN + FN} \tag{2.6}$$

$$accuracy = \frac{TP + TN}{population} \tag{2.7}$$

$$F_{score} = 2 * \frac{PPV * sensitivity}{PPV + sensitivity} \tag{2.8}$$

All these measures will be used in this research to compare our different models and techniques. For the classification of relevant entries in the databases, both the true positive rate as the true negative rate must be high. However, as we do not want to miss relevant results by classifying them as negative, we give extra weight to NPV and the $F_{score}$.

The ProS research centre[9] is a central organization in the analysis and modeling of genomic information systems [46, 32, 49]. In *Smart Data for Genomic Information Systems: the SILE method* [34], they designed a framework to efficiently identify variants related to the risk of suffering a disease with the overall goal of moving from the classical Big Data concepts Volume, Velocity and Variety to Smart Data, which includes Veracity and Value. In this context, Veracity is the different levels of quality your data can have. Value refers to the clinical relevance for practitioners. The SILE method (Section 2.3) was developed and combined it with their proposed Data Quality Methodology (Section 2.4) to ensure a high level of Veracity and Value in the obtained data.

## 2.3 SILE

The SILE method defines four levels that guide the process of information extraction from genomic information systems to provide a systematic approach to answering questions specific to the genomic domain (Table 2.2). The concept of data quality is mainly addressed in the Search and Identification levels.

[9] http://www.pros.webs.upv.es

| Level | Description |
|---|---|
| **Search** | Determination of the information context, required to solve a concrete need, as well as the selection of data sources from which to extract information |
| **Identification** | Determination of a reliable and relevant data set to be used to populate a database which structure is delimited by the CSHG |
| **Load** | Population of the database with the data identified in the previous level |
| **Exploitation** | Extraction of knowledge from the database by using tools to analyse and interpret genomic data |

Table 2.2: Levels of the SILE method as defined by León and Pastor [34].

### 2.3.1 Search

During the selection of sources in the Search level the main research goal has to be taken into account in order to find the most relevant sources. As an example, in the work of León [34, 33] the researchers tested the method on *Early Onset Alzheimer's Disease*, for this reason it was justified to include *AlzForum*[10] in the set of repositories used, aside from more general repositories such as *ClinVar*[11], *Ensembl*[12], *dbSNP*[13], *RefSeq*[14], *NCBI-Gene*[15] and *PubMed*[16], even though no programmatic interface to this database is available. In other work of León two phenotype specific repositories were excluded because they were outdated [32], which is in violation of one of the basic dimensions (Table B.1) in the Search level: *Currency*. Specifically, it violates the criterium of the latest update being not older than one year. Other basic dimensions for search are *accuracy* and *completeness* (see section 2.4).

### 2.3.2 Identification

Identification is the second level of the SILE method. It consists of identifying the available data and corresponding this with the attributes in the Human Genome Database (HGDB) [50, 47]. The identification is a two-part level. At first, the most relevant and accurate data needs to be found. Many of the attributes in the database may also be present in other databases. This can be used to search and cross-reference the found data. The second part of the identification is creating a correspondence between the found data and the HGDB, as this is the destination for the found data. The database is based on the Conceptual Schema of the Human Genome, which is part of the research done by Reyes et al. [50].

[10] https://www.alzforum.org

[11] https://www.ncbi.nlm.nih.gov/clinvar

[12] https://www.ensembl.org/index.html

[13] https://www.ncbi.nlm.nih.gov/projects/SNP

[14] https://www.ncbi.nlm.nih.gov/refseq

[15] https://www.ncbi.nlm.nih.gov/gene

[16] https://www.ncbi.nlm.nih.gov/pubmed

The goal of the Identification level is to construct a reliable and relevant data set for diagnosis and treatment, or research purposes, depending on the target audience (Table 2.2). León and Pastor used the Data Quality Methodology to elect five relevant quality dimensions. Section 2.4 goes into detail on how these dimensions were obtained.

## 2.4 Data Quality Methodology

León and Pastor propose a data quality methodology for genomic data [34]. The methodology consists of five phases: *Dimension Description*, *Metric Description*, *Variable Selection*, *Minimum DQ criteria* and *DQ Assessment*. These phases guide the user through the process of determining which dimensions are relevant and how to ensure these factors.

To determine which data quality dimensions are relevant (Dimension Description) for investigating genomic information systems, León et al. considered established literature [48, 37] and applied this to the genomics domain. Wand and Wang [60] developed a theoretical view of data quality using an ontological approach. They pose that an information system is designed to represent the real world and that looking at the view of the real world that is created by the information system and the view created by looking at the real world directly will expose deficiencies in the quality of the underlying data. By analysing mappings from information system to the real world and back they found four data quality dimensions, namely *completeness*, *lack of unambiguity*, *meaningfulness* and *correctness*. In Wang and Strong [61], another approach was taken. They investigated the view of data consumers on data quality. Practitioners were surveyed in two rounds and a data quality framework was developed on the resulting quality dimensions. Each of the dimension was assigned one of four data quality categories; *intrinsic*, *contextual*, *representational* or *accessibility*.

In earlier research León et al. found the most common errors in different genomic repositories [36]. Using this work they identified the relevant quality dimensions to the genomics domain based on the above literature. They also assigned each step in the SILE method to their relevant dimensions. For the Identification step the relevant quality dimensions are:

- Accuracy: the degree to which data describes an object or event correctly
- Completeness: related to the full representation of data
- Consistency: related to the degree of consistency in the representation of data between systems
- Believability: related to the credibility of data
- Relevancy: the degree of the helpfulness of data to the problem

Table 2.3 shows an example of quality metrics that can be chosen for the different quality dimensions. It is important to note that not every implementation of the SILE method uses the same metrics, they need to be chosen and adapted to the research at hand. León and Pastor [34] determined a set of minimum required and recommended attributes in the different investigated repositories (Minimum DQ criteria). The existence of these attributes ensures a high level of completeness in the data.

## 2.5 Information and knowledge quality

Information quality is defined as *fitness for use* [28, 29]. Ackoff [2] describes information as follows: *information consists of processed data, the processing directed at increasing its usefulness*. Bellinger, Castro and Mills [7] extend this definition as follows: *information embodies the understanding of a relationship of some sort, possibly cause and effect.* Using these descriptions one can easily infer the importance of data quality when aiming for high information quality. Processing and relating data forms information, and the correctness of the data determines in part the correctness of the information.

A similar relationship can be found between information quality and knowledge quality, however, this relationship is less obvious. Bellinger, Castro, and Mills [7] describe knowledge as follows: *knowledge is the application of data and information; answers 'how' questions (...) knowledge is the appropriate collection of information, such that its intent is to be useful (...).* Especially the *appropriate collection* and *usefulness* is important. When a proper selection of data is made, and in turn a proper selection of information, knowledge is gained. The extent to which information serves a specific purpose is called the *usefulness*. When looking at variant extraction, the information is useful when it contains the most amount of relevant entries and the least amount of irrelevant entries. These two metrics are related to the sensitivity and specificity in statistics.

[17] Minimum DQ criteria

[18] http://www.hgvs.org

| **Dimension** | **Quality metric** |
|---|---|
| **Accuracy** | *M1*: Review attributes liable to be error-prone. Syntactic errors must be checked using controlled vocabularies and specific data dictionaries. |
| **Completeness** | *M2*: The minimum information required to be stored in the HGDB is present. These attributes have been determined during Phase IV of the DQ Methodology[17]. |
| **Consistency** | *M3*: The information about the variations is defined by using standard vocabularies and verified ontologies to determine critical attributes such as HGVS[18] expressions, pathogenicity or functional effects. |
| | *M4*: There must be no conflicts in the clinical interpretation of each variation. |
| | *M5*: There must be no conflicts among databases related to the structural characteristics of the variation. |
| **Believability** | *M6*: Each variation must have significant medical or genealogical consequences and be reproducible (e.g. the reported consequence has been independently replicated by at least one group, besides the first group reporting the finding). |
| | *M7*: The relationship between the variation and the disease must have at least one link to a published, peer-reviewed paper with credible statistics and free access. |
| **Relevancy** | *M8*: The Minor Allele Frequency (MAF) of the variation must be less than the frequency of the phenotype in the population. |
| | *M9*: The inheritance pattern, penetrance and mechanism of the variant must be consistent with the disease. |
| | *M10*: The studies provided by the bibliography must have at least 500 participants and it is desirable that they are replicated. |
| | *M11*: For pathogenic variants the Odds Ratio must me greater than 1, and for protective variants the Odds Ratio must be less than 1. |
| | *M12*: For Genome Wide Association Studies (GWAS) the P-value must me less than $5x10^{-8}$. |

Table 2.3: The quality metrics in the Search level of the SILE method used in the work of León [34].

# Chapter 3

# Problem statement

## 3.1 The difficulty of extracting relevant information

Genomic information systems often have differing criteria for including or excluding data and information, which forms the first line of defense for data quality. Often there might also be additional statistics to let the researcher or practitioner decide which data to use and which not to use. This is the second line of defense to ensure data quality. These measures and criteria need to be determined by the user manually. This manual process is time consuming, with the researcher needing to analyse the results multiple times, adjusting the measures and criteria, and re-analysing the results. As León et al. [36] stated:

> The lack of standards and strict enough quality controls to submit information to databases drive to an inefficient management of multiple genome databases and (is) time-consuming for scientists.

The lack of data quality in public genomic information systems is directly causing the wrong results to be obtained from them and wastes time of researchers and practitioners. Because the sourcing of genomic data in these information systems is out of the hands of the users, it is important to properly understand the data quality within is systems. When this is understood, automation can be build on top to reduce the error and time consumed even more. The project aims to define information quality measures that are needed in finding important entries using the SILE method. León performed an extensive study into the quality aspect of genomic information systems [36, 34]. The perspective of this work is mostly top-down, working from data quality definitions and measures defined in literature down to criteria and thresholds for genomic information systems. We will take a bottom-up approach to provide another perspective and discover new aspects of variant retrieval, working from the data contained in genomic information systems, analyzing it and moving towards applicable criteria and automation. We will perform the Identification step of the SILE method on different phenotypes over time to see the changes occurring in relevant information systems (1). The Identification step will also be performed on different phenotypes to compare the differences between them, after which an extensive data analysis will be performed (2). Given the results of the first two parts, the performance of the SILE method will be evaluated under varying criteria to see how they can be improved. If needed, new criteria will be defined. After this, an automated process is developed to adapt the quality criteria and increase the final information quality (3). Section 3.2 will pose the main research question with relevant sub-questions to guide the research.

## 3.2 Research questions

Many different factors potentially influence data and information quality in the genomic domain. It is important to understand the data before we select measures and criteria and build and automation framework. To focus and guide the research we pose the following research question:

> *RQ: How can data quality, and by extension information quality, be improved when identifying relevant entries for different phenotypes from established and public genomic information systems, in an automated way?*

The question consists of three main parts; improvement of data and information quality, identification of relevant entries, and doing so in an automated way. The question is also focussed on the domain *established and public genomic information systems.* As this question consists of several parts we break it up into smaller sub-questions. These questions fit into two of the three steps of the Design Science cycle, which is discussed in section 3.3. SQ1 is related to the definition of data quality and its implementation in a genomic information system domain. SQ2 to 4 seek to answer how we can identify relevant entries in this domain. SQ5 is answered by the design of a treatment to the problems found in the problem investigation using the results from the exploration in SQ2 to 4.

### 3.2.1 Investigation

> *SQ1: What does data quality mean, given a genomic information context?*

SQ1 aims to find out how data quality should be defined in general, and specified for a genomic context. What does it mean to have a high or low data quality and how can it be quantified? It is the main question that drives the investigation of the problem and the domain.

### 3.2.2 Exploration and treatment design

> *SQ2: How much does the amount and quality of data, extracted from established genomic databases, differ over time and between phenotypes?*

SQ2 aims to quantify the problem researchers and clinicians face when selecting target variations. This is done by looking at the amount and quality (as defined in SQ1) that differ over time and between phenotypes. The initial part will focus on the GWAS Catalog and Ensembl, as they are two very well-known sources which have been researched extensively. It is an initial exploration that will push towards the treatment design, together with sub-questions 3 and 4.

> *SQ3: What are the factors that influence the quality of the extracted data?*

To be able to design a treatment for the problem quantified in SQ2, multiple target criteria need to be examined (SQ4). Answering SQ3 will lay the groundwork from a literature perspective by quantifying what types of data are relevant.

> *SQ4: Which criteria for variant selection are relevant?*

To design a treatment for the problem investigated in SQ1 it is important to find out which criteria are relevant. The answer to this question can partly be found in the work resulting from answering SQ3. It is important to note that a selection of the factors found while answering SQ3 might not be implementable with a criterium. Some factors might be very influential, but impossible to test for, or not compatible with current databases.

> *SQ5: How can these criteria be adapted to every context in an automated way to obtain the highest quality genomic data?*

The last part is the actual treatment design and is related to the application in different contexts in an automated way. Very little is known about the aspects related to automizing criteria selection and application, especially in the genomics domain. If it is possible to automate this part of variant selection it will greatly decrease the time it takes researchers and clinicians to reach actionability.

## 3.3 Research methodology

The project revolves around an improvement problem as described by Wieringa [62]. The goal is to improve the design of a method given the context (software APIs, people, processes, etc.), in a research environment. The nature of the project is exploratory. For these reasons the Design Science framework developed by Wieringa is an appropriate methodology. The design cycle will be followed (Figure 3.1). In this cycle, there are three major steps, problem investigation, treatment design, and treatment validation. The problem investigation revolves around creating a clear understanding of the problem and its accompanying concepts. The treatment design aims to develop a treatment to the problem defined during the investigation phase. Once a treatment is designed, a validation is performed. After this validation, there are often adjustments to be made, after which the cycle repeats. In this study, a validation in practice is not performed due to time constraints. As a supplement to the design cycle, the research cycle checklist will be used as a general checklist to ensure all parts of proper research have been addressed.

The Observational Case study was chosen as a methodology for the exploratory analysis. The reason for this is that we have no influence over the studied objects (the public genomic information systems), and that we do not aim to intervene in their operation. The treatment that is proposed works on the results of the exploratory analysis, not on the inner workings of the object of study. Hence, we believe that an Observational Case study is the most appropriate basis to use as a checklist. The following sections run through the checklist, describing each aspect of the research relevant to the Observational Case study. The numbers indicate which step of the checklist (Appendix table A.1) is described. Note that not all criteria are relevant to an Observational Case study, hence some criteria are not addressed (e.g. 8).

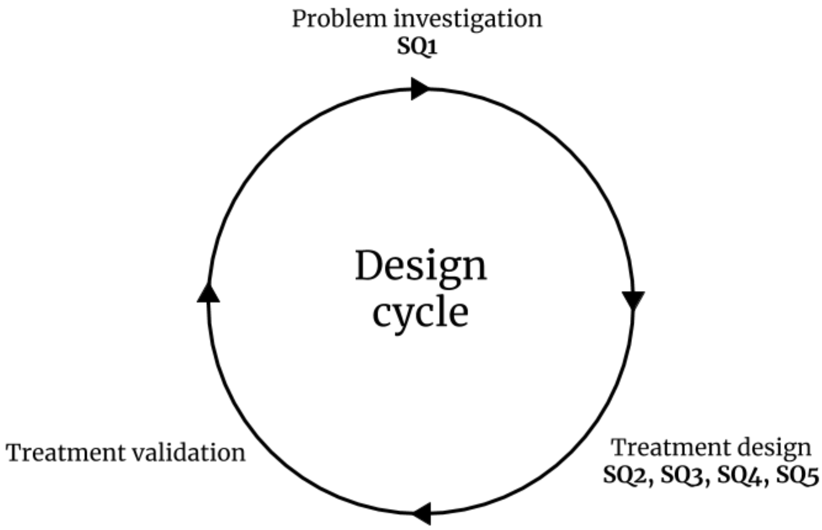


Figure 3.1: Design cycle by Wieringa, with subquestions associated with the three steps.

### Research context

The knowledge goal (1) is described in section 3.1: *to improve information quality measures that are needed in finding important entries using the SILE method.* The research has an improvement goal, which is to be able to provide more relevant research entries for researchers and clinicians. The project works on a part of the SILE method but is not necessarily part of a bigger engineering cycle, it is a standalone project (2). The current knowledge (3) is described in chapters 1, 2, and 3.

### Research problem analysis

To ensure consistency in concepts the research will use the conceptual schema of the human genome [47] as a basis (4). This will serve as a reference and a conceptual framework of the problem context. All sub-questions posed in section 3.2 are open and exploratory and have no hypotheses to freely interpret the results obtained (5). We expect that the questions will be answered descriptively, however, SQ2, 3, and 5 will have large explanatory components. The population consists of the GWAS Catalog and the Ensembl genome browser (6).

**Research & inference design**

The objects of study (7) will be the chosen genomic repositories relevant to the phenotypes investigated. Only established repositories will be used, by using selection criteria like the ones in [34]. The measurement design (9) for the project is fairly straight forward. We measure the number of retrieved entries for sub-question 2. Repeatability is an issue because repositories change over time and a repeat study can find different amounts of entries. The aim is to find generalizations that are also valid for repeat studies. For the sake of repeatability, all relevant resources and results will be shared in an appendix or digitally. For sub-question 1, 2, 3 and 5 different quality criteria will be developed. As this is part of the ongoing research the measures accompanying these criteria are currently unknown. The inference design (10) of this project is also clear. The study is performed on the whole of the population, namely all relevant public genomic repositories. As the study is performed on a per-phenotype basis the results only represent that specific phenotype. However, if similar results are obtained from different phenotype a generalization can be inferred.

## 3.4 Major milestones

To guide the time aspect of the research project, multiple milestones are defined. Table 3.1 outlines the milestones and the original approximate week of around the time of the long proposal. It also states the actual week the work was finished. The description includes reasoning why the implementation differed from the planning. The first milestone (*Scripts*) intends to create a technical basis for the research. Automation of the pipeline is needed to ensure consistency and to avoid human error. It will also speed up the overall process. Scripts will be written in R. It is to be expected that scripts will need adjustment later on, nonetheless the aim is to make them as robust as possible in this initial development. Milestone two (*Data Pulling*) concerns the pulling, storing and cleaning, of all relevant data for the research. As this is the first real use if the scripts created earlier it is to be expected that some adjustments have to be made here. The third milestone (*Data Analysis*) is the least defined, as the research is experimental. Statistical methods are used to find correlations and trend lines. An model will be made of the data. After the analysis of the retrieved data, an automation framework will be proposed and built in milestone four (*Automation*). For this, the analysis done in the third milestone will be used. We will search for patterns and use the model to find different important aspects of the data in genomic information systems. The final milestone (*Writing*) concerns the finishing of all written parts of the thesis project, as well as a preparation for the final presentation and defense.

| **Milestones** | **Week planned** | **Week finished** | **Description** |
|---|---|---|---|
| Scripts (I) | 21 | 21 | This part was finished on-time. Slight changes have been made over the months following. |
| Data Pulling (II) | 25 | 23 | Because of proper scripts, this part took substantially shorter than expected. The pulling would have to be re-run a couple of time in the remaining months because of changing requirements, but because of the script automation (I) this did not induce any more delays. |
| Data Analysis (III) | 28 | 29 | This part took longer than planned, largely because experimentation broadened the scope, which also induced delay in the following part (IV). |
| Automation (IV) | 31 | 34 | The previous part (III) introduced some delay in this part as well. Because of the more extensive experimental part, there were more possibilities for automation. This in turn made practical validation fall outside the scope of this research, something that was considered in the beginning stages of the project. |
| Writing (V) | 40 | 46 | The change of scope in part III and IV naturally changed sections in the final text. Some organisational factors induced additional delay here. |

Table 3.1: Planned and executed milestones of the research, with their original planned week of finishing, as well as the actual week of finishing and a coarse description of the progress.

# Chapter 4

# Exploration and treatment design

To design a treatment that will be valuable in practice we first need to explore the domain, especially the information systems used in for this research, the GWAS Catalog and Ensembl. For sub-question 2 (*amount and quality over time*) an initial exploratory data analysis is performed on the data present in these systems. The goal is to understand the data contained in these databases better and to quantify it. After the exploration of time and phenotypes as differing factors, we focus on other factors, such as risk frequency, P-value, and others, to answer sub-question 3 (*factors that influence quality of extracted data*). For sub-question 4 (*criteria for variant selection*), we also look into which criteria should be set. Furthermore, we figure out how they can be set automatically to answer sub-question 5 (*adapt criteria in an automated way*).

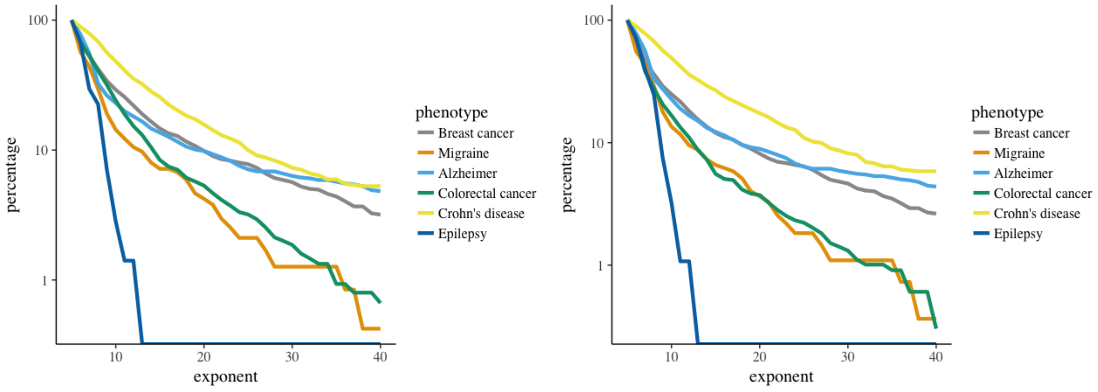
## 4.1 Initial phenotypes

The phenotypes selected for the exploratory part are breast cancer, migraine, Alzheimer's disease, colorectal cancer, Crohn's disease, and epilepsy. The reason for using these phenotypes for the exploratory analysis is that they have been investigated before by the PROS team one way or another, which improves the data set created for this research, as there is already a lot of knowledge about these phenotypes within the research group. Later, additional phenotypes are included.

Table 4.1 shows the number of found entries for the phenotypes included in this part of the data analysis in each database. Already, it can be seen that there is a big difference between the two databases. Ensembl has more entries included for all phenotypes. The difference is especially visible with Alzheimer's disease and Crohn's disease. However, these results also contain non-GWAS genetical studies. The completely unfiltered results lie far apart for most phenotypes, but after filtering the results are more similar. The basic filtering on non-existent values removed 1921 (of 4054, 47%) entries over all the investigated phenotypes in the GWAS Catalog while removing 8504 (of 11320, 75%) of the entries in Ensembl. After filtering there is on average a difference of 36% in the number of entries between the two databases.

| **Phenotype** | **EFO #** | **GWAS Catalog** | **Ensembl** |
|---|---|---|---|
| Breast cancer [10, 11] | 0000305 | 720 (1409) | 782 (2459) |
| Migraine [45] | 0003821 | 235 (237) | 266 (317) |
| Alzheimer's disease [31] | 0000249 | 353 (1098) | 493 (1090) |
| Colorectal cancer [32] | 0005842 | 511 (540) | 666 (4757) |
| Crohn's disease [35] | 0000384 | 264 (717) | 553 (1103) |
| Epilepsy[1] | 0000474 | 50 (53) | 56 (1594) |

Table 4.1: Number of found entries for each phenotype, with a basic filter checking for non-existent values (unfiltered between brackets).



(a) GWAS Catalog

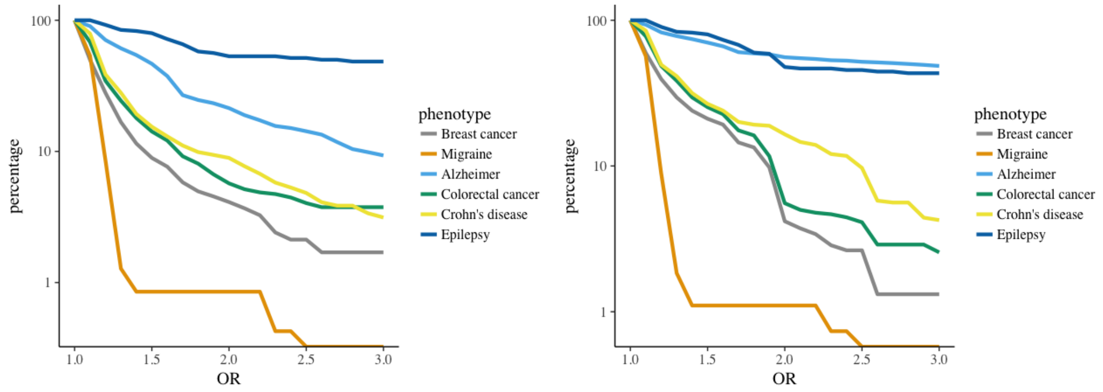
(b) Ensembl

Figure 4.1: Falloff graph of P-value: the percentage of entries remaining when using a varying P-value threshold. On the X-axis we show the exponent of the threshold used ($p < 10^{-x}$). The Y-axis shows the remaining percentage of entries found for the phenotype after applying the filter.

Figure 4.1 visualizes the distribution of entries over P-values thresholds in the GWAS Catalog and Ensembl. Both distributions look quite similar, although not equal. Epilepsy has a strongly declining amount of entries in both databases, which indicates there are no high P-value research results for the epilepsy phenotype. On the other hand, Crohn's disease has a very flat curve in both databases, with about 10% of entries having a $p \leq 10^{-40}$

Then it comes to Odds Ratio, figure 4.2 paints a different picture. It seems that Ensembl contains a higher amount of high OR studies on Alzheimer's disease phenotype. This implies that when looking for entries with a high impact on the development of Alzheimer's disease Ensembl is the more appropriate data source.

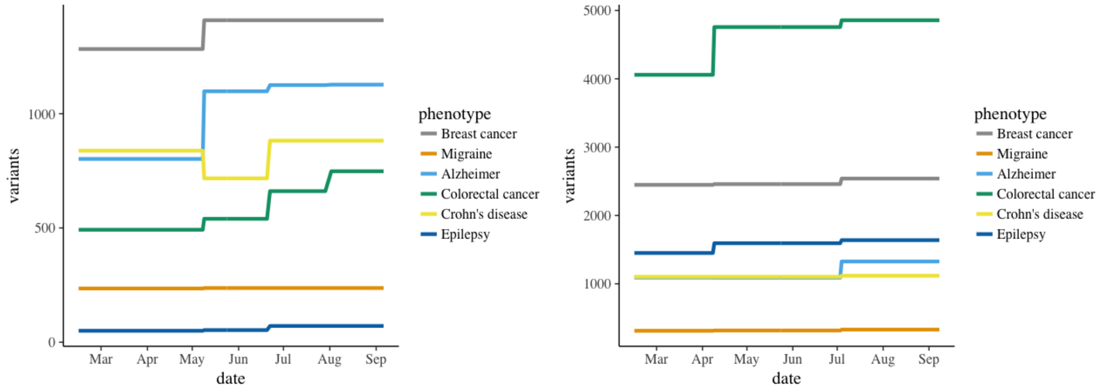
[1] Currently no publication.



(a) GWAS Catalog

(b) Ensembl

Figure 4.2: Falloff graph of Odds Ratio: the percentage of entries remaining when using a varying Odds Ratio threshold. On the X-axis the threshold for the Odds Ratio is shown ($OR \geq x$). The Y-axis shows the remaining percentage of entries found for the phenotype after applying the filter.



(a) GWAS Catalog

(b) Ensembl

Figure 4.3: Retrieved unfiltered entries over 7 months in 2019. Further updates can be seen on *https://sites.google.com/view/variationscanner*.

## 4.2 Entries over time: the volatility of information systems

The amount of entries found is also measured over time. Figure 4.3 shows the number of unfiltered entries in over 7 months. It can be seen that Ensembl performed an update at the beginning of April[2] and July[3] which is in accordance of their three-month update cycle[4]. The GWAS Catalog states that it has a weekly update cycle[5]. However, when looking at the number of entries found in the database the update cycle only seems to be slightly faster than Ensembl. Over the 7 months of tracking entries, only three updates have been observed. These updates were done at the beginning of May, at the end of June, and at the beginning of August.

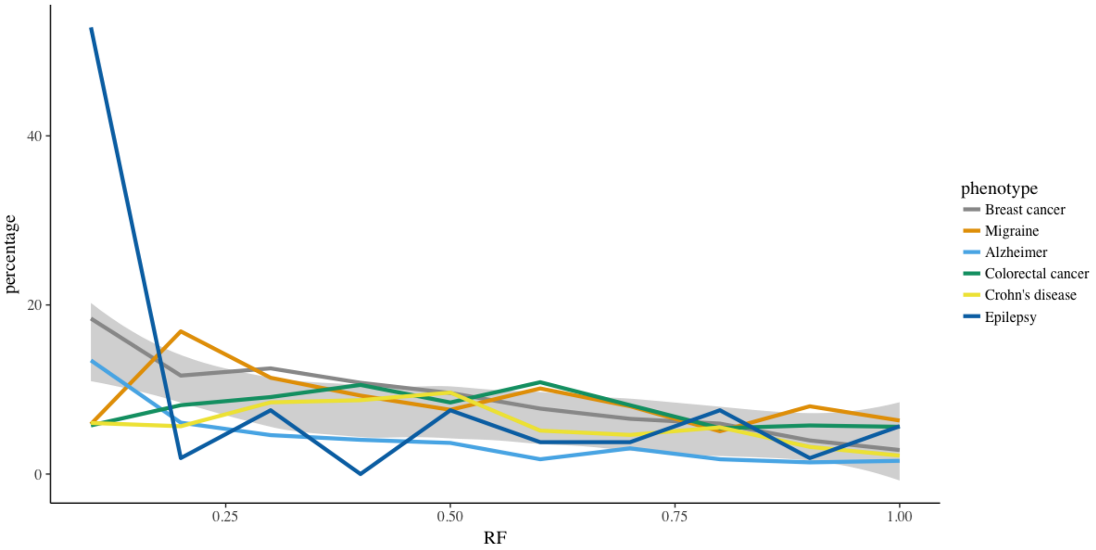


Figure 4.4: Distribution of risk frequency (RF) in the GWAS Catalog on different phenotypes. The amount of entries are binned with a resolution of 0.1 and reported as a percentage of the total amount of non-empty values. The grey band shows the 95% confidence interval.

The GWAS Catalog reports on the risk frequency through the API. To understand the amount of the total population possesses the different stored entries, it is relevant to plot the percentage of entries binned by risk frequency. The results of this operation for the chosen phenotypes can be seen in figure 4.4. In general a slight decrease in remaining entries when increasing the risk frequency bin. This is also shown in the confidence interval. Such a decrease indicates that rare variants are represented more often than more common variants. This can have a biological explanation, but can also be explained by the interest of researchers in less common variants (section 2.1.6).

[2] http://www.ensembl.info/2019/04/09/ensembl-96-and-ensembl-genomes-43-are-out/

[3] http://www.ensembl.info/2019/07/03/ensembl-97-and-ensembl-genomes-43-have-been-released/

[4] https://www.ensembl.org/info/about/release_cycle.html

[5] https://www.ebi.ac.uk/gwas/docs/about, https://www.ebi.ac.uk/gwas/docs/faq#faq-A7

## 4.3 Building a prediction model

For practitioners and researchers, it is important to know how to select relevant research data from these databases. They need to know which factors influence the data quality, and in the end, the information quality. A way in which relevant factors can be found is by creating a data set with relevant and less relevant entries. By finding the difference between these entries on specific measurements one can find which factors have more or less influence on the relevancy (and thus the data quality). A good way of finding these differences is by training prediction models on the data set. Prediction model algorithms are made to find differentiating factors between datapoints. They need this information to predict outcomes on unseen data. Relevant factors are weighed heavier than unimportant ones. The exact way in which this is done is different for different prediction algorithms. We can use this weighting to find more relevant factors for practitioners and researchers if clear factors exist. The predictive nature of these models can also be used to make a relative estimate on how relevant the entries in the database are.

With this end goal in mind, a combined classification and regression model is developed. The base models are created from data retrieved from the GWAS Catalog because of the ease of data extraction and the extent to which the returned data is structured. Initially, two simple models were built to provide an initial starting point to evaluate model performance. These baseline models work with the entries retrieved for the late-onset Alzheimer's disease (LOAD) phenotype as training data. The reasoning behind this is that there is an established set of core target genes used in LOAD research. The Mayo Clinic, for example, reports the following genes as being important in the development of LOAD: APOE, ABCA7, CLU, CR1, PICALM, PLD3, TREM2, SORL1[6]. The relevancy of these genes (among others) is supported by Agrawal [3, 30]. Both baseline models have several variables at their disposal, which can be seen in table 4.2. Section 4.3.1 talks about the first model created using linear regression. Section 4.3.2 shows how a classification model builds on the same data predicts with a higher confidence. Afterward, these two models and the gained knowledge is combined into a single relevancy model.

### 4.3.1 A linear model

The first model to be built on the LOAD data is a linear regression model. Linear regression models are made by fitting lines to the different independent variables in the data and determining their influence on the dependent variable. A reversal of this process can be performed on newly observed data to predict the dependent variable. As the dependent variable is binary (a variant is either on a gene or it is not) a logistical binary regression is used. An advantage of using a regression model is that the resulting predictions come in the form of a numerical value and not just simple classification. This way the results can be ordered according to the likeliness of being a relevant gene.

[6] https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers-genes/art-20046552

| Variable | Type | Description |
|---|---|---|
| $P_{exp}$ | numerical | The exponent of the reported P-value ($\log_{10}(P)$). |
| OR | numerical | Reported Odds Ratio of the variant. |
| RF | numerical | Reported Risk Frequency of the variant. |
| populationNumber | numerical | The total population (initial and replication summed) of the reported study. |
| populationType | class | The type of population (initial or replication, combined into distinct classes). Only used by classification model. |

Table 4.2: Variables at the disposal of the two initial models.

| **Coefficient** | **Variable** | **Value** | **z-value** | **P-value** |
|---|---|---|---|---|
| $\beta_1$ | (intercept) | $-1.553$ | $-1.307$ | 0.191 21 |
| $\beta_2$ | $P_{exp}$ | $-9.364 \times 10^{-2}$ | $-3.253$ | 0.001 14 |
| $\beta_3$ | OR | $-7.328 \times 10^{-1}$ | $-1.066$ | 0.286 29 |
| $\beta_4$ | RF | $4.135 \times 10^{-1}$ | 0.562 | 0.574 08 |
| $\beta_5$ | populationNumber | $-2.216 \times 10^{-6}$ | $-0.223$ | 0.823 60 |

Table 4.3: Logistical model coefficients of model A.

The first trained model (A) takes the following shape, with $\beta$ being the coefficients determined by the training algorithm and $p_i$ being the probability that observation $i$ is a relevant observation:

$$y_i = \beta_1 + \beta_2 * P_{exp} + \beta_3 * OR + \beta_4 * RF + \beta_5 * populationNumber + \epsilon \tag{4.1}$$

$$p_i = \frac{\exp(y_i)}{\exp(y_i) + 1} \tag{4.2}$$

The coefficient values and their associated z-values and P-values can be read in table 4.3. Looking at the z-value for $\beta_4$ and $\beta_5$ it is immediately clear that they have relatively little influence on the dependent variable. They also have a high certainty of being a random occurrence, considering that they have high P-values. By taking these two variables out of the model and retraining it we can simplify it without offering up much power of the model.

Table 4.4 shows the simplified model. As can be seen, this model has higher z-values and lower P-values for all coefficients, which because it is a simpler model as well indicates we found a more appropriate model. By rescoring the entries in the training we can see how much the two models A and B differ from each-other in predictive behavior. Both models categorize the training data in the same way. The results of this categorization can be seen in table 4.5. The True Positive (TP) rate of 13% is quite low, which means the model is not very good at predicting target entries. The True Negative (TN) of 98% looks more promising, but it has to be taken into account that the training data is fairly

| **Coefficient** | **Variable** | **Value** | **z-value** | **P-value** |
|---|---|---|---|---|
| $\beta_1$ | (intercept) | $-1.488\,43$ | $-2.177$ | $0.029\,512$ |
| $\beta_2$ | $P_{exp}$ | $-0.091\,72$ | $-3.558$ | $0.000\,374$ |
| $\beta_3$ | OR | $-0.675\,57$ | $-1.346$ | $0.178\,344$ |

Table 4.4: Logistical model coefficients of model B. Coefficients $\beta_4$ and $\beta_5$ are removed from this model because of the low z-values and high uncertainty.

| Logistic regression | | **True class** | | |
|---|---|---|---|---|
| | | Relevant | Irrelevant | |
| **Predicted class** | Relevant | 5 | 3 | 8 |
| | Irrelevant | 34 | 173 | 207 |
| | | 39 | 176 | 215 |

Table 4.5: Confusion matrix of the prediction by the logistic regression model in table 4.4 (model B). It shows a True Positive rate of 13% and a True Negative rate of 98%. The PPV is 63%, the NPV is 84%, and the $F_{score}$ is 21%.

unbalanced (82% of the data consists of negative/non-target examples). The weakness of the model can be seen in the $F_{score}$ of 21%. It is an interesting model to use as part of a more complicated model though, as the logistic regression model can be used to score the entries. In contrast to a classification model, a (logistic) regression model assigns distinct values to different entries.

### 4.3.2 A classification model

In an attempt to better classify entries as being relevant or irrelevant we build a classification tree on the same data model A and B used. For this, the R package *RPart* is used, with a complexity parameter of 0.05. All other parameters are kept at default. Figure 4.5 shows the resulting tree. In this model C, the $P_{exp}$ is also identified as being of great importance, separating 71% of the entries in the training data at the first split. The remaining 29% is split into 5 separate bins by the variables RF and OR.

When creating a classification on the training data, this model C performs better than model B, as can be seen in table 4.6. The model has a TP-rate of 67% with the same TN-rate as model B, making for an $F_{score}$ of 75%. The PPV is also significantly higher.

## 4.4 Improving our base models

The models already perform quite well on the data we provided, but there are improvement points. In general, the more data is provided to a learning model, the better the model will perform. Therefore, we extend the data set used for training with additional phenotypes and data points. We also balance the data set so there is an equal amount

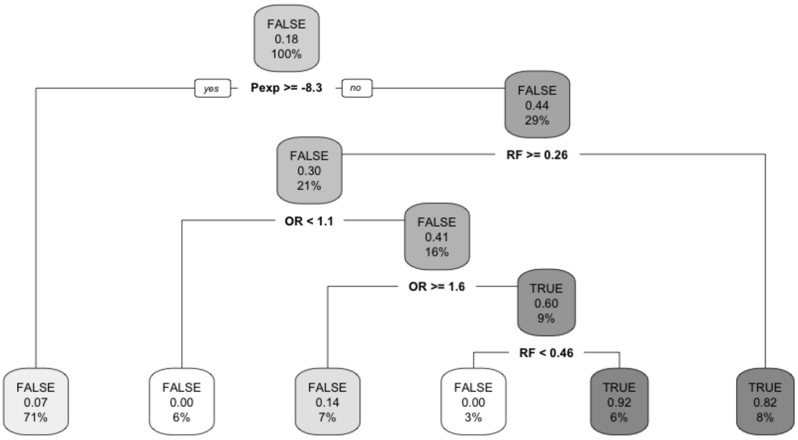


Figure 4.5: Decision tree (model C) created using RPart ($cp = 0.05$) on the cleaned Alzheimer's disease entries (215) extracted from the GWAS Catalog.

| Classification tree | | **True class** | | |
|---|---|---|---|---|
| | | Relevant | Irrelevant | |
| **Predicted class** | Relevant | 26 | 4 | 30 |
| | Irrelevant | 13 | 172 | 185 |
| | | 39 | 176 | 215 |

Table 4.6: Confusion matrix of the prediction by the classification tree (model C) in figure 4.5. It shows a True Positive rate of 67% and a True Negative rate of 98%. The PPV is 87%, the NPV is 93%, and the $F_{score}$ is 75%.

of positive and negative examples. There is an overrepresentation of negative examples. To balance the data set we over-sample the positive data points. The work by Batista, Prati, and Monard [5] shows that random over-sampling performs competitively to more complicated sampling methods. As mentioned before, the entries are also binned in prespecified values by the classification model, which can be disadvantageous when trying to rank the entries. Because this model performs the best on all measurements, this model should be used despite this. To improve this model C, we combine it with model B, the best regression model. As the real test for a model is to perform well on unseen data we will also split the data set in two, using 25% of the data as a test set, with the rest being used as training data. This split is performed at random, with a set seed for repeatability. We do this so the data that we get our performance measures from are not previously seen by the model, hence providing more realistic measurements.

### 4.4.1 Better base data

To increase the effectiveness of the two kinds of base models, more data is needed. In basic terms; the more data a model has to learn on, the less likely it is to overfit and thus behave badly on unseen data. More factors influence the amount of overfitting, but an extensive data set is key. With that in mind, the base data set is extended. Apart from including more Alzheimer target genes, more phenotypes are included; *Parkinson's disease*, *breast carcinoma*, *amyotrophic lateral sclerosis (ALS)*, *colorectal cancer*, *ankylosing spondylitis (AS)*, *prostate carcinoma* and *multiple sclerosis (MS)*. These phenotypes have been chosen as they are all highly debilitating or deadly diseases, which makes them prime targets for relatively expensive genetic research. They are also all very different phenotypes when it comes to protein pathways and the development of the disease.

Table 4.7 shows the phenotypes with the established relevant genes, based on the sources provided. Most of the sources are aimed at the general public because being communicated to the general public through a national health institute indicates that the gene is an established factor in the development of the disease. Parkinson's disease is a disease that affects the nervous system. It develops progressively, which means it worsens over time until death or serious disability. Parkinson's generally starts very mild with symptoms such as slight tremor, impaired balance, or changes in speech. Affected persons have a high chance of developing dementia and depression in later stages. Many late-stage Parkinson's diseased are bed-bound or chair-bound [20]. Where Parkinson's disease can be seriously debilitating, ALS is deadly. Also neurologic in nature, the disease degrades all voluntary muscle movement over time. The patient usually ends up with the inability to eat and breathe, after which they die. The life expectancy is normally between 2 and 4 years after symptoms begin [22]. Breast carcinoma (cancer) is a disease that disproportionally affects females, usually at an age over 50 [53, 56]. It is estimated that 5-10% of cases are hereditary in nature, which makes it an interesting target for genetic research. Prostate carcinoma (cancer), in contrast, is a form of cancer that by definition only affects males. About 15% of diagnosed male cancer cases are prostate cancer, with 7.9% in both sexes combined [63]. Colorectal cancer is a type of cancer that is mostly caused by environmental factors. Genetics play a relatively small part in the chance of developing the disease, it is considered a *lifestyle disease* [64]. This is in contrast with for example breast cancer. It is, however, a very prevalent type of cancer, and thus, every genetic insight is of help to a large group of patients. AS is believed to be an autoimmune or auto-inflammatory disease which is known to be highly genetic in nature [54]. It mainly affects the spine, but other parts of the body can be affected, such as the pelvis, eyes, and bowels [38]. Although genetics are not completely responsible for developing the disease, there is a genetic factor [39]. MS affects nerve cells in the brain and spinal cord, which results in the inability of the nervous system to send signals throughout the body [14]. Because of the all-encompassing nature of the nervous system, symptoms can range widely per patient. Examples of symptoms include eye problems (double vision or blindness), sensing issues and muscle weakness [44]. There is a genetic component in MS, with siblings of affected individuals having a higher risk of developing the disease themselves [16].

| **Phenotype** | **Positive/filtered (unfiltered)** | **Relevant genes** | **Source** |
|---|---|---|---|
| Late onset Alzheimer's disease (LOAD) | 41/186 (1139) | APOE, ABCA7, CLU, CR1, PICALM, PLD3, TREM2, SORL1, APP, PS1, PS2 | [13, 55] |
| Parkinson's disease | 20/126 (263) | SNCA, UCHL1, LRRK2 | [27, 42] |
| Breast carcinoma | 3/621 (1412) | BRCA1, BRCA2 | [25, 40] |
| Amyotrophic lateral sclerosis (ALS) | 2/72 (119) | SOD1, C9ORF72, FUS, TDP43, NEK1, TDP43, UBQLN2, KIF5A | [4] |
| Colorectal cancer | 1/469 (751) | APC, MSH2, MLH1, PMS2, MSH6, PMS1 | [26] |
| Ankylosing spondylitis (AS) | 13/74 (416) | HLA-B, ERAP1, IL1A, IL23R | [39] |
| Prostate carcinoma | 1/510 (714) | BRCA1, BRCA2, HOXB13 | [43] |
| Multiple sclerosis (MS) | 17/187 (422) | HLA-DRB1, IL7R, CYP27B1, IL2RA, TNFRSF1A | [41] |
| **Total** | **98/2245 (5236)** | | |

Table 4.7: Phenotypes used for extended data set. The *unfiltered* entries are the raw entries that come from the GWAS Catalog API for the specified phenotype. These entries are *filtered* for missing values, after which they are split into *positive* and *negative* hits.

# Chapter 5

# Results

As shown in the previous chapter, both a classification and a regression model can help in the selection of relevant entries in a genetic information system like the GWAS Catalog. The amount of properly predicted entries, before optimization of the models, is already above 80%. This can be very useful for fast filtering of entries. There are also improvement points when it comes to the way we test the models, as well as the amount of data they are trained on. This section shows the final models created, as well as the combination of both models into one, capitalizing on the strong parts of both models.

## 5.1 Improved base models

When using the balanced and extended data set to retrain models B and C into the models $B_2$ and $C_2$, we find the following results. Model $B_2$ (Tables 5.1 and 5.2) performs more balanced when looking at the True Positive and Negative rates, as well as the PPV and NPV. This has to do with the balancing of the data set to contain a similar amount of positive and negative examples, which in turn improves the $F_{score}$ drastically, to 48%. It can be seen that the variables $P_{exp}$, $OR$ and $RF$ are included in this model. The reasoning for adding the $RF$ variable back into the model was that it performed with more confidence under this new data set and has quite a substantial influence in the prediction. Model $C_2$ (Table 5.3 and Figure 5.1) also performs more balanced and turns out to have an $F_{score}$ of 84% which is very promising. As can be seen, the model is slightly more complex than the original model. This has to do with the reduced complexity parameter *cp*. It was found that this reduced complexity parameter (which will make the model more complex) makes the model perform better with our new training set.

## 5.2 Combining models

To improve the power of the model on classifying the database entries the two models are combined. The theory behind this is that even though the classification model

| **Coefficient** | **Variable** | **Value** | **z-value** | **P-value** |
|---|---|---|---|---|
| $\beta_1$ | (intercept) | $-0.875\,872$ | $-8.606$ | $2 \times 10^{-16}$ |
| $\beta_2$ | $P_{exp}$ | $-0.025\,588$ | $-9.717$ | $2 \times 10^{-16}$ |
| $\beta_3$ | OR | $0.188\,361$ | $0.033\,578$ | $2.03 \times 10^{-8}$ |
| $\beta_4$ | RF | $0.302\,575$ | $2.213$ | $0.0269$ |

Table 5.1: Logistical model coefficients of model $B_2$.

| Logistic regression | | **True class** | | |
|---|---|---|---|---|
| | | Relevant | Irrelevant | |
| **Predicted class** | Relevant | 201 | 69 | 270 |
| | Irrelevant | 360 | 492 | 852 |
| | | 561 | 561 | 1122 |

Table 5.2: Confusion matrix of the prediction by the logistic regression model, retrained on the extended and balanced data set ($B_2$). It shows a True Positive rate of 36% and a True Negative rate of 88%. The PPV is 74%, the NPV is 58%, and the $F_{score}$ is 48%.

| Classification tree | | **True class** | | |
|---|---|---|---|---|
| | | Relevant | Irrelevant | |
| **Predicted class** | Relevant | 514 | 150 | 664 |
| | Irrelevant | 47 | 411 | 458 |
| | | 561 | 561 | 1122 |

Table 5.3: Confusion matrix of the prediction by the classification tree, retrained on the extended and balanced data set ($C_2$). It shows a True Positive rate of 92% and a True Negative rate of 73%. The PPV is 77%, the NPV is 90%, and the $F_{score}$ is 84%.

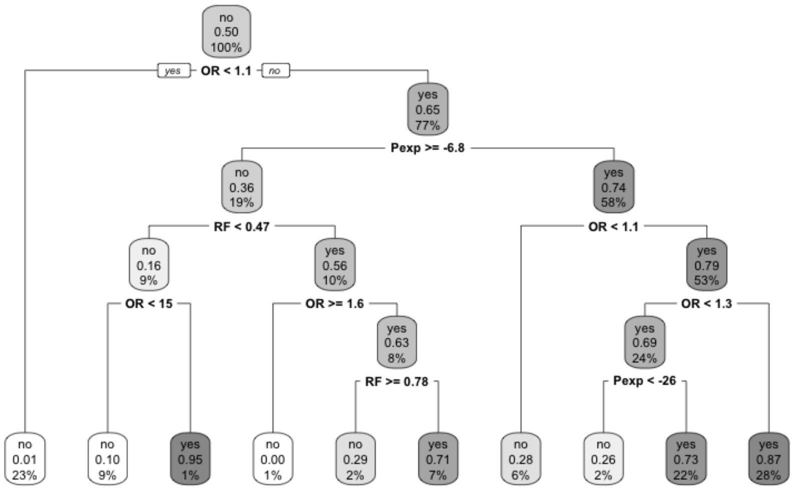


Figure 5.1: Decision tree (model $C_2$) created using RPart ($cp = 0.01$) on the extended and balanced data set extracted from the GWAS Catalog.

performs better on our measurements, all the entries are binned into distinct groups. The regression model can rank the entries within these groups because it predicts distinct values for different inputs. Because the classification models perform better we want to give this model preference in prediction. This can be accomplished by giving this model more weight in the final prediction. Equation 5.1 shows the way in which this is done, with $w_X$ being the weight for model X, and $prediction_{X,i}$ being the prediction by model X on entry $i$.

$$relevancy_i = w_{B2} * prediction_{B2,i} + w_{C2} * prediction_{C2,i} \tag{5.1}$$

Intuition tells us that the weight factors $w$ have to be set so that the dominant factor will be model $C_2$, with model $B_2$ only influencing the entries that are binned together by model $C_2$. The added effect of combining the two models is that entries that are binned distinctly but close together, and are thus very similar, can still be re-ordered by model $B_2$. Naturally, we want to max out all the measures when balancing the two models. However, we give extra weight to the Negative Predictive Value and the $F_{score}$, as maximizing them will insure less, possibly interesting entries will be marked as irrelevant. To proof our intuition we constructed table 5.4, which shows the measurements for different balances of the two base models. The *influence* is the relative weight of model $B_2$ on the final outcome. As described in the table an influence of 50% or less for the $B_2$ model is favorable for our specific use-case.

| **Influence** $B_2$ | **TPR** | **TNR** | **PPV** | **NPV** | **F-score** |
|---|---|---|---|---|---|
| **5%** | **91%** | 78% | **80%** | **89%** | **85%** |
| **10%** | **91%** | 78% | **80%** | **89%** | **85%** |
| **15%** | **91%** | 78% | **80%** | **89%** | **85%** |
| **20%** | **91%** | 78% | **80%** | **89%** | **85%** |
| **25%** | **91%** | 78% | **80%** | **89%** | **85%** |
| **30%** | **91%** | 78% | **80%** | **89%** | **85%** |
| **35%** | **91%** | 78% | **80%** | **89%** | **85%** |
| **40%** | **91%** | 77% | **80%** | **89%** | **85%** |
| **45%** | **91%** | 77% | **80%** | **89%** | **85%** |
| **50%** | **91%** | 77% | **79%** | **89%** | **85%** |
| **55%** | **91%** | 76% | **79%** | **89%** | **84%** |
| **60%** | **92%** | 76% | **79%** | **90%** | **85%** |
| **65%** | **92%** | 75% | **79%** | **90%** | **85%** |
| **70%** | **91%** | 76% | **79%** | **90%** | **85%** |
| **75%** | 85% | 78% | **80%** | 84% | 82% |
| **80%** | 75% | 81% | **80%** | 76% | 77% |
| **85%** | 64% | 84% | **80%** | 70% | 71% |
| **90%** | 46% | **88%** | **79%** | 62% | 58% |
| **95%** | 44% | **89%** | **81%** | 61% | 57% |

Table 5.4: The critical measurements of the combined model under a changing influence of the logarithmic model. For the True Positive rate (TPR), the Negative Predictive Value (NPV), and the $F_{score}$ we can see a sharp drop-off in performance with and influence of over 70%. The Positive Predictive Value (PPV) does not change significantly across all influence percentages and can be ignored when balancing the influence factors (in essence, the models are very similar in this measurement). The True Negative rate (TNR) suffers quite a bit under lower influence factors, although it goes up slightly again below 50%. Seeing these numbers, and taking into account that we give extra weight to the NPV and $F_{score}$ (section 2.2), a balance of 50% or less in favour of model $B_2$ is advisable.

# Chapter 6

# Conclusions & Discussion

The research project conducted for this thesis lays the groundwork for automated variant retrieval. The domain has been explored from a literary standpoint as well as by data analysis. Different solutions to automated data retrieval have been created and tested in an academic environment. This chapter will go into the conclusions to this research and discusses several points and limitations which have to be addressed if a similar solution is to be deployed in more practical settings.

## 6.1 Meaning of data quality

One of the fundamentals of this project is the way we can define data quality. An important aspect is that the context is critical. A question posed at the beginning of this project is:

*What does data quality mean, given a genomic information context?*

The answer to this question consists of two parts. At first, a general definition for data quality is given, after which it is applied to the context of genomic information. As described in section 2.4, the amount of quality in data is related to the amount in which data represents the real world. High data quality means the data describes the real world well. There are multiple ways to measure this descriptiveness, called data quality dimensions. The Data Quality Methodology specifies the dimensions relevant for the SILE method and gives a selection of minimum and recommended quality metrics, all in the context of genomic information systems. The increase in data quality leads to higher information quality, which in turn ensures higher usefulness and an increase in the knowledge we can gain from the data. One way to specify usefulness is the amount of relevant and irrelevant entries in the filtered data set. Expert knowledge is needed to determine if entries are relevant in the reduced data set. Reducing the data set before introducing experts to the process reduces the valuable time spent filtering through heaps of entries.

There is however the complication of missing data. Many entries in the Ensembl database and GWAS catalog miss information on critical measurements such as Risk

Factor and Odds Ratio. There are two practical solutions to this problem when working with both logistic regression models, as well as classification trees. One involves ignoring the missing information by rejecting the entries when training a model. The downside to this is a major reduction of the usable data points. As can be seen in Table 4.7, over half of the entries are expected to be thrown out when using this method. However, these entries can contain valuable data. Another, as of yet untested method, could be to either give sensible values to the missing data points, for logistic regression, or to treat them as a special class when using a classification tree approach. There is a risk in adding information manually this way, as it might introduce biases that are hard to mitigate.

## 6.2 Difference over time and between phenotypes

To fully understand the data stored in the genetic information system, different metrics need to be quantified. The second question posed is related to the differences between different phenotypes and over time.

> *How much does the amount and quality of data, extracted from established genomic databases, differ over time and between phenotypes?*

This second sub-question explores two major dimensions, researched phenotype, and change over time. As can be seen in section 4 the first dimension is of big importance when assessing quality criteria. For some phenotypes, it can be problematic when selecting for high P-value and high Odds Ratio entries, while for other phenotypes it can be essential to weed out less interesting results. A final solution needs to take these (and additional) factors into account.

The second dimension, that of change over time, is a relevant factor. In earlier research, it was concluded that there is at least a substantial amount of change over time. We have shown the frequency of this change. We have also concluded that the stated update-cycle is not always followed by the GWAS Catalog. As long as no clear update schedule is communicated, a solution providing information to researchers and practitioners needs to check for updates frequently, and track which entries have changed. The fact that aside from additions to the database there were also a major amount of deletions (e.g. in the GWAS Catalog, Crohn's disease) means that variants are wrongly categorized at first, or filters were adjusted. Although the retrospective adjustment is good for the overall quality, it also shows the volatility of obtained results. This is something an automated solution needs to deal with.

## 6.3 Relevant factors contained in the databases

For an automated solution to work, we need to find factors that influence the quality of the extracted data, which in turn influences information quality and relevance to practitioners and researchers. Sub-question 3 addresses this need for knowledge:

*What are the factors that influence the quality of the extracted data?*

Of the tested factors the P-value stood out as being a very certain factor of importance, with a fairly high, negative, influence on the resulting value. This means that a low P-value coming from a research entry means the results of that study are more relevant. This is no surprising result in and of itself, but it is interesting when combined with the knowledge that the Odds Ratio and Risk Frequency each have a much lower influence on the relevancy of entries. This means that from the researched factors, P-value has the most influence. However, the Odds Ratio and Risk Frequency remain influential, which can also be seen in the classification trees that were trained. The algorithm chose to include these factors as the second, third, fourth, and fifth split. The population size, in turn, does not seem to have much significance in both models. A possible reason for this is that the population size is already reflected in the P-value and that this variable does not add much additional information.

There is a possibility that there are more influential factors stored in the papers entered in the database, or from different open sources. More research should be performed into extracting these factors in an automated way. Examples of this are; the technology the study was performed on, the background and previous research of the associated researchers, and the amount times the study was cited.

## 6.4 Which criteria are relevant?

We investigated hard criteria for the found factors with sub-question 4:

*Which criteria for variant selection are relevant?*

These hard criteria can best be found using the classification models. They have hard thresholds that represent the difference between relevant and irrelevant results. When looking at model C, it is easy to see that the most influential factor, the P-value has a criterion of $10^{-8.3} \approx 5 * 10^{-9}$. Values higher than that are excluded with high certainty. This is interesting because the P-value threshold for GWAS studies is often set at $5*10^{-8}$. This result can be interpreted in multiple ways. On one hand, the values are fairly similar, especially given that the result was obtained from a model build on a limited sub-set of the available data. One could also say that although a standard is used, research is scrutinized in a way that favors higher P-values, and thus does not follow the standard blindly. In model $C_2$ this distinction is not as clear as P-value is not the initial split here. However, also there the P-value contributes substantially, being the second chosen split.

The Odds Ratio is the initial split on model $C_2$. It sets a threshold of 1.1, entries smaller than this are not considered to be relevant results for most phenotypes. The Risk Frequency plays a large role in model C and a medium-large role in model $C_2$. It is hard to set a hard threshold on this value as it differs much per occurrence in the two trees. It can be seen as more of a conditional requirement under influence by the P-value and OR.

## 6.5 The development of an automated solution

The second part of this research, after the exploratory analysis, consists of the development of an automated solution:

> *How can these criteria be adapted to every context in an automated way to obtain the highest quality genomic data?*

Manual search through all entries for the phenotype researched is becoming less of an option as the amount of entries increases. The solution proposed mitigates this problem by creating a model on the database entries already considered relevant by established researchers and institutes. We created a data set of 2245 entries, spread across eight phenotypes, containing 98 entries marked as relevant. We built two base models using logistic regression and classification which were combined to create a more balanced final model. This final model performed well on Negative Predictive Value, which means it does not mark relevant results as irrelevant often. This is very important when trying to identify possible research opportunities. It performs slightly worse on Positive Predictive Value, which means that a slightly higher amount of irrelevant results are marked as relevant by the model. This is the preferable balance as it is better to perform limited manual filtering after automatic selection than losing valuable entries. Overall, the combined model performs with an $F_{score}$ of 85% given the right balance between model $B_2$ and $C_2$.

There are ways to improve these scores, mainly by extending the data set and tuning the models. We believe that this model is a good basis to use in a practical implementation. It will not replace the practitioners' expert knowledge but can greatly speed up the task of finding relevant research.

## 6.6 The future

We now have a clearer image of what data quality means in the genetics domain and when retrieving research entries. The data contained in the repositories need to be reduced and evaluated according to criteria found in this research project. It is possible to automate this process to a large extent, which allows for more frequent updating of existing knowledge. This will help practitioners and research make more appropriate decisions that are in line with up-to-date biological mechanisms and knowledge. It is also shown that each potential factor and criteria needs to be understood not only from a perspective of theoretical data quality but needs to be investigated from a statistical point of view to determine which criteria will be useful.

Although the proposed solution is grounded in literature, and by performing the exploratory analysis we gained a good understanding of the domain, the solution is not completely validated. An implementation in a practical environment is needed to prove the usefulness to end users. When talking about the improvement that automation can give there are usually two factors to consider; speed and quality. Improving the speed of a process by automation means that humans that operate as part of the project spend less time on menial tasks and can potentially spend more time on tasks that ask for a level of expertise. Some automation projects aim to provide a better quality in the intermediate or final product. Machines and computers can often perform a task more precise than human hands and minds. However, the more complicated a task becomes, the harder this is.

The implementation of automated retrieval methods like the one proposed here can improve the retrieval process on both factors. It already performs many times faster than humans ever could when it comes to learning what measures and entries are relevant. It can perform an analysis of a phenotype within the GWAS Catalog in mere seconds as opposed to weeks or months for humans. There is still room to improve in the quality department, when looking at the measures we evaluated. However, there are no performance measures of humans known. The algorithm might already outperform humans in its current implementation. When combined with a human performing a final check on the retrieved results, an overall improvement in the quality of retrieved entries is inevitable, while also maintaining an advantage of speed. Future research includes validation in a practical setting with both genetics researchers as well as practitioners in the genetics field.

The challenge for future research lies in experimentation with different kinds of models and learning algorithms, as well as tuning the ones proposed in this solution. There will also be plenty of opportunities in the field of natural language processing to extract more information from published research than currently contained within the public databases.

# Appendices

# Appendix A

# Checklist Observational Case Study

| | Step | Check |
|---|---|---|
| 1 | Knowledge goal(s) | What do you want to know?<br>Is this part of an implementation evaluation, a problem investigation, a survey of existing treatments, or a new technology validation? |
| 2 | Improvement goal(s) | If there is a higher-level engineering cycle, what is the goal of that cycle?<br>If this is a curiosity-driven project, are there credible application scenarios for the project results? |
| 3 | Current knowledge | State of the knowledge in published scientific, technical and professional literature?<br>Available expert knowledge?<br>Why is your research needed? Do you want to add anything, e.g. confirm or falsify something?<br>Theoretical framework that you will use? |
| 4 | Conceptual framework | Conceptual structures? Architectural structures, statistical structures?<br>Chance models of random variables: Semantics of variables?<br>Validity of the conceptual framework? Clarity of definitions, unambiguous application, avoidance of mono-operation and mono-method bias? |
| 5 | Knowledge questions | Open (exploratory) or closed (hypothesis-testing) questions?<br>Effect, satisfaction, trade-off or sensitivity questions?<br>Descriptive or explanatory questions? |

| | | |
|---|---|---|
| 6 | Population predicate | What is the architecture of the elements of the population? In which ways are all population elements similar to each other, and dissimilar to other elements?<br>Chance models of random variables: Assumptions about distributions of variables? |
| 7.1 | Acquisition of Objects of Study | How do you know that a selected entity is a case? How do you know it satisfies the population predicate?<br>Validity of OoS. |
| 7.2 | Construction of a sample | What is the analytical induction strategy? Confirming cases, disconfirming cases, extreme cases?<br>Validity of sampling procedure. |
| 9 | Measurement design | Variables and constructs to be measured? Scales, chance models.<br>Data sources? People (e.g. software engineers, maintainers, users, project managers, politically responsible persons, etc.), primary data (e.g. source code, log files, bug tracking data, version management data, email logs), primary documents (e.g., project reports, meeting minutes, organization charts, mission statements), etc.<br>Measurement instruments? Interview protocols, questionnaires, video recorders, sound recorders, clocks, sensors, database queries, log analyzers, etc.<br>What is the measurement schedule? Pretests, posttests? Cross-sectional or longitudinal?<br>How will measured data be stored and managed? Provenance, availability to other researchers?<br>Validity of measurement specification. |
| 10.1 | Descriptive inference design | How are words and images to be interpreted? (Content analysis, conversation analysis, discourse analysis, analysis software, etc.)<br>How are words and images to be interpreted? (Content analysis, conversation analysis, discourse analysis, analysis software, etc.)<br>Validity of description design. |
| 10.3 | Abductive inference design | What possible explanations can you foresee? What data do you need to give those explanations? What theoretical framework?<br>Internal validity. |
| 10.4 | Analogic inference design | What is the intended scope of your generalization?<br>External validity. |

| | | |
|---|---|---|
| 11 | What has happened? | What has happened during selection? Did the cases eventually selected have the architecture that was planned during research design? Have there been any unexpected events during the study? What has happened during analytical induction (i.e. sampling)? Could you study the kinds of cases that you originally planned? What has happened during measurements? Data sources actually used, response rates? |

Table A.1: Checklist when performing an observational case study by Wieringa [62].

# Appendix B

# Quality dimensions search level

| Dimension | Description |
|---|---|
| **Believability** | M1: The information stored in the database must be manually curated or reviewed by experts. This type of database has proved to be less error-prone than those which use algorithms to annotate the information.<br>M2: There are quality controls to ensure the correctness of the submitted information (e.g. submission forms, automated control of HGVS expressions, etc.). |
| **Relevancy** | M3: The database contains enough information and is useful to determine the required data, according to the attributes determined by the CSHG. |
| **Reputation** | M4: The database must be maintained or supported by international or well-known national research centers, institutions or associations. |
| **Currency** | M5: The database must be active and frequently updated as well as provide enough information about it; e.g. the date of the last update and the database version. |
| **Accessibility** | M6: The information must be public and freely accessible.<br>M7: The database must provide mechanisms to download the search results.<br>M8: It is highly recommended that the database provides ways to allow programmatic access to the information stored. |

Table B.1: The quality dimensions in the Search level as defined by León and Pastor [34].

# Bibliography

[1] Jaeger P. Ackerman, Daniel C. Bartos, Jamie D. Kapplinger, David J. Tester, Brian P. Delisle, and Michael J. Ackerman. The promise and peril of precision medicine: Phenotyping still matters most. *Mayo Clinic Proceedings*, 91(11):1606–1616, 2016.

[2] Russell L. Ackoff. From data to wisdom. *Journal of applied systems analysis*, 16(1):3–9, 1989.

[3] Sapeck Agrawal. Alzheimer's disease: Genes. *Material and Methods*, 7, 2017.

[4] ALS Association. Genetics. http://www.alsa.org/research/focus-areas/genetics, Retrieved 11-09-2019.

[5] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.

[6] Sam Behjati and Patrick S. Tarpey. What is next generation sequencing? *Archives of disease in childhood. Education and practice edition*, 98(6):236–238, 12 2013.

[7] Gene Bellinger, Durval Castro, and Anthony Mills. Data, information, knowledge, and wisdom, 2004. http://www.systems-thinking.org/dikw/dikw.htm, Retrieved 28-03-2019.

[8] Lindsey Bever. 'Damaged for the rest of my life': Woman says surgeons mistakenly removed her breasts and uterus., 10 2017. https://www.washingtonpost.com/news/to-your-health/wp/2017/10/24/damaged-for-the-rest-of-my-life-woman-says-surgeons-mistakenly-removed-her-breasts-and-uterus/, Retrieved 25-06-2019.

[9] Buniello, Annalisa, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47, 2019.

[10] Verónica Burriel Coll. Design and development of a genomic information system to manage breast cancer data. In *Sixth International Conference on Research Challenges in Information Science (RCIS)*, pages 1–6. IEEE, 2012.

[11] Verónica Burriel Coll and Óscar Pastor López. Conceptual schema of breast cancer: The background to design an efficient information system to manage data from diagnosis and treatment of breast cancer patients. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 432–435. IEEE, 2014.

[12] William S. Bush and Jason H. Moore. Chapter 11: Genome-wide association studies. *PLOS Computational Biology*, 8(12):1–11, 12 2012.

[13] Amit Chaudhary, Pramod Kumar Maurya, Birendra Singh Yadav, Swati Singh, and Ashutosh Mani. Current therapeutic targets for alzheimer's disease. *Journal of Biomedicine*, 3:74–84, 2018.

[14] Alastair Compston and Alasdair Coles. Multiple sclerosis. *The Lancet*, 372(9648):1502–1517, 2008.

[15] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 10 2005.

[16] Alessandro Didonna and Jorge R. Oksenberg. The genetics of multiple sclerosis. *Multiple Sclerosis: Perspectives in Treatment and Pathogenesis*, pages 3–16, Nov 2017.

[17] European Bioinformatics Institute (EMBL-EBI). What is genetic variation? https://www.ebi.ac.uk/training/online/course/human-genetic-variation-i-introduction/what-genetic-variation, Retrieved 28-05-2019.

[18] Greg Gibson. Rare and common variants: twenty arguments. *Nature reviews. Genetics*, 13(2):135–145, 01 2012.

[19] John H. Gillespie. *Population Genetics: A Concise Guide*, chapter 1, pages 9–11. Johns Hopkins University Press, 1998.

[20] Christopher G. Goetz, Werner Poewe, Olivier Rascol, Cristina Sampaio, Glenn T. Stebbins, Carl Counsell, Nir Giladi, Robert G. Holloway, Charity G. Moore, Gregor K. Wenning, and et al. Movement Disorder Society Task Force Report on the Hoehn and Yahr staging scale: Status and recommendations. *Movement Disorders*, 19(9):1020–1028, 2004.

[21] Anthony J. F. Griffiths, Jeffrey J. H. Miller, and David T. Suzuki. *An Introduction to Genetic Analysis.* W. H. Freeman, 7 edition, 2000.

[22] Esther V. Hobson and Christopher J. McDermott. Supportive and symptomatic management of amyotrophic lateral sclerosis. *Nature Reviews Neurology*, 12(9):526–538, 2016.

[23] Heidi J. Imker. 25 years of molecular biology databases: A study of proliferation, impact, and maintenance. *bioRxiv*, 2018.

[24] National Humane Genome Institute. Deoxyribonucleic acid (DNA) fact sheet. https://www.genome.gov/about-genomics/fact-sheets/Deoxyribonucleic-Acid-Fact-Sheet, Retrieved 28-05-2019.

[25] National Humane Genome Research Institute. About breast cancer. https://www.genome.gov/Genetic-Disorders/Breast-Cancer, Retrieved 11-09-2019.

[26] National Humane Genome Research Institute. About colon cancer. https://www.genome.gov/Genetic-Disorders/Colon-Cancer, Retrieved 11-09-2019.

[27] National Humane Genome Research Institute. About Parkinson's disease. https://www.genome.gov/Genetic-Disorders/Parkinsons-Disease, Retrieved 11-09-2019.

[28] Joseph M. Juran and A. Blanton Godfrey. *Juran's Quality Handbook, Fifth Edition.* McGraw-Hill New York, 1999.

[29] Beverly K. Kahn, Diane M. Strong, and Richard Y. Wang. Information quality benchmarks: product and service performance. *Communications of the ACM*, 45(4):184–192, 2002.

[30] Celeste M. Karch and Alison M. Goate. Alzheimer's disease risk genes and mechanisms of disease pathogenesis. *Biological psychiatry*, 77(1):43–51, 1 2015.

[31] Ana León Palacio, Ignacio Pascual Fernández, and Óscar Pastor López. Genomic information systems applied to precision medicine: Genomic data management for Alzheimer's disease treatment. In *Designing Digitalization (ISD2018 Proceedings)*, 2018.

[32] Ana León Palacio, Alicia García Giménez, Juan Carlos Casamayor Ródenas, and José Fabián Reyes Román. Genomic data management in big data environments: The colorectal cancer case. In *International Conference on Conceptual Modeling*, pages 319–329. Springer, 2018.

[33] Ana León Palacio and Óscar Pastor López. From big data to smart data: a genomic information systems perspective. In *2018 12th International Conference on Research Challenges in Information Science (RCIS)*, pages 1–11. IEEE, 2018.

[34] Ana León Palacio and Óscar Pastor López. Smart data for genomic information systems: the SILE method. *Complex Systems Informatics and Modeling Quarterly*, pages 1–23, 12 2018.

[35] Ana León Palacio, Óscar Pastor López, and Juan Carlos Casamayor Ródenas. A method to identify relevant genome data: conceptual modeling for the medicine of precision. In *International Conference on Conceptual Modeling*, pages 597–609. Springer, 2018.

[36] Ana León Palacio, José Fabián Reyes Román, Verónica Burriel Coll, and Francisco Valverde Giromé. Data quality problems when integrating genomic information. In *International Conference on Conceptual Modeling*, pages 173–182. Springer, 2016.

[37] Felix Naumann. *Quality-driven query answering for integrated information systems*, volume 2261. Springer, 2003.

[38] National Institute of Arthritis, Musculoskeletal, and Skin Diseases. Ankylosing spondylitis causes and treatment, Jun 2016. https://www.niams.nih.gov/health-topics/ankylosing-spondylitis/advanced, Retrieved 20-09-2019.

[39] U.S. National Library of Medicine. Ankylosing spondylitis. https://ghr.nlm.nih.gov/condition/ankylosing-spondylitis, Retrieved 11-09-2019.

[40] U.S. National Library of Medicine. Breast cancer. https://ghr.nlm.nih.gov/condition/breast-cancer, Retrieved 11-09-2019.

[41] U.S. National Library of Medicine. Multiple sclerosis. https://ghr.nlm.nih.gov/condition/multiple-sclerosis, Retrieved 11-09-2019.

[42] U.S. National Library of Medicine. Parkinson disease. https://ghr.nlm.nih.gov/condition/parkinson-disease, Retrieved 11-09-2019.

[43] U.S. National Library of Medicine. Prostate cancer. https://ghr.nlm.nih.gov/condition/prostate-cancer, Retrieved 11-09-2019.

[44] National Institute of Neurological Disorders and Stroke. Multiple sclerosis information page, Aug 2019. https://www.ninds.nih.gov/Disorders/All-Disorders/Multiple-Sclerosis-Information-Page, Retrieved 20-09-2019.

[45] Ana León Palacio and Óscar Pastor López. Infoxication in the genomic data era and implications in the development of information systems. In *IEEE 13th International Conference on Research Challenges in Information Science*, 2019.

[46] Óscar Pastor López, Ana León Palacio, José Fabián Reyes Román, and Juan Carlos Casamayor. Modeling life: a conceptual schema-centric approach to understand the genome. In *Conceptual Modeling Perspectives*, pages 25–40. Springer, 2017.

[47] Óscar Pastor López, José Fabián Reyes Román, and Francisco Valverde Giromé. Conceptual schema of the human genome (CSHG). Technical report, ProS Research Centre, 2016.

[48] Thomas C. Redman and A. Blanton Godfrey. *Data quality for the information age.* Artech House, Inc., 1997.

[49] José Fabián Reyes Román, Ana León Palacio, and Oscar Pastor López. Software engineering and genomics: The two sides of the same coin? In *ENASE*, pages 301–307, 2017.

[50] José Fabián Reyes Román, Óscar Pastor López, Juan Carlos Casamayor Ródenas, and Francisco Valverde Giromé. Applying conceptual modeling to better understand the human genome. In *International Conference on Conceptual Modeling*, pages 404–412. Springer, 2016.

[51] Sue Richards et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in medicine : official journal of the American College of Medical Genetics*, 17(5):405–424, 5 2015.

[52] Reid J. Robison. How big is the human genome? https://medium.com/precision-medicine/how-big-is-the-human-genome-e90caa3409b0, Retrieved 28-05-2019.

[53] National Health Service. Breast cancer in women. https://www.nhs.uk/conditions/breast-cancer, Retrieved 11-09-2019.

[54] Judith A. Smith. Update on ankylosing spondylitis: Current concepts in pathogenesis. *Current Allergy and Asthma Reports*, 15(1), 2014.

[55] Mayo Clinic Staff. Alzheimer's genes: Are you at risk? https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers-genes/art-20046552, Retrieved 11-09-2019.

[56] Mayo Clinic Staff. Breast cancer. https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470, Retrieved 11-09-2019.

[57] Tanya M Teslovich et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466:707–713, 8 2010.

[58] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 10 2015.

[59] E. Van Gijn, Marielle et al. New workflow for classification of genetic variants' pathogenicity applied to hereditary recurrent fevers by the International Study Group for Systemic Autoinflammatory Diseases (INSAID). *Journal of medical genetics*, 55(8):530, 8 2018.

[60] Yair Wand and Richard Y. Wang. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11):86–96, 1996.

[61] Richard Y. Wang and Diane M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.

[62] Roel J. Wieringa. *Design science methodology for information systems and software engineering*. Springer, 2014.

[63] Christopher P. Wild and Bernard W. Stewart. *World cancer report 2014*. World Health Organization, 2014. pages: 17-19.

[64] Christopher P. Wild and Bernard W. Stewart. *World cancer report 2014*. World Health Organization, 2014. page: 395.

[65] Loic Yengo et al. Meta-analysis of genome-wide association studies for height and body mass index in  700,000 individuals of european ancestry. *bioRxiv*, 2018.

[66] Zerbino, Daniel R. et al. Ensembl 2018. *Nucleic Acids Research*, 46, 2018.