CONTENTS

Preface

This book is an outcome of the “Integrated Applied Genetics Training – AppGenEdu” Project which ran between September 2018 and August 2021. The project was financed through the EEA Grants 2014-2021 mechanism, Education, Scholarships, Apprenticeships and Youth Entrepreneurship Programme in Romania, Cooperation Projects in The Higher Education Area with the ID: EY-COP-0029. The Education, Scholarships, Apprenticeships and Youth Entrepreneurship Programme (ESAYEP), the most extensive programme dedicated to Romanian education within the EEA Grants, is implemented by the ANPCDEFP between 2017 and 2024.

The project team consisted of members of "Carol Davila" University of Medicine and Pharmacy, which was the coordinator, and the University of Reykjavik and Exigia Medical, which were the partners.

The project has been initiated acknowledging that in Romania, at the time of the writing, there was an insufficient number of physicians that use medical genetic factors in their practice, creating a discrepancy between the number of available genetic epidemiologists and the demand for genetic data usage. One of the contributing factors was that even the elements of the curricula for applied genetic and epidemiology studies existed in the current educational offer of "Carol Davila" University, there was no course that shows how these elements can be tied together. As a consequence, there are still very few local specialists from the academic world with experience in the applied genetics field.

The project aimed to bring a new perspective of study, in both initial and continuous training, by unifying the existing expertise in a single educational frame. We hope we contributed to the effort to give a formal direction to integrated studies that facilitates the appearance of a personalized or precision-medicine approach in Romania.

The project developed and delivered an integrated curriculum bringing together the applied medical genetic disciplines, in order to provide students with the skills required to work in this field or to introduce applied genetic methods in their professional activities.

This book is based on the integrated curriculum that was developed for the purpose of the project by a team grouping together people from "Carol Davila" University of Medicine and Pharmacy, the University of Reykjavik and Exigia Medical.

AppGenEdu Team

Chapter 1: Epidemiology

1.1. History of Epidemiology

Epidemiology is a rather young discipline, though it does have tenuous antecedents far into the recorded past. The earliest attempts to quantify changes in the size and health of populations extend at least as far back as John Graunt who used the Bills of Mortality kept in England in the 17th century. It was in England in the mid-nineteenth century, however, that the process was first articulated, systematized, applied to a large population, and used to draw implications for health policy.

List of major Epidemiological milestones:

- HIPPOCRATES (400 BC): "On Airs, Waters, and Places" – Hypothesized that disease might be associated with the physical environment, including seasonal variation in illness.
- JOHN GRAUNT (1662): "Nature and Political Observations Made Upon the Bills of Mortality" – First to employ quantitative methods in describing population vital statistics.
- JOHN SNOW (1850): Formulated a natural epidemiological experiment to test the hypothesis that cholera was transmitted by contaminated water
- DOLL & HILL (1950): Used a case-control design to describe and test the association between smoking and lung cancer.
- FRANCES at al. (1950): Huge formal field trial of the Poliomyelitis vaccine in school children.
- DAWBER et al. (1955): Used the cohort design to study risk factors for cardiovascular disease in the Framingham Heart Study.

1.1.1 Successive Eras of Epidemiology

The modern history of epidemiology and public health begins in the mid-nineteenth century. The Public Health Act of 1848 is as good a landmark as any to mark its arrival. The ensuing history of epidemiology is better known than its early origins, and we will deal with it more briefly, focusing primarily on the relation of epidemiology to demography across successive eras. We define three main eras, each of which was characterized by a distinct causal paradigm: sanitary reform, infectious disease, and chronic disease

The era of sanitary reform in England spanned most of the second half of the nineteenth century: from the Public Health Act up to the dominance of infectious disease epidemiology. "The condition of England" had stabilized.

The threat of revolution receded, economic growth progressed to new heights, and the living conditions of working people began to improve.

The dominant causal paradigm was that of “miasma”. Miasma was seemingly a kind of vapour that emanated from decaying organic matter, and produced disease as it spread across the environment. Sanitary reformers believed that the extraordinary filth in the growing urban areas had produced the conditions for the spread of “miasma”. They viewed causation of disease on a broad ecological level, and similarly, the reforms advocated at this time were often on the same broad level.

William Farr was the central figure in this development. He remained in the General Register Office throughout the era. On the other hand, Edwin Chadwick, always a political figure and highly controversial, was deposed from power. John Simon, a man of very different background and character, was then selected to be the chief figure responsible for public health within the government. His study of mortality among the inpatients of asylums of England offers a remarkable example. In what was perhaps the first intimation of a systematic clinical follow-up study using a cohort design, he examined the mortality rate among patients admitted to the innovative Hanwell asylum, and compared it to the mortality rate among patients of other asylums.

In addition, Farr turned his attention to cause specific mortality. He was initially reluctant, but interestingly, having the encouragement of John Simon, he devised a classification of diseases for use in vital statistics. As he compared the mortality patterns of one disease with another, he moved away from describing the size and growth of the population, toward a disease specific epidemiology, a critical shift for our purposes.

This era was ushered in by the discovery of microorganisms and lasted up to the second World War. The dominant causal paradigm changed rapidly from miasma to germ theory. Over a short and dramatic period toward the end of the nineteenth century, microorganisms were discovered, and were established as causes of syphilis, diphtheria, cholera, and other epidemic diseases. The landmark event was the paper of Robert Edward Koch in 1882 on the tubercle bacillus, showing that the paradigm had the power to identify the causative agent for the most important disease of the newly industrialized countries.

With this shift in paradigm, the focus of epidemiologists tended to narrow still further, away from the mortality and fertility of the overall population, and toward the causes, consequences, and potential actions preventing and

treating specific diseases. Following the famous Henle-Koch postulates, epidemiologists sought to establish germs as necessary and sufficient causes of major diseases. Each disease had to be investigated separately, to discover the germ responsible for it. The paradigm for public health intervention, likewise, shifted from ecological measures such as the sanitary reform and improved living conditions, to methods designed specifically to interrupt the path of transmission of a certain microorganism in the population. The increasing focus on transmission patterns of specific diseases was accompanied by a decline in large scale epidemiologic analyses of secular trends, regional differences, and social differences in morbidity and mortality.

The separation of epidemiology from demographic questions reached its height during the succeeding era of chronic disease epidemiology, which roughly extended from World War II up to the end of the twentieth century. Chronic disease epidemiology arose in response to the alarming epidemics of cardiovascular disease, cancer, and peptic ulcer that had become evident in the industrialized countries by the end of World War II. The signal event in the transition was the demonstration that cigarette smoking caused lung cancer, a discovery that depended upon the multiple cause paradigm of chronic disease epidemiology and that could not have been made using the methods of infectious disease epidemiology.

The causal paradigm of chronic disease epidemiology – perhaps more appropriately termed risk factor epidemiology – was the “web of causation”. Under this paradigm, a disease has many causes, each of which may increase the risk of disease but may be neither necessary nor sufficient for the occurrence of the disease. Under this paradigm, the logical approach for epidemiologists is to seek to identify risk factors- exposures or characteristics that confer increased risk- for disease, rather than to look for a one to one relationship between cause and disease. The logical approach for public health intervention is to alter the risk profile of individuals within the population.

The efforts of a risk factor epidemiologist are most often directed to learning why some individuals are at higher risk than others individuals within a given population. The size and growth of the population itself are taken as a given, as a background context, fixed at least for the purposes of the analysis. Thus, in a given study, the dynamic interplay of population change with health, and the comparison of health across populations of different composition, do not usually enter the picture.

Chronic disease epidemiology was continually refined over the next 50 years, and as the risk factor methods became fully established, demography gradually disappeared from epidemiology textbooks and training. In the chronic disease era, epidemiologists were very much focused on the individual level of causation, more so than they were in the era of sanitary reform or infectious disease epidemiology. As demographic questions are first and foremost on the population level, they are not easily incorporated by chronic disease epidemiologists.

1.1.2. The Emergent Era

In the present time, epidemiology is in transition from the chronic disease era. It is still in flux so it is uncertain what the outcome will be. "*Each generation receives its particular impression of epidemic diseases*", Flexner wrote in 1922, reflecting on the devastating influenza pandemic. Although the temporal repertoire of epidemiology is not limitless, its capacity to mix and match, or run alongside, a variety of time frames is not yet exhausted—as developments in planetary health attest. Whereas we talk freely of the eras of epidemiology, conferring on the field a kind of historicity, rarely have we considered carefully and critically the various temporalities implicit in different styles of epidemiological investigation, the canon of chronological technique. One of the main questions highlighted by Warwick Anderson is "*not what is the history of epidemiology, but rather what is the history in epidemiology?"*

1.2. Epidemiology: Basic Concepts

Epidemiology is the study of the distribution and determinants of health-related states or events (including disease), and the application of this study to the control of diseases and other health problems. Various methods can be used to carry out epidemiological investigations: surveillance and descriptive studies can be used to study distribution; analytical studies are used to study determinants.

The objectives of epidemiology include the following:

- to identify the aetiology or cause of disease
- to determine the extent of disease
- to study the progression of disease
- to evaluate preventive and therapeutic measures for a disease or condition
- to develop public health policies.

Epidemiology has traditionally been used to model the spreading of diseases in populations at risk. By applying parameters related to agents'

responses to infection and network of contacts it helps to study how diseases occur, why they spread and how one could prevent epidemic outbreaks. For decades, epidemiology has studied also non-communicable diseases, such as cancer, cardiovascular disease, addictions and accidents. Descriptive epidemiology focuses on providing accurate information on the occurrence (incidence, prevalence and survival) of the condition. Etiological epidemiology seeks to identify the determinants be they infectious agents, environmental or social exposures, or genetic variants. A central goal is to identify determinants amenable to intervention, and hence prevention of disease.

The principles of epidemiology include study designs, testing research hypotheses, exploring possible causal associations between risk factors (i.e., exposures) and outcomes, and addressing disease control and prevention. It plays a fundamental role in medicine and public health.

An epidemiologist determines *What, When, Where, Who, and Why*. The epidemiologist will describe these concepts in the following terms: case definition, time, place, person, and causes.

1.2.1. Case Definition

A case definition is a set of standard criteria for deciding whether a person has a particular disease or other health-related condition. By using a standard case definition we ensure that every case is diagnosed in the same way, regardless of when or where it occurred, or who identified it. We can then compare the number of cases of the disease that occurred in one time or place with the number that occurred at another time or another place.

A case definition may have several sets of criteria, depending on how certain the diagnosis is. For example, during an outbreak of measles, we might classify a person with a fever and rash as having a suspect, probable, or confirmed case of measles, depending on what additional evidence of measles was present. In other situations, we temporarily classify a case as suspect or probable until laboratory results are available. When we receive the laboratory report, we then reclassify the case as either confirmed or “not a case,” depending on the lab results. In the midst of a large outbreak of a disease caused by a known agent, we may permanently classify some cases as suspect or probable, because it is unnecessary and wasteful to run laboratory tests on every patient with a consistent clinical picture and a history of exposure (e.g., chickenpox). Case definitions should not rely on laboratory culture results alone, since organisms are sometimes present without causing disease.

1.2.2. Numbers and Rates

A basic task of a health department is counting cases in order to measure and describe morbidity. When physicians diagnose a case of a reportable disease they send a report of the case to their local health department. These reports are legally required to contain information on time (when the case occurred), place (where the patient lived), and person (the age, race, and sex of the patient). The health department combines the reports and summarizes the information by time, place, and person. From these summaries, the health department determines the extent and patterns of disease occurrence in the area and identifies clusters or outbreaks of disease.

A simple count of cases, however, does not provide all the information a health department needs. To compare the occurrence of a disease at different locations or during different times, a health department converts the case counts into rates, which relate the number of cases to the size of the population where they occurred.

Rates are useful in many ways. With rates, the health department can identify groups in the community with an elevated risk of disease. These so-called high-risk groups can be further assessed and targeted for special intervention; the groups can be studied to identify risk factors that are related to the occurrence of disease. Individuals can use knowledge of these risk factors to guide their decisions about behaviours that influence health.

1.2.3. Descriptive Epidemiology

In descriptive epidemiology, we organize and summarize data according to *time, place, and person.* These three characteristics are sometimes called the epidemiologic variables.

Compiling and analysing data by time, place, and person is desirable for several reasons. First, the investigator becomes intimately familiar with the data and with the extent of the public health problem being investigated. Second, this provides a detailed description of the health of a population that is easily communicated. Third, such analysis identifies the populations that are at the greatest risk of acquiring a particular disease. This information provides important clues to the causes of the disease and these clues can be turned into testable hypotheses.

1.2.4. Analytic Epidemiology

Descriptive epidemiology can identify several characteristics of persons with disease, and we may question whether these features are really unusual, but

descriptive epidemiology does not answer that question. Analytic epidemiology provides a way to find the answer: the comparison group. Comparison groups, which provide baseline data, are a key feature of analytic epidemiology.

Analytic epidemiologic studies measure the association between a particular exposure and a disease, using information collected from individuals, rather than from the aggregate population. Exposure is defined broadly to include behavioural factors such as smoking or diet, environmental pollutants such as asbestos, personal characteristics such as obesity or tendency to sunburn, anthropometric measurements such as body mass index, and genetic traits and other measurable biological factors that may affect cancer.

A simple analytical epidemiological study is usually represented by a traditional cohort study design.. Another group of traditional study designs that belongs to analytical epidemiology are case control studies. Other less traditional analytical study designs include case-case studies and case-crossover design. In each of these analytical studies, observations in one group in the population are compared to another group (also called 'reference group'). Choosing the appropriate reference group is one of the challenging aspects of analytical epidemiology.

1.3. Epidemiological Studies

Epidemiological studies are categorized as either descriptive or analytic. These are the three most common types of analytic epidemiological studies:

- Cohort Study
- Case Control Study
- Cross-Sectional Study

It should be also emphasized that all epidemiological studies are (or should be) based on a particular population (the ‘source population’) followed over a particular period of time (the ‘risk period’). Within this framework, the most fundamental distinction is between studies of disease ‘incidence’ and studies of disease ‘prevalence’. Once this distinction has been drawn, then the different epidemiological study designs differ primarily in the manner in which information is drawn from the source population and risk period

1.3.1. Cohort Studies

In the classic cohort study, the investigator defines two or more groups of people that are free of disease and that differ according to the extent of their exposure to a potential cause of the disease. These groups are referred to as

the study cohorts (from the Latin word for one of the ten divisions of a Roman legion). In such studies, there is at least one cohort thought of as the exposed cohort-those individuals who have experienced the putative causal event or condition-and another cohort thought of as the unexposed, or reference cohort. There may be more than just two cohorts, but each cohort would represent a group with a different level or type of exposure. For example, an occupational cohort study of chemical workers might comprise cohorts of workers who work in different departments of the plant, with each cohort being exposed to a different set of chemicals. The investigator measures and compares the incidence rate of the disease in each of the study cohorts.

Many cohort studies begin with a single cohort that is heterogeneous with respect to exposure history. Comparisons of disease experience are made within the cohort across subgroups defined by one or more exposures. Examples include studies of cohorts defined from membership lists of administrative or social units, such as cohorts of doctors or nurses, or cohorts defined from employment records, such as cohorts of factory workers.

1.3.2. Case-Control

Case-control studies are best understood by defining a source population, which represents a hypothetical study population in which a cohort study might have been conducted. If a cohort study were undertaken, the primary tasks would be to identify the exposed and unexposed denominator experience, measured in person-time units of experience or as the number of people in each study cohort, and then to identify the number of cases occurring in each person-time category or study cohort. In a case-control study, the cases are identified and their exposure status is determined just as in a cohort study, but denominators from which rates could be calculated are not measured. Instead, a control group of study subjects is sampled from the entire source population that gives rise to the cases. The purpose of the control group is to determine the relative (as opposed to absolute) size of the exposed and unexposed denominators within the source population. From the relative size of the denominators, the relative size of the incidence rates (or incidence proportions, depending on the nature of the data) can be estimated. Thus, case-control studies yield estimates of relative effect measures. Because the control group is used to estimate the distribution of exposure in the source population, the cardinal requirement of control selection is that the controls must be sampled independently of their exposure status. In sum, case-control studies of incident cases differ from

cohort studies according to how subjects are initially selected. A cohort study identifies and follows a population or populations to observe disease experience; a case-control study involves an additional step of selecting cases and controls from this population.

1.3.3. Cross-Sectional Studies

A study that includes as subjects all persons in the population at the time of ascertainment or a representative sample of all such persons, including those who have the disease, and that has an objective limited to describing the population at that time, is usually referred to as a cross-sectional study. A cross-sectional study conducted to estimate prevalence is called a prevalence study. Usually, the exposure information is ascertained simultaneously with the disease information, so that different exposure subpopulations may be compared with respect to their disease prevalence. Cross-sectional studies need not have etiologic objectives. For example, delivery of health services often requires knowledge only of how many items will be needed (such as number of hospital beds), without reference to the causes of the disease. Nevertheless, prevalence data are so often used for etiologic inferences that a thorough understanding of their limitations is essential.

Cross-sectional studies often deal with exposures that cannot change, such as blood type or other invariable personal characteristics. For such exposures, current information is as useful as any. For variable exposures, however, current information is less desirable than etiologically more relevant information from before the case occurred. In a study of the etiology of respiratory cancer that compares smoking information on cases and non-cases, the current smoking habits of subjects are not nearly as relevant as their smoking histories before the cancer developed. The cross-sectional approach to such a question could well be viewed as a case-control study with an excessively large control group (because few people in a population would have respiratory cancer), with smoking information from an inappropriate time period, and with biased case ascertainment (short-duration cases are much less ‘likely to be seen than long-duration cases). Of course, the time-period problem could be addressed by asking subjects about their smoking history, rather than about current smoking.

1.3.4. Prospective Versus Retrospective Studies

Studies can be classified further as either prospective or retrospective. We define a prospective study as one in which exposure and covariate

measurements are made before the cases of illness occur. In a retrospective study these measurements are made after the cases have already occurred. The distinction between the classification as cohort or case-control and prospective or retrospective should be firmly drawn, because these two axes for classifying epidemiologic studies have often been confused: early writers referred to cohort studies as prospective studies and to case-control studies as retrospective studies because cohort studies usually begin with identification of the exposure status and then measure disease occurrence, whereas case-control studies usually begin by identifying cases and controls and then measure exposure status. The terms prospective and retrospective, however, are more usefully employed to describe the timing of disease occurrence with respect to exposure measurement. For example, case-control studies can be either prospective or retrospective. A prospective case-control study uses exposure measurements taken before disease, whereas a retrospective case-control study uses measurements taken after disease. Both cohort and case-control studies may employ a mixture of prospective and retrospective measurements; using data collected before and after disease occurred. The prospective/retrospective distinction is sometimes used to refer to the timing of subject identification, rather than measurement of exposure and covariates. With this usage, a retrospective (or historical) cohort study involves the identification and follow-up of subjects, but the subjects are identified only after the follow-up period under study has ended. The identification of the subjects, their exposure, and their outcome must be based on existing records or memories.

Chapter 2: Genetics

2.1. Introduction to medical genetics

2.1.1. Cell cycle

A cell cycle is a succession of an interphase and a cell division (see figure 2.1 adapted from [1])

The interphase is the "preparation for the cell division" phase in which the cell grows in size and doubles its DNA material.

2.1.2. Cell division – Meiosis

Before entering meiosis I, a cell must first go through interphase.

As in mitosis, the cell grows during G1 phase, copies all of its chromosomes during S phase, and prepares for division during the G2 phase.

Each cell division consists of 4 different phases:

1. Prophase
2. Metaphase
3. Anaphase
4. Telophase + Cytokinese

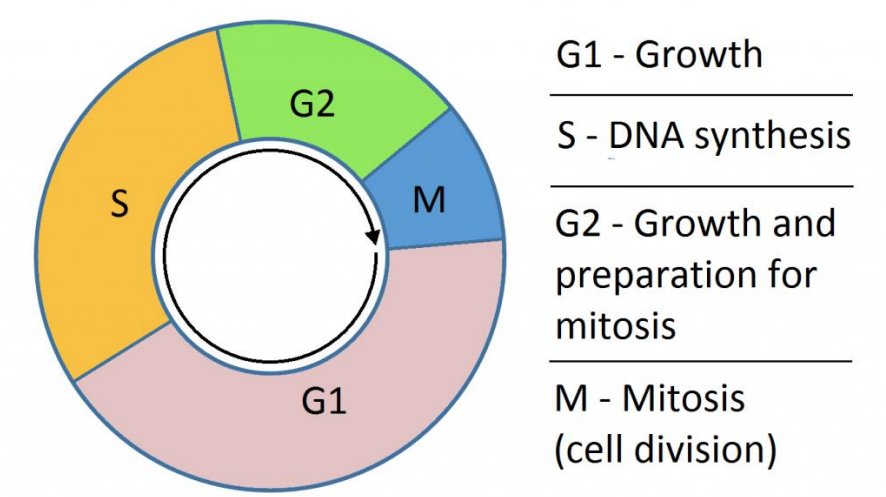


Fig. 2.1 - Cell cycle

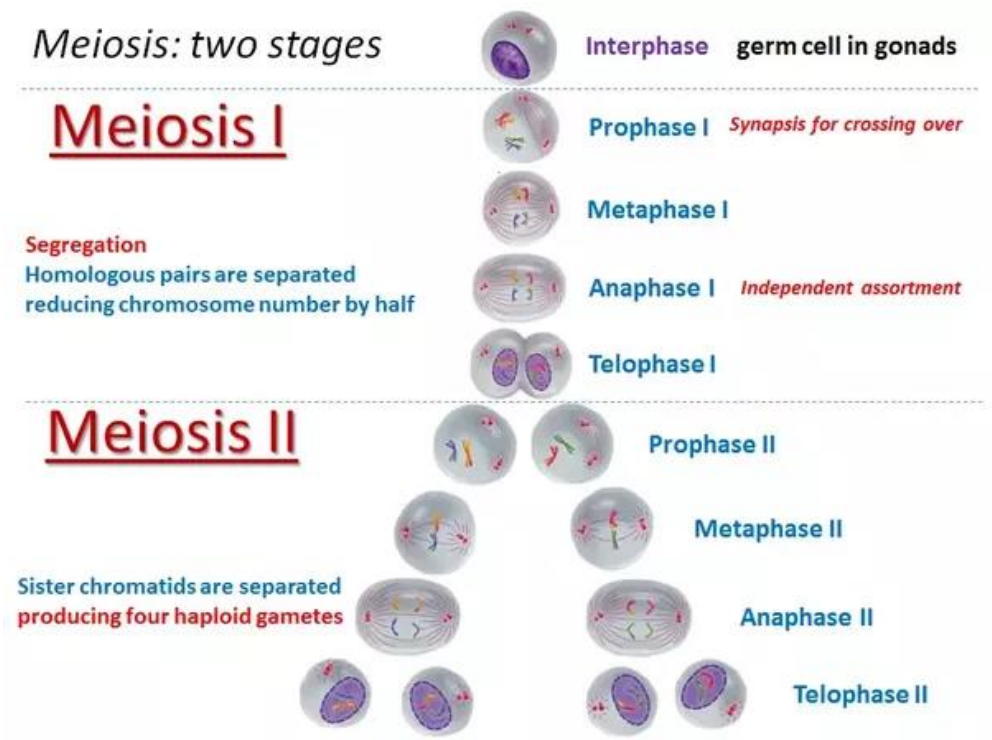


Fig. 2.2 – Meiosis
(https://socratic.org/questions/589fd775b72cff178533a701)

As meiosis has 2 included divisions (meiosis I - reductional - and meiosis II - equational), the whole division will contain 8 phases: Prophase I, Metaphase I, Anaphase I, Telophase I and Prophase II, Metaphase II, Anaphase II and Telophase II.

Although each cell division needs a preceding interphase process, between meiosis I and meiosis II no interphase will happen as the splitting of the sister chromatids will not occur until anaphase II (figure 2.2) [2].

The phases of Meiosis I are:

1. **Prophase I:** The starting cell is diploid, 2n = 4. Homologous chromosomes pair up and exchange fragments in the process of crossing over.
   During prophase I, as in mitosis, the chromatin strings condense for becoming chromosomes, but in meiosis I, they also pair up. Each chromosome is connected through synapses with its homologue so that the two pair up at exactly the same locus along their full length. Crossing-over (intra-chromosomal recombination = equal exchange of genetic material between nonsister chromatids of the homologous

chromosomes) occurs also during prophase I, being one of the most important biological processes for human variability.

2. **Metaphase I:** Homologue pairs line up at the metaphase plate.
   When the homologous pairs line up at the metaphase plate, the orientation of each pair is random (inter-chromosomal recombination).

   This is another important process in variability, allowing for the formation of gametes with different sets of homologues.

3. **Anaphase I:** Homologues separate to opposite ends of the cell. Sister chromatids stay together.
   In anaphase I, the homologues are pulled apart and move apart to opposite ends of the cell. The sister chromatids of each chromosome, however, remain attached to one another and don't come apart

4. **Telophase I:** Newly forming cells are haploid, n = 2. Each chromosome still has two sister chromatids, but the chromatids of each chromosome are no longer identical to each other (fig. 2.3.)
   Finally, in telophase I, the chromosomes arrive at opposite poles of the cell. In some organisms, the nuclear membrane re-forms and the chromosomes decondense, although in others, this step is skipped—since cells will soon go through another round of division, meiosis II

   Cytokinesis usually occurs at the same time as telophase I, forming two haploid daughter cells.

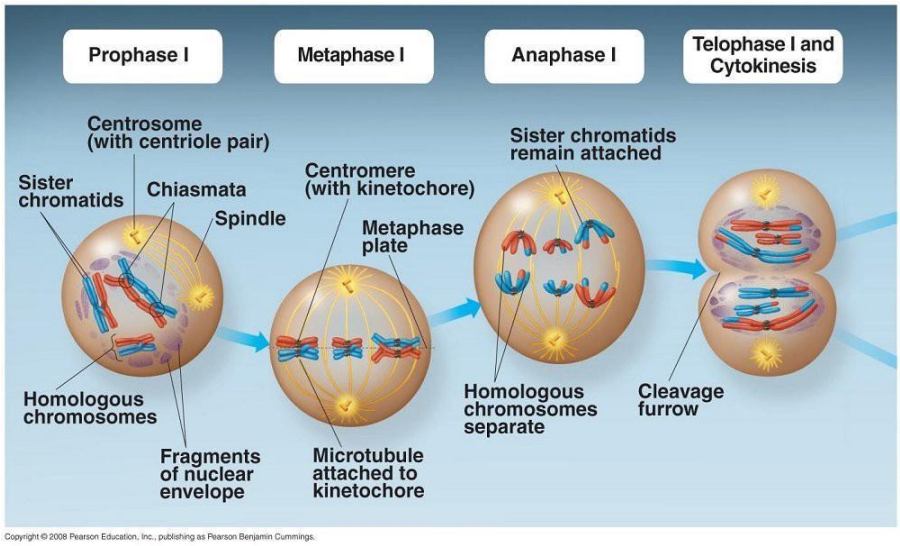


Fig. 2.3 - Meiosis I (adapted from [3])

Phases of meiosis II are:

1. **Prophase II:** Chromosomes recondense, migrate towards the equator of the cell, while the centrioles head towards the poles of the cell.
2. **Metaphase II:** Chromosomes are alligned at the metaphase plate with their centromeres on the equator of the cell. Spindle fibres are connected at the chromosomes` centromeres on each side.
3. **Anaphase II:** Chromatid disjunction - the sister chromatids are pulled apart towards the opposite poles of the cell by the spindle fibres, forming monochromatic chromosomes. During this phase, the cell has 46 monochromatic chromosomes (which actually were the 23 dichromatic chromosomes from metaphase).
4. **Telophase II + Cytokinesis:** The monochromatic chromosomes reach the poles of the cell; the cell membranes are formed, the cytoplasm is being split, the nuclei reform and the chromosomes decondense and become chromatin strings.

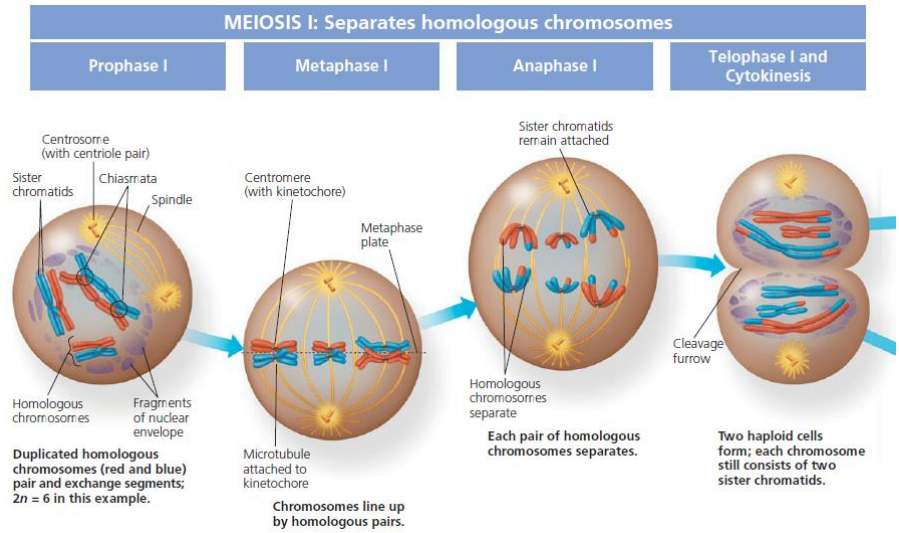


Fig 2.4. - Meiosis II [4]

At the end of meiosis II, 4 halploid cells with monochromatic chromosomes were obtained starting from one diploid cell with dichromatic chromosomes.

2.1.3. Spermatogenesis

The process of formation of sperms is called spermatogenesis. It occurs in the seminiferous tubules of the testes. The seminiferous tubules are lined by germinal epithelium. The germinal epithelium consists largely of cuboidal primary or primordial germ cells (PGCs) and contains certain tall somatic cells called Sertoli cells (= nurse cells). Spermatogenesis includes formation

of spermatids and formation of spermatozoa. At sexual maturity, the undifferentiated primordial germ cells divide several times by mitosis to produce a large number of spermatogonia (Gr. sperma = seeds, gonos-generation). Spermatogonia (2N) are of two types: type A spermatogonia and type B spermatogonia. Type A spermatogonia serve as the stem cells which divide to form additional spermatogonia. Type B spermatogonia are the precursors of sperms.Each type B spermatogonium actively grows to a larger primary spermatocyte by obtaining nourishment from the nursing cells.

Each primary spermatocyte undergoes two successive divisions, called maturation divisions. The first maturation division is reductional or meiotic. Hence, the primary spermatocyte divides into two haploid daughter cells called secondary spermatocytes. Both secondary spermatocytes now undergo second maturation division, which is an ordinary mitotic division to form, four haploid spermatids, by each primary spermatocyte.

Formation of Spermatozoa from Spermatids (Spermatogenesis)

The transformation of spermatids into spermatozoa is called spermiogenesis or spermateliosis. The spermatozoa are later on known as sperms. Thus four sperms are formed from one spermatogonium. After spermiogenesis sperm heads become embedded in the Sertoli cells and are finally released from the seminiferous tubules by the process called spermiation.

Hormonal Control of Spermatogenesis

Spermatogenesis is initiated due to increase in gonadotropin-releasing hormone (GnRH) by the hypothalamus. GnRH acts on the anterior lobe of pituitary gland to secrete luteinizing hormone (LH) and follicle stimulating hormone (FSH). LH acts on the Leydig's cells of the testes to secrete testosterone.

FSH acts on Sertoli cells of the seminiferous tubules of the testes to secrete an androgen binding protein (ABP) and inhibin. ABP concentrates testosterone in the seminiferous tubules. Inhibin suppresses FSH synthesis. FSH acts on spermatogonia to stimulate sperm production.

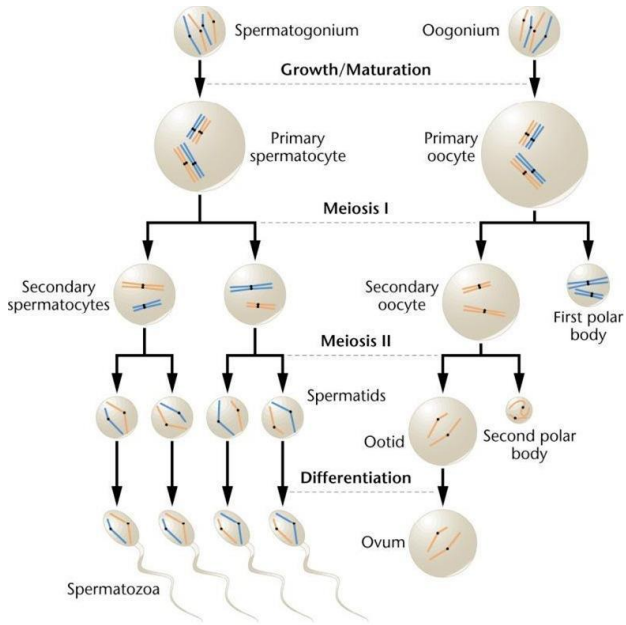


Fig 2.5 - Spermatogenesis vs. oogenesis [5]

2.1.4. Oogenesis

The process of formation of a mature female gamete (ovum) is called oogenesis. It occurs in the ovaries (female gonads). It consists of three phases: multiplication, growth and maturation. In the foetal development, certain cells in the germinal epithelium of the ovary of the foetus are larger than others. These cells divide by mitosis, producing a couple of million egg mother cells or oogonia in each ovary of the foetus. No more oogonia are formed or added after birth. The oogonia multiply by mitotic divisions forming the primary oocytes. This phase of the primary oocyte is very long. It may extend over many years. The oogonium grows into large primary oocytes. Each primary oocyte then gets surrounded by a layer of granulosa cells to form the primary follicle. A large number of these follicles degenerate during the period from birth to puberty. So at puberty only 60,000 - 80,000 primary follicles are left in each ovary. The fluid filled cavity of the follicle is called antrum. Like a primary spermatocyte, each primary oocyte undergoes two maturation divisions, first meiotic and the second meiotic. The results of maturation divisions in oogenesis are, however, very different from those in spermatogenesis. In the first meiotic division, the primary oocyte divides into two very unequal haploid daughter cells— a large secondary oocyte and a very small first polar body or polocyte.

In the second maturation division, the first polar body may divide to form two second polar bodies. The secondary oocyte again divides into unequal daughter cells, a large ootid and a very small second polar body. The ootid grows into a functional haploid ovum. Thus from one oogonium, one ovum and three polar bodies are formed. The ovum is the actual female gamete. The polar bodies take no part in reproduction and, hence, soon degenerate.

In human beings, the ovum is released from the ovary in the secondary oocyte stage. The maturation of the secondary oocyte is completed in the mother's oviduct (Fallopian tube) usually after the sperm has entered the secondary oocyte for fertilization.

In humans (and most vertebrates), the first polar body does not undergo meiosis II, whereas the secondary oocyte proceeds as far as the metaphase stage of meiosis II. However, it then stops advancing any further; it awaits the arrival of sperm for completion of meiosis II.

Entry of the sperm restarts the cell cycle breaking down MPF (M-phase promoting factor) and turning on APC (Anaphase promoting complex). Completion of meiosis II converts the secondary oocyte into a fertilized ovum (egg) or zygote (and also a second polar body).

Hormonal Control of Oogenesis

GnRH secreted by the hypothalamus stimulates the anterior lobe of pituitary gland to secrete LH and FSH. FSH stimulates the growth of Graafian follicles and also the development of the egg/oocyte within the follicle to complete the meiosis I to form a secondary oocyte. FSH also stimulates the formation of oestrogens. LH induces the rupture of the mature Graafian follicle and thereby the release of the secondary oocyte. Thus LH causes ovulation. In brief ovulation in human beings may be defined as the release of the secondary oocyte from the Graafian follicle. The remaining part of the Graafian follicle is stimulated by LH to develop into corpus luteum ("yellow body"). The rising level of progesterone inhibits the release of GnRH, which in turn, inhibits production of FSH, LH and progesterone.

2.2. Chromosomes

Chromosomes are thread-like structures located inside the nucleus of animal and plant cells. Each chromosome is made of protein and a single molecule of deoxyribonucleic acid (DNA). Passed from parents to offspring, DNA contains the specific instructions that make each type of living creature unique.

The term chromosome comes from the Greek words for colour (chroma) and body (soma). Scientists gave this name to chromosomes because they are cell structures, or bodies, that are strongly stained by some colourful dyes used in research.

The unique structure of chromosomes keeps DNA tightly wrapped around spool-like proteins, called histones. Without such packaging, DNA molecules would be too long to fit inside cells. For example, if all of the DNA molecules in a single human cell were unwound from their histones and placed end-to-end, they would stretch 6 feet.

For an organism to grow and function properly, cells must constantly divide to produce new cells to replace old, worn-out cells. During cell division, it is essential that DNA remains intact and evenly distributed among cells. Chromosomes are a key part of the process that ensures DNA is accurately copied and distributed in the vast majority of cell divisions. Still, mistakes do occur on rare occasions.

Changes in the number or structure of chromosomes in new cells may lead to serious problems. For example, in humans, one type of leukaemia and some other cancers are caused by defective chromosomes made up of joined pieces of broken chromosomes.

It is also crucial that reproductive cells, such as eggs and sperm, contain the right number of chromosomes and that those chromosomes have the correct structure. If not, the resulting offspring may fail to develop properly.

The constricted region of linear chromosomes is known as the centromere. Although this constriction is called the centromere, it usually is not located exactly in the centre of the chromosome and, in some cases, is located almost at the chromosome's end. The regions on either side of the centromere are referred to as the chromosome's arms.

Centromeres help to keep chromosomes properly aligned during the complex process of cell division. As chromosomes are copied in preparation for production of a new cell, the centromere serves as an attachment site for the two halves of each replicated chromosome, known as sister chromatids.

Telomeres are repetitive stretches of DNA located at the ends of linear chromosomes. They protect the ends of chromosomes in a manner similar to the way the tips of shoelaces keep them from unravelling.

In many types of cells, telomeres lose a bit of their DNA every time a cell divides. Eventually, when all of the telomere DNA is gone, the cell cannot replicate and dies. (figure 2.6 [6])

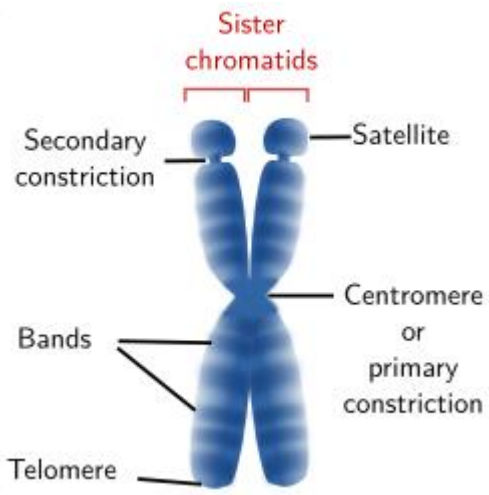


Fig. 2.6. - Chromosome morphology

White blood cells and other cell types with the capacity to divide very frequently have a special enzyme that prevents their chromosomes from losing their telomeres. Because they retain their telomeres, such cells generally live longer than other cells.

Telomeres also play a role in cancer. The chromosomes of malignant cells usually do not lose their telomeres, helping to fuel the uncontrolled growth that makes cancer so devastating.

Karyotyping is used for detecting numerical or large structural chromosomal abnormalities.

It is the testing method of choice for the diagnosis of aneuploidies (e.q. Down syndrome, Turner syndrome, and others).

2.2.1 Chromosomal abnormalities

Almost every cell in our body contains 23 pairs of chromosomes, for a total of 46 chromosomes. Half of the chromosomes come from our mother, and the other half comes from our father. The first 22 pairs are called autosomes. The 23rd pair consists of the sex chromosomes, X and Y. Females usually have two X chromosomes, and males usually have one X and one Y chromosome in each cell. All of the information that the body needs to grow and develop comes from the chromosomes. Each chromosome contains thousands of genes, which make proteins that direct the body's development, growth, and chemical reactions.

Fig. 2.7 - The human karyotype (46,XX) [7]

Many types of chromosomal abnormalities exist, but they can be categorized as either numerical or structural. Numerical abnormalities are whole chromosomes either missing from or extra to the normal pair. Structural abnormalities are when part of an individual chromosome is missing, extra, switched to another chromosome, or turned upside down.

Chromosomal abnormalities can occur as an accident when the egg or the sperm is formed or during the early developmental stages of the foetus. The age of the mother and certain environmental factors may play a role in the occurrence of genetic errors. Prenatal screening and testing can be performed to examine the chromosomes of the foetus and detect some, but not all, types of chromosomal abnormalities.

Chromosomal abnormalities can have many different effects, depending on the specific abnormality. For example, an extra copy of chromosome 21 causes Down syndrome (trisomy 21). Chromosomal abnormalities can also cause miscarriage, disease, or problems in growth or development.

Numerical chromosomal anomalies

The most common type of chromosomal abnormality is known as aneuploidy, an abnormal chromosome number due to an extra or missing chromosome. Most people with aneuploidy have trisomy (three copies of a chromosome) instead of monosomy (single copy of a chromosome). Down syndrome is probably the most well known example of a chromosomal aneuploidy. Besides trisomy 21, the major chromosomal aneuploidies seen in live-born babies are: trisomy 18; trisomy 13; 45, X (Turner syndrome); 47, XXY (Klinefelter syndrome); 47, XYY; and 47, XXX.

Structural chromosomal anomalies

Structural chromosomal abnormalities result from breakage and incorrect re-joining of chromosomal segments. A range of structural chromosomal abnormalities result in disease. Structural rearrangements are defined as balanced if the complete chromosomal set is still present, though rearranged, and unbalanced if information is additional or missing. Unbalanced rearrangements include deletions, duplications, or insertions of a chromosomal segment. Ring chromosomes can result when a chromosome undergoes two breaks and the broken ends fuse into a circular chromosome. An isochromosome can form when an arm of the chromosome is missing and the remaining arm duplicates.

Balanced rearrangements include inverted or translocated chromosomal regions. Since the full complement of DNA material is still present, balanced chromosomal rearrangements may go undetected because they may not result in disease. A disease can arise as a result of a balanced rearrangement if the breaks in the chromosomes occur in a gene, resulting in an absent or nonfunctional protein, or if the fusion of chromosomal segments results in a hybrid of two genes, producing a new protein product whose function is damaging to the cell.

A chromosome's structure can be altered in several ways:

1. Deletions: A portion of the chromosome is missing or deleted.
2. Duplications: A portion of the chromosome is duplicated, resulting in extra genetic material.
3. Translocations: A portion of one chromosome is transferred to another chromosome. There are two main types of translocation. In a reciprocal translocation, segments from two different chromosomes have been exchanged. In a Robertsonian translocation, an entire chromosome has attached to another at the centromere.

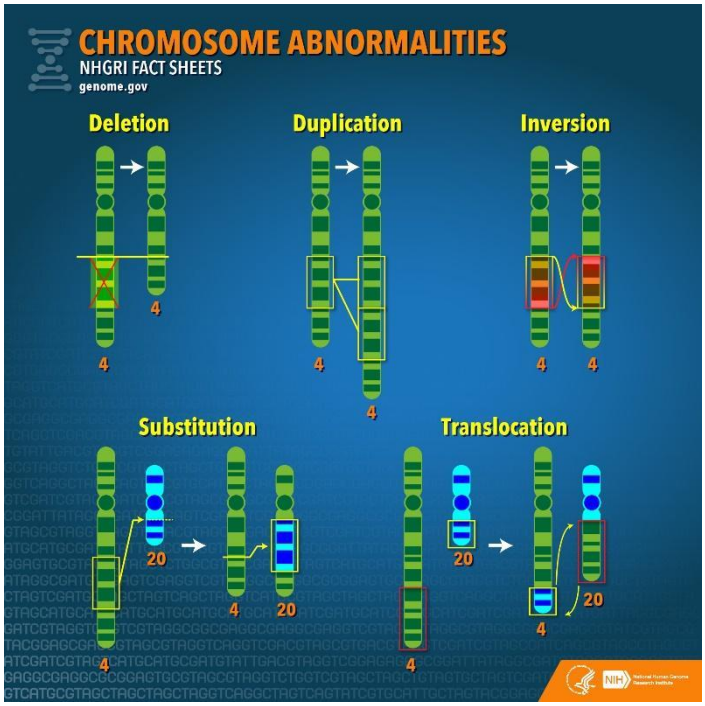


Fig. 2.8. - Chromosomal abnormalities [8]

4. Inversions: A portion of the chromosome has broken off, turned upside down, and reattached. As a result, the genetic material is inverted.
5. Rings: A portion of a chromosome has broken off and formed a circle or ring. This can happen with or without loss of genetic material.

The DNA structure of all chromosomes is a specific alternation of 2 different types of chromatin: euchromatin (loosely condensed chromatin, genetically active) and heterochromatin (tightly condensed chromatin, genetically inactive.

Euchromatin is rich in G-C pairs, while heterochromatin in A-T pairs, euchromatin consists of unique DNA sequences, while in heterochromatin repetitive and highly-repetitive DNA are more frequent.

Based on these differences, special chemical dyes which will connect only to specific elements of chromatin have been produced. Ones will only connect to the A-T DNA pairs (highlighting heterochromatin, as is the Giemsa dye used in G-banding), while other will aim at the G-C pais (R-banding) or the repetitive DNA sequences (Q-banding).

By using this method, chromosome bands are obtained which help geneticists in clearly selecting and pairing up the chromosomes. At the same

time, they are extremely important in observing and diagnosing structural chromosomal anomalies.

Chromosome bands are numbered, as are the different chromosomal regions, starting from the centromere on each arm towards the telomere.

The cytogenetic location (locus) is the address where genes can be found and is set strictly based on the chromosomal bands (e.g.: 4p16.3 means: 4 - chromosome number; q - chromosome arm; 1 - chromosome region; 6 - chromosome band; 3 - chromosome subband - each band consists of a number of subbands which differ in staining intensity)

2.3. Genes

Genes are the working subunits of DNA. Each gene contains a particular set of instructions, usually coding for a particular protein or for a particular function. A gene is a short piece of DNA, a unit of hereditary information that occupies a fixed position (locus) on a chromosome.

There are about 20,000 genes in each cell of the human body. Together, they make up the blueprint for the human body and how it works. Genes instruct how to build specific proteins. Genes achieve their effects by directing the synthesis of proteins.

Genes are made up 3 elements:

1. a promoter region (DNA sequence essential in DNA transcription for the caption of the RNA polymerase which will not be transcribed intro the messenger RNA),
2. the open reading frame (ORF - the actual gene, which will be transcribed intro the mRNA and translated into the final protein), which consists in alternating regions of introns (noncoding sequences) and exons (coding sequences).
3. a terminal region called a Terminator (signalling region for the disconnection of the RNA polymerase and the end of transcription; also not transcribed and not part the of the mRNA)

The production of a functional protein involves the transcription of the gene from DNA into RNA, the removal of introns and splicing together of exons, the translation of the spliced RNA sequences into a chain of amino acids, and the posttranslational modification of the protein molecule (fig. 2.9).

Any 3 nucleotide sequence making up an aminoacid is called a codon.

In total there are 64 codons coding for 20 aminoacids. This is called the genetic code (fig. 2.10).

There is a START codon (AUG) and 3 STOP codons (UAA, UAG and UGA).

The Start codon is always set at the transcription initiation site, at the very beginning of the open reading frame, while one of the 3 Stop codons will terminate DNA transcription.

The genetic code is degenerate, meaning that some aminoacids are encoded by more than one single codon (6 codons for one aminoacid at most).

This allows a certain degree of protection against point mutations (substitutions), the misplacement of a nucleotide not always producing a change in the final encoded aminoacid.

Starting from the genetic code, gene mutations can be understood from the functional-proteic effect point of view.

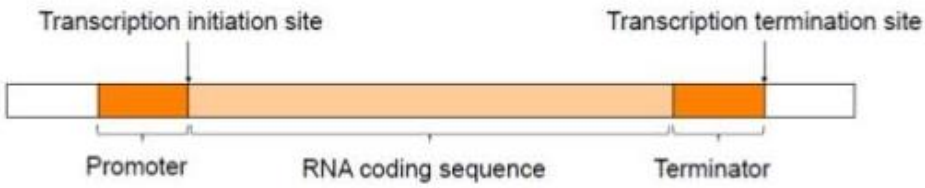


Fig. 2.9. - Gene structure [9]

Some nucleotide changes, insertions or deletions will alter, disrupt or destroy the final proteic product, causing in the end an alteration in the function and the development of a pathology.

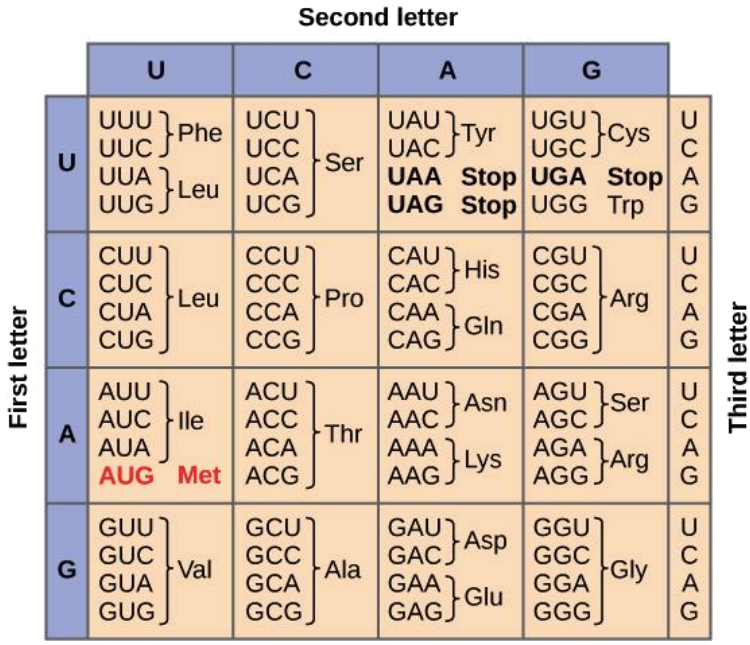
Proteins are composed of a long chain of aminoacids linked together one after another.

There are 20 different aminoacids that can be used in protein synthesis—some must come from the diet (essential amino acids), and some are made by enzymes in the body.

As a chain of aminoacids is put together, it folds upon itself to create a complex three-dimensional structure. It is the shape of the folded structure that determines its function in the body.

Because the folding is determined by the precise sequence of amino acids, each different sequence results in a different protein. Some proteins (such as hemoglobin) contain several different folded chains.

Instructions for synthesizing proteins are coded within the DNA.



First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU, UUC: Phe UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr **UAA Stop** **UAG Stop**	UGU, UGC: Cys **UGA Stop** UGG Trp	U C A G
C	CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg	U C A G
A	AUU, AUC, AUA: Ile **AUG Met**	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn AAA, AAG: Lys	AGU, AGC: Ser AGA, AGG: Arg	U C A G
G	GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp GAA, GAG: Glu	GGU, GGC, GGA, GGG: Gly	U C A G

Fig. 2.10. -The genetic code [10]

2.3.1. Coding

Information is coded within DNA by the sequence in which the bases (A, T, G, and C) are arranged. The code is written in triplets. That is, the bases are arranged in groups of three.

Particular sequences of three bases in DNA code for specific instructions, such as the addition of one amino acid to a chain. For example, GCT (guanine, cytosine, thymine) codes for the addition of the amino acid alanine, and GTT (guanine, thymine, thymine) codes for the addition of the amino acid valine.

Thus, the sequence of amino acids in a protein is determined by the order of triplet base pairs in the gene for that protein on the DNA molecule. The process of turning coded genetic information into a protein involves transcription and translation.

Transcription is the process in which information coded in DNA is transferred (transcribed) to ribonucleic acid (RNA). RNA is a long chain of bases just like

a strand of DNA, except that the base uracil (U) replaces the base thymine (T). Thus, RNA contains triplet-coded information just like DNA.

When transcription is initiated, part of the DNA double helix opens and unwinds. One of the unwound strands of DNA acts as a template against which a complementary strand of RNA forms. The complementary strand of RNA is called messenger RNA (mRNA).

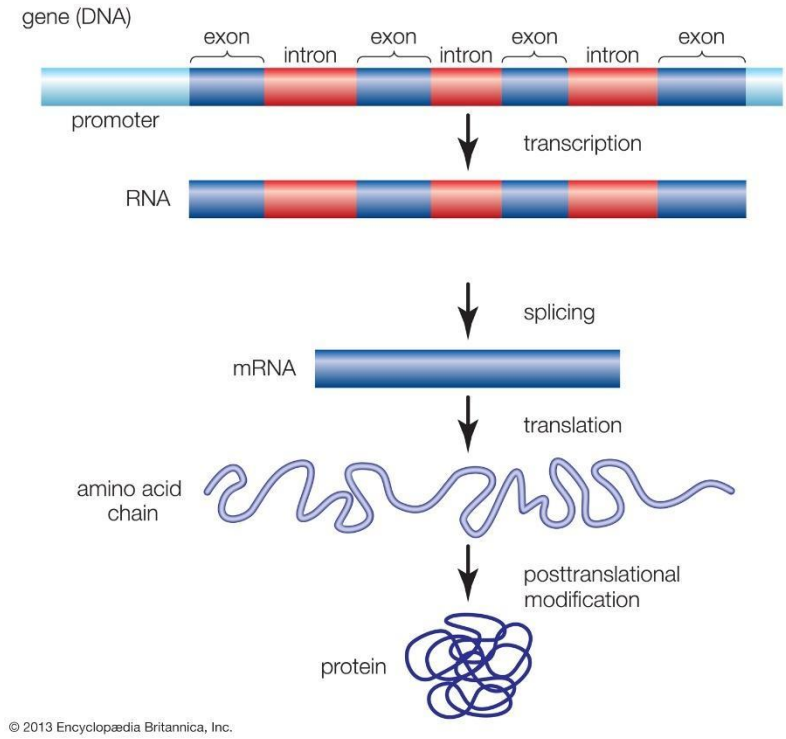


Fig. 2.11 - Protein synthesis – transcription and translation [11]

The mRNA separates from the DNA, leaves the nucleus, and travels into the cell cytoplasm. There, the mRNA attaches to a ribosome, which is a tiny structure in the cell where protein synthesis occurs. With translation, the mRNA code (from the DNA) tells the ribosome the order and type of amino acids to link together.

The amino acids are brought to the ribosome by a much smaller type of RNA called transfer RNA (tRNA). Each molecule of tRNA brings one amino acid to be incorporated into the growing chain of protein, which is folded into a complex three-dimensional structure under the influence of nearby molecules called chaperone molecules.

2.3.2. Gene mutations

A gene mutation is a permanent alteration in the DNA sequence that makes up a gene, such that the sequence differs from what is found in most people. Mutations range in size; they can affect anywhere from a single DNA building block (base pair) to a large segment of a chromosome that includes multiple genes.

Gene mutations can be classified in two major ways:

1. Hereditary mutations are inherited from a parent and are present throughout a person's life in virtually every cell in the body. These mutations are also called germline mutations because they are present in the parent's egg or sperm cells, which are also called germ cells. When an egg and a sperm cell unite, the resulting fertilized egg cell receives DNA from both parents. If this DNA has a mutation, the child that grows from the fertilized egg will have the mutation in each of his or her cells.
2. Acquired (or somatic) mutations occur at some time during a person's life and are present only in certain cells, not in every cell in the body. These changes can be caused by environmental factors such as ultraviolet radiation from the sun, or can occur if an error is made as DNA copies itself during cell division. Acquired mutations in somatic cells (cells other than sperm and egg cells) cannot be passed to the next generation.

Genetic changes that are described as de novo (new) mutations can be either hereditary or somatic.

In some cases, the mutation occurs in a person's egg or sperm cell but is not present in any of the person's other cells. In other cases, the mutation occurs in the fertilized egg shortly after the egg and sperm cells unite. (It is often impossible to tell exactly when a de novo mutation happened.) As the fertilized egg divides, each resulting cell in the growing embryo will have the mutation. De novo mutations may explain genetic disorders in which an affected child has a mutation in every cell in the body but the parents do not, and there is no family history of the disorder.

Somatic mutations that happen in a single cell early in embryonic development can lead to a situation called mosaicism. These genetic changes are not present in a parent's egg or sperm cells, or in the fertilized egg, but happen a bit later when the embryo includes several cells. As all the cells divide during growth and development, cells that arise from the cell with the altered gene will have the mutation, while other cells will not. Depending on the mutation and how many cells are affected, mosaicism may or may not cause health problems.

The DNA sequence of a gene can be altered in a number of ways. Gene mutations have varying effects on health, depending on where they occur and whether they alter the function of essential proteins. The types of mutations include:

1. Missense mutation: this type of mutation is a change in one DNA base pair that results in the substitution of one amino acid for another in the protein made by a gene.
2. Nonsense mutation: a nonsense mutation is also a change in one DNA base pair. Instead of substituting one amino acid for another, however, the altered DNA sequence prematurely signals the cell to stop building a protein. This type of mutation results in a shortened protein that may function improperly or not at all.
3. Insertion: an insertion changes the number of DNA bases in a gene by adding a piece of DNA. As a result, the protein made by the gene may not function properly.
4. Deletion: a deletion changes the number of DNA bases by removing a piece of DNA. Small deletions may remove one or a few base pairs within a gene, while larger deletions can remove an entire gene or several neighboring genes. The deleted DNA may alter the function of the resulting protein(s).
5. Duplication: a duplication consists of a piece of DNA that is abnormally copied one or more times. This type of mutation may alter the function of the resulting protein.
6. Frameshift mutation: this type of mutation occurs when the addition or loss of DNA bases changes a gene's reading frame.
7. The reading frame consists of groups of 3 bases that each code for one aminoacid. A frameshift mutation shifts the grouping of these bases and changes the code for amino acids. The resulting protein is usually nonfunctional. Insertions, deletions, and duplications can all be frameshift mutations (fig. 2.12)
8. Repeat expansion: nucleotide repeats are short DNA sequences that are repeated a number of times in a row. For example, a trinucleotide repeat is made up of 3-base-pair sequences, and a tetranucleotide repeat is made up of 4-base-pair sequences. A repeat expansion is a mutation that increases the number of times that the short DNA sequence is repeated. This type of mutation can cause the resulting protein to function improperly.

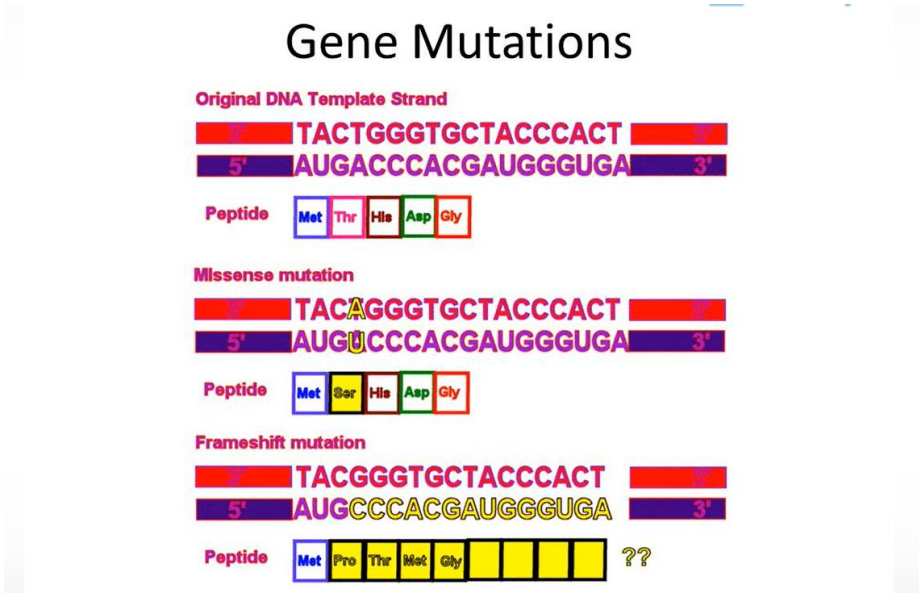


Fig. 2.12 - Gene mutations [12]

2.4. Population Genetics

2.4.1. Introduction

Alleles are versions of our genes. Each gene (except for most of the genes on the X and Y chromosomes in males) comes in 2 copies, one on each homologous chromosome of a pair, at the exact same position (position called locus). The combination of the two alleles is called a genotype and is responsible for a specific trait (phenotype).

Through population genetics scientists try to comprehend how the presence or absence of certain alleles and the frequencies in which alleles are found in different populations modulate our understanding of human evolution and of the evolution of life in general. Population also genetics allows geneticists to predict the predisposition for various complex disorders, susceptibility risen from the presence of certain risk alleles.

2.4.2. Population genetics and natural selection

Natural selection and some of the other evolutionary forces can only act on heritable traits, namely an organism's genetic code.

Alleles are inherited from parents to their children. Certain alleles that bring upon beneficial traits may be selected for, while deleterious alleles may be selected against. Most of the acquired traits are not inherited.

Heritability is the fraction of phenotype variation that can be attributed to genetic differences, or genetic variance, among individuals in a population.

The greater the heritability of a population's phenotypic variation, the more susceptible it is to the evolutionary forces that act on heritable variation.

The Mathematical Theory of Natural Selection

Natural selection is the name given to an evolutionary process in which "adaptation" occurs in such a way that "fitness" increases.

Under certain conditions, this results in descent with modification.

> *If:* variation exists for some trait, and a fitness difference is correlated with that trait, and the trait is to some degree heritable (determined by genetics),
>
> *Then:* the trait distribution will change over the life history of organisms in a single generation, and between generations.

The process of change in the population is called adaptation.

The General Selection Model

Evolution and Natural Selection can be modelled genetically.

> variation = variable p & q
>
> fitness = differential phenotypes of corresponding genotypes
>
> heritability = Mendelian principles

Natural Selection results in change of allele frequency (q) (delta q) in consequence of differences in the relative fitness (W) of the phenotypes to which the alleles contribute.

> Fitness is a phenotype of individual organisms.
>
> Fitness is determined genetically (at least in part).
>
> Fitness is related to success at survival AND reproduction.
>
> Fitness can be measured and quantified (see below) i.e. the relative fitness of genotypes can be assigned numerical values.

Alleles. Mendel`s laws and genotypes

An allele is a version of a gene, a heritable unit that controls a particular feature of an organism.

For instance, Mendel studied a gene that controls flower colour in pea plants. This gene comes in a white allele, w, and a purple allele, W. Each pea plant

has two gene copies, which may be the same or different alleles. When the alleles are different, one—the dominant allele, W—may hide the other—the recessive allele, w. A plant's set of alleles, called its genotype, determines its phenotype, or observable features, in this case flower colour.

2.4.3. Gene linkage

Meiosis is a 2 steps process which involves 2 successive cell divisions - meiosis I (reductional) and meiosis II (equational). Each of these divisions is comprised of the 4 phases (prophase - P, metaphase - M, anaphase - A and

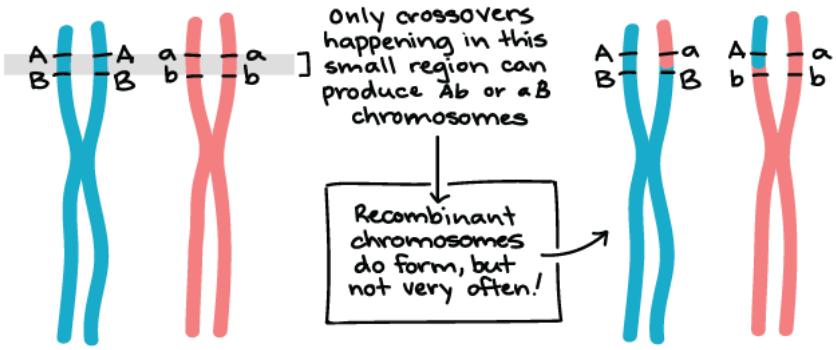


Fig. 2.13 - Gene linkage and linkage disequilibrium (LD) [13]

telophase with cytokinesis - T), meiosis including a total of 8 phases in the end (P1, M1, A1, T1 and P2, M2, A2, T2).

During prophase I of meiosis an extremely important process for genetic variability happens: crossing-over (intra-chromosomal recombination). Crossing-over is the mutual equal exchange of DNA fragments between the homologous chromosomes.

Depending on the distance between two genes situated on the same chromosome, they can or cannot be inherited together. Genes situated close on the same chromosome arm are more likely to be passed on together to the offspring and are called linked genes. Genes on separate chromosomes are never linked. Genes that are farther away from each other are more likely to be separated during a process called homologous recombination.

Genetic linkage refers to the probability that 2 genes situated close to each other on the same chromosome arm would be separated by crossing-over and not inherited together. Two genes linked together are not disrupted by recombination (crossing-over), the chances for this happening being extremely low. Two genes not being inherited together create linkage equilibrium, while two linked genes will be in linkage disequilibrium (LD).

2.4.4. Recombination and recombination fraction

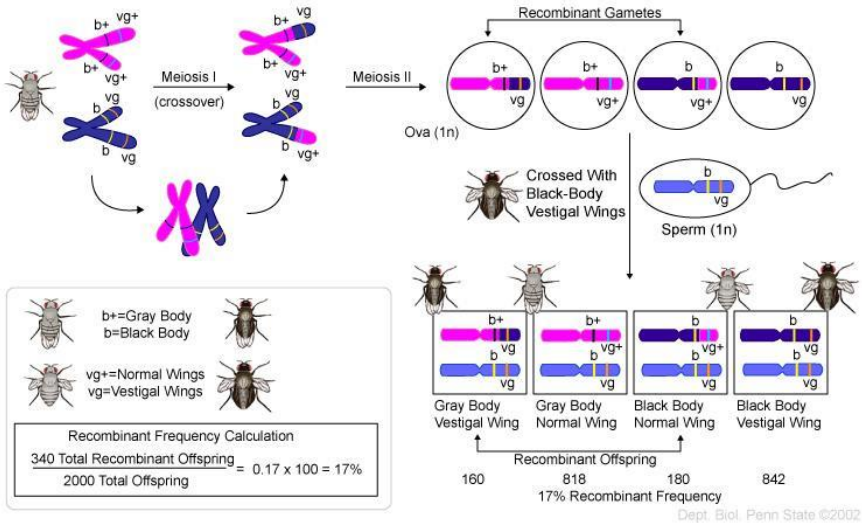


Fig. 2.14 - Recombination fraction (frequency) [14]

Recombination (meiotic crossing-over) is the tool through which we are able, through genetic analysis, to determine whether two genes are linked.

The gametes obtained at the end of meiosis (therefore after crossing-over) are called recombinant. So one can compare the recombination rate by comparing the parental and the offspring's genotypes. (figure 2.14)

Geneticists use linkage to identify the location of a gene on a chromosome by determining the frequency in which different genes are inherited together, and thus creating maps of the relative distances between them.

Considering the fact that each gamete will inherit one of the two possible versions of a chromosome, two unlinked genes will be inherited together 50% of the time randomly. The closer these genes will be one to the other, this percentage will rise as the crossing-over chances diminish. Unlinked genes may be on different chromosomes, or so far apart on the same chromosome that they are often separated by recombination.

Physical and genetic chromosome mapping

Genetic and physical maps illustrate the arrangement of genes on a chromosome.

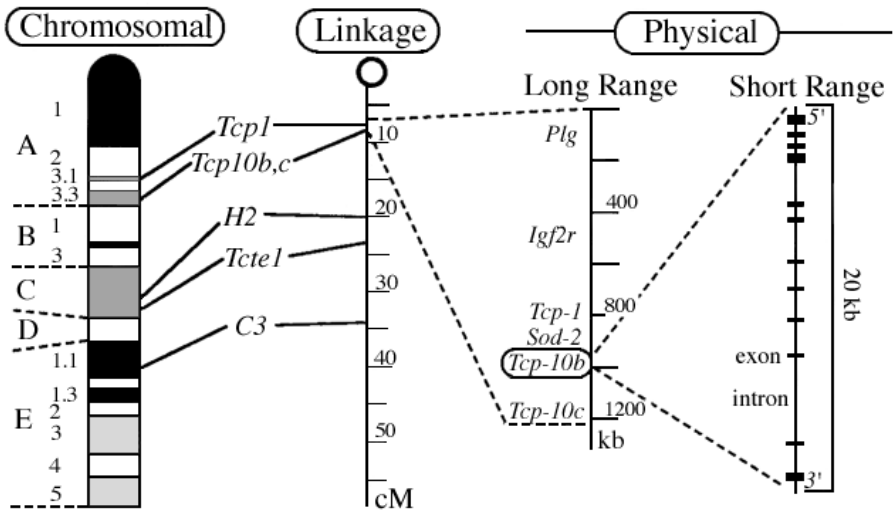


Fig. 2.15 - Physical vs. genetic chromosome mapping [15]

The relative distances between positions on a genetic map are calculated using recombination frequencies, whereas a physical map is based on the actual number of nucleotide pairs between loci. These maps are a key resource for understanding genome organisation.

2.4.5. Allele frequencies

Allele frequency is the percentage in which a particular allele appears in a certain population. Apart from allele frequencies, genotype frequencies are important tools in understanding the genetic make-up of a population and also in clinical practice for determining the risk of individuals or populations for certain complex disorders.

Considering the fact that there are four possible nucleotides in each position (A, T, C, G), in many cases there are more than two alleles in a population. Allele frequencies can help us understand the evolution of populations throughout history. Also, in today's medicine, the frequency of single nucleotide variants / polymorphisms (SNVs / SNPs) is used widely in variant interpretation and genetic counselling.

2.4.6. Hardy-Weinberg equilibrium

Hardy Weinberg equilibrium (HWE): in an infinitely large, randomly mating population in which selection, migration, and mutation do not occur, the frequencies of alleles and genotypes do not change from generation to generation. The genotype frequencies are related to the allele frequencies by the square expansion of those allele frequencies.

In other words, the Hardy-Weinberg Law states that under a restrictive set of assumptions, it is possible to calculate the expected frequencies of genotypes in a population if the frequency of the different alleles in a population is known.

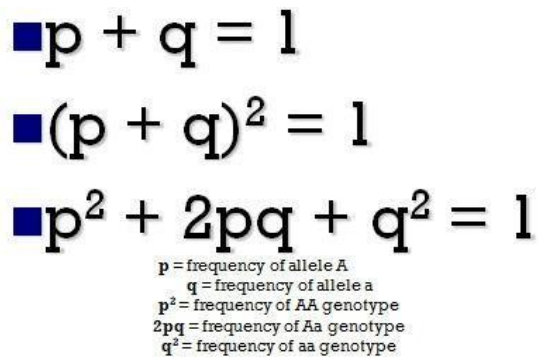


Fig. 2.16 - Hardy-Weinberg Equilibrium
(if there are only 2 alleles, p and q, for a specific trait in a population) [16]

HWE states that evolution will not occur in a population if the following conditions will be met:

- no mutation
- no natural selection
- an infinitely large population
- all members of the population breed
- all mating is totally random
- everyone produces the same number of offspring
- no migrations

2.5. Clinical interpretation of genetic variants

Next Generation Sequencing has brought to light a massive number of genetic variants possibly associated with most of the known human disorders.

Interpreting the association between the presence of a certain mutation and the development of a genetic disease has been a great challenge for researchers and clinicians alike, a challenge that has not yet been surmounted.

The clinical interpretation of genetic variants is essential in the clinical diagnosis of genetic disorders, on which the future therapy and management of the patients will be set upon.

2.5.1. Types of genetic mutations

Approx. 1.5% of the entire genome is represented by genes, the rest being non-coding DNA. Genetic mutations can occur anywhere throughout the genome, with the highest probability (more than 98.5%) of happening outside of genes. More than that, genes themselves are an alternation of coding (exons) and non-coding (introns) sequences.

Most genetic variants causing disorders (pathogenic variants) are located inside the coding sequences of genes. Depending on the position of the mutated nucleotide, gene mutations can be either inside the reading frame of the gene (exonic / inframe mutations), inside the non-coding regions of the gene (the 5` and 3` untranslated regions - 5`UTR and 3`UTR - or intronic variants) or outside, but in the vicinity of the gene (upstream or downstream variants).

2.5.2 Effects of mutations on codon and aminoacid sequences

Gene mutations (substitutions, deletions, insertions, indels) can cause alterations in the codon sequence of the gene and/or in the final protein product.

Depending on the aminoacid change, point mutations (single nucleotide mutations) can be harmless in general (synonymous variants/silent mutations) or might determine drastic changes in the codon sequence and the production of truncated proteins (as are non-sense or frameshift mutations).

2.5.3. Clinical interpretation of genetic variants

Starting from the afore-mentioned types of mutations and types of genetic variants, researchers and clinicians have been having the difficult task of understanding the clinical effect on different pathologies.

The clinical interpretation and classification of genetic variants is a complex process in which changes at every level of protein synthesis (gene, RNA, aminoacid sequence), molecular pathways, gene expression, gene-gene interractions and gene-environment interactions are being taken into account.

In clinical diagnostics, the interpretation of mutations / genetic variants has to meet certain criteria. Clinical guidelines recommend that genetic testing should be performed on patients which are at a clinical risk for a genetic disease, risk which includes a personal pathological medical history

(documented or clinically suspicioned) and/or the existence of a genetic disorder in the family (positive family history).

After genetic testing, a variant will be associated with the patient's phenotype only if the mutation affects the protein expressed by the gene and the respective gene is known to cause a disorder matching the clinical symptomatology of the patient.

2.5.4. ACMG classification

A mutation is defined as a permanent change in the nucleotide sequence, while a polymorphism is defined as a variant with a frequency above 1%.

The terms SNP or SNV are used for defining variants produced by point mutations (SNP - Single Nucleotide Polymorphism; SNV - Single Nucleotide Variant).

The terms “mutation” and “polymorphism”, which have been used widely, often lead to confusion due to incorrect assumptions of pathogenic and benign effects respectively. Thus, it is recommended that both terms be replaced by the term “variant” with the following modifiers:

1. Class 1 - pathogenic variants (P),
2. Class 2 - likely pathogenic variants (LP),
3. Class 3 - variants of unknown/uncertain significance (VUS),
4. Class 4 - likely benign (LB),
5. Class 5 - benign (B).

To better standardize this complex process, the American College of Medical Genetics and Genomics (ACMG) published an updated set of variant interpretation guidelines in 2015. These guidelines consider more than 20 lines of evidence which combine together in different patterns to classify variants into five categories: pathogenic, likely pathogenic, uncertain significance, likely benign and benign.

Having access to a tool that enables the geneticist to select the relevant lines of evidence and then automatically computes the score that is supported by the selected lines of evidence makes the scoring process more rigorous and reliable.

2.5.5. Variant interpretation databases

A number of centralized comprehensive databases exist nowadays for the clinical interpretation of genetic variants. One of the most important databases is the Clinical Variation Database (ClinVar) accesible at: https://www.ncbi.nlm.nih.gov/clinvar/

Another example of a clinical interpretation database is the Leidein Open Variation Database (LOVD): https://www.lovd.nl

dbSNP is an online comprehensive single nucleotide variants database connected to ClinVar. In dbSNP, the mutation will be defined by its SNP reference number (rs) (e.g. rs4994). The DB is accessible at: https://www.ncbi.nlm.nih.gov/snp/

Other online hubs useful in variant classification:

1. Gene information:
    a. Online Mendelian Inheritance in Men (OMIM): www.omim.org
    b. Genecards: https://www.genecards.org
2. Mutations and alleles:
    a. Ensembl: https://www.ensembl.org/index.html
    b. SNPedia: https://www.snpedia.com/index.php/SNPedia
3. Gene expression:
    a. GTex: https://gtexportal.org/home/
    b. gnomAD: https://gnomad.broadinstitute.org
4. Genome:
    a. Genome Browser: https://genome.ucsc.edu
5. Proteins:
    a. UniProt: https://www.uniprot.org

2.5.6. Variants of Unknown/Uncertain Significance (VUS)

The variants posing most problems when interpreting and classifying are the variants of unknown/uncertain clinical significance (VUS, class 3 variants). Two types of mutations are included in this class of variants.

1. The first type is represented by mutations which have never been described in any studies so far and of which their protein effect cannot be predicted (Unknown).
2. The second type of VUS variants are mutations which have been described in previous studies but with contradictory results (e.g. a variant classified as likely benign in one study and as likely pathogenic in other) - these are the Uncertain significance variants.

The American College of Medical Genetics and Genomics (ACMG) previously developed guidance for the interpretation of sequence variants.

In the past decade, sequencing technology has evolved rapidly with the advent of high-throughput next generation sequencing. By adopting next generation sequencing, clinical laboratories are now performing an ever increasing catalogue of genetic testing spanning genotyping, single genes, gene panels, exomes, genomes, transcriptomes and epigenetic assays for genetic disorders.

2.5.7. VUS reclassification criteria

The allele frequency is an important piece of information when classifying a variant as pathogenic or benign. Initially stated that variants with frequencies higher than 5% in the general population are benign, laboratories around the world have dropped that percentage to 2% nowadays.

Information about how often a variant has been observed in the general population provides valuable context for curated (phenotype associated) and predicted variants alike. Variants associated with rare diseases and phenotypes are expected to be found in the general population very rarely – otherwise the diseases would be much more prevalent.

Access to integrated data from resources such as gnomAD, which as of this writing aggregates information from >123,000 exomes and >15,000 genomes from many large-scale sequencing projects, provide such contextual information.

Certain portals have been developed for predicting the possible pathogenicity of mutations. Examples of such portals are:

- SIFT: https://sift.bii.a-star.edu.sg
- PolyPhen: http://genetics.bwh.harvard.edu/pph2/
- Mutation Taster: http://www.mutationtaster.org

These hubs are mostly for synonymous or missense variants which could not be previously classified as pathogenic or benign.

Conservation score

The genome has specific regions that have been highly conserved during the development of life on Earth. These regions are expected to play an essential role in the organism, for which they usually are not affected by mutations and are not involved in adaptation. The higher the conservation score a region has, it will be more expected for a mutation in that DNA sequence to have a pathogenic effect.

Other criteria are being used when trying to reclassify a genetic variant:

- When a VUS is identified in a gene associated with the patient's phenotype but at the same time a reported pathogenic variant has already been detected in the same individual for the same phenotype, the VUS is most likely to be actually a benign or likely benign variant
- Detected unknown/uncertain variants causing phenotypes much more severe than the symptomatology described in the patient are more probable to be class 4 (LB) or 5 (B) variants

- Variants which are reported in low frequencies in the general population but are found in high frequencies in the studied (local) population are benign (probably population-specific mutations)

2.6. Genetic Variation

The last two decades have seen extensive efforts to catalogue human genetic variation and correlate it with phenotypic differences. The first complete human genome sequence, 17 years ago [17], opened the possibility of investigating the various forms of human genetic variation. Based on the frequency of the minor allele (MAF) in the human population, human genetic variants are classified as common (MAF>5%), low frequency (MAF 1-5%) or rare (MAF<1%).

In terms of nucleotide composition, variants in the human genome can be separated into two different classes: single nucleotide variants and structural variants [18].

Single nucleotide variants or SNPs are the most frequent class of genetic variation observed in the human genome. Sequencing results have estimated that the human genome contains at least 20 million SNPs and 1.5 million insertions-deletions [19]. Also, in each human generation, there are a large number of rare and novel or *de novo* single nucleotide variants, in some cases present in only a family or a single individual [20].

Structural variants are characterized as multiple base pairs that differ between individuals and that are not single nucleotide variants. The most common structural variants are insertion-deletions, inversions of DNA sequences and copy number differences. Structural variants represent around 1% of the human genome and influence genome organization, contributing to human disease [21], [22]. Despite the observed genetic variation, the DNA sequences of any two individuals are 99.9 percent identical. The variations, however, may substantially affect an individual's disease risk.

Large numbers of single nucleotide variants on the same chromosome are inherited in blocks; these blocks define haplotypes. Blocks may contain a large number of SNPs, but a few SNPs are enough to identify the haplotypes in a block. The HapMap project was the first effort to define a map of these haplotype blocks and the specific SNPs that identify the haplotypes. The HapMap project provided valuable information, by reducing the number of SNPs required, to examine a large part of the genome for association with a phenotype from 10 million SNPs to roughly 500,000 tag SNPs. The HapMap project helped genome scan approaches to find regions with genes that affect diseases in a much more efficient and comprehensive way than was possible before.

Launched in 2008 and defined as an international research consortium aiming to sequence the genomes of at least 1000 individuals, the 1000 Genomes Project aimed to create a deep catalogue of human genetic variation [23]. This project sequenced genomes from more than 1,000 volunteers worldwide ensuring representation of African, Asian and European populations. The 1000 Genomes Project characterized over 95% of variants that are in genomic regions accessible to high throughput sequencing technologies and that have an allele frequency of 1% or greater in one of five major population groups [24].

The development of next-generation sequencing (NGS) technologies since 2005 has brought revolutionary benefits to medical genetics studies by reducing costs and increasing yield by several orders of magnitude [25]. A comprehensive collection of published germline mutations in nuclear genes that underlie, or are closely associated with, human inherited disease are collected by The Human Gene Mutation Database. In May 2018, the database contained more than 224,000 different gene lesions identified in over 8,000 genes manually curated from over 2,600 journals. The Human Gene Mutation Database represents the central unified gene/disease-oriented repository of heritable mutations causing human genetic disease. It is used worldwide by researchers, clinicians, diagnostic laboratories and genetic counsellors, and is an essential tool for the annotation of next-generation sequencing data [26].

Despite all these improvements, understanding the relationship between genotype and phenotype is still one of the central goals in medical genetics and genetic epidemiology. Genome-wide association studies (GWAS) have evolved over the last fifteen years into a tool for investigating the genetic risk factors for human disease. They have identified new genetic risk factors for many common human diseases and have forced the genetics community to think on a genome-wide scale(Bush & Moore, 2012). Integrating these types of studies in the Romanian medical genetics landscape is crucial for future genetic studies of the population. The data from HapMap allows for the comparison between the LD blocks structure observed in the Romanian population and other European populations. The dataset provided by the full 1000 Genomes Project allowed more accurate imputation of variants in the Romanian GWAS and thus more accurate localization of disease-associated variants. Combining all these databases was pivotal for completing large-scale GWAS using Romanian data.

2.7. Genetics of Cancer

Cancer is characterized by a diversity of genetic and epigenetic alterations occurring in both the germline and somatic genomes [27]. Reflecting these two types of genetic alterations, there are two types of approaches to the genetics of cancer. The first approach is examining the inherited risk of cancer defined as susceptibility or predisposition and the second one, the somatic approach, refers to the actual carcinogenic processes on a genetic level. There is an expanding interest in identifying germline genomic variants associated with different types of cancer, the past decade has seen a dramatic increase in the identification of germline variants that associate with the disease.

2.7.1 Germline variants - inherited risk of cancer

The advancement of genome-wide association studies and genome sequencing techniques has led to improvements in the processes of estimating risk of germline mutations in cancer susceptibility genes and assessing risks of cancer based on personal and family histories. Family history has been examined extensively as a risk factor for cancer. Neoplastic disease serves as a useful model for studying heritability since the familial contribution to this kind of disease risk is usually high in the general population [28]. One of the most comprehensive studies on familial risk and heritability of cancer was performed using 80,309 monozygotic and 123,382 same-sex dizygotic twin individuals within the population-based registers of the Nordic countries [29]. The study reported significant excess familial risk for cancer overall and for specific types of cancer, including prostate, melanoma, breast, ovary, and uterus.

Familial clustering of cancer has proven to be relatively common and is likely to be due to a combination of environmental factors, rare gene mutations with high penetrance, and more common lower penetrant gene variants acting together to alter disease susceptibility. Only a small proportion of cancers are due to highly penetrant inherited mutations in genes [30]. As the field of genetics of cancers has matured, alternate methods to assess familial risk have been developed. There are good reasons to expect that common genetic variants explain a large fraction of the inherited risk of the common cancers [31].

2.7.2. Somatic variants

Cancers appear in individual cells that have accumulated one or more mutations in the DNA sequence, mutations that lead to malignant transformations. Most of these mutations affect the genes involved in signaling for cell proliferation, cell cycle control, apoptosis, and DNA repair. Mutations that activate protein function in signal transduction, advance cell cycle progression or inhibit apoptosis are found in dominant oncogenes that affect cellular phenotype despite the contralateral presence of the normal allele [32]. Suppressor proteins are affected by the loss or disruption of genes involved in cell cycle control, apoptosis, or DNA repair. Genes encoding cell membrane surface molecules involved in adhesion or growth inhibition may act as tumor suppressors. In some cases, the genes may be haploinsufficient, and the loss or inactivation of a single allele is sufficient to influence the pathogenesis of cancer [33].

2.7.3. Oncogenes

Oncogenes are variants of normal (proto-oncogenic) genes that have undergone mutations. Proto-oncogenes control the cell cycle by transmitting to the cell information related to the time and frequency of cell divisions. A gain-of-function mutation in a proto-oncogene that has a dominant cellular effect can be enough to initiate oncogenesis [34].

The first information on the influence of viruses on the occurrence of cancers was obtained from studies by Peyton Rous in 1911. In the 60's and 70's, the molecular mechanisms of virus action were identified, so it was demonstrated that Rous sarcoma virus (RSV) is a retrovirus whose RNA genome is reverted into DNA that will be incorporated into the host cell genome [35].

In 1977, Michael Bishop and Harold Varmus showed that normal human cell DNA contains sequences similar to those of retroviruses. These genes have been called proto-oncogenes. Inappropriate activation of these may cause the process of tumorigenesis to occur. About 100 oncogenes have been identified so far [36]. They encode various cellular proteins with essential roles in cell structure and cell function. Activation of oncogenes is involved in various stages of progression of cancer, together with the loss of tumor suppressor gene function.

Activation mechanisms of oncogenes:

1. Oncogenic point mutations (SNP) - these changes often occur in functionally important parts of genes, resulting in the continuous and uncontrolled activity of proteins involved in intracellular signalling pathways.

The RAS gene is an example, but mutations in this gene occur in ~ 15% of human cancers. The mutations lead to a protein that is constitutively active driving cell growth.

2. Gene amplification. This mechanism results in overexpression of the protein encoded by the gene which then leads to overactivation of the respective biological pathways. A good example is the c-Myc proto-oncogene which is amplified in a large fraction of cancers.

3. Chromosomal translocation - acting through two mechanisms:

a) Translocation brings a gene that controls cell growth near a strong gene promoter, resulting in over-expression of the gene. The oncogenic effect of this mechanism is due to a normal protein that has an increased level of expression.

b) Translocation to obtain a hybrid protein with new properties.

4. Insertion of retroviruses into the proto-oncogene sequence (e.g., HTLV1 virus, HPV).

2.7.4. Tumor Suppressor Genes

Tumor suppressor genes control cell division, slowing down this process. Additionally, proteins that are encoded by these genes play a role in repairing DNA damage or control apoptosis. Mutations occurring in tumor suppressor genes lead to loss of protein function and usually have a recessive character in the cell phenotype.

Chapter 3: Applied Statistics in Genetic Data Analyses

3.1. About Bioinformatics & Genomics

Bioinformatics can be defined as the application of computational tools to capture, organize, analyse, understand, visualize and store biological data information associated with biological macromolecules. [37]; [38]. It is involved in the following areas [39]:

- development of computation algorithms for analysis and interpretation of biological data
- development of methods for management of data in modern biology and medicine
- for example its toolbox includes computer software programs such as BLAST and Ensembl, which depend on the availability of the internet
- analysis of genome sequence data, particularly the analysis of the human genome project - one of the main achievements of bioinformatics to date
- functional understanding of the human genome, leading to enhanced discovery of drug targets and individualized therapy.

Genetics is the science of heredity, or how the characteristics of living organisms are transmitted from one generation to the next via the DNA, the substance that comprises genes, the basic unit of heredity. Genetics dates back to Augustinian friar and scientist Gregor Mendel, whose studies of pea plants in the mid-1800s established many of the rules of heredity. Modern genetics also involves the study of specific and limited numbers of genes, or parts of genes, that have a known function. In biomedical research, scientists try to understand how genes guide the body's development, cause disease or affect response to drugs.

Genomics, by contrast, study the entire set of an organism's genes – called the genome. The term genomics was first coined in 1986 by Jackson Laboratory scientist Tom Roderick [40]. Using high-performance computing and mathematical methods developed by bioinformatics and statistics, genomics researchers analyse enormous amounts of DNA-sequence data to find variants that affect health, behave like risk factors for a disease, control the drug response, or various traits, or possibly protects against a disease. In humans that means searching through about 3 billion bases of DNA across 23,000 genes.

Genomics research is a much newer field than genetics and became possible only in the last couple of decades due to technical advances in DNA sequencing and computational biology. Genomics is the study of all of a

person's genes (the genome), including interactions of those genes with each other and with the person's environment [41]. It also includes the understanding of regions that, although do not contain genes, may still have a biological role, as some studies have demonstrated.

The *Human Genome Project (HGP)* was an international scientific research project with the goal of determining all the base pairs that make up human DNA, and of identifying and mapping all the genes of the human genome both from a physical and a functional standpoint. It remains the world's largest collaborative biological project. The planning started after the idea was picked up in 1984 by the US government, the project being formally launched in 1990, and was declared complete on April 14, 2003 [42].

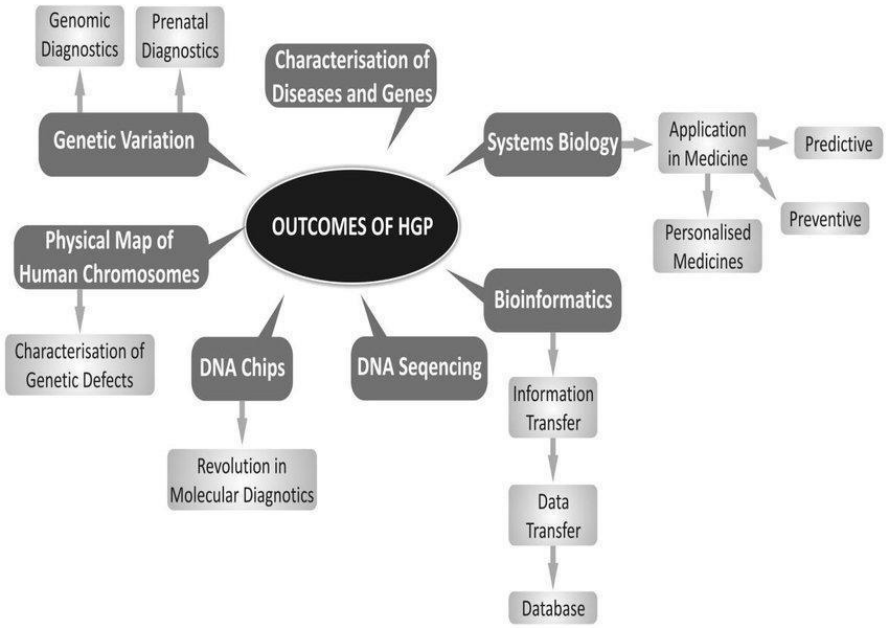


Fig. 3.1. –The outcomes of the Human Genome Project [43]

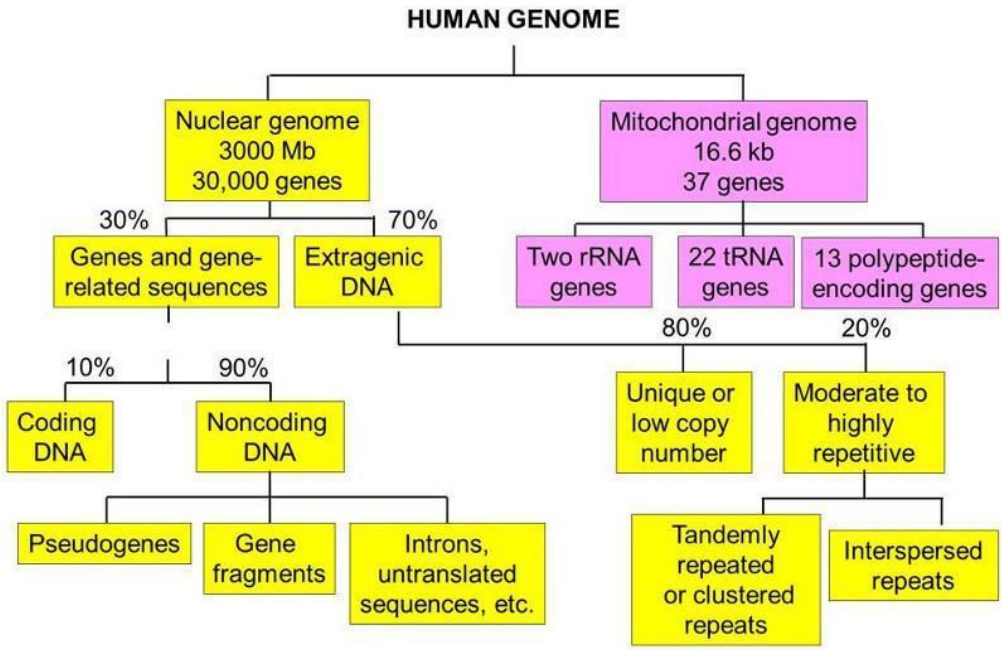


Fig. 3.2 - Organization of the human genome [44]

3.2. Key concepts in genetic analysis

3.2.1. DNA structure, chromosomes and alleles

The DNA is a molecule formed by two polynucleotide chains coiled around each other through complex physical mechanisms which leads to a double helix. Each polynucleotide chain contains 4 types of nucleotides, adenine (A), cytosine (C), thymine (T), and guanine (G), each of which forming base pairs with the nucleotides from the opposite chain. The base pairs are connected by hydrogen bonds. According to the Watson-Crick rules there are only two possible pairings: adenine with thymine and guanine with cytosine. Hence for a succession of A,T,G or C on one strand of the double helix, the order of the paired bases on the other strand would be T, A, C and G. Thus, if we know the sequence of a strand, we know the sequence of the other strand, based on the Watson-Crick pairing. A complete turn of a DNA double helix has ten base pairs.

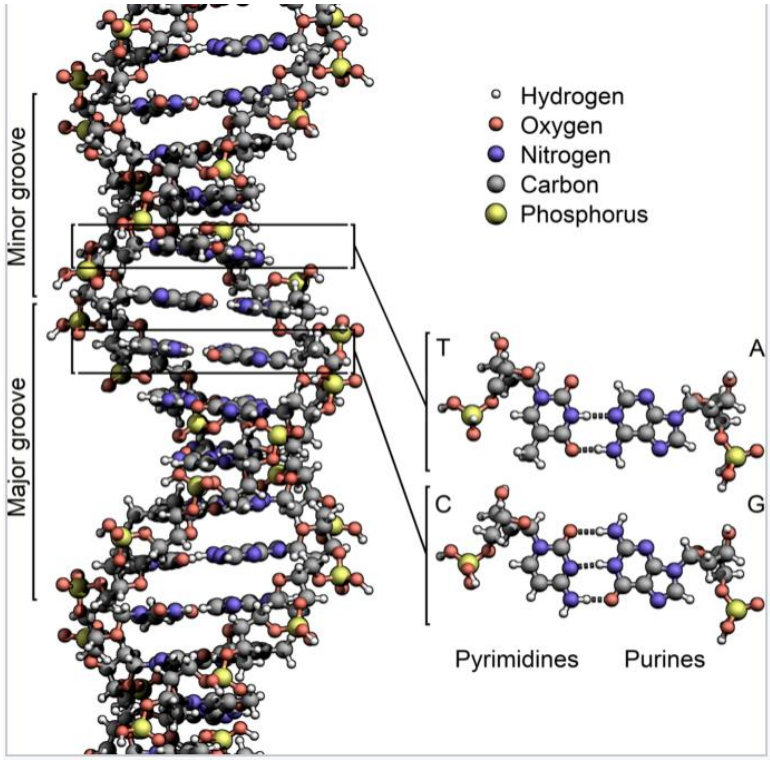


Fig. 3.3 The structure of the DNA double helix. The atoms in the structure are colour-coded by element and the detailed structures of two base pairs are shown in the bottom right. [Wikipedia; [45]

Humans inherit DNA with a total length of one meter from each parent. The DNA is transferred in segments, which are tightly coiled, and are called chromosomes. The genetic information of any organism is stored inside chromosomes and the physical location of a certain DNA sequence on a chromosome is known as a locus (pl. loci). For example each gene within a chromosome has a locus. Somatic cells found in most human tissues are diploid cells which contain two copies of genetic information stored in 46 chromosomes, i. e. 23 chromosomes from each parent, whereas the gametes, which are haploid cells, contain only one set of 23 chromosomes. The genes are segments of the DNA which encode for protein synthesis, situated on each chromosome. Notably, each chromosome contains a different unique set of genes. The maternal and paternal copies of a particular gene differ slightly by the exact base pair sequence, but they are supposed to have the same biological role.

The chromosomes are numbered from 1 to 22, according to the physical size in the packed form. Chromosome number 1 is the largest and chromosome

number 22 is the smallest. Chromosomes 1-22 are called autosomes. Each parent transmits to the offspring one out each pair of homologous chromosomes, which are 1&1, 2&2, 3&3, …, 22&22. The last pair, number 23, are the sex chromosomes, and they make the difference between females, who have two similar chromosomes, called X, and males who have an X and a Y. The size of the X chromosomes is somewhere between the size of chromosomes 7 and 8, and the Y chromosome is about three times smaller.

The coiling process of the DNA is not well understood. It is clearly a physical process for which the molecular forces are responsible. These forces are essentially electrostatic forces between positive and negative electric charges. Let's not forget that the genomic DNA is situated inside cells, where there is a complex solution of water with a lot of atomic ions, molecular ions, and proteins. The ions have a deficit of electrons, and thus have positive electric charge. The DNA is an acid, with an excess of electrons on the sides, and those regions attract the positive ions. The ions surround the DNA and influence its shape. The coiling of the DNA begins around spherical large molecules called histones.

The exchange of DNA between chromosomes during cell division or reproduction is termed genetic recombination. The exact mechanism behind this process is also not well understood. The correct exchange of genetic material requires a perfect recognition between opposing chromosomes, or else the interaction will either not occur, or the recombined chromosome may miss multiple genes, may have a duplicated region, or may have a shorter length etc. It is probable that this exact recognition between chromosomes also involves electrostatic forces. However, the charge distribution is very complex, possibly related to the twist of the DNA and the relative orientation of the base pairs. This mechanism implies that the DNA of a chromosome needs to unpack and stretch out during the recombination process.

Large parts of the genome are similar or even identical between different individuals so that less than 1% of the human DNA is polymorphic and contributes to individual variation. This hereditary variation between different individuals is in fact the result of allelic variation. Each gene can display several variants called alleles, which differ by the exact DNA sequence. Given that somatic human cells contain 2 sets of chromosomes, each cell will contain two alleles for each gene.

The combinations of alleles corresponding to a specific gene form the genotype of that individual at that specific locus. If it happens that the two

alleles are identical, the genotype is called homozygous, and otherwise it is called heterozygous. Note that the maternal or paternal origin of the alleles of an individual are usually unknown unless some information on the parental genotypes is available.

Fig. 3.4 - NextSeq 500 System

The genotypes are obtained with specialized machines that evolved very fast over the last decades. In some cases only genotypes at selected loci on the genome are detected, in other cases the whole genome can be observed. Delivering the power of high-throughput sequencing with the simplicity of a desktop sequencer, the NextSeq 500 System transforms exome, transcriptome, and whole-genome sequencing (WGS) into an everyday research tool. This is a versatile and flexible equipment which easily switches from one application to another.

To characterize the genomic DNA we need to look at the configuration of the bases. Historically, the complete sequencing of the human genome has been achieved in 2003, by the US company Celera. The surprise was to find out less polymorphism than initially estimated. And only about 25000 human genes, compared to about 100000 expected at that time. Hence, it is difficult to explain even the difference between humans and dogs using the DNA. On the other hand the DNA is still the basic molecule of life, but definitely the other molecules included in the cells have a fundamental role too. But still in interaction with the DNA.

The complete sequence of the human genome is now possible, and its costs dropped down dramatically, from millions of USD in the beginning, to less than 1000 USD today, due to the spectacular progress in the DNA technology. Still, many facilities still employ the analysis of local configurations (smaller DNA segments) to identify correlations with certain phenotypes. One reason is that local configurations are easier to detect and cheaper than the whole genome sequence. Another reason is that the whole sequence will be 99% identical for all people.

Fig. 3.5. - Sequencing the whole human genome

3.2.2. Genetic markers

To understand the inheritance of a specific gene, one requires a genetic marker, namely a DNA sequence with a known chromosome location and which can later serve as a reference point for further analyses. For the purpose of genetic analysis, the DNA molecule is viewed like a linear chain, as a row of letters, each of which referring to one of the four nucleotides (A – adenine, C- cytosine, T- thymidine, G- guanine). The physical interactions between bases, the helix shape, and the coiling, are ignored. Only one DNA strand (the positive strand) is mentioned, the other one can be inferred using the Watson-Crick rules.

Depending on the length of this DNA sequence genetic markers can refer to:

- single nucleotide polymorphisms (SNP) – which involves the variation of a single base within a fully conserved sequence, occurred through a mutation.
- multibase markers, i. e. longer DNA sequences, like microsatellites, insertions or deletions of small DNA fragments, etc.
- microsatellites, for example, are repeated sequences such as GAGAGAGA; the number of repetitions defines the alleles displayed by each chromosome.
- ultimately, the complete sequence of the whole genome can be considered a huge collection of markers.

SNPs have been estimated to occur on the average as one in about 1000 nucleotides. SNPs can be located within genes, as part of coding or non-coding regions which regulate the gene expression. Or they can be outside a gene, but linked to the gene sequence, and although such SNPs do not exert an effect on protein synthesis, they give information on locus haplotype.

SNPs have been used in multiple studies such as association studies, which analyse whether a certain SNP allele is associated with a specific phenotypical trait, or in studies of linkage-disequilibrium. In the later example, close SNPs located on the same chromosome can be inherited together in patients with a certain disease. Thus a mutation thought to affect the disease segregates together with a known marker (the SNP whose inheritance pattern is known in the studied individuals), and based on the transmission to the offsprings, researchers can identify correlations between the mutation and a disease.

3.3. Biostatistics concepts and terminology

Biostatistics is the science that uses observed data and deals with the statistical processes and methods applied to the analysis of biological phenomena. From the genetic perspective, biostatistics is involved in:

- characterizing the population, for example to distinguish between groups of individuals, using genetic data
- predicting the implication of the genomic variants in the evolution of the individuals (for example the genetic risk to diseases).

One basic idea is to correlate the biological information represented by the phenotypes with genomic variants. If the correlation is very strong, like in the case of the blood groups, the statistics is not very useful, unless we are interested in some fine details.

Statistical data is useful in cases where the data has a certain level of complexity. For example in complex diseases. Simple diseases, meaning in fact Mendelian characters, are those which can be associated with a single gene. That disease or that trait can be explained by a mutation of that gene. For example the progeria, the disease when the child's body ages fast. A complex disease instead, like cancer, is influenced by hundreds or thousands of genes, and the statistical analysis becomes very complicated, and must be based on a large amount of data, and sometimes on advanced statistical methods.

Statistics uses mathematical methods to characterize data, to predict the occurrence of various events and to identify the factors which influence the data and the relationships established between these factors and the outcome. Statistical genetics refers to the statistical methods employed for the analysis of genetic data, namely the analysis of phenotypes and genotypes.

3.3.1. Probabilities. Carrier and allelic frequencies

The concept of probability is fundamental, but sometimes it is interpreted in a philosophical manner. It is not the case in statistical genetics, where the probabilities are inferred strictly from the observed data.

Examples:

- What is the probability that tomorrow will rain? The answer depends on the available information: in what location, in what season, etc.
- What are the chances of team A to win against team B? The answer could lie in the results obtained in previous matches or in other information such as where the match might take place.

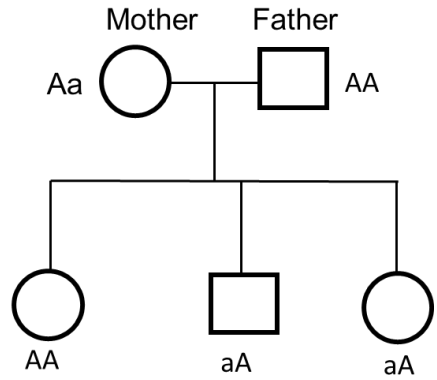


Fig. 3.6. - Example of genetic inheritance on a locus with two alleles A and a

Hence, the probability is an estimate about future events. Depending on the pre-existent information probabilities can thus be derived using theoretical assumptions or calculated from real-life data. In the absence of information the probability does not have much practical meaning.

Probabilities are assigned values between 0.0-1.0 (or 0-100%) with 1 indicating that an event will certainly occur and 0.5 that there are equal chances for it to occur or not.

In order to genetically characterize a population, we need to introduce the concept of genotype and allele frequencies. Consider a location in the genome where we observe an allele A. Here "A" (upper case) is a notation for a specific DNA configuration with an unspecified number of bases, not an adenine base, although in particular it could also be an adenine. Let us denote by "a" (lower case) the complement of allele A, i.e. anything that is not A at the same location.

An individual can have one out of three possible combinations of alleles: (A,A), or (A,a), or (a,a). The combination of the two alleles at one locus, each one inherited from one parent, is called genotype. Considering now the genotypes of multiple individuals, say a population sample, we can calculate the fraction (or frequency) of those individuals carrying a specific genotype, or the fraction (or frequency) of carriers of a specific allele A or a (see below).

In Figure 3.6 we see a small pedigree consisting of mother, father, and three children. In such a pedigree females are represented with circles and the males with squares. Suppose the mother has genotype (A,a) and the father has (A,A). Each child inherited one of each parent's allele, and possible results are shown. Note that no child can be homozygous (a,a) since only one parent has the allele a. In this example we can see the maternal and

paternal inheritance. However, in general, the maternal and paternal alleles of an individual cannot be identified in the absence of information about the parental genotypes. So in general the genotypes (A,a) or (a,A) cannot be distinguished.

Suppose now we consider a population sample of 200 individuals with the genotypes for an autosomal location given in Table 5.1. Given the observed genotypes of each individual, i. e. (A,A), (A,a), (a,a), we can calculate the genotype frequencies, as

$$\textbf{Genotype frequency} = \frac{number\ of\ individuals\ with\ a\ specific\ genotype}{total\ number\ of\ individuals}$$

For example the frequency of the homozygotes (A,A) is 60/200=0.3.

Genotype	Number of individuals	Genotype frequency
AA (dominant homozygous)	60	0.3
Aa (heterozygous)	100	0.5
aa (recessive homozygous)	40	0.2
Total	200	1

Table 3.1. Example of genotype data in a typical genetic study

The allelic frequency, or allelic probabilities, indicate the expected frequency of a specific allele. Because the alleles are defined by chromosomes, their frequency must be evaluated by counting chromosomes, and not individuals.

$$\textbf{Allele frequency} = \frac{number\ of\ chromosomes\ with\ a\ specific\ allele}{total\ number\ of\ chromosomes\ in\ the\ sample}$$

The total number of chromosomes is simply two times the number of individuals. The number of chromosomes with a specific allele, for example A, is two times the number of homozygous genotypes (A,A) plus the number of heterozygous genotypes, (A,a). The allele with the maximum frequency is called the major allele, and the one with the smallest frequency is called the minor allele.

Example: Using the data from Table 3.1, the number of chromosomes with allele A is: 2×60 + 100 = 220 and the total number of chromosomes is 2×200 = 400. The frequency of allele A is thus 220/400=0.55. The frequency of the complementary allele a can be calculated as (2 × 40 + 100)/400=0.45 or as 1 – frequency of allele A. Here A is the major allele and a is the minor allele.

The carrier frequency of a specific allele, say A, is the frequency of individuals carrying at least one copy of A. In this case we look at genotypes, not at chromosomes, and we have in our example (60+100)/200 =0.8. The frequency of carriers of the minor allele a is (40+100)/2000.7. Note that the carrier frequencies do not add up to 1, because the individuals can carry both the major and the minor alleles at the same time.

The *Hardy-Weinberg equilibrium* indicates the relationship between allele and genotype frequencies and it is commonly used in genetic studies to verify the presence of inbreeding or selection or population mixture (or stratification). This principle states that chromosomes belonging to each individual are independent, meaning that they are inherited from genetically independent parents. In other words, in an idealized large population, the allele frequency remains constant with each generation of chromosomes.

We saw that given the genotypes of a population sample we can calculate the genotype and the allele frequencies. Note that the genotypes refer to individuals, and the alleles refer to chromosomes. Now, using the allele frequencies, we can recalculate the genotype frequencies, assuming the genotypes correspond to independent chromosomes. We thus have two ways to calculate the genotype frequencies: a direct way from the genotype data, and an indirect way from the allele frequencies. If the population sample is in Hardy-Weinberg equilibrium, the genotype frequencies should be the same either way.

Let's denote by p the frequency of allele A and by q=1-p the frequency of allele a. Under the assumption that the chromosomes are independent, we can calculate the probability of genotypes (or the genotype frequencies):

- the probability of homozygous AA individuals = p × p
- the probability of heterozygous Aa individuals = p × q + q × p
- the probability of homozygous aa individuals = q × q

Example:
If p=0.55 and q= 0.45 as in the example above,

the frequency of genotype AA = p^2=0.55^2=0.3025,

the frequency of genotype aa= q^2= 0.45^2=0.2025

the frequency of genotype Aa= 2pq = 2 × (0.55 × 0.45) = 0.495

Observe that the resulting genotype frequencies are slightly different from the results obtained in Table 3.1. And in general the genotype frequencies calculated in the two ways are not expected to be exactly equal. In this case the small difference is (most likely) due to the small sample size. But the difference can be much larger, and then we should conclude that the population sample is not in equilibrium.

Suppose for example we consider a set of homozygous individuals, half (A,A) and half (a,a), and no one (A,a). In this case the allelic frequencies are p=q=0.5. The reconstructed genotype frequencies will be 0.5^2 = 0.25 for the homozygotes (A,A) or (a,a), and 0.5 for the heterozygotes (A,a), which are different proportions than in the original data, where we had 0.5 for each homozygote and 0 for the heterozygotes. Obviously, this set of individuals is not in Hardy Weinberg equilibrium. It is not a random population sample, it was selected by excluding homozygotes.

In general, to verify accurately if a population is in genetic equilibrium, one needs to compare the observed and the expected genotype frequencies by performing a chi-square statistical test (with one degree of freedom). The sample size is then automatically included, and the test gives a p-value which measures if the difference is significant or not. If it is significant the population sample does not obey the Hardy-Weinberg equilibrium for a reason that needs to be understood from case to case.

3.3.2. Population sampling

Population genetics aims at describing the population and genetic features or changes which occur and may affect the general population or certain groups of people, such as the prevalence of a disease, the transmission of a trait causing a disease, the rate of mutations or recombination, etc. Depending on the tested hypotheses, researchers can opt for a case-control study design, a family based study, or a cross-sectional study. Genetic studies usually employ a case-control design, namely a retrospective analysis which includes

individuals with a specific phenotypic trait/disease and healthy controls without this specific characteristic/disease.

The prevalence of a disease in a population is the proportion of people from that population having that disease. Or, in terms of probabilities, the chances that a randomly selected person from that population has the disease. To determine the prevalence, researchers analyse a sample of the population (e.g. the people registered at a given health centre).

A sample of a population selected for genetic studies is commonly called a cohort. Ideally, a good sample should be unbiased, meaning that it should be a good representation of the entire population. Depending on the type of study and on the research hypothesis, a sample of individuals diagnosed with a disease may not be representative due to a potential selection bias. Still, in genetic studies where we are interested in a genetic location not related to that disease, we may be able to use those individuals as an (approximately) unbiased population sample.

In a case-control study the cases are usually collected from hospitals. The relative random factor for cases may not be “perfect”. For example, people may originate from the same geographic region, possibly with a genetic background slightly different than the rest of the population, and the group of cases is genetically biased. Meaning that if we want to find a genetic variant carried by the cases that explains the disease, we may confuse it with a genetic variant that is common in that group only because they may be related people.

A case – control study also needs a group of population controls. In general, the most difficult work in such studies is to collect the biological samples. The patients from the hospitals are easier and naturally recruited by doctors, but the population controls are typically neglected. A small set of population controls reduces the statistical power as much as a small set of patients. Therefore, is very important to collect an adequate, preferably large, number of controls.

One particular situation is when a gender specific disease is studied, like ovarian or prostate cancer. In that case all patients are only females or only males, respectively. Can we use as controls a mixed group of females and males? The answer is yes, as long as we are not involving the sex chromosomes in the study. If the X chromosome has to be analysed, we ought to carefully take into account that females carry two copies and the males carry only one. The reason to ignore the sex of the controls is that the chromosomes themselves are not different between genders. In other words,

Further reading

Depending on the study hypothesis**,** sampling methods in genetic studies can be divided in non-population and population-based sampling designs. Non-population based-sampling methods refer to the collection of geographical data or of other spatial coordinates tested in anthropology or archaeologic studies. By comparison, medical genetic studies commonly use a population-based sampling method and can record either individual phenotypical data or the entire pedigree of the subject.

The sample size should be determined beforehand, taking into consideration the study hypothesis and the power of the study. Conventionally, a larger sample size ensures a higher statistical accuracy and a lower risk of false positives (type I error).

Regarding the study hypothesis, the sample size depends on the frequency of the study allele (rare alleles require large sample sizes) and on the methodology (studies of linkage disequilibrium also require large samples). GWAS sample size can be determined using online calculators

E.g.: https://omictools.com/power-sample-size-calculation-category

Fig. 3.7. - Sampling methods in genetic studies

we compare chromosomes of the sick people with the chromosomes of the population, hoping to find a location where the two sets of chromosomes have different allele frequencies, eventually explaining the disease.

3.3.3. Descriptive Statistics

Depending on the research question, the studied variables can be divided into quantitative or qualitative data.

Quantitative data refer to numerical data and can be further divided into:

- *Discrete data* (finite, countable events) such as the number of students in a class, the number of heads in 50 coin tosses, the number of malign lymphocytes in one microscope field, the number of infected erythrocytes in one microscope field in patients with malaria, etc.
- *Continuous data* (measurements which cannot be counted with integer numbers) such as the weight, height, age, temperature, the serum level of cholesterol, etc.

Continuous data can be further classified into:

- Interval data such as the temperature, where a reference point is selected, and each 1°C difference defines an interval.

- Ratio data such as height or weight, divided by a reference value. A meaningful reference value must be defined from case to case. For example a baby with 4000g is twice the weight of a baby with 2000g, whereas a temperature of 50°C is not twice as hot as a temperature of 25°C. In the temperature case 0°C still indicates some heat, so in this case perhaps the Kelvin degrees are better units.

Qualitative data refer to descriptive data which cannot be quantified, and include:

- *Nominal data* – gender, eye colour, pass/fail, positive/negative, etc.
- *Ordinal data* – satisfaction levels, pain assessment.

The summary statistics of a data set is the information on:

- the frequencies for categorical data.
- central tendency of quantitative data, such as mean, median, module.
- statistical dispersion of quantitative data, such as standard deviation.
- the shape of the distribution of quantitative data, such as skewness, kurtosis.

Let's considering the numeric random variable X and a sample of these numbers $x_1, x_2, x_3 \ldots, x_n$. The sample mean is defined as:

$$\underline{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

The sample variance, which measures the dispersion of the data around the sample mean, is:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}$$

It should be seen as a mean value of the squared deviations of each data point from the mean, but divided by n-1 instead of n. The reason is that there are only n-1 independent numbers summed up, because the sample mean itself includes all data points. The standard deviation is the square root of the sample variance:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}}$$

R example:

An example with the software R on how to define a data set and how to calculate the mean value and standard deviation is seen on the figure above.

Use the RStudio console to test them:

```
x=c(1,2,5,9,2,5,6,9)    # define the data set as a vector
mean(x)                 # mean value
sd(x)                   # standard deviation
```

You should get a mean = 4.875 and a standard deviation = 3.090885

Fig. 3.8. - Mean and standard deviation computation in R

GWAS implication

Summary statistics plays an important role in the quality control and in the analysis of GWAS data. Examples of summary statistics in GWAS studies include genotype counts and missing genotypes rates, allele frequencies, heterozygosity rates, Hardy-Weinberg equilibrium failures etc.

3.3.4. Random variables

A random variable is the outcome of an experiment or process which cannot be predicted.

Discrete case: The variable is a discrete number (like 1,2,3...) or possibly a category (like white, yellow, blue – i.e. not countable numeric values)

Continuous case: The variable is any number in a certain interval

We denote the random variable with capital X and a particular outcome with small x. P(X = x) or P(x) is the probability that the outcome of X is x, and it is called the probability distribution function (or the probability mass function), or simply the distribution of X.

3.3.5. The normal distribution

The normal distribution of a random variable x with mean value μ and standard deviation σ is a continuous probability distribution, symmetrical around the mean (bell-shaped distribution), characterized by the following probability density function:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The normal distribution has multiple implications in statistics due to the central limit theorem. The central limit theorem states that given many samples of independent random numbers, the average values (sample means) of each sample will obey the normal distribution. Hence, the normal distribution is practically the distribution of the mean values of random data sets.

Considering that X has a normal distribution, with the mean value μ and the standard deviation σ:

- The total area under the normal curve is equal to 1. In fact this is a general property valid for all distribution functions, because the sum of all probabilities that X has any unspecified outcome x must be 1.
- Depending on the standard deviation, if σ is small, the shape of the curve is taller and narrower, whereas if σ is large, the height of the curve reduces, but it spreads more along the x axis
- 68% of the area under the curve is included within the mean ± 1 standard deviation
- 95% of the area under the curve is included within the mean ± 1.96 standard deviations (or 95.5% within ± 2).
- 99.7% of the area under the curve is included within the mean ± 3 standard deviations

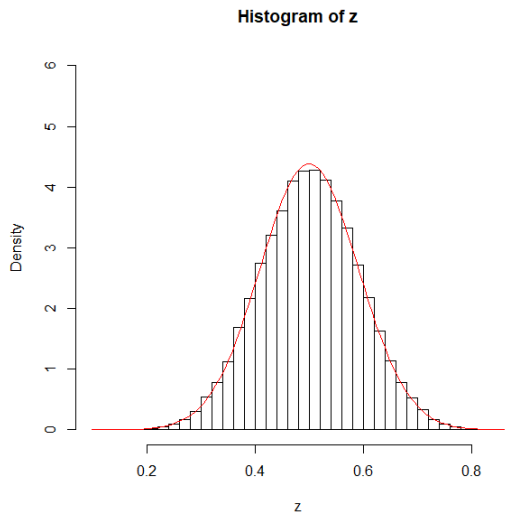


Figure 3.9 – An example of normal distribution

R example on how the mean values of 10 random numbers with uniform distribution have a normal distribution

```
z=vector()
k=100000; n=10
for (i in 1:k) { z[i]=mean(runif(n,0,1)) }
m=mean(z); s=sd(z)
hist(z,prob=T, breaks=30, ylim=c(0,6))
curve(dnorm(x, mean=m, sd=s), col="red", add=T)
remove(z)
```

Fig. 3.10. - The normal distribution example of Figure 3.9 using k=1000000 sets of n=10 random numbers with uniform distribution.

Suppose we have a set of n random numbers with values between 0 and 1. Suppose our random numbers have uniform distribution, meaning that any number between 0 and 1 can be in our set of n. The mean value of our set, let's call it *x*, should be something close to 0.5, but not exactly, because we have only n numbers. The example of Figures 3.9-3.10 we used n=10, but is we would use much more numbers, say on million, the observed mean value should be much closer to 0.5 (and the distribution much sharper). Suppose now we have another set of n such random numbers, and another, and another, and each time we store their mean values x. These numbers x are also random numbers, but with a normal distribution, with two parameters, μ and σ, which represent the expected mean values of the x's, and their expected standard deviation, respectively. In our case population mean (μ) is 0.5, and standard deviation (σ) is $1/\sqrt{12n}$.

3.3.6. Hypothesis testing

One of the most important subjects in statistics, for practical applications, is hypothesis testing. The problem consists in evaluating if the data set indicates a regular situation, called the null hypothesis H_0, or a special situation called the alternative hypothesis H_1.

The evaluation is based on a number called "the statistic", which is derived using the regular expectation defined by the null hypothesis, and which has a known probability distribution. That number reflects the deviation of the

observed data from the expected values, based on the null hypothesis. If the deviation is not large, only due to regular random effects, that number is likely small. If, however, that number has a large, unexpected value, then we may conclude that the null hypothesis is not correct, and instead the alternative hypothesis could be the correct one. Meaning that the data shows something special not expected under the null hypothesis. For example, a particular allele of a gene is carried in excess by the patients of a disease, compared to a random population sample, meaning that this allele has a role in the disease.

Technically, the acceptance or rejection of the null hypothesis is decided with a number called p-value, which indicates what is the probability that the statistic number is even greater than observed. Typically, if this probability is below 0.05 we are inclined to reject the null hypothesis and believe in the alternative. In this later case we would say that the excess of the studied alleles in patients is "significant". The 0.05 value is pure conventional, and lower values can also be considered.

Note that significant does not necessarily mean true, but rather should be interpreted as a tendency of the observed data not to obey the expectation. Sure, if the p-value is much lower, significant may be associated with almost true. But in general the "truth" can be established after the tendency observed is confirmed or reproduced in similar, but independent experiments.

One of the most important goals of statistics is to observe differences between data sets, in order to decide whether assumptions about the origin of these differences are valid, and possibly if the experiment merits to be replicated with another data set. Examples of statistical hypothesis testing include tests of the population mean, teste of variances, differences between several subgroups or categories, etc.

3.3.7. Genetic Association studies

An example of hypothesis testing in genetics is whether an allele A is related to a specific phenotype. Therefore, one would conduct a case-control study (individuals with/without that phenotype) and record the genotypes of each individual. Therefore, two possibilities emerge for allele A:

- the null hypothesis H_0: allele A is present both cases and controls, in similar proportions
- the alternative hypothesis H_1: the proportions (or frequencies) of allele A in cases and controls are different

Before applying a statistical test, one would decide upon the threshold for declaring that allele A is more common in individuals with that phenotype.

This statistical threshold is the significance level, commonly denoted as α and usually set to 0.05.

The result of the hypothesis testing is a p-value. Hence:

- if $p < 0.05$ (less than α) we consider the difference significant, we reject H_0, and accept H_1.
- if $p > 0.05$ the difference is clearly a result of chance and we accept H_0.
- if $p \approx 0.05$ we have to be careful, the case is marginally (borderline) significant and we may need more information for a proper conclusion.

3.3.8 An example of data

A typical genetic study analyses the number of carriers of a specific allele in two groups of individuals, one group with a specific phenotype (such as disease) and another group of healthy individuals (controls). The table shown below is a small 2x2 matrix, called a contingency table.

	Carriers of allele A	Non-carriers of allele A
Patients	18	4
Controls	40	30

Table 3.2 – 2x2 contingency table for a typical genetic study

We see that the number of carriers exceeds the number of non-carriers for both groups of individuals. We have 18+4=22 patients and 40+30=70 controls. Let's denote with p the carrier frequency in patients and with c the carrier frequency in controls. According to the data p=18/22=0.82 and c=40/70=0.57 We see that $p > c$. But can we say that this difference is statistically significant? Or that is simply a result of chances? Observe that the sample size is not large, and it is easy to understand that a small sample can have large fluctuations. Meaning that if we were able to have a second data set for a separate group of 22 patients and 70 controls, we would probably see substantially different proportions of carriers and non carriers. Therefore, a correct statistical test must include the sample size. Let us see how we can perform such a test.

3.3.9. A possible genetic test based on the normal distribution

We have to realize that the carrier frequencies are in fact some mean values. If we associate to the carriers the number 1 and to the non-carriers the

number 0, the carrier frequency is simply the sample mean of these numbers. We can, of course, also consider the number of alleles A carried by each individual, and then we would use the numbers 0,1, or 2, but we will not do that now.

Since the carrier frequencies are mean values they should have a normal distribution. Considering the controls as a reference group the (expected) mean value of this normal distribution is c=0.57 and its standard deviation is $\sigma = \sqrt{\frac{c\,(1-c)}{70}}$ = 0.059 (which is the standard deviation of the sample mean of the binary random variable associated with carriers and non-carriers). With these two parameters we obtain the normal distribution shown in Figure 3.11, which illustrates our expectation about probable values of the carrier frequency in a healthy group of individuals.

The carrier frequency c=0.57 is calculated on our sample, but it can be interpreted as an estimated carrier frequency for all individuals in the population who do not have the disease that it is studied in. But we have to realize that the true carrier frequency, that we do not exactly know, may be more or less different. How can we describe how different it can be? As we mentioned in Subsection 3.3.5, 95% of the area under the normal distribution function corresponds to the interval covered by the mean value ±1.96 standard deviations, which in the present case is [c-1.96σ, c+1.96σ] = [0.57-1.96×0.059, 0.57+1.96×0.059] = [0.45,0.69]. This is called the 95% confidence interval for the true population mean value, and it tells us that there are 95% chances that this true (unknown) value is within these limits.

The carrier frequency in the patients, 0.82, is situated in the right tail of this distribution, outside the 95% confidence interval for the population mean value. That first tells us that it would be hard to believe that this difference occurs simply by chance, due to random sample effects. The probability for observing an even larger frequency is the area under the curve to the right of 0.82, which is 1.5×10^{-5} . This is the p-value, and it is very small, suggesting that the difference between the carrier frequencies in the patients and controls is very significant.

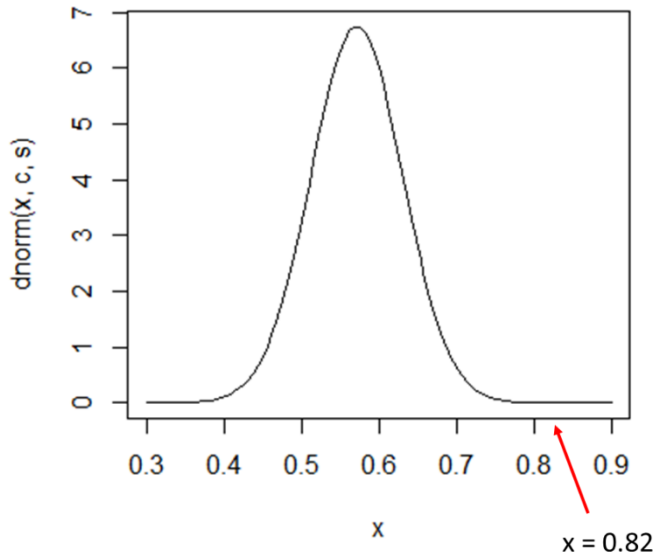


Fig. 3.11 – The p-value is the area under the curve to the right of x=0.82

R example

```
c=40/70; p=18/22
s=sqrt(c*(1-c)/70)
curve(dnorm(x,c,s),xlim=c(0.3,0.9))    # gives the graph
Pval = 1-pnorm(p,c,s) = 1.5e-05        # area to the left of p
```

Fig. 3.12. - Normal distribution test example in R

This is the basic idea of a statistical test: build the expected distribution of the carrier frequency based on data from healthy individuals, place the observed frequency in patients on that distribution, and calculate the p-value which is the probability that the frequency in patients is even larger than observed. The normal distribution is a convenient choice for the reference group. However, in this method we did not include the sample size of the group of patients. If the group of patients is smaller or larger, we may expect different fluctuations, and our p-value and maybe our conclusion are not correct. Next, we will go over some details related to the quality or reliability of a statistical test.

3.3.10. Quality aspects of a statistical test

The statistical theory on hypothesis testing also includes the concept of *statistical error*. The result of a test can be positive or negative from the perspective of the null hypothesis. If the result of the test does not correspond to reality then an error occurred. Since, the very nature of the test is statistical, the results, even significant, are possibly wrong. We can distinguish two types of errors:

- Type I or false positive: is the false rejection of the null hypothesis. The p-value < 0.05 and one rejects H_0, accepts H_1, but in reality H_0 is true (e.g.: telling a patient who doesn't have cancer that he is actually sick).
- Type II or false negative: is the false acceptance of the null hypothesis. The p-value > 0.05 and one rejects H_1, and accepts H_0, but in reality when H_0 is not true (e.g.: telling a patient that he is cancer-free, when he is actually sick).

Usually both types of errors appear during testing and the in order to reduce of them the test must be carefully designed, for example by collecting accurate data and by using a good mathematical model of the null distribution. Ultimately, for a specific test, the only way to reduce the errors is to increase the size of the sample.

3.3.11. Power of a statistical test

The power of a statistical test is the capacity of the test to identify the correct hypothesis: H_0 or H_1. Mathematically, the power is the probability of rejecting the null hypothesis when it is false. (Or 1 minus the probability of a false negative result.) In general we cannot evaluate the power of a test with real data, because real data are unique, and also expensive. There are however methods based on computer simulations, where – for example – genotype data for fictitious patients and controls are simulated and the statistical test finds a significant difference between these groups or not.

Here are some practical methods to increase the power of a statistical test:

- Make sure too use all available information.
- Maximize or increase the sample size if possible.
- Correctly formulate both H_0 or H_1.
- Prepare the data carefully, remove outliers if they are dubious.
- Avoid unnecessary approximations of the distributions if possible.

Our earlier proposed test based on the normal distribution is not efficient for our needs if the sample size is too small. In principle the normal distribution is

valid for large samples only. And we also saw that we did not properly include the total number of patients.

If the statistical test is well designed, then we have to be aware that a very significant result, giving a very low p-value, in fact may be observable without statistics. It may be obvious that the null hypothesis cannot be true without calculations. Indeed, a calculation will confirm that. But in practice the statistics intends to show the significance when it is not obvious. For this reason a lot of statistical work dedicated to new genetic discoveries is spent around p-values which are not very far from 0.05. Again, for this reason, it is crucial to increase (or maximize if possible) the statistical power. That is what we are going to do next, by replacing the test based on the normal distribution with a test based on the chi-square distribution.

3.3.12. The chi-square test

The chi-square (χ^2) test is a statistical test commonly used in genetic association studies. It tests the probability that the observed data can deviate from the expected data by chance, based on the chi-square distribution function. It is a non-parametrical test which can be used for two or more categorical variables, with several assumptions, as follows:

- The tested variables are independent of each other.
- The categories of each variable are mutually exclusive.
- The expected values for each cell should be at least 5 in 80% of cells and no cell should have an expected value of zero.

The statistic is calculated using the formula

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \ ,$$

where the labels *i,j* denote the cell from row *i* and column *j*, O_{ij} is the observed counts (number) in cell *i,j* and E_{ij} is the expected counts in cell *i,j*.

	Carriers of allele A	Non-carriers of allele A	Total
Patients	18 (13.87)	4 (8.13)	22
Controls	40 (44.13)	30 (25.87)	70
Total	58	34	92

Table 3.3 – The previous example of data set, including now the total numbers on raw of columns (the margins) and the expected numbers in each cell (in brackets).

In order to calculate the χ^2 we must first calculate the sums on each row and column as above. These row/column sums are often called margins. The summation of the margins on the rows or on the columns gives the total number of individuals, N=92. The expected values in each cell are calculated based on the assumption that allele A makes no difference between patients and controls. Thus, regardless of whether a subject is a carrier or not, the expected cell counts are calculated as simple proportions considering the margins fixed:

$$E_{ij} = M_i \frac{M_j}{N},$$

where M_i and M_j are the margins on raw *i* and column *j*, respectively. In our example the null hypothesis H_0 assumes no difference between groups, apart from pure random effects. Hence the expected data depends only on the total numbers. Based on the previous formula we obtain $\chi^2 = 4.37$.

The number χ^2 can be seen as the second power of a variable with normal distribution. The number of degrees of freedom is – by definition – the number of independent numbers in the contingency table, considering the margins fixed. Observe that if we want to change the observed numbers in the table it is sufficient to change the number in only one cell. The numbers in the other three cells will be determined by the fixed margins. For this reason the data has only one degree of freedom. In a more general case, when the contingency table has n rows and m columns, the number of degrees of freedom is (n-1)(m-1). Obviously, the expected counts in each cell remain constant if we change the observed counts, because the expected numbers depend only on the margins. Therefore one can build a distribution function of the number χ^2, which is in fact a family of distribution functions where the number of degrees of freedom is a parameter.

3.3.13. Chi-square distributions

Consider a random variable x with a normal distribution. If so, the random variable x^2 has a chi-square distribution with one degree of freedom. Consider now k random variables, independent, each one with a normal distribution. The sum of their squares is a random number with a chi-square distribution with k degrees of freedom. Depending on the number of degrees of freedom, the chi-square distribution can have several shapes, as shown in Figure 3.13. The mathematical expression of the distribution can easily be found online [https://en.wikipedia.org/wiki/Chi-square_distribution].

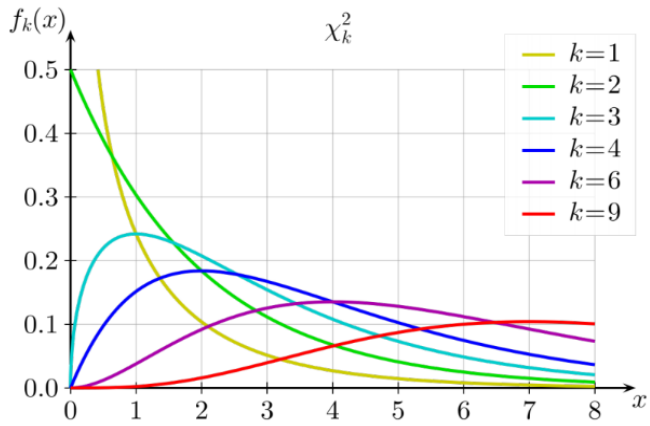


Fig. 3.13 – Chi-square distribution functions with k degrees of freedom [46].

To represent graphically a chi-square distribution you can use the RStudio console like in the examples of Figure 3.14.

R example

```
>curve(dchisq(x,1), xlim=c(0,5))
>curve(dchisq(x,2), xlim=c(0,5))
>curve(dchisq(x,3), xlim=c(0,5))
>curve(dchisq(x,4), xlim=c(0,5))
>curve(dchisq(x,6), xlim=c(0,5))
>curve(dchisq(x,9), xlim=c(0,5))
```

Fig. 3.14 - Examples of commands in R to obtain graphs of chi-square distributions with 1 to 9 degrees of freedom

3.3.14. The chi-square test in practice

The practical way to perform a chi-square test is with a computer software. Almost any software for numerical calculations includes it, as well as other software such as Excel. In R the basic command is *chisq.test*. In addition there are many online tools where all you need is to plug in the observed data. See for example the chi-square calculator at https://www.socscistatistics.com/.

In our case the calculated $\chi^2 = 4.37$ must be placed on the horizontal axis of the chi-square distribution with one degree of freedom, and the p-value is the area to the right of this number. One obtains the p-value 0.03648. Some

software, like R, has implemented the Yates continuity correction, which is a technical detail that takes into account the small sample size, and with this correction one obtains a larger p-value, 0.06601. Anyway, we see a much larger p-value than in the preliminary test proposed, which was based on the normal distribution. With the chi-squared method we increased the statistical power by improving the statistical method (for example properly including the sample size of the patients subgroup), and now we see that the difference between the null and alternative hypotheses is at most marginally significant.

R example

The output of the chi-square test command was shown using italics

```
> data=rbind(c(18,4),c(40,30))   # here is the data matrix


> chisq.test(data)    # with Yates correction
Pearson's Chi-squared test with Yates' continuity correction
X-squared = 3.3795, df = 1, p-value = 0.06601


>chisq.test(data,corr=F)    # without Yates correction
Pearson's Chi-squared test results:
X-squared = 4.3745, df = 1, p-value = 0.03648
```

Fig. 3.15. – The chi-square test performed in R with and without Yates correction. That correction is included by default, to reactivate it one needs to use correction=FALSE (corr=F)

3.3.15. The Fisher test

Another popular test, and even more rigorous in our case, is called the Fisher exact test. Given the fixed margins we need to specify only one single number in one cell, and all the other numbers of the other cells are known. Say we have k patients carriers (the top left cell). We can calculate how many combinations of k patients we can obtain out of the total of 22, and also how many combinations of the 58-k controls carriers can we have out of all 70 controls. With these numbers we can figure out the probability that we see k individuals in the top-left cell. This is called the hypergeometric distribution:

$$P(k) = \frac{(22\ k\ )(70\ 58-k\ )}{(92\ 58\ )} = \frac{22! \times 70! \times 58! \times 34!}{k!\,(22-k)!\,(58-k)!\,(12+k)!\,92!},\ k = 0,\ 1,\ 2, \ldots, 22.$$

Now, we can evaluate the p-value, as the probability that $k \geq 18$, which is 0.030.

	Carriers of allele A	Non-carriers of allele A	Total
Patients	k	22-k	22
Controls	58-k	12+k	70
Total	58	34	92

Table 3.4 – The data table corresponding to k patients carriers.

The Fisher test does not rely on a distribution model, like the chi-square test, and that is why it is called exact. It is applicable to both small and large samples, and it does not need small sample size corrections. The p-value can be calculated from the hypergeometric distribution, but also directly with a software command or with an online calculator.

R example

```
> k=c(0:22);
> plot(dhyper(k,22,70,58))  # shows the graph of the hypergeometric distribution
> pvalue=sum(dhyper(18:22,22,70,58))  # gives 0.02997532

> data=rbind(c(18,4),c(40,30))    # data matrix
> fisher.test(data,alt="g")    # one sided Fisher test p-value = 0.02998
> fisher.test(data)            #  two sided Fisher test p-value = 0.04427
```

Figure 3.16. – Hypergeometric distribution and Fisher test with R

In any statistical test of hypotheses we have a null hypothesis H_0, and an alternative hypothesis H_1 . Here the null is that the fraction of patient carriers (18/22) and control carriers (40/70) are statistically equal. Meaning that if they are numerically different, that is only an apparent fact, due to randomness. Instead, the alternative is that this frequencies are truly different. But this

alternative hypothesis can have two forms: the fraction of carriers is *larger* in patients than in controls (which is called the one-sided alternative), or the fraction of carriers is *different* than in controls (the two-sided alternative). The difference between these two alternative hypotheses is that the second leads to a larger p-value. One typical attitude (though generally incorrect) is to consider the one-sided H_1 simply because that is what the observed numbers seem to suggest. This is a biased attitude, and not "healthy" for an objective conclusion. Unless we have a very strong argument that the allele should be in excess in patients, we should go for the two-sided alternative, i.e. leave open the possibility that the allele A is either in excess (larger frequency) or in deficit (smaller frequency) in patients than in controls. In fact, if the allele A is in a significant deficit in patients, it means the complementary allele (that was denoted by the small letter a) is in excess, and that one can be called a risk allele for the disease, whereas A can be seen as having a protective effect.

The chi-square test is by nature two-sided, because the χ^2 number is calculated using the squares of the differences between the observed and expected counts, and thus the information on which one is larger and which one is smaller is lost. This is not the case for the Fisher test, where the p-value depends on whether we test the excess or the deficit of the allele A in patients. In the software implementation of the test this option is present. In R the option alt="g" means testing the excess (alternative greater) and alt="l" means testing the excess (alternative lesser), and in the absence of such option the two-sided test is performed, as shown in Figure 3.16.

We see now that the p-value given by the Fisher test (which is the most correct for our data) is 0.03 (one sided, greater version) and 0.044 (two-sided version). These numbers indicate that the effect of the allele A is marginally significant. A final conclusion is not really possible, and the result should be considered rather suggestive than significant. But it could be a basis to proceed with collecting more data and increasing the sample size, which is a common situation in real projects.

Chapter 4. Elements of Data Science – Infrastructure

4.1. The Processing of Data

4.1.1. Why Data Processing?

There is little to be argued about the usefulness of data processing and data based decisions in medicine and life sciences. For the last couple of decades, collecting data and using the results of the processing contributed massively to the quality increase of everything that is connected with the medical act. The large amount of data that are generated and used by the healthcare industry makes the traditional methods slightly obsolete. A Ponemon Institute survey found that 30% of the world's data storage resides in the healthcare industry, with a single patient typically generating close to 80 megabytes each year in imaging and Electronic Health Record (EHR) data. McKinsey estimates that this data, efficiently used by innovative processing pathways, can bring significant value, enabling, for example, a reduction of 300 billions USD in US only [22].

A great set of arguments about the importance of data analytics for healthcare has been published by Harvard Business Review in 2017 [23]. The key focus is on discovering and integrating new methods of data processing and analysis in existing operational procedures.

In Romania, the National Institute of Statistics estimated that in 2017 there were ~4 million patients admitted into Romanian hospitals with an average of 7 days of hospitalization [24]. In order to increase the healthcare system efficiency and make the correct choices, we have to be prepared to collect and process a volume of ~300 TB of new data every year.

To achieve the goal to increase the quality of care-delivery while achieving better efficiency and resource utilisation, we should be prepared not only to use the existing data for *Descriptive Analytics* (answering to the "What happened?" question [25]), which shows the existing status by describing and interpreting the historical data, but to perform *Predictive Analytics* (answering to the "What will happen next?" question [26]) using the same data to make predictions about the future. Then, the next step will be *Prescriptive Analytics* (answering to the " What should be done?" question [27]) that will suggest to us not only what options we have in terms of decisions, but also what will be their consequences.

This approach will lead to fast and accurate operational decisions not only in the clinical area (e.g. diagnosis and customized treatment for a patient illness), but also in the areas of healthcare resource administration and public health.

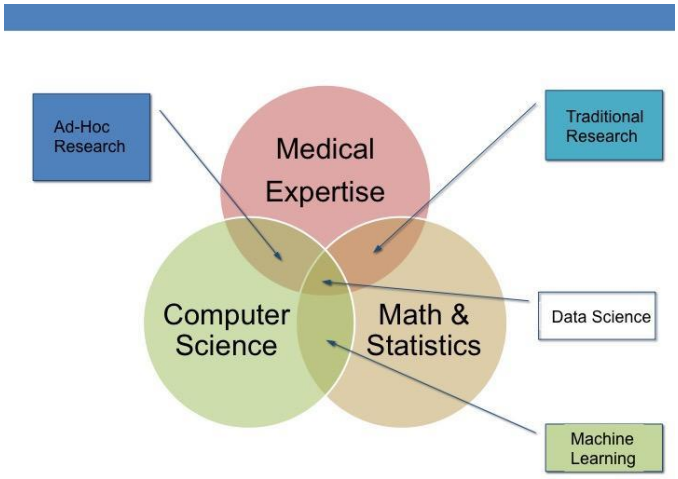


Fig. 4.1. – The Data Science Space

Processing the data requires the existence of multiple capabilities. In our case, this covers the medical domain, statistics, mathematics and computer science. The intersection of all the above creates the premises for a complete and efficient analysis.

As can be seen from the figure above, this is an evolution of the previous models that allow pairs of capabilities for traditional and ad-hoc research. An important innovative element brought by this approach is the focus of the process to make the research reproducible, enabling other groups to replicate the results and improve upon them.

The beauty of the model is that these capabilities can be seen at the individual or at the team level; someone working in this area can and should have knowledge from all the fields, but naturally the focus will be in one of them. In order to expand the area of knowledge, groups of scientists are congregating together for a common goal. This focus of the model on teamwork is another important step ahead, as it became clear that it is impossible to run sophisticated analysis without data scientists from different fields. This makes interdisciplinary training even more important than it was before, because now it becomes the glue element that keeps the team aligned to the same goal.

This organisational model has created new roles over the time. If initially, when the need for data processing exploded a decade ago, the term Data Scientist was coined to describe the unification of domain expertise with knowledge of statistics and computer science elements, later the term of Data Engineer was invented to describe the need for special focus on infrastructure and tools needed for big data.

4.1.2. The Data People: Scientists and Engineers

There are multiple roles that are involved in any data processing activity, the main ones being "Data Scientist" and "Data Engineer". As a matter of fact, the data science team is much larger, including Mathematicians, field experts, System Administrators, Database Administrators, Network Engineers and many others, but the first two roles are really specific.

What is a Data Scientist?

Briefly, a "Data Scientist" can be defined as being a scientist that works with data to get answers to industry or domain questions. Given this, this activity domain is for people that regularly:

- Are doing research to answer to industry-specific questions
- Use large volumes of data to provide the required answers to specific problems
- Prepare the data for use in inferential and predictive studies
- Explore the data to find hidden patterns
- Automates the processes for statistical studies
- Presents results to decision-makers

Their goal is to create innovative models and explain the benefits of applying them in their scientific domain. But in order to do so, they need a whole infrastructure to support their research, and this is where Data Engineers come in.

What is a Data Engineer?

A Data Engineer creates, designs and builds the infrastructure for data analysis. The activity domain for these people involves:

- Developing, building, operating and maintaining architectures and solutions for data processing and storage
- Aligning architectures and solutions to the requirements of data processing processes
- Discovering new ways to acquire data
- Developing and implementing processes for data cleansing, data modelling, data mining etc.

- Recommending procedures for improving the quality, efficiency and security of the data

There are multiple proposals about how these two functions work together inside a specific workflow, and there are multiple workflow proposals. Using as a reference the one proposed by Berkeley School of Information [28], we can observe the tight cooperation between the two functions and how much involvement each one has in different stages of the life cycle of a data science project. The figure below shows how responsibilities are split (by representing them in different colours: blue for Data Scientist and orange for Data Engineer) between the functions.

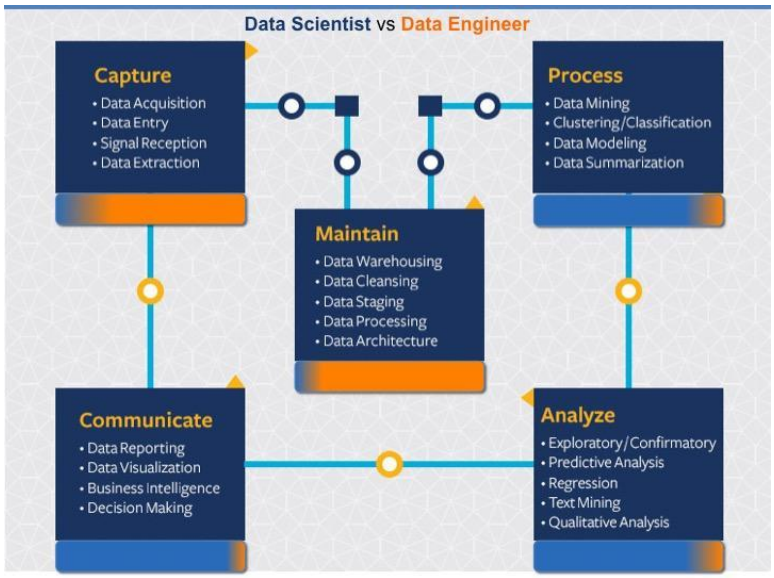


Fig 4.2. - Data Scientist and Data Engineer cooperation in a project

The data science premise is there, but the healthcare industry has not yet embraced the model on a wide scale. One reason this has not happened is the healthcare industry's requirement for stability and regulations above all. These two requirements will only be fulfilled when decisions resulting from a data science project are fully transparent and results are reproducible across multiple organisations with different underlying processes.

Another factor is the lack of internal demand, different organisations preferring to launch ad-hoc projects and to use existing experience and skills in making the decisions, instead of bringing together new, interdisciplinary teams. The lack of trained specialists and their relative unavailability on the labour market further increase the difficulty of organising new teams and starting data analysis as a sustainable process.

However, more and more regulation bodies are starting to take data processing seriously and to make recommendations for its usage. The American Hospital Association has already introduced for a while the data processing as one of the “must-do strategies” and “core competencies” that hospitals must adopt in order to survive in a value-based economic model [29].

4.2. Computing, Storage & Communication Systems

In any information technology system we will find elements from three different groups, computing, storage and communication, which contribute together to the realisation of that specific system. Their roles are clearly defined comprising processing, storage and transportation of data between different systems or locations.

4.2.1. Elements of Computing Systems

The computing part of the infrastructure is built with special elements, unanimously called *servers*. There are always two components, sometime seen as being different entities, which build a server:

1. the hardware server: the physical machine that runs one or more applications
2. the software server: the application that makes data available to other entities (e.g. users, computers etc.)

The fundamental three parts present inside of any computing device are:

1. CPU: the Central Processing Unit (often shortened as *Processor*) responsible for calculations
2. Memory: the entity responsible to hold the operands needed for the calculation and to keep result of the operation performed on them
3. I/O: Input/Output, the entity responsible to bring the data inside the system for processing and deliver the results to the outside world.

Even though it follows the same internal Von Neumann architecture and it is not different in principle from any other (desktop-format) computer, the hardware part of a server is a highly specialised piece of technology in terms of:

a. *size:* the standard way of mounting the servers is in 19-inch wide metal cabinets, called *racks*, with height expressed in Rack Units (RU, or simply U; 1 RU = 1.75 inches or 44.45 mm), the standard height

being 42 RU. However, multiple other dimensions exist, between 7 and 44 RU. Racks can be installed on the floor or on the wall, depending on the specifics of the site. The depth of a rack is between 600 and 1200 mm, depending again on the specific usage.

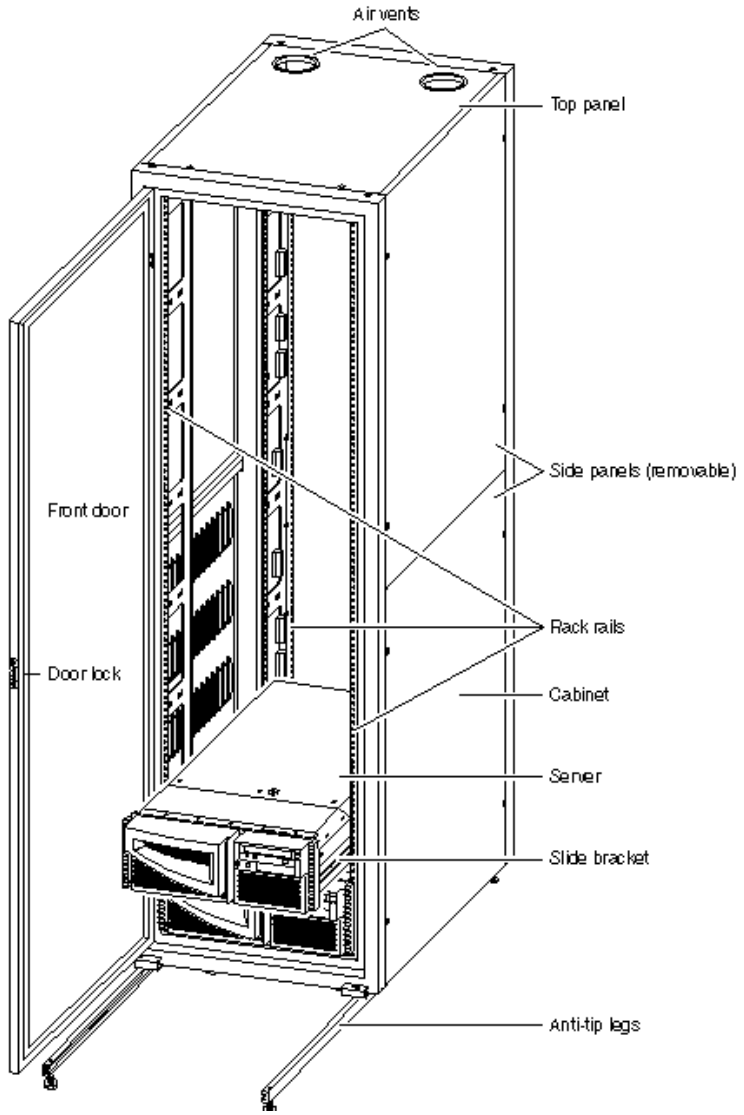


Fig. 4.3. - Rack structure

b. *form-factor:* usually, servers come in one of the following 3 form factors:

   1. tower: less utilised, the server looks like a (bigger) desktop, being built to sit on the floor, potentially in a dedicated space

2. rack-mountable: very popular, they come in standard 19" width, 1 to 4 RU height with different depths, depending on the internals of the server. Sometimes they are colloquially called "pizza-box" servers.
3. blade server: specialised packaging with 12 to 18 servers with different vendor specific dimensions, inserted into a single common chassis, of 19" and 4 to 7 RU height. This type of server is specially built for large data centres, in order to deliver a very efficient usage of existing space.

Examples of the rack-mountable and blade form factors can be seen in the figures below.

Fig. 4.4. – Rack-mountable hardware server

Fig. 4.5. – Blade Center server

c. *performance:* the servers are built to deliver high performance for the applications running on top of them. This refers to how fast the data can be extracted from memory, processed by the CPU and the result written back. The assumption is that the data is already stored in memory, leaving aside the important process of ingesting the data from storage, which itself brings an additional layer of complexity.

Delivering high performance is influenced not only by the type and quantity of CPUs and memory available but also by the way these elements are connected together.

The continuous increase in the quantity of data that needs to be analysed further increases the need for computing performance. As an example, consider the case of running genome-wide associations for a sample of 5000 patients. Suppose that we will use imputed genotypes that need less space than a full genome (which usually uses 80–100 GB per individual). The size depends on the number of individuals and imputed markers; for 5000 individuals the data size will be around 100 GB. For the computation, for 5000 individuals, a typical case-control association for a single phenotype needs 150 CPU-hours. Running on a single-core server, it would take approximately 6 days to complete the run, but running the same analysis on a typical computer that has 4 cores, it will take about 40 hours. In the ideal case, running it on 10 nodes, each one with 4 cores, will give a result in 4 hours.

Based on this need there is a whole new research area in developing systems that deliver high performance, called High Performance Computing (HPC).

The following two features becomes visible when many servers are installed together in a Data Centre:

1. *power consumption:* this is a point where the difference between servers and other general-purpose computers is significant. Servers are optimised to convert as much of the electrical power absorbed from the power grid into useful computation power. When many systems are put together, the power consumption becomes significant and the optimisation of each individual element for power efficiency becomes important.
2. *availability:* servers are built to stay running for as long as possible. As more than 90% of critical defects come from the power sources and overheating, they have redundant power supplies and redundant fan modules, many of them replaceable without shutting down the machine.

On the facility level (Data Centre) the metric that tracks power efficiency is called *Power Usage Effectiveness* (PUE) and is defined as the ratio between the total energy absorbed and the total energy used for IT purposes. This happens because part of the energy is consumed for other purposes (e.g.

cooling the machines to keep them in a temperature range that gives maximum life span). This metric is expressed as a number, higher than 1. The closer to 1 the PUE is, the better efficiency we have on the facility level.

We can also compute the *Availability* of the system, calculated as the ratio between the up-time (the time when the server is functional) and the total time. Availability is usually expressed as a percentage, for a specified interval of total time. As an example, an availability of 99% per year means that the system is available for ~361 day, which leaves only 4 days per year of unplanned interruptions.

4.2.2. Elements of Storage Systems

Storage is the subsystem that takes care of keeping the data ready for computation. Despite the simple definition, it involves the usage of different storage media for keeping the information available for the required period of time, which varies by application. This means that, depending on the requirements, we need to use different types of memories, built with different technologies, which will deliver the performance needed, described by access time and capacity.

From a functional point of view, the two main types of memory are:

1. *Volatile* memory is the computer storage that only maintains its data while the device is powered (e.g. RAM).
2. *Non-volatile* memory is the type of storage technology that does not require power to retain data (e.g. HDD, SSD).

Memory of the first type is usually the faster one, used directly on the motherboard, with very high-speed connectivity with CPU. Here the speed is measured in *cycles-per-second.* For example, if a RAM module is rated at 2400 MHz, it performs 2.4 billion cycles per second. The bigger the number is, the faster data can be stored and read. The base memory package is the RAM module, usually with a capacity of 8–64 GB, but a server can accommodate many modules, adding up to 1 TB of RAM.

Memory of the second type is slower, but has bigger capacity. The memory speed is usually limited by the connection bus and the size is limited by technology. The basic element is the disk (HDD = Hard Disk Drive or SSD = Solid State Drive) with usual capacities between 128 GB and 10 TB. A storage system can accommodate many disks (SSD for speed, HDD for

capacity), allowing it to store up to 1 EB (1 Exabyte = 1000 Petabytes = 1 million Terabytes or 1 billion Gigabytes).

The following diagram shows the relation between size and speed for different types of memory, including the atomic organisational unit for data storage at each level.

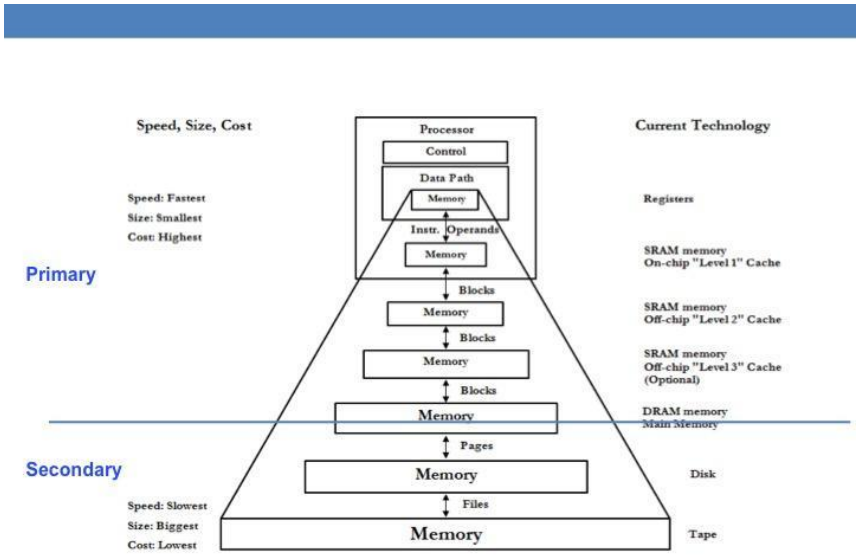


Fig. 4.6. – Memory hierarchy

Memory is a critical sub-system, as it holds all the data for processing. It is crucial that the data not be altered, intentionally or not.

Keeping the data safe from intentional tamper is a major concern of the industry and a lot of measures are taken on different levels. Encryption is typically used when data is at rest (e.g. stored on a disk or tape) or when it is transmitted between systems. In other, special cases encryption of data is even more pervasive.

Another concern is to keep data safe from unintentional change (e.g. bit flipping due to natural cosmic radiation). Depending on the technology used, different methods of error detection and correction, in the form of special coding of data, are used.

The above approaches are applicable to the basic units of a data storage system (RAM modules and HDD or SSD disks). In order to keep data

available and recover it from faulty elements, higher-level data-integrity solutions have been developed. In one of those, many disks are connected together in a Redundant Array of Independent Disks (RAID), either increasing the total capacity (e.g. RAID0) or creating copies of the data on the fly, mirroring the disk (e.g. RAID1). Other, more complex RAID levels have been invented to simultaneously provide an increased capacity and data redundancy.

The technology used to build the storage elements is also an important factor for data safety. If the purpose is to store the data for a very long time (e.g. in longitudinal studies, for years or decades), then proper physical support and proper technology have to be chosen:

- HDDs have moving parts (platters and magnetic heads), so it is expected they do not have a sufficiently long life span. Different studies from the industry (e.g. the Backblaze 2012 study [30]and the 2018 update [31]) suggest a median lifetime of 6 years. (in 6 years, 1/2 of the drives are dead.) Of course, this depends on a lot of factors (vendor, temperature, usage patterns), so, in some conditions, the lifetime of a HDD can be much shorter (or longer). A study [32] about the reasons for HDDs failing that has been published by USENIX leads to the same conclusion.
- SSDs do not have any mechanical moving parts, using flash memory instead to store persistent data. This characteristic gives faster speed of access to data and more resistance to shock. On the other hand, using NAND flash memory means that each cell is slightly worn down on each write operation, which gives an upper limit to the number of times a cell can be used reliably. Cell wear is much higher on writes than on reads, so memory usage patterns in applications affect drive's lifetime. Additionally, the temperature and the type of cell (single- or multi-level) are also important factors in establishing the lifetime of an SSD. Even if wear factors are different, the expected life span should be similar with HDDs (see the Backbalaze SSD comments [58]).

The above data refers to disks in use, i.e. powered-up and connected to a live system). When the disks are not in use, their lifetime can change significantly. As an example, an SSD left without power, due to its mechanism of storing data based on electrical charges in cells, starts losing data after ~2 years. Newer technology may improve this time interval, but this may come at a significant cost.

However, the real concern is data storage for long time periods (a.k.a. data archival). In the long term, due to natural causes (e.g. cosmic radiation) data at rest may (and will!) be altered. The most well known phenomenon is the one called *bit rotting*. HDDs are affected by this data degradation issue when small regions of the disk lose their magnetic orientation, causing the stored bits to flip (change from 0 to 1, or from 1 to 0). Similar things happen to SSDs when the electrical charges in a cell leak.

An example of how an image may be altered by 1, 2 or 3 bit flips can be seen in the figure below (authored by Jim Salter [34]) where the image alteration become very visible when only 1, 2 or 3 bits are changed, respectively (the leftmost one is the original image).

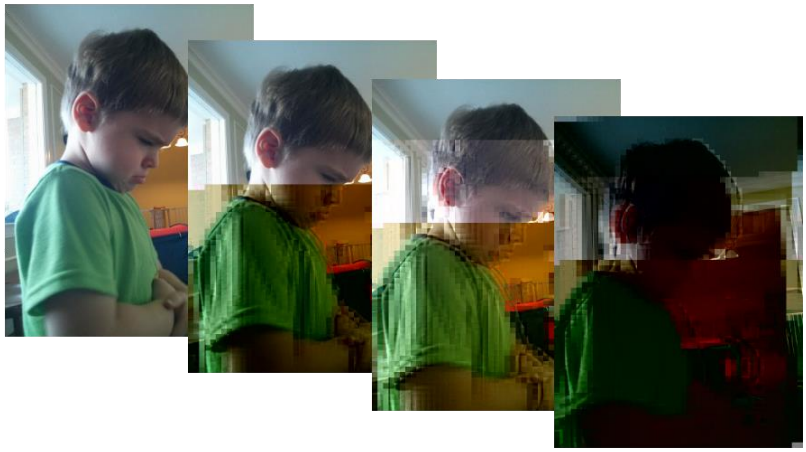


Fig 4.7. Bit errors effect on picture

Optical disks (CD-ROM or DVD) are also prone to data degradation, although the life span is better than HDD and SSD, reaching tens of years on M-disks.

This suggests that *all* storage media are intrinsically unreliable on the long term (in terms of keeping data unchanged), due to various physical processes. Therefore, the discussion moves to compensating for this degradation. There exist methods to compensate for errors that appear in generic data streams, and some of the same techniques can be applied to storage. But keeping data available for long periods (archiving) is much more complicated than simply protecting it from isolated errors. This involves physical security of the media, data format, media physical format and other special measures, which will not be discussed in this course.

Archiving is a useful operation for longitudinal studies running for tens of years, but *backups* are more important for day-to-day activities. A backup is usually defined as a copy of a data set created to be used for the recovery of the data from any corruption incident. Typically they are made in a regular fashion, on fixed time intervals or when the data is changed. As they are copies of data, there are different management processes to organise backups.

On live data, backup operations, whatever policy may govern them, have and intrinsic flaw: data created between the current moment and the snapshot of the last backup is not saved, hence not protected. This is so important that a special metric to keep track of it has been created - Recovery Point Objective (RPO) - defining the length of interval of time in which data can be lost.

In special cases, for important data, there are ways to ensure that every time a chunk of data is written to a disk, a copy of it is also written on other disks. This enables having a copy of the most recent data, so RPO = 0. One way to accomplish this is through using RAID (Redundant Array of Independent Disks). In fact, RAID is a mature, complex technology that can do much more than simple copies of data in real time, and good overviews of it can be found relatively easily.

4.2.3. Elements of Data Communication Systems

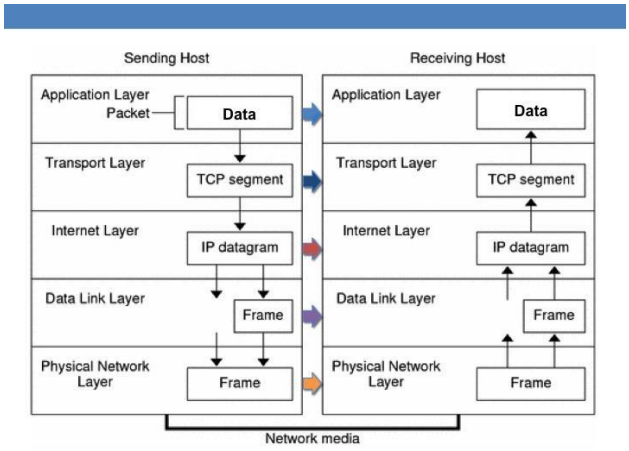
In the modern world, useful data needs to travel between sites, unchanged. Data communications technologies have evolved constantly in the last fifty years, mainly in terms of speed (total number of bits transmitted per second) and quality (number of erroneous bits transmitted).

In order to achieve good performance, a structured approach has been taken, developing through the years different *protocol stacks*, i.e. sets of rules for each entity involved in the communication process.

The most prevalent communication stack is TCP/IP. It started as a Department of Defence project, it was later used in more and more universities and companies, until it was adopted ubiquitously with the Internet era.

The name comes from the 2 most important protocols of the stack: *TCP* = *Transmission Control Protocol* and *IP* = *Internet Protocol*. These handle the data in transit from one system to another over a network. For the purpose of this discussion you can view the network as multiple nodes connected together with links, like a graph.

The key concept is to split data into manageable chunks on the source and send the data from node to node up to the destination, as you can see in the figure below (adapted from Oracle).



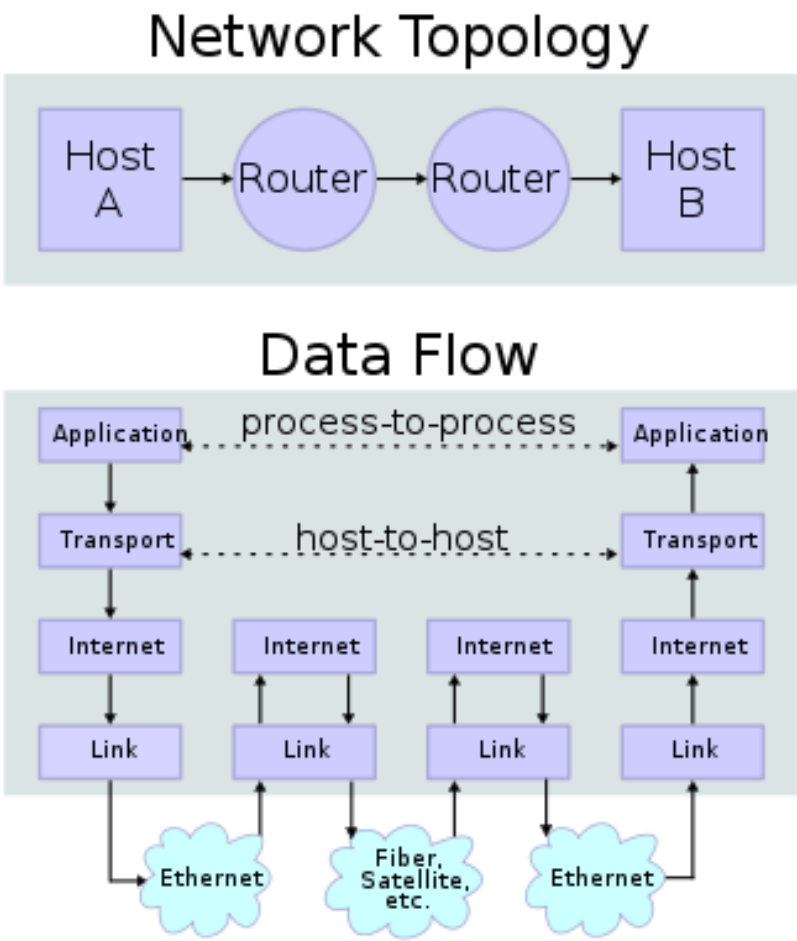
4.8 - The TCP-IP Model

The layering shown in this model is needed to *mask* the differences that may exist between the elements present on both ends (e.g. different connectivity media), as well as the usage of an intermediary node (multiple locations).

We need to use intermediary nodes (usually called routers) because it is impossible to have direct connections between all the systems on the planet.

The trick is to see each level of the stack "talking" with the corresponding level on the next node, be it an intermediary node or the destination. An illustration of this concept is in the figure below (source: Wikipedia).

As an example, assume that the system on the left (a Source email server) has a message that has to be delivered to the system on the right (a Destination email server). They need to establish a direct communication between them in order to deliver the message from one to the other. Because the systems may be different (e.g. different software application, running on different hardware systems with different Operating Systems) a special set of rules (the protocol) has been agreed by standardisation bodies (as The Internet Engineering Task Force - IETF) to make the communication possible on the highest level, the *Application* one.



4.9. IP stack connection between two applications

The protocol uses special messages and additional data to allow intermediate nodes to route the application data (in our case, the mail message) correctly. So, the email is "packed" in protocol units and passed to the level below as a raw stream of bytes.

The next level, the *Transport* one, takes care of the correct transmission (all the parts reached the Destination and are assembled in the correct order) of the message bytes to the destination system, unaware of the actual paths and any intermediate nodes on the way there. It splits the stream into segments of various lengths, adds its own metadata and pushes each segment to the level below.

The next level down, called *Internet* or *Network*, takes each data segment and tries to send it to the destination, finding a suitable route (defined as a set of intermediary nodes) for this. Different segments can take different routes from the Source to the Destination, so the role of this layer is only to have them delivered, letting the *Transport* layer assemble them in the right order. If needed, the data segments can be further split into smaller packets, padded with additional, specific metadata, and then pushed down to the level below.

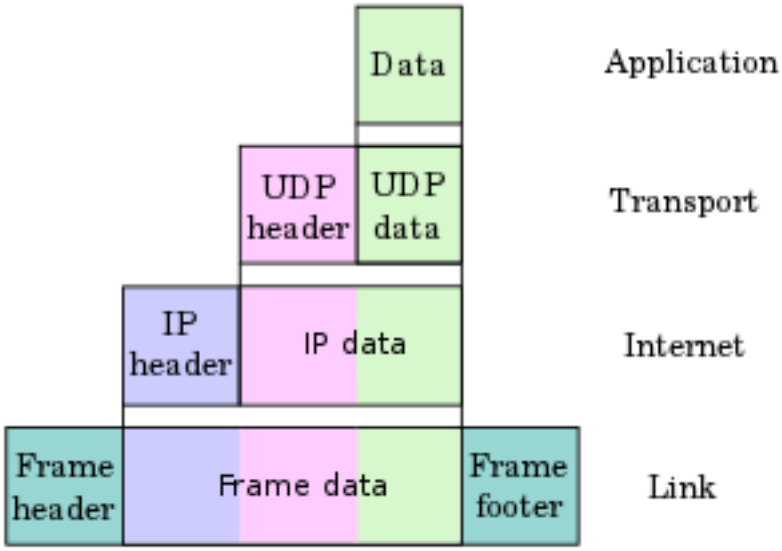
The *Link* layer is the one responsible to use whatever physical connection media is available on site (like a Fiber Optical link, or an Ethernet, electrical one, or a wireless/radio link) to transmit the data received from the *Network* layer in an appropriate way (e.g. as electrical signals on copper wires, or as modulated light beam on FO) to the next node. As you may have guessed, it uses its own metadata, added to the data received from the layer above, to signal to the next node how to interpret it.

Once the data is received on this next node, the stream of bytes received by the *Link* layer from the previous node is assembled, interpreted in accordance to the specific layer metadata and the result passed to the layer above. The layer above does the same (interprets its own layer metadata and processes the data received accordingly) and passes the result above again, up to the last level present in the node.

The process is repeated on every intermediary node up to the *Network* layer, passing the user data from node to node up to the Destination. On the destination, the data is processed up to the *Application* level, where the original user data (the email) is supposed to arrive.

The communication between layers of any adjacent nodes, including the metadata significance and the processing rules, is described by specific protocols of that layer. A representation of the metadata used by each layer, on top of the user data (represented in green), can be seen in the figure below (source: Wikipedia).

A really good document explaining in detail the TCP/IP stack is one of the IBM Red Books, *TCP/IP Tutorial and Technical Overview*. It has 1004 pages, so it is either for the very passionate, or for those that need a reference from time to time. A much shorter overview of TCP/IP stack can be found on Wikipedia [60].



4.10 UDP Encapsulation

3.2.4. Distributed Systems

The true power of data communication is actually seen when 2 or more compute nodes (including here the associated storage) are linked together, forming a distributed system.

A distributed architecture includes computing elements (nodes) that reside in different locations (physical or virtual systems) and coordinates application actions with messages that are transmitted over a data network to accomplish a common goal.

Notable characteristics of a distributed system are:

- It does not depend on location
- Each node has its own memory
- Nodes interact together by exchanging messages
- The system tolerates individual defective components

The most common architectures of distributed systems are:

1. Peer-to-peer
2. Client-server
3. N-Tier

Peer-to-Peer (often abbreviated P2P) systems allocate no special roles to any of the participating compute nodes. All the responsibilities are equally split between nodes, each one being, simultaneously, a provider and a consumer of data (e.g. file sharing networks like BitTorrent). Each entity makes available to the community a part of its own resources, contributing equally to the process. The nodes are connected together, in a partial mesh fashion, with virtual links (incorporating one or more physical links), as can be seen in the figure below (source: Wikipedia).

Fig. 4.11. - P2P network

The *Client-Server* architecture is a different type of distributed system where the provider of a resource (called *Server*) is separated from the consumer/requester of that resource (called *Client*). The resources are made available via services, and the server is the entity concerned with making the resources available to the client. The Client and the Server communicate one with each other over a data network, potentially running on separate hardware, in different locations. However, it is not uncommon to have both residing on the same system.

The hardware that runs the server application(s) is sometimes called *host* and shares its resources (through the server applications) with the clients. The client does not share any of its resources, but uses the server resources or services, by initiating the communication with it and making the appropriate request. On the other hand, the server expects these requests in a specific manner, defined by a specific protocol.

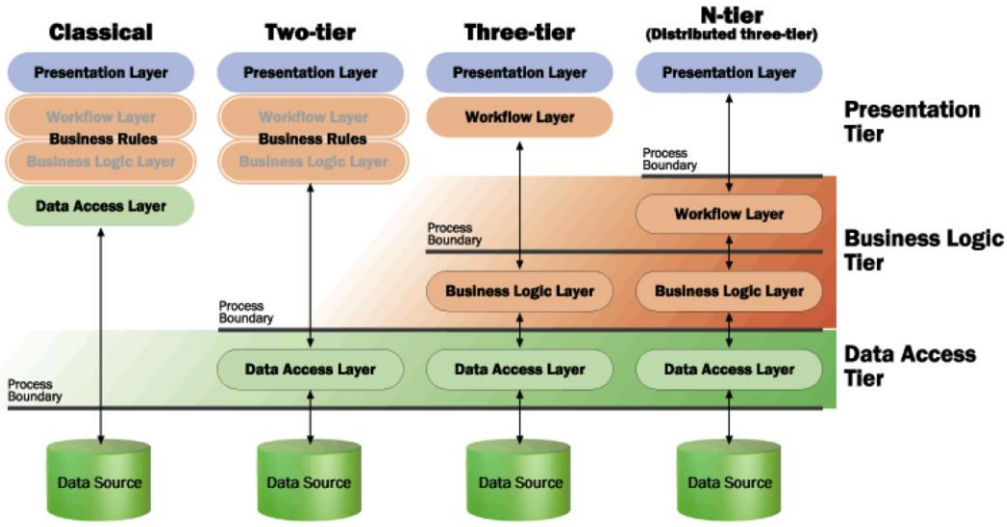
Whether a computer acts as a Server or a Client is entirely dependent on the nature of the application that requires the service (e.g. mail server, web server, file server, etc.). To accomplish the functions, clients connect to the server using a request-response model, by sending request messages to the server and receiving responses from it. A generic model is figured below (source: Wikipedia), where many clients are connected to a single, central server.

4.12. Client-Server network

The N-tier model (or multi-tier architecture) is a form of Client-Server architecture in which functions are separated and split between the client and the server. Most of the time these functions are organised, by purpose, in presentation, business/applications and data management, forming what is known as "3-tier architecture". A graphical presentation of a generic N-tier architecture is in the figure below.

You can see how the split of the functions between different systems is done in different N-tiers systems. The concept of tier is more connected with the physical aspect of the infrastructure, potentially equivalent with the hardware node. So, we can assume that in a 3-tier system we have 3 different hardware nodes that run different functions needed for the application.

This flexibility and the potential to reuse functions for different tasks have made this model popular. Even further, the upgrade and maintenance of the application is easier and faster, as developers can act on specific layers of the application, instead of rewriting everything.



4.13. N-tier system models

A special subclass of distributed systems is *clusters*. These are tightly connected (low latency, large bandwidth), mostly homogeneous (using the same hardware and software) array of computers, located usually in a single site and viewed as a single system. The members of a cluster are called *nodes,* each one being a server running its own operating system.

Clusters are dedicated to a small number of well-defined tasks, in solving of which the cluster acts as one single entity. Each node in a cluster performs the same task, as instructed and controlled by the software, mainly for performance increase and for high availability (see figure below).

Remind yourself of the previous example where we considered a typical case-control association for a single phenotype for 5000 individuals that takes 150 CPU-hours. The task can be completed in 150 hours (~1 week) on a single CPU, or it can be done in 1 hour on a cluster with 150 CPUs.

4.3. Operation Systems

We define the *Operating System (OS)* as being a software set administrating the resources in a computer system and making them available to applications.

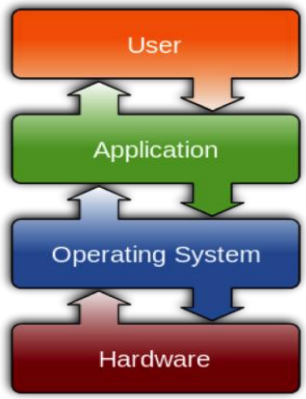
As we have discussed already, the usual hardware resources are:

- Memory
- CPU
- I/O (devices)

For all of these, the OS acts as an intermediary entity between them and the applications, managing:

- the file system
- the applications running on the system
- network communication

The figure below gives a graphical representation of the relationship between the user, the application, the OS and the hardware resources (source: Wikipedia).



4.14. Operating system's stack

Operating systems are complex and an in-depth discussion is beyond the scope of this course. For reference purposes, see below the evolution of the Unix systems (source: Wikipedia).

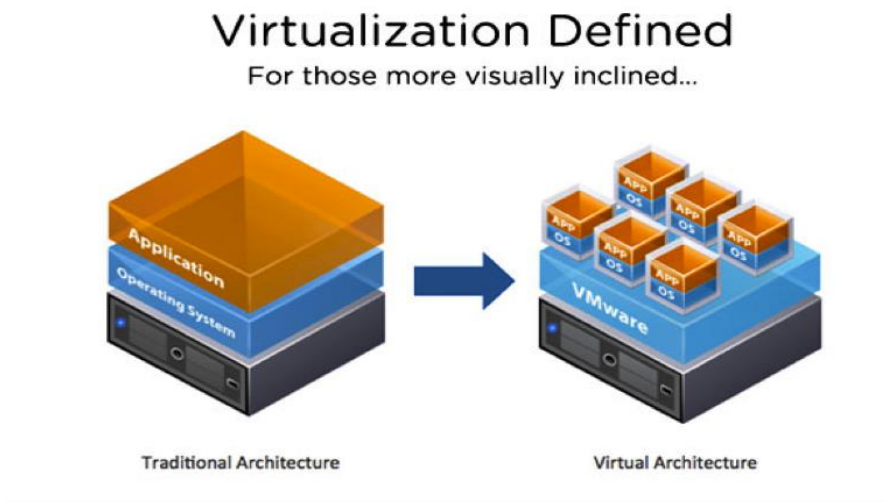
4.15. Unix Operating System map

Coming back to the first figure, the position of the *OS* in the software stack suggests that it makes invisible to the *Application* the details of the *Hardware.* This is indeed one of the roles of the Operating System, to mask to specific (set of) applications the characteristics of the hardware they are running on. This makes it possible, for example, to run Microsoft Word (an application) on a wide range of different hardware platforms that support Windows (an OS).

This observation leads us to the conclusion that we can expand this decoupling between the existing, real resources on the hardware level and what an application receives. The mechanism that allows us to do that is called *Virtualisation.*

Virtualisation is about grouping and abstracting the resources and services in such a way that the real nature and physical limits are hidden from the users/applications. In fact, the real resources are hidden from the Operation System itself. Virtualisation happens using an additional layer, placed between the OS and the Hardware, called the *Hypervisor*. A similar effect can be obtained by using a full host OS, instead of the hypervisor, but the efficiency of the overall construction will be lower.

So, the *Hardware (full) virtualisation* allows one to build a “virtual machine” seen by the OS as a real computer, with its own resources, partitioned from the physical resources of the real hardware that runs it, as can be seen in the figure below (source: Vmware).



4.16. Server Virtualization Technology

As you can see, on top of the hypervisor that runs directly on the hardware, there are many VMs (virtual machines), each one running its own OS and its own applications.

This approach has 2 important implications:

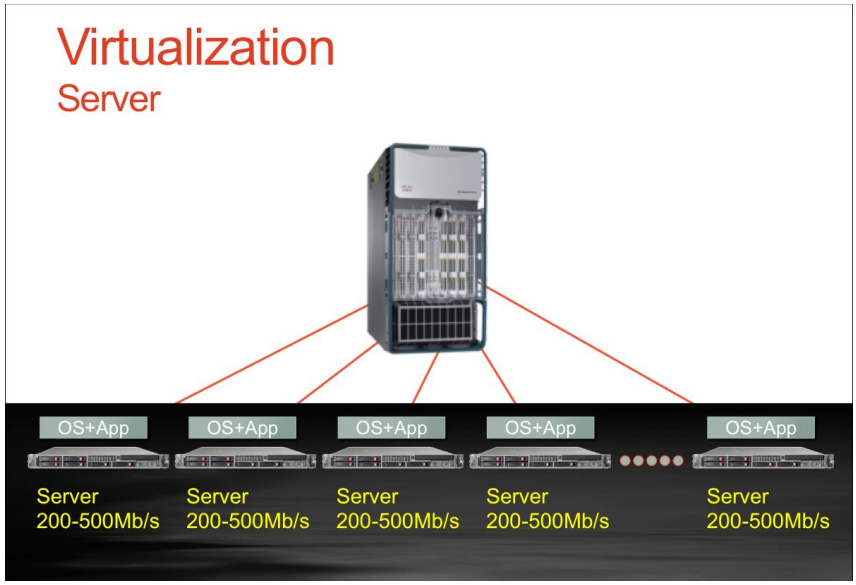
1. on the same hardware we can run multiple VMs, using the hardware (CPU, memory) more efficiently. Some sources cite a doubling of the CPU utilisation, from 40% to almost 80%.
2. there is no coupling between the VM and the actual hardware, allowing us to move the VM to another (potentially different) hardware system if required.

The observations above pave the way to a completely new way of managing the workloads of an application. Now, with the virtualisation model, the application can be moved to the location where it runs most efficiently and it can be started (and shut down) only when needed. This means that we can have an application (a server) running wherever and whenever we need it, independent of some fixed hardware and its physical location.

4.4. Virtualisation

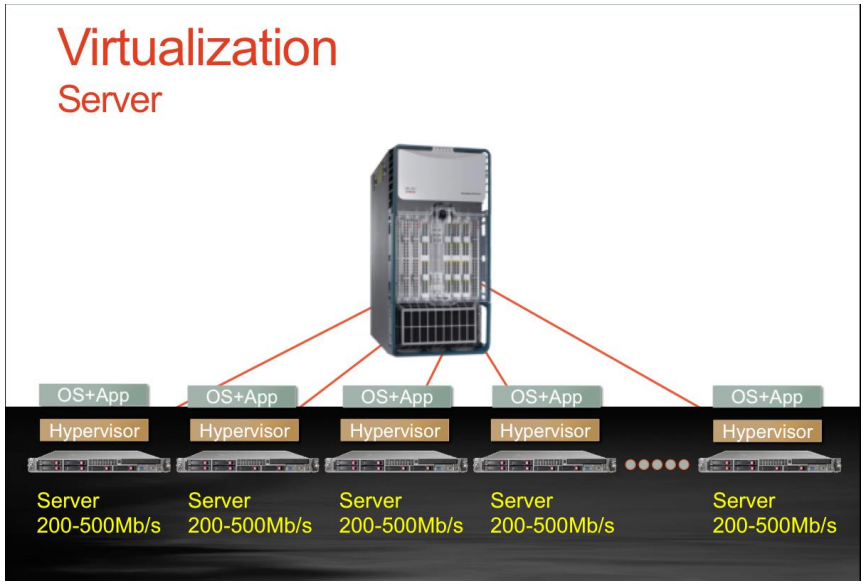
Virtualisation is a very important concept that has lead to the profound transformation of the ICT industry in the last decade. The decoupling between the Virtual Machine (the ensemble of an application and the underlying OS) and the hardware that runs it opens the door to a completely different type management of the server workloads.

Before the virtualisation era (or even today when no virtualisation is involved), a physical server can run one or more applications, and it is connected to other servers by a data communication system in a structure similar to the one figured below.



3.17. Traditional servers (bare metal)

As you have seen before, the virtualisation operation means the addition of a *hypervisor* between the hardware and the Operating System (see below), creating the abstraction called Virtual Machine (VM). The purpose is to create the premises to manage the VMs as independent objects, having no relation with the hardware, making it possible to run them on any available system.



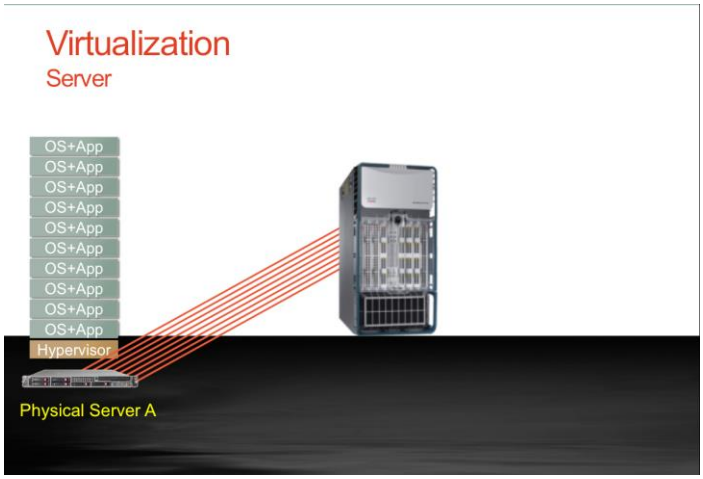
4.18. Servers with Hypervisor installed

The hypervisor takes a small amount of the system's resources for itself (CPU and memory), but modern systems have special CPU architectures built for virtualized systems and enough memory to make this a non-issue. However, each VM needs dedicated CPU and memory, so the hardware systems hosting VMs should be properly equipped with both.

As the virtual machines are now independent from the hardware they are using, it means that we can move them around on whatever server has the resources (usually memory and CPU) available for it. Let's assume that, for our set of systems described above, we add CPU and memory to the leftmost server. Then, instead of running one VM per server, we move all of the VMs on this one server, removing the rest of the hardware, as in figure 4.19.

Now we have a bigger server that runs all the applications we had before distributed on different systems, with the advantage of a simpler management and a better efficiency.

As a side note, the efficiency is measured through a dedicated metric PUE (Power Usage Effectiveness) that shows how much energy is consumed outside the computation (e.g. in cooling).



4.19. Single physical server hosting multiple VMs

Each one of the virtual machines has its own share of resources (CPU, memory, data network) and functions as if it were the only one installed on the hardware server.

This type of system is typical for all virtualized environments: efficient hardware with plenty of resources shared between multiple VMs. It allows better usage of electrical power (because less cooling is needed) and easier management of the applications, powering them up and down as necessary.

Now consider the case when we have 2 similar systems as the one described above, each one running its own set of VMs, as figured below.

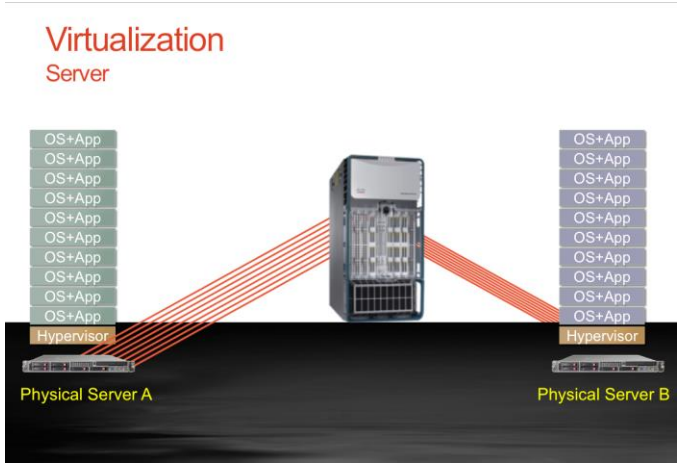


Figure 4.20. Two hardware servers systems with multiple VMs

In this setup, we have 2 similar systems, running their own sets of applications, each one in a separate VM.

We said earlier that the main advantage of virtualisation is breaking the relation between the application and the hardware, allowing us to run multiple VMs on the same system, completely independent from each other. But virtualisation gives us even more flexibility: as the VM is independent from the hardware, it means that we can take VMs from the left side server and move them to the right side one.

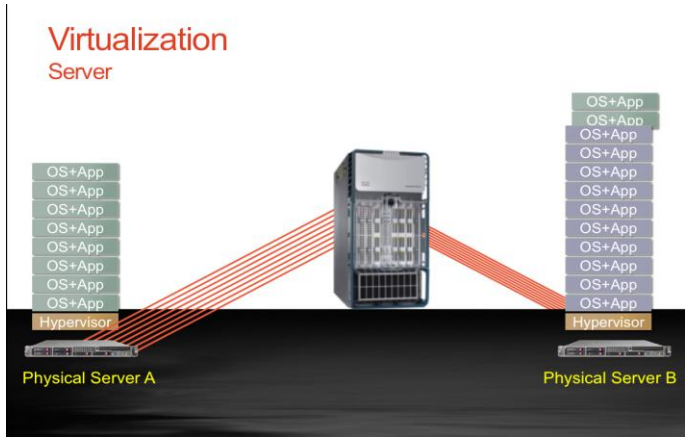


Figure 4.21. Moving VM between physical servers

Moving VMs around the hardware has multiple use-case, the main ones being listed here:

- high availability: move the VM to another server when the original one has failed, keeping the application functional from the viewpoint of an external user
- better efficiency: move the VM to the least loaded system
- management of energy consumption: move the VM to the cheapest energy source
- automation of VM management: VMs can be started, stopped or moved as needed by the applications.

From this perspective, the virtualized systems (hardware and software) are usually installed (hosted) together in special environments where we can take advantage of all the benefits. These special locations are called Data Centres.

4.5. Data Centers

By definition, a Data Centre is a facility hosting computing, storage and communication elements with the goal to offer ICT services to clients. A set of mandatory regulations exists (specified by standards such as TIA-942, GR-3160) in respect of data connectivity, electrical power sources, internal cabling, heating, cooling and many others. The aim is to create an environment where the applications can be run efficiently, as needed.

In this context, aspects of cooling or electrical power consumption that might not be obvious for a single system become clear when talking about thousands of systems installed together in a Data Centre. Spectacular examples of how a data centre is organised can be seen in the virtual tours published by Google or other providers.

A example of a typical physical topology for a Data Centre includes:

- 4–6 Zones with 6–15 MW consumption per DC
- 5,000–8,000 m^2 per zone with 1–3 MW per zone
- 200–400 racks/cabinets per zone
- 8–48 servers per rack/cabinet with 1–1.5 KW per cabinet
- cooling and power per pod (per pair of rack rows)
- pair of cold isles and hot isles (the regions where the cold air, of about 20°C, is inserted and hot air is extracted for cooling), as the servers have front-to-back air flow

This set-up is represented in figure 4.22.

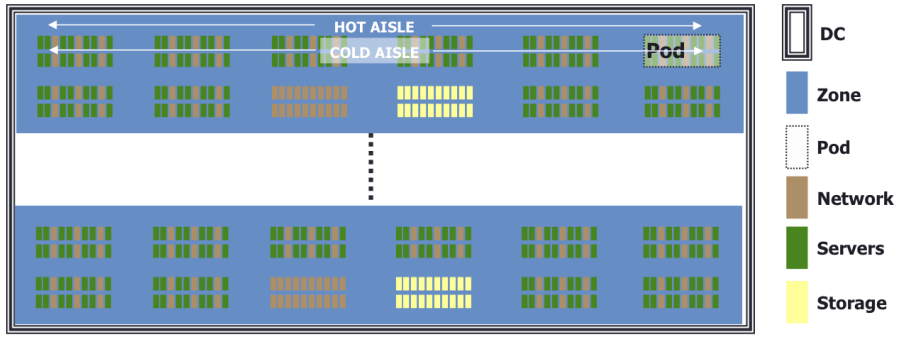


Figure 4.22. Physical topology for a Data Centre

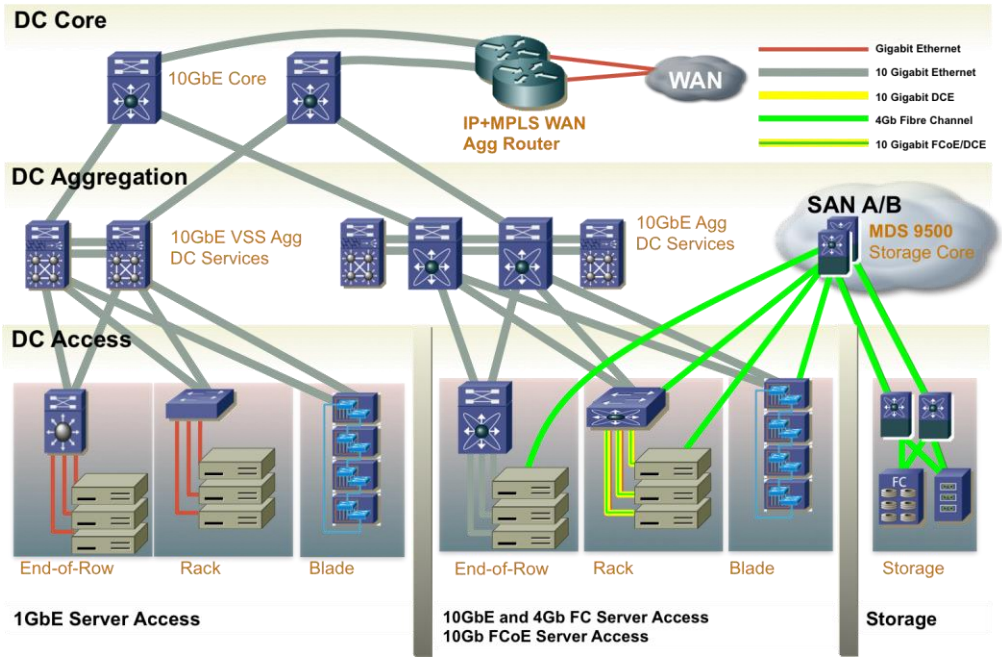


Figure 4.23. Data Center Logical Structure

There exists a viewpoint different from the physical one (presented above): the logical structure of the DC, where the computing, storage and data communications are presented. For such a view, see the figure 4.23. Each element has its own role, being connected to the others following specific rules, all to create a functional and manageable structure.

A Data Centre has the single purpose of keeping safe and functional the systems inside, which in turn keep applications alive. The owners of the hardware and the software in the infrastructure need not be the same as the users of the services offered by the applications running in the DC.

From this perspective we can have privately owned data centres, running applications used by the owners of the DC (e.g. a bank DC running their own banking apps), or data centres that offer services to external entities, either free of charge (e.g. Google Docs) or for a fee (e.g. Microsoft Office365).

4.6. Cloud Computing

As we saw above, the software applications running on systems in Data Centres implement services with a specific purpose, for specific clients (e.g. mail service on Google or Yahoo).

Cloud Computing is a consumption model of ICT resources and services, decoupled from the physical infrastructure needed for delivery and offered *by request*, in an *elastic* way, to a *multi-user* environment, on a *pay-per-usage* basis, i.e. only for the amount used.

There are some essential characteristics for any cloud service:

1. it is a *measured* service – all resource consumption is measured with an accuracy suitable for charging a fee
2. it is an *elastic* service – it can be scaled up (more resources) or down (less resources) as needed by the user
3. it is *on-demand* **–** the user can start the service when needed and can stop it when it is no longer needed
4. it is *resource-pool-based* – any new instance takes the resources needed from the corresponding pool (e.g. CPU, memory, storage, Input/Output) and returns them there when the instance is destroyed. The same happens when the instance grows or shrinks.
5. it has a *broad network* connectivity – it can be accessed via communications network (e.g. Internet, for the public services) from virtually anywhere

A cloud service has 3 different service models:

1. *SaaS (Software-as-a-Service)* offers a software product as it is (e.g. Microsoft Office365) and the user may use the product without any possibility to modify it outside its regular settings
2. *IaaS (Infrastructure-as-a-Service)* offers a virtual server (VM) with the desired hardware configuration and OS (e.g. Amazon EC2) and the user has total control over it
3. *PaaS (Platform-as-a-Service)* offers a virtual machine with additional elements that allows complex services (e.g. Google AppEngine) that are configurable by the user according to specific needs.

In all these models the user does not own, control or manage the hardware infrastructure beneath the service, but only uses it within the limits of the model. The infrastructure belongs to the organisation that offers the services.

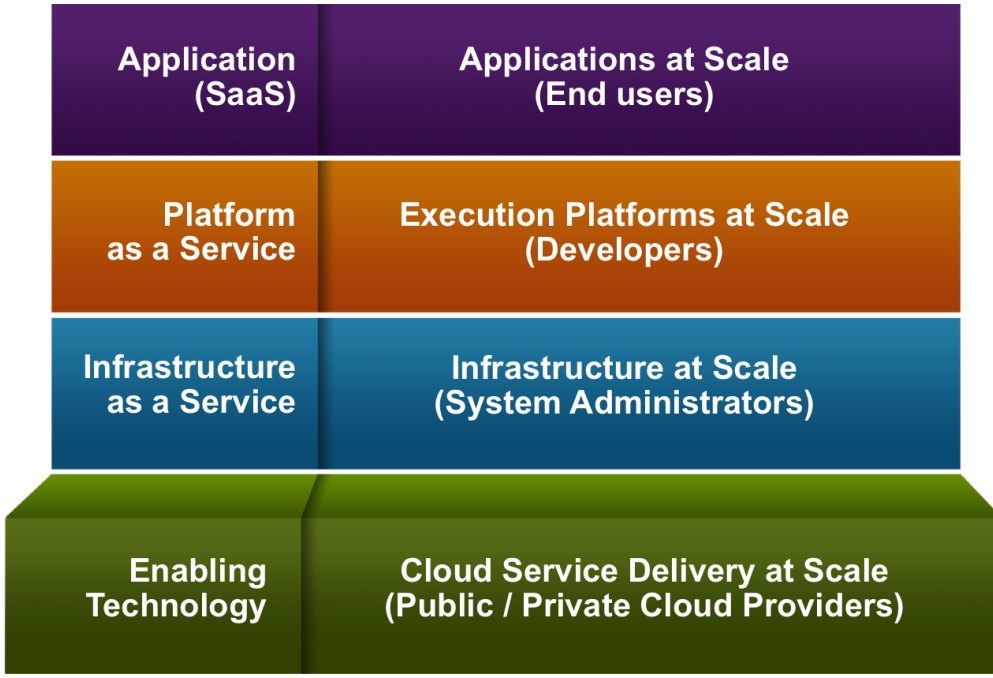


Figure 4.24. Cloud computing stack of service models

Any cloud service can be deployed as:

1. *private cloud*: the services are dedicated exclusively for the members of the organisation providing the services and owning the infrastructure
2. *public cloud*: the services are available to anyone, free of charge or for a regular fee (e.g. hourly/daily/weekly/monthly/yearly)
3. *hybrid cloud*: the services are available only to a private community, but the infrastructure is owned and managed by another entity, the one that provides the service.

Sometimes, an additional deployment type can be found in literature, namely *community cloud*, defined as a service available to a group of users with common interest, but from different organisations (e.g. police cloud). You can read more details about this taxonomy on the NIST 800-145 document [36].

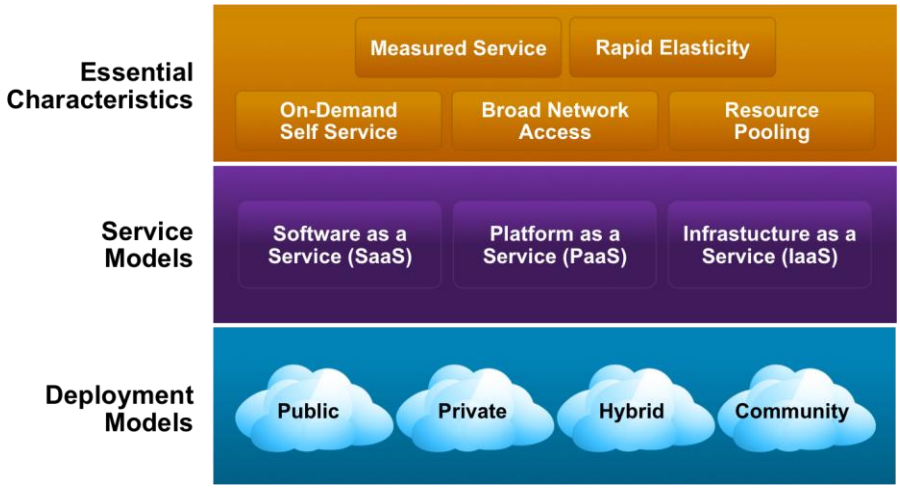


Figure 4.25. Cloud computing taxonomy

All cloud services—a *cloud service* is nothing else but a *consumption model*, in which you pay for resource *usage,* not resource ownership—are based on an infrastructure that includes all the elements we discussed above: computing, storage, communications. The equipment that delivers these resources is hosted in a Data Centre and managed by the organisation that owns it. Services are offered according to their delivery and deployment models, and monetised according to their business rules.

Remember a cloud service:

- is self (auto) provisioned
- creates the illusion of infinite resources
- is paid for per-usage

Chapter 5: Elements of Data Science – Databases

5.1. Introduction to Database

The healthcare industry needs very large quantities of complex data that have to be easily used and managed.

A database (DB) is a data collection with a specified *structure* and a well-defined *purpose*. Usually, the structure is given by *Records/Observations* (defined as collections or lists of elements) and *Fields/Variables/Properties* defined as the elements or attributes of a record).

Other types of data can be collected and organised in a database for easy access, as the phenotype, demographics, etc. The source of this data can be specialised lab equipment, patient conversations captured in EHR (Electronic Health Record) or, more and more present during the last years, wearable technology.

There are many advantages of the database usage, the most important being:

- Storage of arbitrarily large quantities of data
- Easy creation, retrieval, updating, and deletion of data
- Quick manipulation (e.g. sorting)
- Efficient (fast and easy) user interaction
- Data sharing
- Data security
- Data protection and availability

Depending the way the information is organised within, the main types of data structures are:

- Flat
- Relational
- Unstructured (Data Lakes)
- Hierarchical
- Object-Oriented

The last 2 items from this list are very specialised and will not be discussed at all in this course. For the other three, we will discuss (similar with what you have seen in the 1st part of this module) only the very few basic things, just to give you an idea about the topic. We consider this to be necessary

because medical data (e.g. patient health records) are often a combination of one or more of these fundamental data types.

5.2. Type of Data Structures–Flat Files

A flat file can be a simple text with one line of text corresponding to each record. This arrangement may have some advantages:

- Works well for simple data or for data with a very clear structure (e.g. genomic data)
- Large majority of existing software include easy access to flat data files (e.g. .csv)
- No performance penalty on operations over the whole database (no query or filtering is needed).

However, there are also some disadvantages:

- Waste of storage space keeping information not directly accessible via logical operations (e.g. measurement units)
- Not friendly for complicated queries; however, these are possible via programming languages, but they are tied to the a fixed structure of the file (on any change, like adding a new field, the query will likely need to be rewritten).

See below an example of a flat file including a table with phenotype data:

ID	Hospita	Sex	Age	Height	Weight	HTA	Study
94456383	JHP	1	67	178	76	0	3
89348270	ACR	1	61	167	98	0	1
66481148	JHP	1	84	175	90	1	3
52567168	ACR	0	73	154	67	1	1
11553474	JHP	1	73	180	84	1	2
35666103	JHP	0	52	170	82	1	3
17102233	ACR	0	46	157	108	0	1

Figure 5.1. Example of file with phenotype data

As you have observed, the organization of a flat file database is the following:

- Data records are stored as lines
- The records have a rather simple structure with (usually) a fixed number of fields
- There is no intrinsic mechanism to create relations between different files

- Data can be managed by special applications written specifically and exclusively for a certain purpose and certain structure of the file.

Each one of these can be seen as an advantage or a disadvantage depending on the context, showing the relativeness of any taxonomy.

One of the most common examples of domains of science that consistently use flat files is genomics, where, for a specific purpose, the structure of the database is fixed and the data processing mechanisms are so specialised that the connection between data and processing programs is more a feature than a disadvantage. See below and excerpt from a FASTQ file [37], very common for storing genetic sequences:

```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

Fig. 5.2. Example of a FASTQ file

The FASTQ file has the specific property to store not only the nucleotide sequence, but also a quality score for each reading. By convention there are 93 levels of quality represented using 1 byte, each corresponding to an ASCII character. The scale runs from 0x21 (lowest quality; '!' in ASCII) to 0x7e (highest quality; '~' in ASCII). Here are 93 values of the quality characters, in the increasing order of quality, lower on left, higher on right:

!"#$%&'()+,-*
./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdef
ghijklmnopqrstuvwxyz{|}~

In the example from the figure above, the first 4 bases have a quality score of 28, 30, 27 and 3, respectively. Being the de-facto standard for sequencing instruments, more details about FASTQ files can be found on the sites of vendors of such machines (an example is Illumina, that provides a good set of articles and references [63].

This is a very good example of a flat-file database, where the genetic structure of a specific sample is encoded together with the quality index of each reading, giving the scientist the reliability of the reading.

The main characteristics for flat-file databases are:

- Data is stored in files
- Each file can have a different format
- Data access (directly or programmatically) depends on precise understanding of (different, but specific) file formats
- Programs for data processing are written for each data set structure in particular (which may not be a big problem if the file uses a commonly agreed format).

However, there are also some challenges:

- Lack of standards, even in narrow areas as genomics
- Lack of efficiency (duplicate or not useful data)
- Hard dependency on the data's physical format
- No ad-hoc query/interrogation mechanism
- Reduced security
- Reduced shared access.

5.3. Type of Data Structures–Relational Databases

A relational database is a collection of tables representing entities and relationships between them. The name of the table gives the name of the relationship, the columns represent the features/properties of the entity and the rows represent the actual data.

Relational DB uses a simple but powerful organisational model, allowing simple interrogation via high level programming languages and an efficient implementation in terms of storage and access speed. An important aspect of relational databases is *data integrity*, which refers to the maintenance and assurance of the accuracy and consistency of data over its entire life-cycle.

The structure of this type of DB is:

- Data is organised in tables: rows & columns
- Each row (record) represents an instance of an object
- Each column represents an attribute or property of the object
- Each column is described by associated metadata describing, for example, the types of the data (alphanumeric, numeric, boolean, etc.)
- The union of all tables forms the database
- Relations between entities are represented by the values stored in the columns that make the correspondence between the tables (keys)
- In order to access a specific record keys are used. Each record can have one or more keys; the key that uniquely identifies the

record is called the primary key. Keys are part of the logical level of the database

- Access to data is done via a high level programming language called Standard Query Language (SQL)
- Data from one or more (existing) tables can be combined in virtual tables, composed of parts/subsets of the real tables, called views.

Relations can be expanded between tables, connecting fields from one table to another using the keys. There are 3 types of relations:

- One-to-one: one row from Table A can be connected (via a key) with exactly one row from (a different) Table B. The relationship is true in both directions (A → B and B → A). An example can be the relation between a gene (Table A = genes table) and the protein (Table B = protein table) it encodes.
- One-to-many: one row from Table A can be connected (via a key) with multiple rows from Table B, but each one of the rows in Table B can be connected with only one row in Table A. A good example is the relation between a mother (Table A = mothers table) and her children (Table B = children table): the mother is connected with (potential) multiple children, but each child is connected with only one mother
- Many-to-many: one row from Table A is connected with multiple rows from Table B and each row in Table B is connected with multiple rows in Table A. Usually, this type of relationship is modelled using multiple one-to-many relations. An example could be the relationship between hospitals (Table A = hospitals table) and patients (Table B = patient table).

In the figure below you will find an example from real life, an excerpt from an EHR (Electronic Health Record) showing two encounters of a specific patient (identified through "Patient_ID").

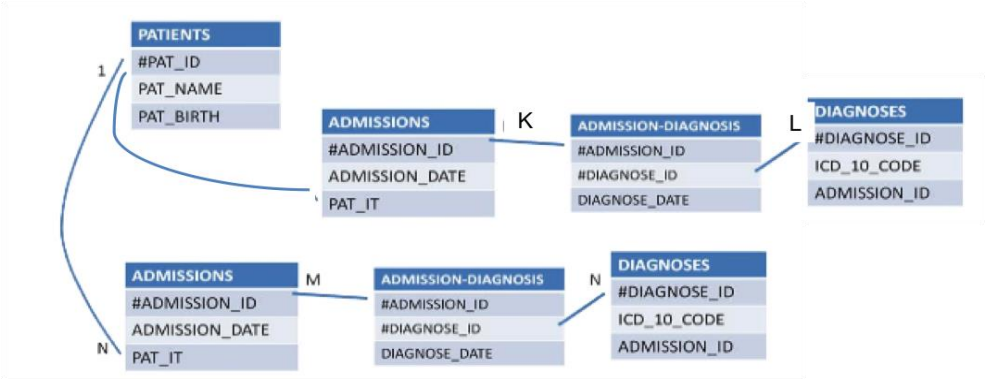


Fig.5.3. Example of the Records in EHR Database

You can observe a one-to-many relationship between the *Patients* and *Admissions* Tables, modelling 2 visits of the same patient, on 2 different occasions. Then you can observe a series of one-to-one relationships between the *Admissions* and *Admission-Diagnosis* Tables, corresponding to each of the hospital admissions of the patient, potentially with different diagnosis.

The fields used to make these associations are known as *keys*. A key is an attribute with a unique value in each record (or a set of attributes with a unique combined value), used for identification of that record. Please note that sometimes an attribute can have the value of *NULL* (sometimes marked *N/A* or *NA*), which is a special value used for an attribute that is "unknown" or "undefined".

We can use the keys to select parts (subsets) of different tables in a query (or set of queries) and the result of these operations can be assembled in a new table called a *view*. Unlike other tables, a view is a virtual table (in the sense of not being part of the physical, stored tables of the database) that is assembled dynamically from the physical tables, on a specific request. This approach allows us to extract from existing data only those pieces that we need at a specific moment in time. In the figure below, you can see an example:

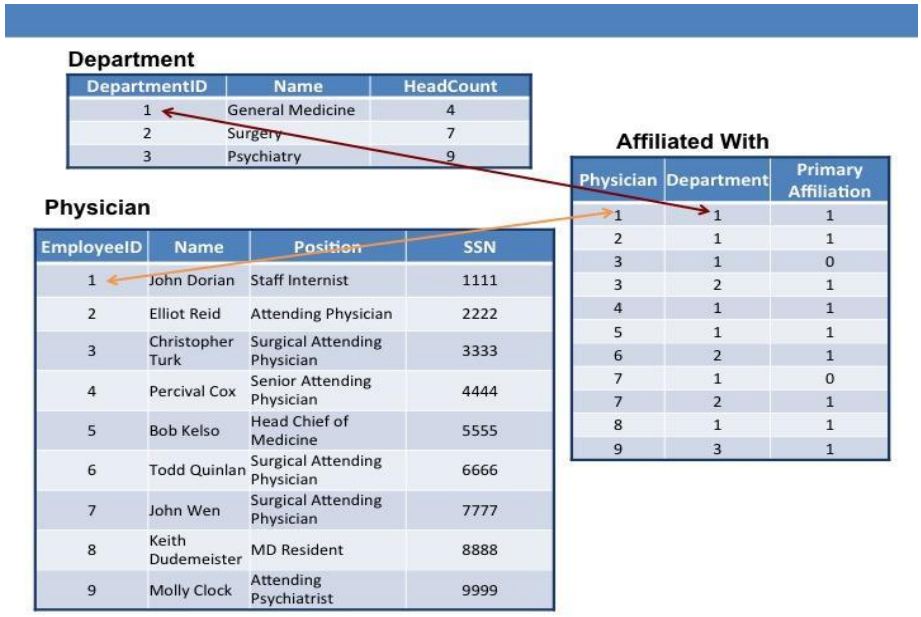


Fig.5.4. A table called a "View" example

Above, you can see 3 physical tables, stored permanently in the databases of a HIS (Hospital Information System), holding data necessary for the hospital to operate:

1. Physician: stores data about the physicians working in the clinic
    - Name of the MD
    - Position in the clinic
    - Unique social ID
    - Unique internal employee ID
2. Department: describes each department
    - Unique internal department ID
    - Name of the department
    - Number of doctors activating in that department
3. Affiliated_with: keeps the affiliation of MDs to departments
    - Internal ID of the physician
    - Internal ID of the department
    - Existing affiliation (1=TRUE, 0=FALSE)

Using the keys indicated by double arrows, you can build a view that will show you only the information you need (name of the MD, name of their associated Department and their internal IDs), as in the figure below:

EmployeeID	Name	DepartmentID	Name
1	John Dorian	1	General Medicine
2	Elliot Reid	1	General Medicine
3	Christopher Turk	1	General Medicine
4	Percival Cox	2	Surgery
5	Bob Kelso	1	General Medicine
6	Todd Quinlan	1	General Medicine
7	John Wen	2	Surgery
8	Keith Dudemeister	1	General Medicine
9	Molly Clock	3	Psychiatry

Fig.5.5 - View created from multiple tables

This example should give you an idea about the power behind relational DBs, which can give you, programmatically, the information you need, at the moment you need, disregarding the physical construction (a.k.a. tables) of the DB. This capability to create custom views using any existing information in the database gives you the freedom to organise the tables by other rules (e.g. efficiency, performance, etc.) and is built on the property of relations DBs called *data independence*.

Data independence is the property of different architectural layers of a database to be completely independent (abstracted) of the changes performed on other layers. From the user perspective, there are 3 architectural layers of a DB:

1. *physical layer*: describes the way the data is physically stored on storage media (e.g. files on disks, number of disks, etc.). Physical data independence means that any change here (like file renaming or disk changing) should not be visible on any layer above.
2. *logical layer*: describes how the data is organised (e.g. tables, row, columns). Logical data independence means that any change (e.g. adding a feature/property to a new column) should not change any existing view. Any program used to create previous views should still function without changes
3. *view/visual layer*: this is the top-most layer, the one presenting the data to the user, where only the needed data should be visible, describing how data is visualised by the user, hiding the unnecessary details or data that have been marked as invisible (masked). No change below this layer should influence what is presented here.

The structure of data on each layer is called the *schema*. Schema is a structural description of the organisation of data, described in a formal language by a set of sentences (formulas) called integrity constraints, part of the data integrity efforts. For relational DBs, the schema includes the tables, fields, relationships, views, indexes, and other elements, all stored in a metadata repository, often called the *data dictionary*. Even though the schema is defined in a text language, very often the representation is graphical, as you can see below:

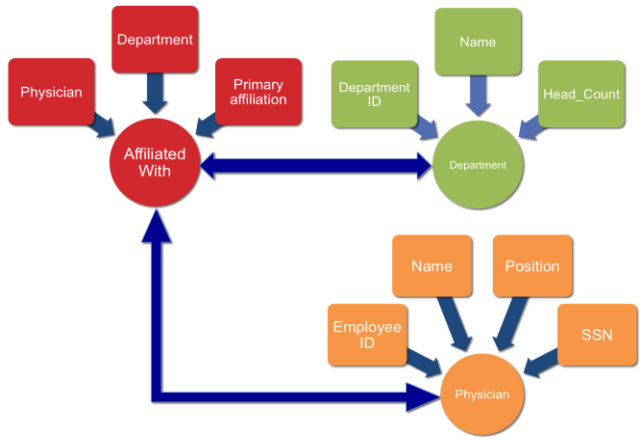


Fig. 5.6. Relational DBs schema for MD activity

This is a graphical representation of the example given above, related to MD activity in clinical departments. Now, let's see below the same schema represented in a more detailed way (including the metadata):

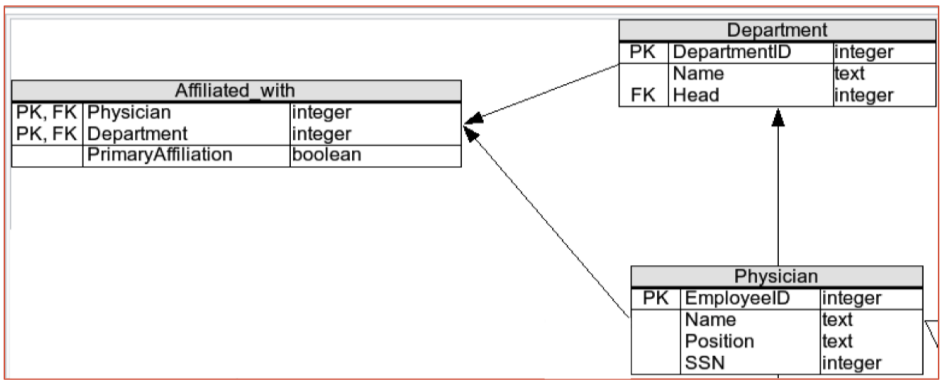


Fig. 5.7. Relational DBs schema for detailed MD activities

Metadata defines data that describes the properties or characteristics of useful data ("data about data"). It shows properties, but it does *not* include the actual data (sample data). The reason is to provide designers and users with a way to understand the meaning or the actual data.

To illustrate, find below an example of metadata and sample data from a database of bacteria genome:

Metadata

Name	Type	Max Length	Description
Name	Alphanumeric	100	Organism name
Size	Integer	10	Genome length (bases)
Gc	Float	5	Percent GC
Accession	Alphanumeric	10	Accession number
Release	Date	8	Release date
Center	Alphanumeric	100	Genome center name
Sequence	Alphanumeric	Variable	Sequence

Sample data

Name	Size	Gc	Accession	Release	Center	Sequence
Escherichia coli K12	4,640,000	50	NC_000913	09/05/1997	Univ. Wisconsin	AGCTTTTC ATT…
Streptococcus pneumoniae R6	2,040,000	40	NC_003098	09/07/2001	Eli Lilly and Company	TTGAAAGA AAA…
…						

Fig. 5.8. - Metadata and sample data from a database of bacteria genome

Please observe how the metadata table defines the properties of the real (sample) data, shown in red arrows.

All the sampled data and metadata are stored, accessed and maintained by a Database Management System (DBMS). The DBMS is a collection of programs (a software system) for access, development and management of databases (data collections). The main goal for any DBMS is fast query/retrieval of particular data from all stored data, in a controlled manner. A DBMS has the following main functionalities:

- *Defines* a database in terms of data types, structure and constraints
- *Loads* the initial content on a storage media
- *Manages* data:
  - Retrieval: querying, reporting
  - Modification: creating, deleting and updating content
  - Accessing: via applications
- *Shares* data between many users and applications simultaneously

Very often in common language you will find the analogy:

database = DBMS (Database Management System)

In the figure below you can see a generic representation of a DBMS:

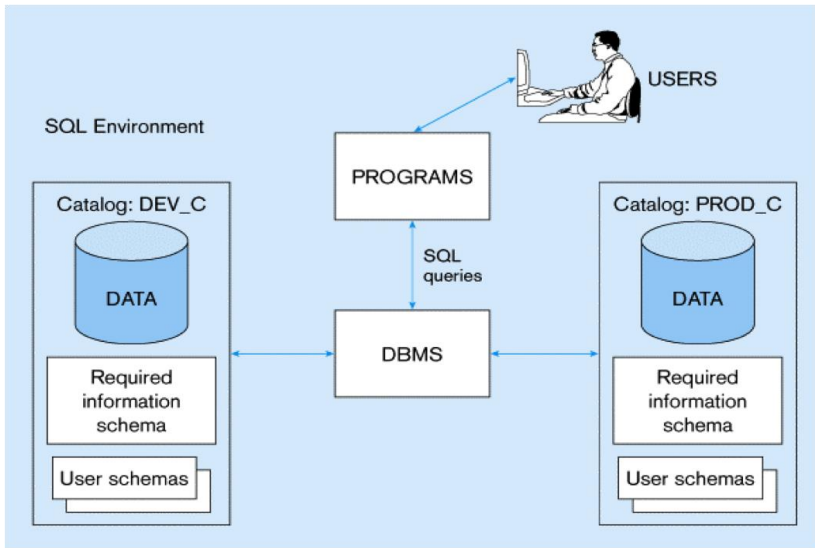


Fig. 5.9. - General representation of a DBMS

5.4. Relational Databases–SQL

In the previous part we said that the access to a database (within the framework of a DBMS) is done programmatically.

Typically the interaction with a software component implies the existence of an Application Programming Interface (API) that is nothing else but a set of software routines, protocols, and tools used to build software applications. An API expresses a software component in terms of its operations, inputs, outputs, and underlying types.

However, a relational DBMS includes a specialised computer language (Domain-Specific Language) that understands the relationship between entities and is specifically designed for managing databases. This programming language is named SQL (Structured Query Language). Just as a side note, major DBMS support a standard API for access to them, called Open Database Connectivity (ODBC), that allows building SQL queries to the databases.

SQL is standardised by ANSI (American National Standards Institute) and it is supported by all major commercial/non-commercial systems. Based on relational algebra it uses transactions (non-breakable units of work that induce a change in the database) for retrieving and updating records.

Data applications that build a specific view use SQL to specify “what” but not “how” the data is handled. SQL is independent of data applications. SQL includes few specialised sub-languages:

1. *Data Definition Language (DDL)*: defines databases by managing tables and index structure, including creation, alteration, and deletion of tables and associated constraints
2. Data Manipulation Language (DML): maintains and queries the database
3. *Data Control Language (DCL)*: controls the access to data, including administrative privilege control

Data Definition Language holds the syntax elements used for definition of the structure of the database. See below an example of defining the tables used in the previous example:

```
CREATE TABLE Physician (
     EmployeeID INTEGER PRIMARY KEY NOT NULL,
     Name TEXT NOT NULL,
     Position TEXT NOT NULL,
     SSN INTEGER NOT NULL
  );
CREATE TABLE Department (
     DepartmentID INTEGER PRIMARY KEY NOT NULL,
     Name TEXT NOT NULL,
     Head INTEGER NOT NULL
     CONSTRAINT fk_Physician_EmployeeID REFERENCES
Physician(EmployeeID)
  );
 CREATE TABLE Affiliated_With (
     Physician INTEGER NOT NULL
     CONSTRAINT fk_Physician_EmployeeID REFERENCES
Physician(EmployeeID),
     Department INTEGER NOT NULL
     CONSTRAINT fk_Department_DepartmentID REFERENCES
Department(DepartmentID),
     PrimaryAffiliation BOOLEAN NOT NULL,
     PRIMARY KEY(Physician, Department)
  );
```

Data Manipulation Language holds syntax elements used for selection, insertion, deletion or updating data. Below is an example of inserting data into the tables created:

```
Table "Physician":
    INSERT INTO Physician VALUES(1,'John Dorian','Staff
Internist',1111);
    INSERT INTO Physician VALUES(2,'Elliot
Reid','Attending Physician',2222);
    INSERT INTO Physician VALUES(3,'Christopher
Turk','Surgical Attending Physician',3333);
    INSERT INTO Physician VALUES(4,'Percival Cox','Senior
Attending Physician',4444);
```

```
    […] - continued up to the end of all records

Table "Department":
    INSERT INTO Department VALUES(1,'General Medicine',4);
    INSERT INTO Department VALUES(2,'Surgery',7);
    INSERT INTO Department VALUES(3,'Psychiatry',9);
    INSERT INTO Affiliated_With VALUES(1,1,1);
    INSERT INTO Affiliated_With VALUES(2,1,1);
    INSERT INTO Affiliated_With VALUES(3,1,0);
    INSERT INTO Affiliated_With VALUES(3,2,1);
    […] - continued up to the end of all records
Table "Affiliated_With":
    INSERT INTO Procedure VALUES(1,'Reverse
Rhinopodoplasty',1500.0);
    INSERT INTO Procedure VALUES(2,'Obtuse Pyloric
Recombobulation',3750.0);
    INSERT INTO Procedure VALUES(3,'Folded
Demiophtalmectomy',4500.0);
    […] - continued up to the end of all records
```

Data Control Language holds the syntax elements used for access control to database, meaning allowing to users to have access to a specific database, as in the generic example below:

- ```
  GRANT SELECT
      ON [database]
      TO user_one;
  ```

- ```
  REVOKE SELECT
      ON [database]
      TO user_two
  ```

5.5. Type of Data Structures–NoSQL Databases

Despite the large adoption and significant development of the relational databases, some processes generate and consume data in ways that cannot be easily captured in a predefined schema (an organisational structure that can be defined prior to sampling the process data). These cases need a database that provides a mechanism for storage and retrieval of data that is

modelled in means other than the tabular relations used in relational databases. Such a DB is called NoSQL ("Not Only SQL").

These are next generation databases that model the data in a different way than relational DBs, dealing with challenging points of relational DBMS: non-relational, distributed, large databases.

Usually, NoSQL refers to databases with key-value pairs, document stores (JSON, XML), graph-based databases or even object-oriented databases.

NoSQL databases have few distinctive characteristics:

- contain huge volumes of data
- there is no generic data model; data is stored in files
- the user needs specific functions to interact with the DB
- the system provides access and data processing capabilities using built-in functions and deals with error handling, scalable replication and data distribution
- potentially installed on thousands of servers
- potentially installed in any geographical location
- built for performance criteria (fast answers to queries, fast data insertion, etc.)
- the large majority of operations are queries (very few are updates)
- there is no predefined associated schema.

The most significant differences between SQL and NoSQL databases are:

- Scalability – SQL does not allow massive parallel processing, which bring the needs for large single systems (scale up) for processing large data. NoSQL systems are good at dealing with distributed data, so they can take advantage of numerous small systems connected together or cloud instances (scale out).
- Modelling – SQL databases are normalised, so they require a predefined data model before introducing the actual data (schema). In contrast, NoSQL does not need this kind of model, allowing the definition of a schema on the fly
- Quality – the lack of a single schema for all the data allows different perspectives over data quality
- Consistency – for short periods, in large distributed systems the consistency may be altered, as price for fast access and response
- Access – needs specialised tools & languages for access and interfacing, which are potentially different for different NoSQL DBMSes. Each NoSQL provides an API for different programming languages and the user has to write his own tools.

Given the distributed nature of the NoSQL DBs, which are built specifically to accommodate large quantities of data located in different sites, there are

some additional issues that have to be solved. One of the most important is known as Brewer's CAP Theorem: a distributed system can support *only two* of the following characteristics:

- Consistency: clients perceive that each set of operations occurs all at once, and they all receive the most current data
- Availability: every operation must terminate with a well defined response
- Partition tolerance: operations will complete, even if individual components become unavailable.

This means that in the case of a connection error between distant members of the database cluster you should decide if the operation is cancelled, affecting Availability, but preserving Consistency, or it is continued, preserving Availability, but affecting Consistency (there is no guarantee that a read will receive the most current data, from the latest write). Different choices of 2 out of 3 CAP properties have been (consciously) made by different NoSQL developers for their products, serving special application areas (see figure below).

Please note that the same principles above apply also to relational DBs that are distributed in multiple locations. However, relational DBMS make special efforts to enforce the ACID properties (atomicity, consistency, isolation, durability) [64] on each transaction, inducing limitations on the distributed clusters. NoSQL databases adopt a different model, called BASE (Basically Available, Soft state, Eventual consistency), [40] to serve their native distributed computing nature.

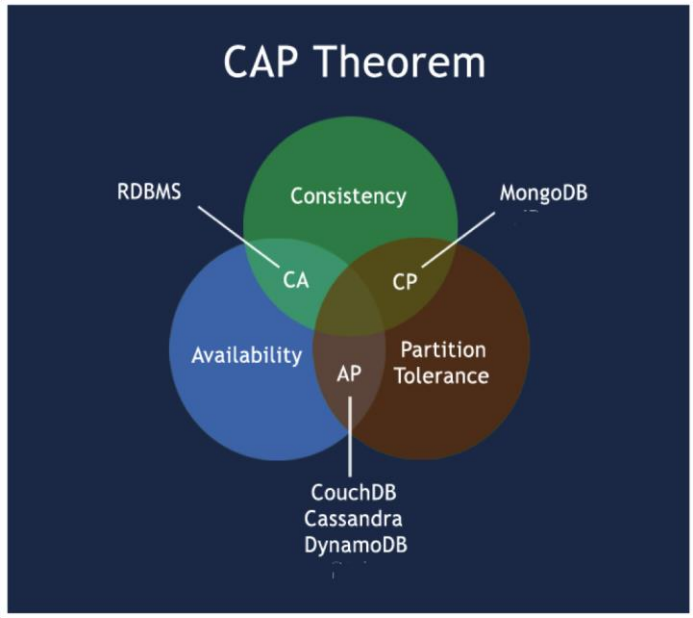


Figure 5.10. – CAP Theorem representation for NoSQL databases

There are different types of NoSQL databases, each one serving a different type of application:

1. Distributed Key-Value Systems – Lookup a single value for a key
    - Amazon's Dynamo
2. Document-based Systems – Access data by key or by search of "document" data.
    - CouchDB
    - MongoDB
3. Column-based Systems – Structure data as a set of columns
    - Google's BigTable
    - Facebook's Cassandra
4. Graph-based Systems – Use a graph structure
    - Google's Pregel
    - Neo4j

The trade-off is usually on the complexity of the data stored vs. the size of the data that can be stored and managed. As a generic rule of thumb, key-value databases can handle the largest amounts of data, but relationships need to be simple, whereas Graph databases are used for complex data relationships, but can only handle smaller amounts of data.

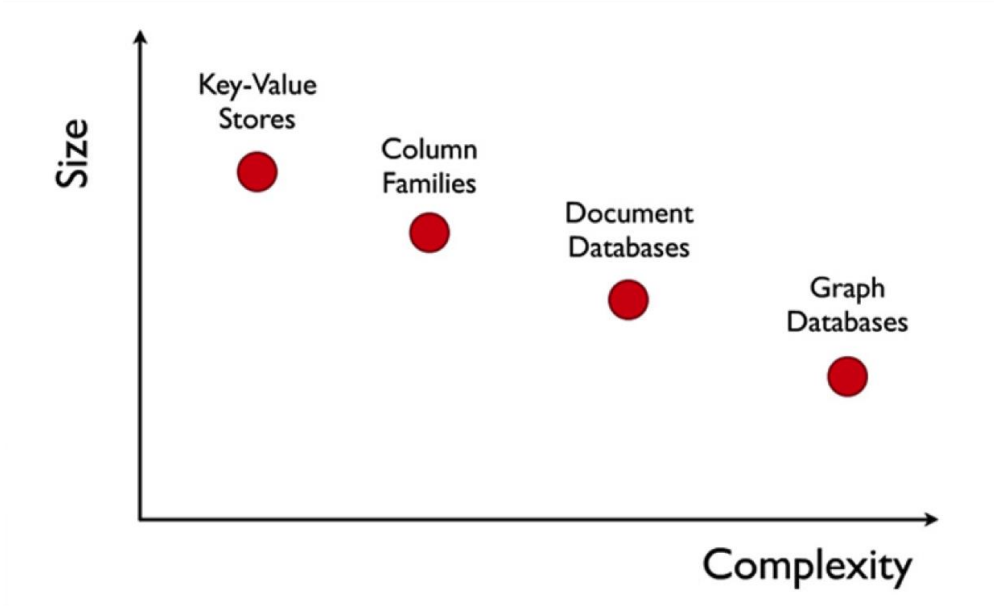


Figure 5.11. Complexity of the different types of data stores

5.5.1. NoSQL Databases – Key-Value Pair Stores

Data representation for a KV Store is straightforward as Key-value NoSQL databases access data (values) by strings (keys):

```
KEY:VALUE
```

Data has no required format, which means that data may have any format, the data model being simple *(key, value)* pairs. The value is stored as a blob, without caring or knowing what is inside. In simple terms, a NoSQL Key-Value store is a single table with two columns: one being the (Primary) Key, and the other being the Value, with each record having its own schema.

They have an extremely simple interface with few basic operations:

- Insert(key,value)
- Fetch(key)
- Update(key)
- Delete(key)

The implementation of key-value pair stores is targeted for efficiency, scalability and fault-tolerance. There is an easy replication mechanism,

records being distributed to nodes based on key. The application is solely responsible for understanding the data.

An example of an unstructured data is shown below:

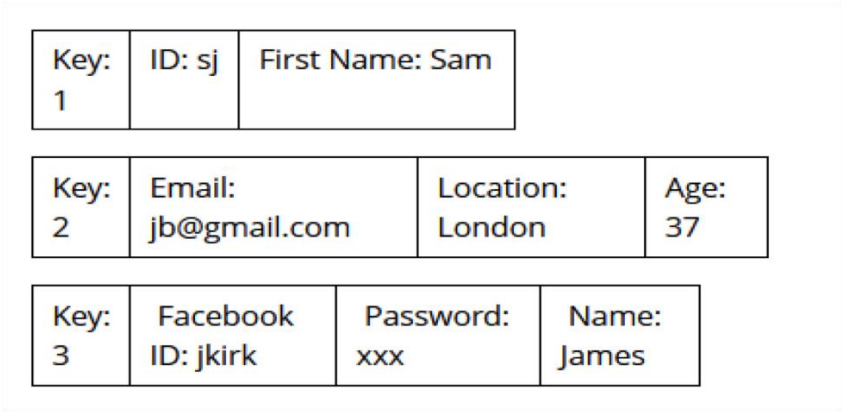


Figure 5.12. - Example of an unstructured data

5.5.2. NoSQL Databases – Document Stores

Document Store NoSQL databases are similar to Key-Value Stores, except the Value is now a *Document*, which stores arbitrary/extensible structures as a "value". Please note that even the records are called "documents", they are not documents in the sense of a word-processing document.

The data model is (key, document) pairs, where the Document is formatted as JSON, XML, or other semi-structured formats. Unlike the simple key-value stores, the value column in document databases contains semi-structured data with specifically attribute name/value pairs

The basic operations for a Document data store are:

- Insert(key,document)
- Fetch(key)
- Update(key)
- Delete(key)
- Fetch() based on document contents.

A single value column can hold hundreds of such attributes, and the number and type of attributes recorded can vary from row to row. Unlike simple key-value stores, both keys and values are fully searchable in document databases. This type of DB is good for storing "sparse" (semi-structured) data that would require an extensive use of "nulls" in a relational DBMS.

The structure of any document can be modified on the fly by adding and removing members from the document, either by reading the document into your program, modifying it and re-saving it, or by using various update commands.

A record from a Document based store looks like:

```
{
  "_id" : ObjectId("4fccbf281168a6aa3c215443"),
  "Patient_ID" : "2790151167",
  "Control" : TRUE,
  "Gender" : "Male",
  "Clinic" : "SFM",
  "DateOfBirth" : {
        "day" : "31",
        "month" : "March",
        "year" : "1940"
    }
}
```

A much more complex example of a Document based store, holding genomic data, can be seen below (example from Mongoworld, 2017):

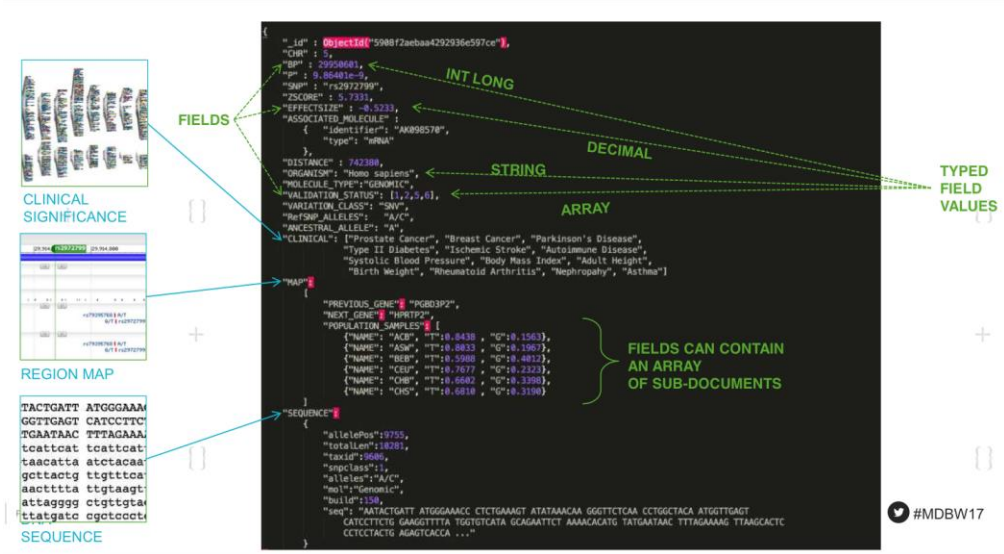


Figure 5.13. - Document based store, holding genomic data

5.5.3. NoSQL Databases – Column Stores

Data tables are stored as sections of columns of data, rather than as rows of data (see picture below). Columns are logically grouped into *column families* (the equivalent of tables in the relational DB) that can contain a virtually unlimited number of columns. Reading and writing from the database is done using columns rather than rows.

In comparison to most relational DBMSs, which store data in rows, the benefit of storing data in columns is two-fold:

- fast query/access: relational databases store single records (all the cells corresponding to a row) as a continuous disk entry (successive disk sectors), with different rows being stored in different places on disk. Column databases (often called "Columnar") store all the cells corresponding to a column as a continuous disk entry, thus making the access faster, as the data corresponding to a property (a column) can be read sequentially in a single operation, instead of jumping on the corresponding places on the disk, as in the case of a row-oriented database.

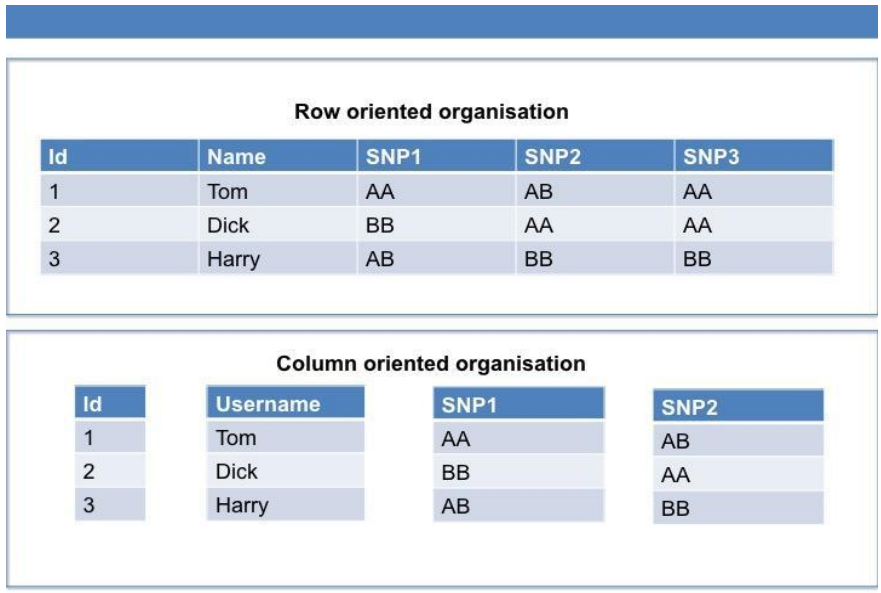


Figure 5.14. - Example of data tables stored as sections of columns

- data aggregation: aggregate functions, being data combinations or analysis functions (like counting, summing, calculating the average, etc.), need all the values for a specific feature/property, which can now be easily retrieved, getting the entire column at once.

For example, suppose that you have 1 million genetic samples, from 1 million patients, and you want to know how many of them have a specific set of SNPs in different loci. Organising the data in a Column database lets you do the counting by getting the corresponding column with a single disk operation, instead of reading all the corresponding places of all rows, in 1 million disk operations. If the operation is more complex, involving the combination of different tables (called joins in RDBMS), the performance of Column stores is even higher than relational DBs.

5.5.4. Database Usage

Usually, we differentiate between two different types data processing that use databases:

1. OLTP (Online Transaction Processing) defines a type of data processing oriented on transactions, where the system responds immediately to requests, modifying the state of database recordings. They are usually:
    - Short transactions
    - Simple queries
    - Work with small quantities of data
    - Frequent updates
    - Performed on Row-based stores
2. OLAP (Online Analytical Processing) defines complex operations with data for reporting or analysis purposes. They are usually:
    - Long transactions
    - Complex queries
    - Operations on large quantities of data
    - Frequent reads (rare updates)
    - Performed on Column-based stores

To get the data you need, you have to look at different databases/data stores and extract from there what you need for your research, performing a procedure called ETL (Extract-Transform-Load).

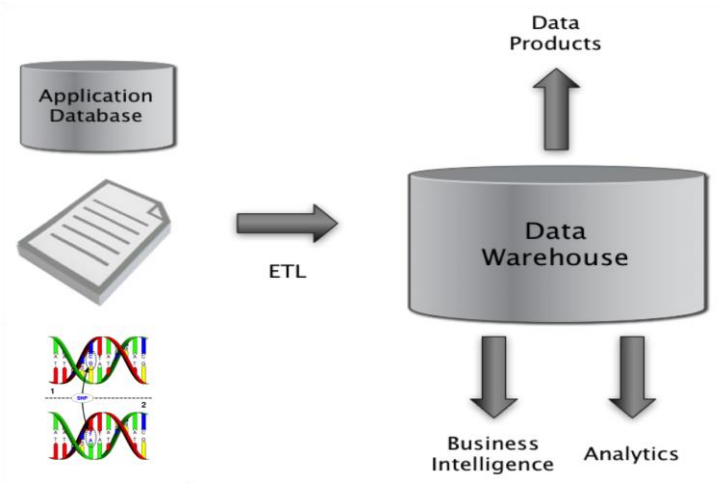


Fig. 5.15. - Extract-Transform-Load data from a database

A *Data Warehouse (DW)* is defined as a system used for reporting and analysing data. DW is the central data storage area integrating one or more different sources. The analysis process means that data is brought from operational sources (OLTP) into a single "warehouse" for analysis (OLAP).

As DWs are quite large, we usually use a simpler form of DW, focused on a single topic or one with local relevance, called *Data Mart (DM)*. The DW/DM plays a central role in historic data analysis, but a similar mechanism may use DW/DM data for prediction. The results from both types of analysis may be used to enhance operational decisions, in which case DW is considered as being part of a Decision Support System (DSS).

A typical DW and its relation with other data stores is presented in figure 5.16.

The DW/DM binomial structure served very well for data analysis with relational databases for a long time. However, since the data started having high volumes, high velocity and high variety (Big Data), the situation started to change. The main problem is the predefined schema that a DW is supposed to have. This single schema is both difficult to build in advance to match all the data types needed and difficult to match all data processing required. DW also implies some data pre-processing, at least for quality purposes, if not more, activity that changes the original data in a way that makes it suitable for a certain type of analytics (e.g. decimal precision for numbers). The problem is this "suitability", which only works for a limited number of processing types, but not for all.

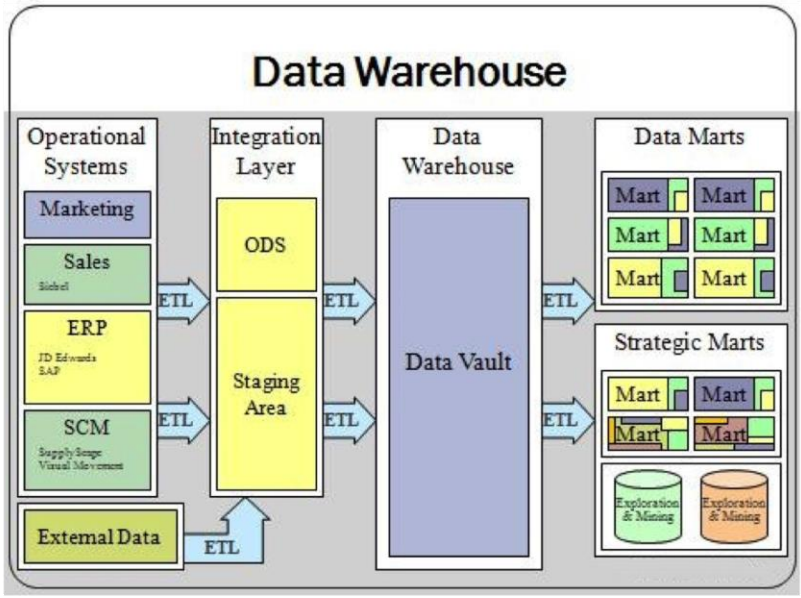


Fig. 5.16. Data Warehouse in relation with other data stores

To adapt to the new characteristics of data, the Data Warehouse concept evolved into the *Data Lake* (DL) one. A Data Lake is just a raw data repository, holding the original values. The advantage is that the data quality validation is left to be done at the moment of analysis and so preprocessing (cleansing) is adjusted to the very scope of this activity. Different thresholds, or even different methods, are used when different analyses are performed, making data more suitable for the aim of the process and results better fitting reality.

The figure 5.17 (adapted from Martin Fowler) shows the main difference between a DW and DL.

Data Warehouses usually would have the data pre-processed (cleansed in a specific way, or even aggregated) to make it easier to analyze further. However, this is not necessarily good because aggregation implies losing some data. In contrast, the Data Lake should contain all the data because you don't know what will be valuable/useless in the near future.

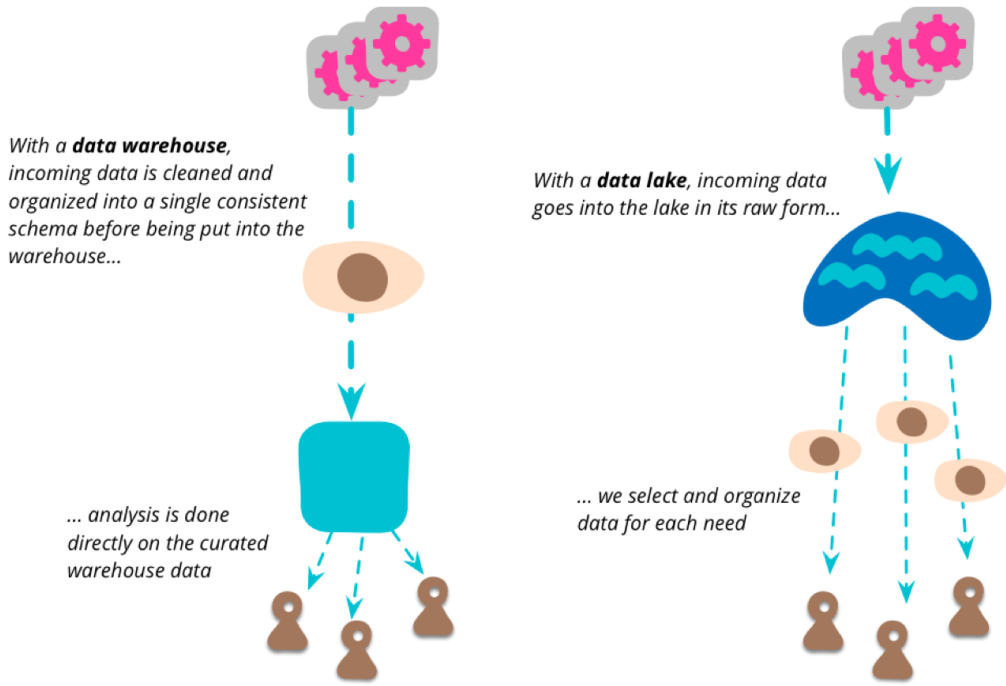


Fig. 5.17. - The main difference between a Data Warehouse (DW) and the Data Lake *(DL)*

It is important that all data should have a clear provenance, both in place and in time, making possible to trace it back to the system that generated it and find when the data was produced.

Do not forget that the Data Lake is schema-less, that means it will accept the data with whatever schema the source systems decide to use. It is the responsibility of the data processing app to deal with it. Also, keep in mind that the source system may change its data schema in time and this change is also left to be addressed on processing time. Despite the complexity, this is an assurance that you have all the available data for your analysis, and nothing has been lost in generic cleansing operations (e.g. outlier removal).

The Data Lake shouldn't be accessed directly very often. Specialised teams/processes get data out of DL and create Data Marts that offer a useful view of data, each one tied to a well-defined context. These DMs can be considered authoritative sources for that specific context by all the subsequent users.

5.5.5. NoSQL Databases – Comments

NoSQL databases appeared as a reaction to relational databases' limitation regarding their administrative complexity and their (lack of) ability to scale on very large numbers of records. Unpredictable RDBMS response times become “dangerous” at the scale required by Big Data.

In order to achieve their potential, NoSQL relaxed a subset of relational DBMS properties (ACID) to achieve the desired scalability, becoming "Eventually consistent".

NoSQL databases use Light-to-no schema enforcement, allowing for difficult and complex queries, without joins, aggregates, or other tedious operations needed in RDBMS.

NoSQL databases make use of a simple programming and administration model, but different NoSQL types need different tools for the same operation. Also, custom application code is required for complex queries, as there is no universal query language like SQL.

NoSQL have already proved to be good in terms of performance for workloads at scale, when very large numbers of records (millions) are used.

Significant development, both in implementation and modelling, should be expected in the future, as the NoSQL era has just begun (40+ years for relational DBs versus ~ 6 years for NoSQL).

Chapter 6: Ethical Concerns in Research on Human Subjects

Theories of confidentiality and privacy of identifiable health data are featured in the earliest conceptions of medical ethics starting from Hippocratic oath until Declaration of Helsinki, which is one of the most important international ethical guidance for human subjects research since 1964.

6.1. Privacy and confidentiality in research on human subjects

6.1.1. Privacy

The notion of privacy is a constitutional right in every constitution and the Human Rights Act and expresses that intimal privacy must be protected and refers to a person and their interest in controlling the access of others to identifiable health information acquired or that may be used in research. The participant in any research will disclose as much information as it considers necessary and will not be constrained in any form to share information that does not wish to be revealed.

6.1.2. Confidentiality

Confidentiality is an extension of privacy and refers to the duty of an individual (physician, researcher and investigator) to safeguard entrusted information such as identifiable data or information about the person who is subject of human research, collected in the process of research. This information will not be divulged except in ways that have been previously agreed upon.

In research context, confidentiality means not to disclose information obtained from a participant to others, take additional measures to ensure confidentiality of participants, and presenting findings in ways that ensure individuals cannot be identified [66].

The need to protect privacy and confidentiality is stipulated by ethical and legal norms that concern research on human subjects.

6.1.3. Ethical Norms

Article 24 from the Declaration of Helsinki specifically mentions the need to respect subjects' privacy through data confidentiality protections: "*Every*

precaution must be taken to protect the privacy of research subjects and the confidentiality of their personal information" [67].

International Ethical Guidelines for Biomedical Research Involving Human Subjects (CIOMS and WHO) also state the need to respect subjects' privacy and confidentiality: "*researchers must establish secure safeguards of the confidentiality of subjects' research data. Subjects should be told the limits, legal or other, to the investigator's ability to safeguard confidentiality and the possible consequences of breaches of confidentiality*" [68].

One of the main sources of risk in medical research is that information obtained by investigators could harm participants if disclosed outside the research setting.

Belmont Report: Privacy and confidentiality are supported by two principles of the Belmont Report [69].

- Respect for persons – each person should be threatened as an autonomous person, capable of making their own decisions and choices. In cases of those with diminished autonomy additional protection should be taken in consideration
- Beneficence – „do not harm" and maximize the potential benefits. Maintaining promises of confidentiality, monitoring the data to ensure the safety of participants and assessing reasonable risks to participants are requirements based on beneficence.

6.1.4. European Laws

The General Data Protection Regulation (GDPR) is a European privacy law in effect as of May 25th, 2018, which protects the personal data of individuals located in the European Area. EU regulation imposes a unique set of rules on the protection of personal data, replacing Directive 95/46/EC and implicitly the provision of law no 677/2001.

GDPR defines personal data as "*any information that relates to an identified or identifiable natural person*" and it is based on the concept of privacy as a fundamental human right [70].

GDPR has six general data security principles [71]:

- Fairness and lawfulness
- Purpose limitation
- Data minimisation
- Accuracy
- Storage limitation
- Integrity and confidentiality

The rights underlying GDPR are:

- The right to be informed – GDPR gives individuals the right to be informed about collection for their personal data and future use. Every person needs to be informed about the duration of storage, the rights of the data subject.
- The right to withdraw consent
- The right to be forgotten – the right to delete data if they are no longer needed for their original purpose or if the subject has withdrawn his consent and there is no other legal ground for processing.

6.1.5. Types of personal information

In the context of research any type of information about participants can be collected [72].

- Directly identifying information - name, personal health number, health insurance number, personal identification number;
- Indirectly identifying information - information that can lead to a person identification through a combination of indirect identifiers (date of birth, address)
- Coded information – direct identifiers are removed and replaced with a code. when there is a list that still maintains a link between the code and the person's identification data, re-identification can easily be done
- Anonymized information – direct identifiers are totally removed without allocating a code; a future possible re-identification is minimal based only on indirect identifiers.
- Anonymous information - information about participant were collected through an anonymous survey, the information never had any type of identifiers.

When all direct and indirect identifiers are removed, and the anonymization is irreversible, data are no longer under GDPR requirements.

Protocols should be designed to ensure safeguards of confidentiality and limit possible consequences of breaches of confidentiality. Security measures should take into account the nature of data, if the data are recorded on paper or electronic form, content, mobility and vulnerability [73].

Ways to protect confidentiality are [74] [75]:

- Collect only the necessary identification data
- Remove any direct identifiers from survey instruments or medical records
- Destroy subjects identifiers when they are no longer needed
- Instead of recording identifying information, use codes for each information in part and keep a separate document with information that can connect the study code to the subject`s data. Restrict access to this document and to identifiable information. If it is necessary to

link participants with their data, use a study ID assigned for each participant before collecting data and insert this ID on their data documents.
- Encrypt identifiable data
- Securely store data documents within locked locations and assign security codes to computerized records.
- Avoid storing subject data on portable devices

Maintaining confidentiality is necessary during and after the study. Breaches of confidentiality involving human subjects can cause potential harms including psychological harms, stigma or distress, social harms and in some cases civil liability.

During the informed consent process subjects must be informed about the precautions that will be taken to protect confidentiality and their limitations. Subjects need to know who or may have access to their data and decide about the adequacy of the protection and the possible release of the private information.

Ethical guidelines recommend anonymisation of any identifiable data. Data is anonymised when any identifier disappears. If an item may serve, alone or with others, allowing re-identifying the person, the anonymization was not made and the data should not be retained.

6.2. Informed Consent In Research Involving Human Subjects

6.2.1. Short History of Informed Consent

The need for informed consent is regulated in all codes of ethics, declarations and ethical guidelines, the most important of which being the Nuremberg Code, the Declaration of Helsinki and the International Ethical Guidelines for Health-related Research Involving Humans (CIOMS and WHO).

Although the Nuremberg Code in 1947 is considered to be the first document in requiring ethical regulation in relation to informed consent, first claims of informed consent dates back to 1881 when the Prussian Minister of Interior issues a directive addressed to prisons stating the need to respect the patient's will to treat tuberculosis by administering tuberculin "*must in no case be used against the patient's will* " [76], expressing the first time the voluntariness.

In 1900 the Prussian Minister for religious, educational and medical affairs issues the second directive addressed to medical directors stating the need of consent after correct information in relation to non-therapeutically interventions: Non-therapeutic medical interventions are excluded "*if the*

human subject is minor or incompetent for any reason" or the subject did not give "*unambiguous consent*" after the preliminary and "*correct explanation of the possible negative consequences*", expressing the first time the mandatory assessment of legal competency.

The Circular of the Reich Minister of the Interior, issued in 1931 the guidelines for new therapy and human experimentation, expressed clear demands regarding informed consent making distinctions between therapeutic and non-therapeutic research: „*new therapy may be applied only if consent or proxy consent has been given in a clear and undebatable manner following appropriate information. New therapy may be introduced without consent only if it is urgently required and cannot be postponed because of the need to save life or prevent severe damage to health*". Non-therapeutic research was "*under no circumstances permissible without consent*" [77], expressing the first time the ethical non-therapy v. therapy procedures, the „*consent*" expression [78] and the appropriate information on the subject.

The Nuremberg Code established in august 1947, was a response to atrocities committed by physicians during World War II medical experiments and it was designed to protect the rights and welfare of research subjects. One of the strengths of the Nuremberg Code is the fact that the first statement is making it clear that voluntary, competent and informed consent is absolutely essential, meaning that the person involved in the study must have legal capacity to decide free from coercive influence and should have "*sufficient knowledge and comprehension of the elements of the subject matter as to enable him to make and understanding and enlightened decision*." [79]

The Declaration of Helsinki was adopted by the World Medical Association in 1964 and amended 9 times from then, last time in 2013, contains in its last version eight paragraphs on informed consent strengthening the ideas of freely given informed consent, voluntariness in the study, informed consent from the legally authorised representative in the case of incompetent persons.

International Ethical Guidelines for Health- related Research Involving Humans prepared by the Council for International Organizations for Medical Sciences (CIOMS) in collaboration with the World Health Organization (WHO) states that informed consent is consent given by a competent person who receives the information, understands the information, decides in the

interest of his/her own person without being subjected to coercion, intimidation or in a state of vulnerability.

Informed consent is mandatory in any research involving human subjects as an expression of respect for the individual's autonomy, ethical principle of respect for persons.

The goal of informed consent is to ensure that potential subjects of research are fully informed about any risks and benefits that can be obtained as a result of participation in the study [80] .

6.2.2. What autonomy means?

Autonomy is a concept recognised in medical ethics, bioethics, medical practice and national laws and refers to the right of an individual to make free decision that fits their values and interests.

Full autonomy requires three capacities: [81]

- Capacity to understand received information
- Capacity for voluntariness and desire to express his decision regarding the participation in the research procedure
- Capacity for reasoning and deliberation in order to express a decision a.k.a. make a reasoned judgment about the effect of participation in the research which include the communication of this decision

A valid informed consent is typically understood to have three distinct components:

- consent must be preceded by information,
- given on a voluntary basis,
- by a competent person (legally competent).

In general, informed consent can be only given by competent adults who are able to make decisions for themselves. For children and for persons with impaired mental status informed consent can be given by parents or by the legal guardians.

Informed consent is not just a written and signed acceptance of the participant but a continuous process whose main purpose is to ensure that the subject is voluntarily enrolled and that the decision belongs totally to the subject and has been taken in full knowledge of the case.

The process of informed consent does not come to an end after the subject decide to enroll in a study but when the participant reaches the objective of the study or decide to cease participation [82].

Different from the informed consent in medical treatment which addresses a specific moment of the recommended treatment (i.e. a surgical procedure, etc.), the informed consent for a participation in a research is for a longer period of time, as long as the participation and the research itself is lasting.

Another major difference is that informed consent in research requires full disclosure of the data and information in research whereas the treatment procedures only adequate disclosure or as needed as the patient may have full information in order to have a decision in his best interest.

Another important difference is that informed consent in research may not go over the major risks while in medical treatment may go over the risks enabling the only chance the person has (i.e. in emergencies if the indication still exists despite all risks the recommendation is for treatment).

6.2.3. How to inform?

The investigator has the responsibility to give appropriate and pertinent information using an easy-to-understand language, so that the participant fully understands the risks and benefits of participating in the study taking in account that for those who test positive for the targeted mutation the benefit depends on how well they can use the knowledge of their status to prevent the disease itself.

National Human Genome Research Institute, US, recommends involving genetic counselors in the informed consent process because of the complexity of the information to be transmitted, which must be able to answer any questions the participant may have.

Research investigators who are often also physicians has to ensure that prospective participants in the study understand how their relationship will evolve during and after the study.

6.2.4 Informed consent elements

Informed consent form is a written summary of the research project and has some basic elements that should be included:

- Assurance that participation in the study is voluntary
- General description of the study, design of the study, the purpose, objectives and goals of the study, the condition for which the study is being performed, who is sponsoring the study
- How the test will be performed, what samples will be collected and what will happen to them short and long-term, the expected duration of

the subject`s participation in the study; when appropriate, the process of collecting personal/family health history.

- Whether enrolling in the study will allow researchers to have access to participant's medical records and if so, by whom and to what process data
- Information about sharing and the duration of storing should be described; any possible future use in secondary research.
- How the participant will benefit from the research and if it's possible for the subject of research to find out the individual research results.
- Any physical or emotional risks associated with participation in the study; What will happen in cases of incidental findings or information that could extend to relatives or identifiable population or groups
- How the confidentiality will be maintained, if their samples will be attached to their identity or not; any possible privacy breach due to possible re-identification, what information will be made available along with the sample
- Who will have access to the collected samples or data? How and to whom the results of the test will be reported and under what circumstances
- Clear explanation about genomic studies that involve bio-banked samples, individual genomic, health data, if a complete withdrawal of samples and data may or may not be possible after distribution to other laboratories.
- Financial reimbursement or expected compensation provided to participant or any possible expenses incurred by the participant if enrolled in the study
- An explanation to whom to contact for answers to questions about the research
- Ways of future contacts.

No form of informed consent may include additional parts of exculpatory language by which the participant waives any legal rights or releases the investigator or sponsor from liability for negligence.

6.3. An example of research project with human subject – ROMCAN

The RomCan project ("The genetic epidemiology of cancer in Romania") proposes a systematic evaluation of the genetic risk factors associated with breast cancer in women (BrCa), colon and rectal cancer (CRC), prostate (PrCa) and lungs (LuCa), representing almost half of the total number of new cancer cases in the country, in the Romanian population, following the definition of high risk groups for which they are specific, preventive measures can be implemented. Also, the project wanted to verify the fact that there is any modification of the genetic risk effect according to ethnicity, with emphasis on the Roma ethnic group. At the end of the project, a unique

database and a biobank for cancer research in the Romanian population were created, including the assembly of a case-control study of Roma ethnic groups.

Collectively, these studies of different types of cancer have led to a better understanding of the genetic susceptibility to cancer in Romania and will have an immediate use for screening in high risk families. The results of the project allowed to catalog the population frequencies of the common genetic variations in Romania, which are useful for other geneticists in the country. Last, but not least, the project raises the profile of cancer research in Romania internationally and increases the possibility of attracting additional funding. In the long term, this research on the epidemiology and genetic causes of cancer will be useful for policymakers and may affect strategies to reduce the burden of cancer on public health.

The data collection in the RomCan project involved a close collaboration between several institutions. Thus, in several clinics in Bucharest addressing malignant pathologies (prostate cancer at "Prof. Dr. Th. Burghele" Hospital) phenotypic data but also genetic material were collected from cancer patients and from patients with other pathologies that were included in the control group. The activity of these clinics was directed by the Institute of Public Health (INSP), which also functioned as a link between the Romanian institutions involved in the project and the Icelandic partner. It is worth mentioning that the phenotypic data and the genetic material from Romania were sent to Iceland by DeCode Genetics who carried out a Genome Wide Association study to detect SNPs specific to the studied Romanian population that can be incriminated for the appearance of neoplasms and especially of the forms. aggressive cancer.

The medical staff involved in the research work in each clinic was numerous and was not limited to the researcher who comes into direct contact with the patients whose data are collected. First of all, it must be said that the admission of patients in a university clinic also implies the signing of a consent stipulating that they may be approached by the medical staff for the purpose of being included in research activities, whether we are talking about clinical studies or other projects.

After the start of the RomCan project, the doctors involved also carried out an activity of disseminating the information regarding their research work. It was thus desired to establish a connection with the other doctors in the hospital in order to obtain an informal agreement from them in order to involve patients diagnosed by them. This was necessary because the doctors directly

involved in RomCan had to include in the project for the collection of genetic material hundreds of patients with prostate cancer during only 3 years, which was almost impossible without using the other colleagues in the clinic. .

Also very important is the relationship with the Department of Pathological Anatomy. It was agreed that all histopathological findings after puncture positive prostate biopsy for malignancy and prostatectomy should be reported to RomCan researchers. After the patient presents for the removal of the histopathological result, a doctor involved in RomCan is announced by the Department of Pathological Anatomy. After the patient presents himself, he travels together to the attending physician who discusses with the patient the anatomopathological result, its significance but also the need for everyone to get involved in the research work. After obtaining the verbal agreement from the patient, the attending physician guides the patient in a specially arranged space where the researcher from the RomCan project questions him and takes his genetic material.

For the beginning, the researcher discusses again with the patient diagnosed with prostate cancer about his disease, the significance for him but for the general population and will seek to obtain a verbal agreement from him to continue an interview that will not bring him any direct benefit. Moreover, he occupies 20-30 minutes of time, discussing the cancer with which he has just been diagnosed and which will change his whole life. After this agreement has been obtained, the researcher briefly describes the RomCan project and its importance, and then the patient gives his consent (see the informed consent for participation in the study regarding the degenerative diseases in relation to the environmental and genetic factors in Appendix 1).

After reading and signing the patient's consent, the researcher shows him the two mouth swabs with which he uses to take saliva to study the genetic material. The patient is instructed to open his mouth and then, with a swab, he will react with gentle but determined movements inside each cheek. It is intended that the paper of each swab be moistened as a greater amount of saliva leads to a greater chance of isolating the genetic material. The swabs are then labeled with the patient's name and then inserted in an envelope labeled with the investigator's name and the date on which the sampling was performed. It is important to mention that the success of the RomCan project was probably due to the fact that the genetic material was extracted from the saliva and not from the blood, as in the past, which made the procedure easier for patients to accept.

After this time, the actual interview can begin, in which the patient is asked for data on personal, heredocolaterale, environmental, toxic, etc. The document containing the patient's data consists of 5 sections, from A to E and it was completed by the investigator in the presence of the subject.

Section A is the personal data section and contains 21 questions. These refer to identification data such as name, surname, personal identification number (CNP), address, telephone numbers, belonging data, educational status but also social status. It should be noted that at question number 15 the patient is asked what ethnicity is identified with, followed by question number 16 by the investigator to note whether the subject is Roma or ethnic, often ethnic Roma being reluctant to identify themselves as Roma. This is important because a very sensitive point of the RomCan project was the study of the Roma population from a genetic point of view.

Section B refers to medical data. Here, the patient is questioned about the most common pathologies, the exposure to anti-inflammatories and especially to aspirin but also the exposure to X-rays during the most usual radiological investigations. The 10th page is dedicated exclusively to patients and within the RomCan was completed to the patients included in the control group. This page contained 9 questions regarding the patients' gynecological and obstetric history.

In Section C, the cardinal question is whether anyone in the patient's family has suffered from cancer, so the patient's malignant heredocolaterale history is reviewed. The blood relatives referred to in the questionnaire are father, mother, paternal grandparents, maternal grandparents, paternal uncles, maternal uncles, brothers, sisters, sons, daughters. Some neoplasms have a genetic determinism, and prostate cancer is no exception.

Sections D and E are called Tobacco, respectively Alcohol, Coffee, Tea. As the name suggests, it refers to exposure to the above substances.

After the completion of all sections, in which the personal data of the patients are entered, they proceed to complete a different document which is reserved for the case group, not for the witness group. This is a form for clinical and anatomical-pathological data. This form is filled with data from the observation sheets and the anatomopathological reports issued by the Pathology Department by the investigating doctor.

The data will be completed as accurately as possible based on the medical documents and not on the subjective opinions of the doctor conducting the

interview with the patient. In order to avoid errors and for greater ease in digitizing these data these data will be passed using capital letters.

The hospital or clinic in which the patients are interviewed will be identified according to some acronyms previously established (ex: STB = Th. Burghele Hospital).

Also, data on possible treatments of the patient are noted. Then, one proceeds to complete the prospective anatomical-pathological data and the PSA value (prostate specific antigen) in a specific form.

In the documentation there are two columns for TNM staging: clinical staging, specific for prostate biopsy score positive for malignancy, respectively pathological staging, for radical prostatectomy piece, when complete staging can be done.

After all personal data and clinical data are completed by the investigator, the patient is registered in another different document.

It should be mentioned that within RomCan it was encouraged to enroll patients in the control group for subjects with prostate cancer, because it is clear that female patients have zero chance of developing this pathology. In "Prof. Dr. Th. Burghele" hospital there is a section dedicated to patients which was very useful when setting up the control group.

For each patient enrolled in the research project is reserved a number of six labels that present on their surface a number and a bar code. These barcode labels are very useful because once the genetic material and the attachments with the personal and clinical data are delivered they are anonymised (for easier manipulation and subsequent identification).

The investigators involved in research projects are always looking for new subjects and witnesses. The actual action of interviewing the patients and collecting biological material is seconded by the one of filing and labeling. This is considered wrong by the lower rank and unfortunately we often see that many errors can arise from the value of a project. The advice that can be given to future investigators is to give as much importance to this step, which is good to be done immediately after the interview and harvest. It is tempting to let "gather" more cases and then be listed in the annexes of the registry type. This will lead to many different types of errors like confusion between subjects, loss of swabs or labels, incorrect labeling, etc. It should also be noted that investigators are primarily doctors and the research activity conflicts with the daily activity.

The envelopes containing the biological material and the annexes are stored in a special place for the research project, to which only the investigating doctors have access. This is important for two reasons: first of all, avoid the inconvenience caused by the loss of annexes or registers, and secondly, in the case of a control or audit, the investigators must provide strict evidence of the patients' personal data.

The biological material collected, in the back of the patients' saliva swabs, is introduced in small envelopes containing desiccant material. In turn, the envelope with swabs, filled with the name of the investigator and the date of harvest is included in the envelope A4. The A4 envelopes, filled with the patient's name and labeled, are kept in a cabinet specifically designed for this purpose, at room temperature.

A safety measure, which can be very useful in case of confusion or further clarification, after the interviewed person leaves the clinic, is to photograph all the completed and labeled materials. These photos can then be stored on a cloud server or on a hard disk.

The delivery of the finished material, of the A4 envelopes with the collected genetic material and with all the annexes completed is done to the coordinating center staff, in the case of RomCan being the National Institute of Public Health. This process is done at pre-set time intervals, or in agreement with the INSP staff. Once the cases are handed over, the records are also handed over.

Chapter 7. Ethics of scientific publications

As declaration of Helsinki states researchers have the obligation to publish and disseminate results of their studies, including negative or inconclusive results [83].

A scientific paper is a written and published work describing original research results. Scientific Integrity and Ethical Issues in Publishing 2010, International Society of Addiction Journal [84] editors recommends measures to prevent or solve possible ethical issues regarding scientific publication

- Ethical review
- Risk -benefit analysis
- Informed consent
- Peer review
- Uniform requirements for attributing authorship credits
- Professional ethics: Ethical Guidelines

Ethics of scientific publications - „*is a system of standards of professional conduct in relations between authors, reviewers , editors, publishers and readers in the creation , dissemination and use of scientific publications*" [85].

In Romania the legal regulations of publishing ethics are integrated into legislation of research ethics. The most important legislative norms are:

- Law no. 206 of May 24, 2004 (updated) on good conduct in scientific research, technological development and innovation
- Law No. 8 of March 14, 1996 on Copyright and Neighboring Rights
- Education Law No. 1 of January 5, 2011
- Order 5735/2011 regarding Regulation for the organization and functioning of the National Council of Ethics for Scientific Research, Development and Innovation.

According to law no 206/2004, rules of good conduct in research and development includes:

I. Rules of good conduct in scientific activity;
II. Rules of good conduct in the activity of communication, publication, dissemination and scientific popularization, including within the framework of the financing applications submitted in the framework of public funded project competitions;
III. Deviations from the norms of good conduct in the scientific activity includes:
    i. Making results or data and presenting them as experimental data, such as data obtained from computer calculations or

simulations, or as data or results obtained by analytical calculations or deductive reasoning;

ii. Falsification of experimental data, data obtained by computer numerical calculations or simulations or data or results obtained by analytical calculations or deductive reasoning

IV. Deviations from the norms of good conduct in the activity of communication, publication, scientific dissemination includes:

i. Plagiarism
ii. Self-plagiarism
iii. The inclusion in the list of authors of a scientific publication of one or more co-authors who did not contribute significantly to the publication or the exclusion of co-authors who contributed significantly to the publication;
iv. The inclusion in the list of authors of a scientific publication of a person without his consent;
v. The unauthorized publication or dissemination by the authors of unpublished results, hypotheses, theories or scientific methods;
vi. Entering false information in grant or grant applications, in applications for empowerment, for academic teaching positions or for research and development positions.

Kerstin Stenius has set up a list of seven serious deviations from the ethics of publishing and called it "*the seven deadly sins in scientific publishing*"

1. Carelessness

   Failure to adequately review the literature, secondary citation, citation bias, selecting only the sources that support a particular point of view, self-citation, understatement

2. Redundant publication

   Same tables or literature review reported without noting prior source; if falls under copyright infringement

3. Unfair authorship

   Honorary/gift authorship, exclusion of eligible authors

4. Undeclared conflict of interest
5. Human/animal subjects violation
6. Plagiarism
7. Other frauds.

7.1. Plagiarism

Plagiarism is, according to the provisions of art. 4 par. (1) lit. d) of Law no. 206/2004: "*the exposure in a written work or oral communication, including in electronic form, of texts, expressions, ideas, demonstrations, data, hypotheses, theories, results or scientific methods extracted from written works, electronically, by other authors, without mentioning this and without referring to the original sources*".

National Ethics Council for Scientific Research, Technological Development and Innovation from Romania has developed its own definition of plagiarism, in accordance with Law 206/2004 stipulating that "*Plagiarism is when an author appropriates certain elements from the intellectual creation of another author and presents it in the public space as part of their own work. Plagiarism is the result of the action to plagiarize and it refers to the work that is generated by unlawful appropriation, whether intentional or not, from a deontological point of view*".

7.1.1. Common types of plagiarism

Intentional plagiarism occurs when the author knowingly and deliberately copies entire paragraphs or assumes someone else's ideas and passes them off as its own, without any citation.

Unintentional plagiarism is a result of disregarding citation rules or failing to interpret them correctly. Occurs when the author

- does not cite a source that is not common knowledge,
- does not use quotation marks if he indicates the source
- poorly paraphrases

Copy paste or word to word plagiarism occurs when an author takes complete sentences from another source, failing to credit the authors through omission of full text citation and without using quotation marks. (Plagiarism: Types, Causes and How to Avoid This Worldwide Problem; Yam Bahadur Roka; Review Article Nepal Journal of Neuroscience 14:2-6, 2017).

Incomplete or incorrect citation appears even if the source is cited, incomplete or wrong reference makes it difficult to identify the source (Types of plagiarism, https://www.scribbr.com/plagiarism/types-of-plagiarism/).

Paraphrasing plagiarism means rephrasing a text of another author using your own word, changing both the structure and the words. Paraphrasing itself is not plagiarism as long the writer cites properly the source.

Mosaic plagiarism or patchwork plagiarism occurs when the writer uses a text but changes the sequence of words by reordering and replacing with synonyms while keeping the structure of the original text without acknowledging the author

Plagiarism by translation means translating a text from another language without citing the original source

Plagiarism of non-textual sources appears when introducing in a paper a graphic figure, an image, a scheme, a table without citing the source or without obtaining a copyright agreement.

Self-plagiarism is defined by the provisions of Art. 4 par. (1) lit. e) of Law no. 206/2004 as "*the exposure in a written work or oral communication, including in electronic form, of texts, phrases, demonstrations, data, hypotheses, theories, results or scientific methods extracted from written works, including in electronic format, of the same or the same authors, without mentioning this and without referring to the original sources*".

7.1.2. How to avoid plagiarism

In order to avoid suspicion of plagiarism follow the few advices below

1. Correct use of common knowledge
   - Well-known facts
   - Ideas and interpretations are not common knowledge if they are not widely debated and recognized
   - At least 5 different sources
2. Properly quote and paraphrase
   - Quoting- involves copying short sentences from the original text and place it within „quotation marks"; includes an in- text citation, stating the last name of the author, year of publication, page number
   - Paraphrasing- rephrasing the original text, keeping the original meaning of the text maintaining the same ideas of the original text but rephrased with own words; includes an in-text citation stating the last name of the author and the year of publication
   - Summarizing- the original meaning may be changed, the author may maintain or not the same ideas of the original text or may summarize them in only one sentence using the writer's own words, usually in a much shorter sentence/sentences than the original text
3. Properly citing of sources
   - Citing twice- all the information must be cited both in text and in references list. The citation is in the text exactly in the place where the idea that belongs to another person is used by paraphrasing or summarizing. The reference list presents the way to the citation

reference as the author of the present work may study the original creation. Therefore there is not a double citation but only one citation within the text and its reference in the reference list.

7.1.3. Assigning authorship

Who deserves authorship or which is the order in which the authors should be listed may lead to ethical issues at the time of submission of the manuscript. The International Committee of Medical Journal Editors (ICMJE) has developed a set of four criteria to define the authorship:

1. significant contribution in conception and design, data acquisition, analysis and interpretation of the results
2. drafting the manuscript or revising it for intellectual content
3. final approval of the version to be published
4. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

All these conditions must be fulfilled simultaneously for a person to be placed on the list of authors. The authorship's condition necessarily implies that any author can identify and take responsibility for at least one component of the work (which belongs totally) and be able to identify which co-authors are responsible for specific other parts of the work.

ICJME recommends that the group decide who will be an author before starting the work and confirm who is an author before submitting the manuscript for publication.

In the author list, the first and last positions are the most sought-after ones. There are no guidelines regarding first position, as a common rule, usually the first author is the person who had the highest contribution to the study design, acquiring and analysing data from experiments, the one who writes the draft of the manuscript.

The name of the first author has the highest visibility, especially when the citation rules reduce all other names at ”et al.”.

Traditionally, the last author position is reserved for the senior author, the supervisor, the principal investigator who had a significant contribution in coordinating the work, usually the one with the highest contribution following the first author.

Corresponding author is the person who has the responsibility of submitting the manuscript to the publication. The corresponding author is the one who

receives all notifications from the journal including manuscript status, reviewers' comments, and the final decision. All three positions are principals authors in medical Journals.

The sequence of authors should be determined by the relative contribution to the manuscript but other methods are common:

- multiple first authors: if the scientific work has more than one first author, each name will be marked with a symbol, followed by an explanatory note
- alphabetical list;
- in the order of the institutions they belong to.

Several methods have been developed to quantify the contribution of the authors and ease the sequence of authors. Stephen M. Kosslyn proposes 6 criteria with different point values, from a total of 1000 points:

- Idea of the study – 250 points
- Design of the study – 100 points
- The implementation of the study – 100 points
- Conducting the experiments -100 points
- Data analysis – 200 points
- Writing – 250 -points.

To achieve the quality of the author, you must get at least 10% of the total points and the order of the authors is determined by the number of points [86].

The Ahmed method [87] assigned 1(minimal), 3(some) or 5 (significant) points for each contribution and the author list will be in the descending order of the total score obtained by each author.

	First author	Second author	Third author
Conception	5	3	1
Design			
Implementation			
Data analysis			
Drafting the article			
Revising the article			
Public responsability			
Total score			

Table 7.1.

Any other person who assisted the study but does not meet the criteria for authorship should be listed as a contributor in the "Acknowledgement" section, giving their names and specific roles. In the list of contributors can be included persons who helped with technical editing, proofreading, technical assistance, assistance in lab work, provided research space, obtained

financial support, participate in the clinical trial, contributes to casuistry but did not write the paper (elsevier, icjme,

https://www.enago.com/academy/authorship-and-contributorship-in-scientific-publications/)

7.1.4. Unethical authorship practices

Among the most notorious unethical authorship are:

- Including in the list of authors a person who has not contributed in the study,
- Including an author without his/her permission and/or without asking
- Exclusion from the list of authors of a person involved in the study
- Removing names of contributors.

Honorary or gift authorship

This involves naming in the list of authors of a person who has made no contribution in the study and does not meet any criteria for authorship. Is often practiced as a form of respect or gratitude of a person who holds a leading position or has a high notoriety in the academic world.

Guest authorship

This involves naming a person, generally a senior, as an author, hoping that his name will increase the chances of a manuscript being published although his contribution in the study may be minimal.

Ghost authorship

This practice involves exclusion from the list of authors as well from the "Acknowledgement" section, of a person who has a significant contribution in the study. These individuals can be those who might be perceived having conflicts of interest or professional medical writers often paid by sponsors.

World Association of Medical Editors and COPE recommends several methods to combat ghost and guest authorship:

- Clear authorship criteria
- Requirement for a statement signed by all authors about their contribution to the work and explicitly state the contribution of each.
- disclose all contributors, regardless of author status, and their specific individual contributions and affiliations
- Authors should complete a checklist if they received help from a medical writers

Bibliography

[1] Kamashi Pandirajan. (2021, January) TeachMe Physiology. [Online]. https://teachmephysiology.com/biochemistry/cell-growth-death/cell-cycle/

[2] P Mandira. (2017, November) Socratic. [Online]. https://socratic.org/questions/589fd775b72cff178533a701

[3] Benjamin Cummings. (2008) https://www.pearson.com. [Online]. https://medium.com/@biologynotes/meiosis-16a31fec3838

[4] Lisa Urry, Michael Cain, Steven Wasserman, Peter Minorsky, and Jane Reece, *Campbell Biology in Focus, Books a la Carte Edition*, 2nd ed.: Pearsons, 2016.

[5] D Johnson. (2013, June) University of Samford. [Online]. http://www2.samford.edu/~djohnso2/jlb/333/division.html

[6] University of Vigo. (2019, July) Atlas of Plant and Animal Histology. [Online]. https://mmegias.webs.uvigo.es/02-english/5-celulas/ampliaciones/8-cromosomas.php

[7] Charles Mallery. (2007, March) Introductory Biology. [Online]. http://fig.cox.miami.edu/~cmallery/150/mendel/karyotype.htm

[8] NHGRI. (2020, August) National Human Genome Research Institute. [Online]. https://www.genome.gov/about-genomics/fact-sheets/Chromosome-Abnormalities-Fact-Sheet

[9] Kayeen Vadakkan. (2018, June) St. MaryCollege Thrissur. [Online]. https://www.slideshare.net/kayeenvadakkan/gene-structure

[10] Connie Rye et al. (2016, October) Open Stax. [Online]. https://openstax.org/books/biology/pages/15-1-the-genetic-code

[11] Enciyclopaedia Britannica. (2013) Enciyclopaedia Britannica. [Online]. https://www.britannica.com/science/gene

[12] Emerald Robertson. (2015) Gene Mutations. [Online]. https://slideplayer.com/slide/6188442/

[13] Khan Academy. (2016) Khan Academy. [Online]. (https://www.khanacademy.org/science/biology/classical-genetics/chromosomal-basis-of-genetics/a/linkage-mapping

[14] Glen Wolkenfeld. (2017) Learn Biology. [Online]. https://learn-biology.com/ap-biology/module-18-meiosis/understanding-sordaria/

[15] Lee Silver. (2008, January) Mouse Genome Informatics. [Online]. http://www.informatics.jax.org/silver/figures/figure7-1.shtml

[16] Jessica Joyce. (2018) Genetics & Genomics. [Online]. https://ro.pinterest.com/pin/825636544157667753/?nic_v1=1aZUD%2BgWvUutwEqhw9eK0lJiIOsteusac3wq0M1wYLDi55A2PELMiREpkS3olsq%2BDO

[17] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, February 2001. [Online]. https://www.nature.com/articles/35057062

[18] The Human Genome Structural Variation Working Group, "Completing the map of human genetic variation," *Nature*, vol. 447, pp. 161–165, May 2007.

[19] Daniel Gudbjartsson, Patrick Sulem, and Hannes Helgason, "Sequence variants from whole genome sequencing a large group of Icelanders," *Scientific Data*, vol. 2, no. 150011, March 2015.

[20] Nicholas Schork, Sarah Murray, Kelly Frazer, and Eric Topol, "Common vs. Rare Allele Hypotheses for Complex Diseases," *Current opinion in genetics & development*, vol. 19, no. 3, pp. 212–219, June 2009.

[21] MacNamara Á. Collins D, "Much Ado about? A response to Hardy et al. ," *Prog Brain Res.* , vol. 232, pp. 149-154, Prog Brain Res. 2017, .

[22] Murray SS, Schork NJ, Topol EJ. Frazer KA, "Human genetic variation and its contribution to complex traits," *Nat Rev Genet.*, vol. 10(4), pp. 241-51, 2009.

[23] The 1000 Genomes Project Consortium, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, October 2015.

[24] Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA Abecasis GR, "A map of human genome variation from population-scale sequencing. 1000 Genomes Project Consortium," *Nature*, pp. 1061-73, 2010.

[25] Metzker ML, "Sequencing technologies - the next generation," *Nat Rev Genet.*, vol. 11(1), pp. 31-46, 2010.

[26] Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, Cooper DN. Stenson PD, "The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation

data for medical research, genetic diagnosis and next-generation sequencing studies," *Hum Genet.*, vol. 136(6), pp. 665-677, Jun 2017.

[27] Goddard KA, Hollombe C, Tingle SR, Gillanders EM, Mechanic LE, Nelson SA. Feigelson HS, "Approaches to integrating germline and tumor genomic data in cancer research.," *Carcinogenesis*, vol. 35(10), pp. 2157-63, 2014.

[28] Freedman ML Pomerantz MM, "The genetics of cancer risk," *Cancer J.*, vol. 17(6), pp. 416-22, Nov-Dec 2011.

[29] Lorelei Mucci, Jacob Hjelmborg, and Jennifer Harris, "Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. Nordic Twin Study of Cancer (NorTwinCan) Collaboration.," *JAMA.*, vol. 315, no. 1, pp. 68-76, Jan 2016,. ; 2016 Jan 5;. doi: 10.1001/jama.2015.17703. PMID: 26746459.

[30] Hodgson S., "Mechanisms of inherited cancer susceptibility," *J Zhejiang Univ Sci B*, vol. 9(1), pp. 1-4, Jan 2008.

[31] Tomlinson I. Bodmer W, "Rare genetic variants and the risk of cancer," *Curr Opin Genet Dev*, vol. 20(3), pp. 262-7, Jun 2010.

[32] Howard MS Abreu Velez AM, "Tumor-suppressor Genes, Cell Cycle Regulatory Checkpoints, and the Skin," *N Am J Med Sci*, vol. 7(5), pp. 7(5):176-88, May 2015.

[33] Sherr C, "Principles of tumor suppression," *J. Cell.*, vol. 116(2), pp. 235-46, 2004, 2004 Jan 23. doi: 10.1016/s0092-8674(03)01075-4. PMID: 14744434.

[34] Sun Y, Ehrlich M, Lu T, Kloog Y, Weinberg RA, Lodish HF, Henis YI. Liu X, "Disruption of TGF-beta growth inhibition by oncogenic ras is linked to p27Kip1 mislocalization.," *Oncogene.*, vol. 19(51), pp. 5926-35, Nov 2000, .

[35] Vogt PK Weiss RA, "100 years of Rous sarcoma virus. ," *J Exp Med.*, vol. 208(12), pp. 2351-5, Nov 2011.

[36] Varmus H., "A Prize for Cancer Prevention," *Cell.*, vol. 171(1), pp. 14-17, Sep 2017, 2017 Sep 21; doi: 10.1016/j.cell.2017.08.020. Epub 2017 Sep 6. PMID: 28888324.

[37] D Greenbaum, M Gerstein N M Luscombe 1, "What is bioinformatics? A proposed definition and overview of the field," *Methods Inf Med* , vol. 40(4), pp. 346-58, 2001.

[38] S Zhang and S Liu. (2013) Science Direct. [Online]. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/bioinformatics

[39] Ardeshir Bayat, "Science, medicine, and the future Bioinformatics ," *BMJ*, vol. 324(7344), pp. 1018–1022, Apr 2002. [Online]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1122955/

[40] The Jackson Laboratory. Genetics vs. genomics. [Online]. https://www.jax.org/personalized-medicine/precision-medicine-and-you/genetics-vs-genomics#

[41] The National Human Genome Research Institute. (2020, August) NHGRI. [Online]. https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics

[42] Wikipedia. (2020, September) Wikipedia. [Online]. https://en.wikipedia.org/wiki/Human_Genome_Project

[43] Prakash S Bisen. Impact of HGP on Molecular Diagnostics. [Online]. https://www.researchgate.net/figure/The-outcomes-of-the-Human-genome-project-allows-us-to-unravel-some-of-the-mysteries-of_fig2_283700575

[44] Ashwin Dalal Divya Pasumarthi. Pseudogenes: Implications in Disease and Diagnostics. [Online]. http://iamg.in/genetic_clinics/full_text.php?id=290

[45] (2020, May) Wikipedia. [Online]. https://en.wikipedia.org/wiki/DNA

[46] Wikipedia. (2020, March) Wikipedia. [Online]. https://en.wikipedia.org/wiki/Chi-square_distribution

[47] Basel Kayyali, David Knott, and Steve Van Kuiken. (2013, April) McKinsey Insights. [Online]. http://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care

[48] Sanjeev Agrawal. (2017, October) Harvard Business Review. [Online]. https://hbr.org/2017/10/why-hospitals-need-better-data-science

[49] Andoria Ionita, Steluta Radoi, Delia Gusicov, Dana Stanciu, and Cornelia Sireanu, "Activitatea unităților sanitare în anul 2017," Institutul National de Statistica, Bucharest, Annual Report 2018. [Online]. http://www.insse.ro/cms/sites/default/files/field/publicatii/activitatea_u

nitatilor_sanitare_anul_2017.pdf

[50] Dan Vesset. (2018, May) IBM Business Analytics Blog. [Online]. https://www.ibm.com/blogs/business-analytics/descriptive-analytics-101-what-happened/

[51] Dan Vesset. (2018, May) IBM Business Analytics Blog. [Online]. https://www.ibm.com/blogs/business-analytics/predictive-analytics-101-will-happen-next/

[52] Dan Vesset. (2018, May) IBM Business Analytics Blog. [Online]. https://www.ibm.com/blogs/business-analytics/prescriptive-analytics-done/

[53] Berkeley School of Information. (2019) What Is Data Science? [Online]. https://ischoolonline.berkeley.edu/data-science/what-is-data-science/

[54] American Hospital Association, "Hospitals and Care Systems of the Future," Guide 2011.

[55] Brian Beach. (2013, November) Backblaze Blog. [Online]. https://www.backblaze.com/blog/how-long-do-disk-drives-last/

[56] Andy Klein. (2018, May) Blackbaze Blog. [Online]. https://www.backblaze.com/blog/hard-drive-stats-for-q1-2018/

[57] Eduardo Piheiro, Wolf-Dietrich Weber, and Luiz Barosso, "Failure Trends in a Large Disk Drive Population," in *Proceedings of the 5th USENIX Conference on File and Storage Technologies*, 2007.

[58] Roderick Bauer. (2019, February) Blackbase. [Online]. https://www.backblaze.com/blog/how-reliable-are-ssds/

[59] Jim Salter. (2016, October) Wikipedia. [Online]. https://commons.wikimedia.org/w/index.php?curid=64047258

[60] Wikipedia. (2019, May) Internet Protocol Suite. [Online]. https://en.wikipedia.org/wiki/Internet_protocol_suite

[61] Peter Mell and Tim Grance. (2011, September) NIST - Information Technology Laboratory. [Online]. https://csrc.nist.gov/publications/detail/sp/800-145/final

[62] Wikipedia. (2019, May) Wikipedia. [Online]. https://en.wikipedia.org/wiki/FASTQ_format

[63] Illumina. (2019, December) Illunima Basespace. [Online]. https://help.basespace.illumina.com/articles/descriptive/fastq-files/

[64] Wikipedia. (2019, Spetember) Wikipedia. [Online]. https://en.wikipedia.org/wiki/ACID

[65] Wikipedia. (2019, September) Wikipedia. [Online]. https://en.wikipedia.org/wiki/Eventual_consistency

[66] Rose, Crow, Graham, Heath, Sue and Charles, Vikki Wiles, "The Management of Confidentiality and Anonymity in Social Research'," *International Journal of Social Research Methodology*, vol. I, 11:5, pp. 417-428, 2008.

[67] (2018) WMA Declaration of Helsinki - Ethical Principles for Medical Research Involving Human. [Online]. https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects

[68] Council for International Organizations of Medical Sciences, "Research involving Humans," International Ethical Guidelines for Health-related Research involving Humans, 2016.

[69] The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, "Belmont Report," Department of Health, Education, and Welfare, The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research,.

[70] Intersofgt Consulting. (2019, June) General Data Protection Regulation. [Online]. https://gdpr-info.eu/

[71] Michelle Goddard, "The EU General Data Protection Regulation (GDPR): European Regulation that has a global impact," *International Journal of Market Research*, vol. 59, no. 6, pp. 703-705, Nov 2017.

[72] Justice Institute of British Columbia. (2019, Jun) Privacy and Confidentiality of Research Information. [Online]. www.jibc.ca/procedure/3404-008

[73] (2019) Panel on Research Ethics, Government of Canada. [Online]. www.pre.ethics.gc.ca

[74] Institutional Review Board. Protecting confidentiality and anonymity. [Online]. www.irb.vt.edu

[75] UCI Office of Research. Data security. [Online]. www.research.uci.edu/compliance/human-research-protections/researchers/data-security.html

[76] J., & Winau, R. Vollmann, "Informed consent in human experimentation before the Nuremberg code.," *BMJ (Clinical research ed.)*, vol. 313(7070), pp. 1445–1449, 1996.

[77] Lazar O., Purcaru A., Rogozea L. Purcaru D., "Evolutia consimtamantului informat in cercetarea clinica," *Istoria Medicinii*, 2012.

[78] Helsinki Declaration, "Informed consent," *WMA*, 1964.

[79] Leo Axexander and Andrew Ivy, "The Nuremberg Code," United States Counsel for War Crimes, Nuremberg, December 1947.

[80] Beskow LM McGuire AL, "Informed consent in genomics and genetic research," *Annu Rev Genomics Hum Genet*, vol. 11, pp. 361–381, 2010.

[81] Emmanuel Ezekiel, Christine Grady, and Robert Crouch, *Philosophical Justification of Informed Consent in Research*.: The Oxford Textbook of Clinical Research.

[82] Owonikoko T. K., "Upholding the principles of autonomy, beneficence, and justice in phase I clinical trials.," *The oncologist*, vol. 18, no. 3, pp. 242–244, 2013.

[83] World Medical Association. (2019, May) The World Medical Association (WMA). [Online]. https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects

[84] European Association of Science Editors. (2019, May) International Society of Addiction Journal editors. [Online]. https://www.iasociety.org/Web/WebContent/File/AIDS2010_Scientific_Integrity_Wor kshop_presentation.pdf

[85] University of St. Petersburg. (2019, May) Vestnik. [Online]. http://vestnik.spbu.ru/ENG/Editorial/ethics/s03.pdf

[86] John Lindsley Stephen M. Kosslyn. Criteria for authorship. [Online]. https://kosslynlab.fas.harvard.edu/files/kosslynlab/files/authorship_criteria_nov02.pdf

[87] C A Maurana, J A Engle, D E Uddin, K D Glaus S M Ahmed, "A method for assingning authorship in multiauthored publication," *Fam Med*, vol. 29:, no. 1, pp. 42-4, 1997.

[88] Wikipedia. (2019, October) Prescriptive Analitics. [Online]. https://en.wikipedia.org/wiki/Prescriptive_analytics

[89] Wikipedia. (2019, October) Predictive Analytics. [Online]. https://en.wikipedia.org/wiki/Predictive_analytics

Appendix 1

Informed consent for participation in the study regarding the degenerative diseases in relation to the environmental and genetic factors - ROMCAN

Patient Name ....

Study information:

The study investigates in detail the environmental factors, lifestyle, genetic predisposition and their effect on tumor degenerative diseases in the population.

What it means for you to participate in the study:

You will be contacted to complete a detailed questionnaire regarding your lifestyle and working conditions. Completing the questionnaire can take 30-45 min. All information will be confidential. You will be asked to agree to the collection of two saliva samples. Saliva will be examined to determine genetic markers, which may induce an increased predisposition for degenerative tumor diseases. The results are anonymous and will not be given to any study participants, as their clinical utility (for diagnosis or treatment) is not yet clear.

The risks of the study:

There is no direct risk for participating in this study, only in the collection of saliva can there be some small inconveniences due to the procedure. Participation or non-participation in the study will have no influence on your therapy and medical care.

Benefits of the study:

The study is meant to increase understanding of multiple factors and mechanisms that cause different diseases in humans and to help develop effective protective measures to prevent them. Saliva samples will be used to investigate genetic risk factors, which increase the likelihood of the disease occurring. The unused biological samples will be kept in the biobank for further analysis.

This study will not have a direct benefit for you.

Popularizing information:

Your samples will be completely anonymous, meaning the researchers will not be able to correlate the results with your name. The anonymous results of the genetic tests will be popularized for scientific purposes only.

I was informed about the study and had the opportunity to ask questions. All information will be strictly confidential and my name will not be used in the study.

| __ | *I agree with the interview*

| __ | *I agree to participate in the study, with the collection of saliva and with their use in identifying risk factors for degenerative diseases.*

| __ | I agree with the examination of my medical sheet by the doctor participating in the study, in order to obtain the clinical information, as well as other medical records.

| __ | I agree with the popularization, at the level of other scientific centers, of the information regarding the genetic and cellular characteristics.

If you need to contact us in the future for other information, will you agree?

| __ | Yes | __ | No

Date: ..... Patient's signature .....