1

Biobanks, a Source of Human Samples and Health Data

Precision medicine is forcing modern biobanks to shift their focus from sample-driven to data-driven strategies.

[KOZ 19]

In October 2018, the scientific journal *Nature* provided an international platform for the UK Biobank, a biobank of biological samples. This biobank collected biological samples from more than half a million individuals between 2006 and 2010. All volunteers and spread throughout the United Kingdom, 500,000 individuals gave their time, their medical data and their samples.

In this issue of *Nature*, and for the first time, the publications gave a description of the entire cohort, including the genetic data of all individuals [BYC 18]. Then, in a second study, the researchers showed the brain imaging results of 10,000 individuals. They revealed the genetic influence on brain structure and function and showed correlations between personality and neurodegenerative and psychiatric traits [ELL 18]. These results are valuable because they contribute to the study of a population cohort and will be used in other genomic medicine projects.

The UK Biobank is the first project to carry out a large-scale collection. It successfully demonstrates, on a population scale, that clinical, genetic and

physical information are linked. Other countries have also launched biobanking projects such as Estonia, Japan, Canada, Finland and Iceland. However, the UK has overtaken them. In 2010, the UK Biobank was building a prospective cohort of some 500,000 individuals aged 40–69 years.

In the October 2018 *Nature* publications, Bycroft [BYC 18], Elliot [ELL 18] and their teams described an immense resource of genomic data linked to clinical data from 500,000 individuals. This large-scale biobank promises to uncover relationships between genomic variations and diseases and help better understand their mechanisms.

As described by Bycroft et al., the 500,000 UK Biobank participants gave urine, saliva and blood samples, which were used for genetic analysis and *biomarker* testing of known diseases. Participants were between the ages of 40 and 69 between 2006 and 2010 when they were included in the study. This age range corresponds to the age of onset of disease in adults. Volunteers filled out questionnaires that included information on several factors, such as family history of disease and lifestyle. They also gave the researchers access to their medical records.

While Bycroft and his team detailed the generation of genomic data from the biobank, Elliot et al. provided insight into how these data could help discover genetic associations of disease. The authors explored brain imaging data from more than 8,400 participants. These data were processed to generate a list of thousands of image-derived phenotypes (IDPs), which are features related to brain structure and function that can be identified from medical imaging images. Elliot and his team studied the associations between IDPs and genetic variants.

In many population-cohort studies, access to data is not possible until the results are published. In some cases, some researchers do not even make all the data available. The UK Biobank, funded by the Medical Research Council and the Wellcome Trust, has made the complete data available to researchers.

The size of this resource is unprecedented in terms of the number of samples and the volume of data collected. From the beginning, the premise of the project was based on sharing all data with any human health researcher or scientist. As a result, thousands of scientists around the world

have been able to conduct their research projects on this data since July 2017.

Is the UK Biobank unique? Are there other biobanks? What are their activities?

In this first chapter, we define the concept of the human sample bank by providing examples of population and clinical biobanks. Then, we present their mapping according to geographical distribution by locating the major international biobanks. Finally, we describe the management of the biological samples and processes of a biobank, which combine technical expertise and the infallible maintenance of the quality of the human sample within a legal and ethical framework.

1.1. From the collection of biological samples to the concept of a biobank

According to the principles of data sharing and access to samples for all researchers, biobanks must be equipped with communication tools to facilitate the exchange of material and knowledge. The work of Fransson et al. has highlighted the priority of a common language shared among all actors in the multidisciplinary field of biobanking [FRA 15].

Biobanking activities mobilize the resources of biology to describe the properties of samples, of medicine for clinicopathological annotations, of informatics to manage the data associated with the samples and of law to define the regulatory framework for sample donation and to protect the personal data of patients. The terminology used by the biobanking community is found in the lexicons of biology, medicine and law. A preliminary explanation of the words used by biobank users and scientists using the associated data is a necessary preamble to entering the biobanking world.

Fransson et al. analyzed the language used in biobanks and showed that there is no universal definition of the word “biobank”. They selected a dozen terms related to biological sample management activities, which they submitted to biobank users. The co-authors – an epidemiologist, a lawyer and a computer scientist – asked about 100 people working in European biobanks to choose the most appropriate definition from a list of proposals

[FRA 15]. This study identified a wide range of definitions. Two answers came out on top in defining a (human) *biobank*.

The first definition explains the concept of a biobank as "an organized collection of human biological material and associated information stored for one or more research purposes". This definition comes from the lexicon of the international consortium called the Public Population Project in Genomics and Society (P3G). This non-profit consortium brings to the international scientific community expertise, resources and innovative tools for medical and social science research.

The second definition comes from the European biobanking infrastructure called the Biobanking and Biomolecular Resources Research Infrastructure – European Research Infrastructure Consortium (BBMRI-ERIC). Biobanks are defined as "collections, repositories and distribution centres of all types of human biological samples, such as blood, tissues, cells or DNA and/or related data such as associated clinical and research data, as well as biomolecular resources, including model- and micro-organisms that might contribute to the understanding of the physiology and diseases of humans". The authors state that two-thirds of the responses came from biobank staff involved in clinical trials, which explains the development of the activities specified in the second definition. The remaining third were from a population biobank.

In this study, Fransson et al. show that some terms overlap, such as sample and specimen. The *sample* is the unit of material derived from a specimen according to the Organization for Economic Cooperation and Development (OECD). The *specimen* is a specific tissue, such as a blood or urine sample, taken from a participant at a specific time. According to the National Cancer Institute's thesaurus, the term *biospecimen* defines any material sample taken from a biological entity for scientific research.

Other terms discussed, such as "sample collection", "aliquot" and "personal data", focus on the research orientation and exclude clinical applications. Ideally, definitions should be universally accepted across sectors to limit confusion among the biobanking community.

The number of publications mentioning the term "biobank" is a reliable indicator of the activity's development. Over the last 10 years, this number has increased by a factor of more than 10 according to the PubMed NCBI

bibliographic database. There were 223 publications in 2010 and 2687 in 2020 and 2021. Among the publications of the year 2020, the term "biobanking" was cross-referenced with each of the following terms: health, quality, ethics, economics and informatics. It was found that human health was by far the most important objective of (human) biobanks, followed closely by sample quality and informatics resources. Two disciplines appeared in the ranking with a small but growing number of publications each year: bioethics and economics.

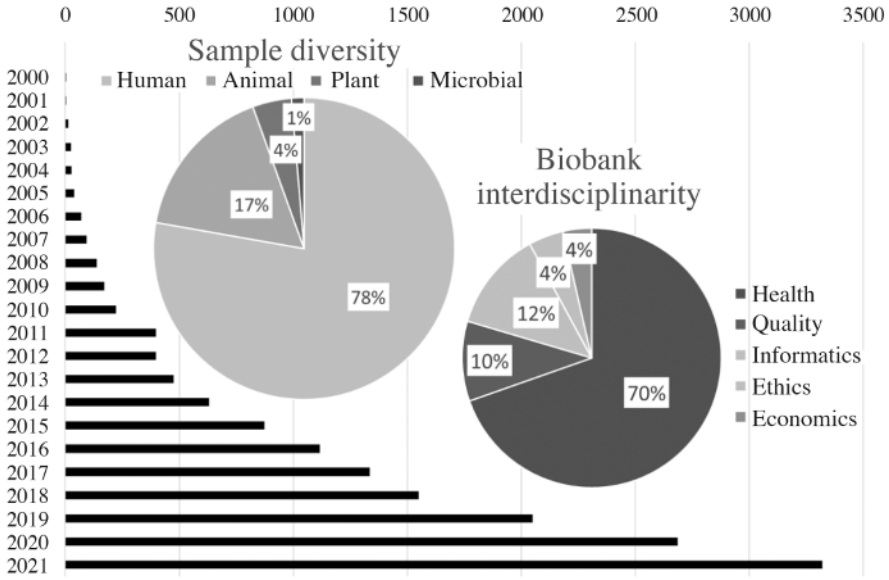

Figure 1.1. *Annual number of publications with the term "biobanking" from 2000 to 2021 according to the PubMed NCBI bibliographic database. For a color version of this figure, see www.iste.co.uk/arrighi/biobanks.zip*

COMMENTARY ON FIGURE 1.1.– *The various biological samples are from human, animal, plant and microbial organisms. In the publications from 2020, the disciplinary fields concerned were health, quality, informatics, ethics and economics*.

Bioethics highlights the fact that in genetic databases the obligations of confidentiality and donor consent are indispensable. The work of Godard et al. raised the issue in 2004 [GOD 04]. They showed how public consultations have raised awareness of sample donation in Iceland, Estonia,

the United Kingdom and Quebec, the countries where the first large-scale biobanks were located. They communicated about the ethics of their project, and the researchers gained the trust of the population. Patient volunteers entrusted their personal and medical data to the researchers to build the first cohort-populations. In 2005, Baird and Frome addressed the cost of storing hundreds of samples at very low temperatures [BAI 05]. The investments in infrastructure (building and freezer equipment) and operating costs represent a significant financial commitment, hence the need to design a business model for large-scale biobanks in order to ensure their longevity. Publications on this topic are becoming more frequent.

1.1.1. *The biobank concept*

The development of medical research requires an ever-increasing number of qualified human tissues and samples. Researchers are faced with an eternal need for human material for the validation of new therapeutic tools. They consult with surgeons and establish collaborations with clinicians who will take specimens from patients during surgical procedures. Samples are mainly reserved for diagnostic purposes and are rarely used for scientific research. Access to samples is therefore limited.

Once the sample has been identified and authorized (date and time of the procedure, mode of transport of the sample), the researcher must be able to be sure of its conservation as it guarantees the sample's quality. Part of the tissue or tumor is analyzed and the remaining part is stored at very low temperature or used in another project.

In a hospital department, the storage of samples for a long period of time is impossible due to lack of space. The researcher must therefore diversify their sources of specimens, especially for rare diseases (multicenter cohort). From the need to access samples and considering the investment in time devoted to the search for samples – and not to the experiments – the idea emerged to create a structure specific to the scientific community with:

– hundreds of samples;

– organized collections;

– high-quality specimens and samples.

Thus, the concept of the biobank was born. According to Hewitt and Watson, the term biobank is applied to biological collections of samples of human, animal, plant and microbial origin [HEW 13]. It refers not only to collections of samples and their associated data, but also to the management of samples according to professional standards. The term biobank is general, regardless of the size of the biobank, the number and the purpose of the samples. Generally speaking, biobanks are professional "repositories" of biological samples.

1.1.2. *The first biobank described in the Framingham study*

The first biobank was described in the Framingham Heart Study. Begun in 1947, this epidemiological approach to heart disease was carried out with the population of Framingham, a town located about 30 km from Boston in Massachusetts [DAW 66]. The 28,000 inhabitants of Framingham constituted a representative sample of the American population. The cohort was large, with 5,209 individuals aged 30–60 years in 1947. It has two advantages over other studies conducted at the same time: it was made up of the general population, i.e. the population was not selected, which eliminated the bias associated with the selection of participants; and, unusually, it included 2,873 women, which would allow for later tracking of the onset of heart disease by sex.

Thomas Dawber, director of the study for 15 years, from 1949 to 1965, described the study design and its implementation mobilizing medical personnel from nearby hospitals. Using this test population, the objective of the study was to improve understanding of the development of cardiovascular disease and to detect the early signs of disease. The clinical examination record was taken every two years. Interpretation of the participants' diagnostic and exposure data was standardized according to the recommendations of a technical committee of cardiologists. Through the rigor of its results, the Framingham survey has made it possible to identify the main risk factors for cardiovascular disease, such as high blood pressure, high blood cholesterol, smoking and obesity. It is still active today with its third generation of participants; it is a prospective cohort study model [SPL 07].

What samples and data are collected in this pioneering biobank?

The scientific committee defined the descriptive, medical and biological examinations for the participants [DAW 51]. Each individual was subjected to a thorough medical questionnaire about their medical, family and lifestyle history. Next, the personal examination recorded each participant's anthropometric parameters (weight, height, chest and waist circumference, eye and hair color, etc.), health status (presence of tumor, liver hypertrophy, femoral pulse, etc.) and, more specifically, cardiac constants (heart rate, blood pressure, etc.). The blood samples and clinical examinations dictated by the committee were attended to in nearby hospitals.

Biological constants were monitored and interpreted by the same committee based on:

– blood samples for biochemical (glucose, hemoglobin, blood cholesterol, serum phospholipids, uric acid) and serological (syphilis test) assays;

– urine samples (routine analysis);

– clinical examinations (X-ray and electrocardiogram).

Dawber et al. invented the *population biobank* by collecting patient information from medical history, physical examinations and laboratory tests. They associated the characteristics of each individual with the potential development of cardiovascular disease. They classified the population into two groups: the first with symptoms of arteriosclerosis or cardiovascular disease and the second without them, the healthy population.

1.1.3. *Classification of human sample biobanks*

In 2001, the OECD formalized the role of human sample collections, historically called BRCs. For the OECD, BRCs are "centers for the preservation of living cells, the genome of various organisms, and information on the heredity and function of biological systems. BRCs must meet the high standards of quality and expertise demanded by the international research community and industry for the dissemination of biological information and materials" [OEC 01].

The European Commission Joint Research Centre (EC-JRC) lists the five general missions of a biobank:

– collect and store medically and sometimes epidemiologically annotated biological samples;

– conceive of sample collections as dynamic – not static – projects that evolve over the long term;

– associate the specimens with current and future research projects;

– code donations to ensure confidentiality of data, while allowing the patients' identities to be traced if clinical information were to be transmitted to the patient;

– respect the protection of the rights of the donors and the biobank managers.

There are several types of biobanks that meet these criteria. A first distinction can be made according to the mode of financing, which distinguishes between public and private biobanks. Public biobanks have mainly been developed within university hospitals and cancer centers by benefiting from recurrent public funding. Private biobanks are set up by pharmaceutical companies and contract research organizations (clinical research organizations). Their activity is focused on clinical trials. Private funds can be invested in public biobanks, thus creating a mixed model of public–private partnership. The second distinction is the type of biobank, which is the most common designation. It concerns the donor group and the purpose of the biobank, depending on whether it is for the study of diseases that will eventually be developed in a population or the specific study of a single disease, such as cancer.

The three main groups are population biobanks, clinical biobanks and tumor biobanks, as developed below.

1.1.3.1. *The population biobank*

Population biobanks, such as national biobanks, collect samples from the population without selecting participants, with the goal of reflecting the general population. Typically, blood samples are collected at the beginning of the study, along with associated data (medical and family history, lifestyle, environmental exposures, etc.), and then the tests are followed up for several years. The strength of these studies lies firstly in the observation

of the occurrence of the disease in a healthy population, secondly in the identification of genetic variants predisposing to the disease and finally in the identification of environmental risk factors. This prospective approach is essential in preventive medical programs. Thanks to population biobanks, it is possible to study predictive biomarkers when the participants have not yet developed the disease. The only drawback is that the study must be conducted over a long period of time (10–15 years), which is usually the time frame in which the disease develops.

The best example of a population-cohort is the Estonian genome center of the University of Tartu (EGCUT). The project was launched in 2018 with the goal of sequencing the genome of the entire Estonian population of 1.3 million [LEI 15]. Currently, the cohort size is 200,000 adult donors, reflecting the population distribution by age, gender and geographic location. The biobank has DNA, plasma and white blood cells for each donor. The aim of this biobank is to collect (genetic) information from the Estonian population and to use it as a preventive tool for personalized medicine. The originality of this biobank is that it is a public biobank for large-scale DNA collection and analysis.

1.1.3.2. *The clinical biobank*

Clinical biobanks dedicated to a pathology have a selection criterion. The specimens collected come from patients suffering from a particular disease. Collected during the patient's medical diagnosis or treatment, these pathological samples may be biological fluids (blood and urine) or tissues. They are often stored at the point of collection, i.e. in hospital biobanks, and stored for research purposes.

Integrated in the University Hospital of Nice, the Côte d'Azur biobank (Laboratory of Clinical and Experimental Pathology, LPCE06) is dedicated to lung cancer. Washetine et al. analyzed the organization of this biobank since its beginning in 2006 using a quantitative and qualitative approach [WAS 18].

Quantitatively, the collection currently includes nearly 4,000 biological specimens from thoracic surgeries performed at the University Hospital of Nice. The main epidemiological data are listed: histological type, molecular characteristics of lung adenocarcinoma and tumor stage. Five tumor tissue samples and five non-tumor tissue samples are frozen for the same specimen. Formalin-fixed paraffin-embedded (FFPE) blocks were also

collected for the majority of patients (94%). Blood samples are obtained for 67% of patients.

Qualitative data is optimized by reducing ischemia and transport time to the laboratory. Thanks to the transport of the specimens by pneumatic tube, cold ischemia lasts about 30 minutes. Blood samples, taken before surgery, are prepared immediately through centrifugation, separation of circulating blood cells, sampling and freezing at –80°C for plasma and serum and in liquid nitrogen for blood cells in the presence of DMSO.

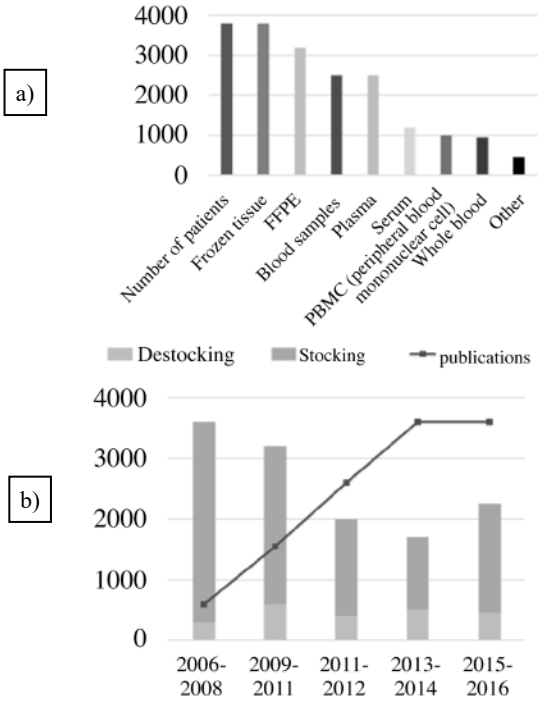

Figure 1.2. *Activity indicators of the Côte d'Azur biobank. a) Quantitative data for numbers of samples stored in the biobank. b) Storage and withdrawal of samples according to the given periods and number of citations mentioning the Côte d'Azur biobank. For a color version of this figure, see www.iste.co.uk/arrighi/biobanks.zip*

COMMENTARY ON FIGURE 1.2.– *The data are taken from the work of Washetine et al. with their kind permission [WAS 18]*.

The analysis of the biobank's functioning has highlighted weaknesses such as the absence of specific collections (feces, saliva), the low number of stage III and IV tumors and the accumulation of samples without request for inclusion. One of the strategic actions to be implemented as a priority is gaining access to samples from clinical trials, after authorization from the local ethics committee, in order to systematically obtain samples from stage III/IV patients.

The standardized quality procedures enabled the NFS 96-900 to be obtained in 2010 and renewed in 2013 and 2016. The samples were collected by the ISO 15189 accredited hospital laboratory in 2013. These standards guarantee the biobank's partners high-quality samples that are perfectly annotated clinically.

1.1.3.3. *The tissue biobank or tumor biobank*

The French National Cancer Institute (*Institut National du Cancer* (INCa)) defines *tumor biobanks* as "infrastructures for the cryopreservation of tumor samples from cancer patients, in response to medical and health obligations and scientific objectives". In France, there are 58 tumor biobanks in the main university hospitals and cancer centers. Twelve of them are located in Paris within the *Assistance publique des hôpitaux de Paris*, which is the university hospital system operating in Paris, the largest hospital system in Europe.

Tumor libraries respond to the need for genomic classification of tumors and the search for molecular markers in cancerology. By providing samples and associated medical data, tumor libraries allow researchers to develop large-scale transcriptomic, proteomic or metabolomic analyses, making it possible to identify new prognostic or diagnostic *biomarkers and new therapeutic targets*. The identification of these biomarkers has also allowed the development of "*companion tests*" for the use of targeted therapies. The French National Authority for Health (*Haute Autorité de Santé* (HAS)) defines a companion test as "a diagnostic test that makes it possible to select, based on their status for a predictive biomarker identified by this test, only those patients for whom treatment is likely to be beneficial, among those diagnosed with a given disease. For this reason, the test is referred to as a 'companion' to the treatment".

Figure 1.3. *Population and clinical biobanks. For a color version of this figure, see www.iste.co.uk/arrighi/biobanks.zip*

COMMENTARY ON FIGURE 1.3.– *Depending on the purpose of the collection, biobanks are population-based if they keep specimens from the entire population for a long time. If they select pathological samples, they are clinical biobanks dedicated to a given pathology in order to understand the molecular and cellular mechanisms and discover therapeutic leads*.

The discovery of new candidate biomarkers leads to retrospective studies, enabled by large sample collections accompanied by epidemiological, clinical, histological and genomic data. These data can be shared in a national or international network of several biobanks. In the following section, we locate the major biobanks by presenting their objectives and the first results of large-scale cohorts.

1.2. Mapping of biobanks

The coordination center for European biobanks (the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI)) is located in Graz, Austria. Ten years ago, Professor Kurt Zatloukal, a pathologist at the Medical University of Graz, had the idea to connect European biobanks. He

was the coordinator of the BBMRI project, funded by the European Strategy Forum on Research Infrastructures (ESFI) under the 7th European Framework Program for Research and Development (FP7). This project started with a dozen countries in 2008 and today there are more than 30. The number of institutions that have joined the BBMRI project reaches nearly 280 biobanks. The objective of this network is to create a platform that connects researchers and allows scientists to share samples and their associated data. In a first step, the goal of this infrastructure was to list the biobanks in Europe and to methodically identify all the collections. Then, in a second step, identical and standardized operating procedures were imposed in order to meet harmonized routines when centralizing data in this unique network.

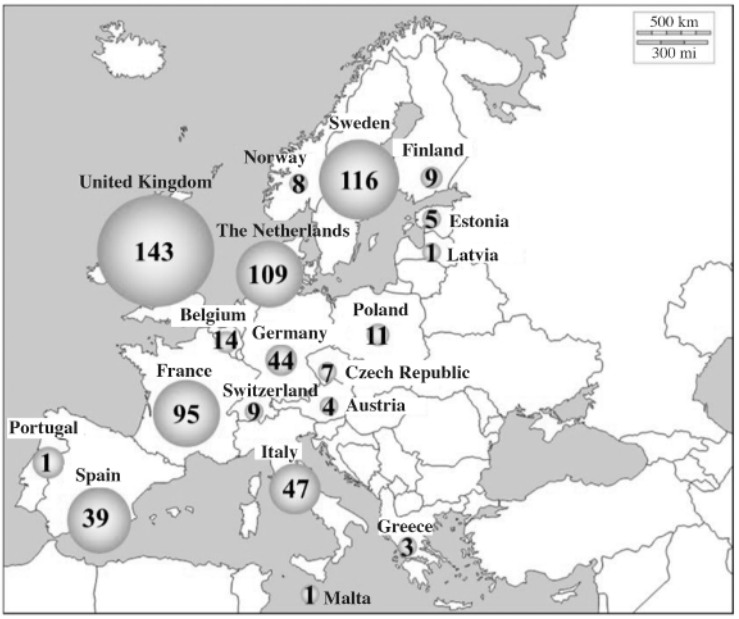

Figure 1.4. *Map of biobanks in Europe. The number of biobank members of the BBMRI-ERIC infrastructure is indicated by country*[1]

Between 2014 and 2020, the research and innovation program of the European Union Horizon 2020 planned to provide Europe with world-class research infrastructures. It is in this context that *the research infrastructure*

1 Portal data: www.bbmri-eric.eu/, May 2021.

for biobanking, named the Biobanking and Biomolecular Resources Research Infrastructure–European Resource Infrastructure Consortium (BBMRI-ERIC), has been funded. The objective of this infrastructure is to bring together all biobanking stakeholders – researchers, biobankers, industry and patients – to accelerate medical research and the discovery of new treatments.

It is the largest European infrastructure dedicated to human health research. It helps researchers find the sample and data they need for their research project while respecting the quality and regulatory constraints imposed on biobanks. BBMRI-ERIC also facilitates interdisciplinary collaboration (informatics, legal and ethical).

1.2.1. *The large catalog of European biobanks*

The first version of the BBMRI-ERIC catalog, released in July 2015, consisted of a directory of information on biobanks. A few months later, in December 2015, this information tool was enriched with additional data on biobank collections and networks. In practice, the BBMRI-ERIC catalog is used to locate samples or services that are essential to the advancement of research, thus identifying biobanks that hold these samples or are able to offer the expected services.

Note that this is not personal data. All data are public and can be shared freely. Thanks to this interface, all biobank actors, as well as researchers and patients, can access the general data for each biobank:

– *biobank users* publish information on the site about a biobank's activity and gain visibility. Informed of the composition of the collections and their location, they can initiate collaborations with these biobanks;

– *researchers* identify the samples and associated data and contact the biobanks that hold the collections they need to conduct their research project;

– *patients and donors*, made aware of the needs of research through communication campaigns, can participate by giving a biological sample to the biobank of their choice.

This database describes the samples and general data of each biobank; it is an information tool between biobanks. Contrary to existing systems, the

architecture of this tool centralizes an online and offline information flow with a possible connection to the biobank management system. This interoperability makes the database sustainable while providing an interactive interface and automatic updates.

The web application is built on the MABIS 2.0 (minimum information about biobank data sharing) standard, an open source software adopted by biobanks and research communities. The work of van de Velde et al. has shown that this bioinformatics data software is aimed at non-bioinformaticians. It collects, manages, analyzes, visualizes and helps to share extensive and complex biomedical data without requiring advanced bioinformatics skills [VAN 19].

The development of the database was defined to integrate three features:

– storage capacity for biological materials and derived products;

– data storage capacity;

– the integration of the expertise of biobank users.

The work of Holub et al. detailed the contents of this massive catalog of European biobanks [HOL 16]. In 2016, the authors identified 515 biobanks and estimated the total number of samples in 136 clinical and 189 population biobanks in Europe to be over 60 million. In 2020, the BBMRI-ERIC brought together 19 member countries of the European community and an international organization, the International Agency for Research on Cancer (IARC). They cover a population of more than 500 million individuals in Europe. This large catalog of European biobanks inventories 100 million samples [BBM 21]. In addition to the ability to select data by country, by biobank name or by pathology, other selection criteria facilitate the search for resources – sample or biobank – such as the nature of the sample, the type of collection, the type of data and the quality labels of the biobank and the collection.

Thus, the biobank repository provides the researcher with access to cohorts of individuals with a specific phenotype and guarantees that these biological samples have been collected, processed and stored according to standard quality procedures.

<table>
<tr><th colspan="2">Material</th><th>Collection</th><th>Data</th></tr>
<tr><td rowspan="4">Blood</td><td>Whole blood</td><td rowspan="10">Healthy controls
Birth cohort
Twins
Cross-sectional study
Longitudinal study
Population
Hospital
Specific disease
Rare disease
Image collection
Non-human</td><td rowspan="10">Biological sample
Genetic investigation
Imaging data
Medical records
National registries
Physiological and biochemical analyses
Surveys and studies</td></tr>
<tr><td>Blood cells</td></tr>
<tr><td>Plasma</td></tr>
<tr><td>Serum</td></tr>
<tr><td rowspan="2">Tissue</td><td>Freezing</td></tr>
<tr><td>Paraffin blocks</td></tr>
<tr><td>Nucleic acids</td><td>DNA
cDNA or mRNA
RNA
microRNA</td></tr>
<tr><td colspan="2">Cell line</td></tr>
<tr><td colspan="2">Saliva</td></tr>
<tr><td colspan="2">Urine</td></tr>
<tr><td colspan="2">Feces</td></tr>
<tr><td colspan="4">Quality: certification of the biobank/collection by an accreditation body and/or BBMRI-ERIC audit of the biobank and the collection</td></tr>
</table>

Table 1.1. *Criteria for locating a collection and identifying a biobank in the BBMRI-ERIC database*

COMMENTARY ON TABLE 1.1.– *Selection is made via a drop-down menu. By choosing a blood material, for example, the sub-menu proposes several derived products, such as circulating blood cells, plasma or serum*[2].

1.2.2. *International biobanks*

As soon as the samples (blood, tissue, urine, etc.) arrive at the biobank, they are received and the donor's personal information, such as their name, surname, age, gender and primary diagnosis, is entered into the database. Then, the sample is identified with a barcode linking it to the patient's

2 Portal data: https://directory.bbmri-eric.eu/menu/main/app-molgenis-app-biobank-explorer/biobankexplorer, March 2020.

medical record, and the sample undergoes a processing step before storage. It is divided into several fractions or transformed into derived products, such as nucleic acids, for example, after an extraction protocol. All products derived from the initial specimen are stored and made available to researchers.

The sum of the numbers of specimens (or fraction of specimens) and derivatives is an indicator of the total amount of samples stored and thus of the size of a biobank. However, there is a second indicator previously estimated during a collection project, the target size of the biobank. It aims at a maximum number of participants, knowing that a participant can provide several samples and that a sample can generate several derived products. These two indicators are publicly available because they indicate the order of magnitude of the size of the biobanks. According to the information they publish on their websites, the three largest biobanks in the world are located in Europe, China and the United States. While the size of the biobank and the number of samples are important, the purpose of the biobank is crucial. Is it a public, not-for-profit organization? Is it a population biobank? All of these questions need to be asked in order to understand how these infrastructures work.

1.2.2.1. *The Graz Biobank as a hub for European biomedical research*

Founded in 2007 and located in Austria at the Graz University Hospital (Landeskrankenhaus-Universitätsklinikum Graz), this biobank was conceived as the research center of the Medical Faculty of Graz. This public non-profit organization is partly financed by the medical faculty for the expenses related to staff salaries and overheads. The development of the infrastructure is supported by the Austrian government, the city of Graz and the BBMRI network through the European Union's Horizon 2020 research and innovation program.

Since 2009, the biobank has been ISO 9001:2008 certified; it has developed a rigorous quality management system based on standard operating procedures. It has undertaken pioneering work in the automation of sample storage for two reasons: firstly, to reduce transfer times during retrieval and ensure optimal sample quality, and secondly, to adopt a professional organization and reduce errors during sample retrieval. All ultra-low-temperature freezers (–80°C) used for the storage of biological

fluids such as serum, plasma, blood cells and nucleic acids, as well as all liquid nitrogen tanks used for the storage of tissues and certain cell lines (–196°C) are under continuous monitoring. Each piece of equipment is equipped with a temperature recorder with an alarm signal in case of a power failure (freezers' power supply stop) or a decrease in the volume of liquid nitrogen in the tanks (release of nitrogen gas from the container).

Currently, the Graz Biobank has combined the collections from the Graz Medical School and the Institute of Pathology. These FFPE samples have been stored since 1984 – thus for more than 30 years – and represent more than 1 million patients and more than 20 million samples of various types, such as paraffin-embedded (FFPE) or cryopreserved tissues and biological fluids (whole blood, plasma, serum, urine, cerebrospinal, follicular or seminal fluid, etc.), and all these samples are strictly linked to the patient clinical information [BIO 21]. The Graz Biobank has done a great deal of work to ensure that the patient medical and personal information is accessible and confidential. Every sample is linked to the clinical information in the medical record, diagnostic and treatment results, as well as histological, medical imaging, immunohistochemical and molecular genetic characterization data.

The Graz Biobank is a clinical biobank in the broadest sense. Any information from a patient registered in the clinical database of the Graz University Hospital can be linked to the sample deposited in the biobank provided that the patient is of age and able to give consent (no samples from psychiatric patients). All patients and donors sign a free and informed consent form that authorizes the use of their samples for medical research in the form of extended consent. Biospecimens are routinely collected at diagnosis or during therapeutic interventions, which are performed on an annual population of about 400,000 outpatients and 85,000 inpatients, according to data from Huppertz et al. [HUP 16].

Since 2012, all samples are coded in the form of a QR code (2D datamatrix code) that aggregates the minimum patient information: identification number, age, gender and primary diagnosis. In addition, detailed information is entered into the hospital's clinical information system in connection with the sample identification.

During clinical diagnosis, all results from the Graz Medical University's Institute of Pathology are made available to the biobank for fresh frozen and

paraffin-embedded samples. In the case of tumor tissue sampling, a sample of normal tissue close to the tumor is taken together with the affected tissue. This serves as a non-tumor control tissue (fresh frozen or FFPE). In addition, healthy donor samples are collected in parallel and serve as control samples of serum, EDTA plasma and buffy coat (white blood cells and platelets). As the samples are collected by the hospital during surgery without prior selection, this collection reflects the Austrian population without any filter. In addition, the biobank also has targeted collections, such as cohorts for cardiovascular diseases, fertility disorders, febrile diseases and osteoporosis.

Huppertz et al. examined the activities of the Graz Biobank using both a quantitative approach, involving the increase in the volume of samples collected over the last 30 years, and a qualitative approach, involving the automation of sample labeling and tracking [HUP 16]. They see this infrastructure, which is in permanent interaction with hospital services and scientific research, as the hub of European biomedical research.

1.2.2.2. *China National Genebank (CNGB)*

The first Chinese biobank was created in 1994 when Chu's team discovered cell lines specific to Chinese ethnic groups [CHU 98]. Under the impulse of the Chinese Academy of Sciences, which was anxious to preserve these precious cell lines, the program for the creation of human biobanks was started in China. The Chinese government and the academy perceived the importance of the conservation of biological resources, especially human resources. They have massively funded precision medicine research projects requiring the construction of huge biobanks backed by biomedical research. According to the history of the study provided by Gan et al., the number of new biobanks has exploded in the last 10 years [GAN 15]. Clinical biobanks have been set up in large cities such as Beijing, Shanghai and Guangzhou near medical research centers and large hospitals.

More recently, China's first national genebank, inaugurated in 2016, has been described as one of the largest and most modern. Named the China National Genebank (CNGB), it is a public non-profit organization located in Shenzhen in southern China [CHI 21]. It consists of a bioresource bank, a bioinformatics data network and an operational system deployed in five interactive units, including three banks and two platforms:

– a biorepository where more than *10 million samples* of human, animal, plant and microbial origin are stored;

– the bioinformatics center for the storage and interpretation of transomics data (genomics, transcriptomics, proteomics and metabolomics);

– the "living" bank intended to cryopreserve living biological material, and more specifically to safeguard more than 300,000 plant species and millions of animals and microbes;

– the digital sequencing platform;

– the synthetic biology and genome editing platform.

The CNGB's mission is to collect and store genetic and biological samples from human beings and other living organisms. Its major activities focus on sequencing and genetic engineering applied to health, agriculture, microbiology and marine biology. The CNGB works in partnership with about 100 international organizations, including biobanks, for example, the Svalbard global seed vault in Norway, and international networks, such as the International Society for Biological and Environmental Repositories (ISBER).

The CNGB is part of the Beijing Genomics Institute (BGI), the number one genetic sequencing company in China and one of the world's leading genomic companies, along with the Sanger Institute in the UK and the Broad Institute in the US. The BGI group is a complex semi-private structure made up of various branches:

– public non-profit organizations such as the China National Genebank, the Gene Research and Training Center (BGI Research, BGI College) and Gigascience Scientific Publishing;

– subsidiaries dedicated to genetic sequencing (BGI Genomics) and agriculture (BGI Agro).

Created in 1999 in Beijing by the Chinese Academy of Sciences, the BGI set up its main office in Shenzhen and began its activities with participation in the human genome project. According to David Cyranoski's analysis, Chinese life science research is mainly oriented towards massive sequencing. In the scientific journal *Nature*, Cyranoski analyzes the role of the BGI, which could become "the world leader in genome sequencing" [CYR 10]. According to David Cyranoski, the BGI is becoming a "genomics factory" equipped with hundreds of high-throughput sequencers that could theoretically sequence 10,000 genomes per year. The Chinese group has

4,000 scientists and technicians, including 1,500 bioinformaticians, according to the firm's 2011 data.

The BGI has published its sequencing work in the journals *Nature* and *Science*, decoding the genomes of several plant (rice [YU 02] and cucumber [HUA 09]) and animal species (silkworm [XIA 04], chicken [INT 05] and giant panda [LI 10]), as well as the human genome (for example, in a prehistoric human preserved in the permafrost of Greenland for 4,000 years [RAS 10]).

In 2003, the BGI sequenced the genome of the coronavirus [QIN 03] responsible for severe acute respiratory syndrome (SARS). In 2019, a new species of coronavirus appeared, causing a viral pneumonia first localized in Wuhan, China, which then spread to all continents, triggering a global pandemic. Its genome, sequenced by the BGI, confirms that it is a virus previously unknown in humans, officially named Covid-19.

The Covid-19 virus genome was sequenced from bronchoalveolar lavage fluid samples from nine hospitalized patients in Wuhan. It is close – with 88% identical sequences – to the bat-hosted SARS coronavirus found in 2018 in the Zhoustan region of eastern China. It is more distant from the human coronavirus SARS (79% identical sequences) and MERS (50% identical sequences) [LU 20]. Lu et al.'s phylogenetic analysis suggests that bats are the original hosts of the virus, which was transmitted to intermediate hosts, animals sold in the Wuhan market, before crossing the species barrier and infecting humans. The molecular structure of the virus has shown that it binds to the human angiotensin-converting enzyme 2 receptor, which is the preliminary and essential step of viral amplification.

Based on a *reverse transcriptase PCR* (RT-PCR) approach, the BGI developed a molecular detection test for Covid-19. The primers designated and synthesized by the BGI were tested on the nine patients in the study. The results were then validated with a different protocol by the China disease control and prevention center (CDC).

This real-time fluorescent RT-PCR kit for detecting SARS-2019-nCoV includes the reagents and controls necessary to test approximately 40 patients in a few hours. It has been approved by the Chinese authorities, certified in Europe, and the US Food and Drug Administration (FDA) has approved it for emergency use in the US clinical market [BGI 21].

1.2.2.3. *The American genetic biobank 23andMe*

The French daily newspaper *Le Monde* profiled Anne Wojcicki, who runs the company 23andMe, in the article *Anne Wojcicki et 23andMe, les tests génétiques à portée de clic*, which appeared in the *Talents du net* series on July 19, 2018. After studying biology at Yale University, Anne Wojcicki worked for a decade as a financial analyst in the biotech industry. In 2006, she and Linda Avey, a biologist like herself, founded 23andMe, the leading Californian company in the genetic testing market, "23" being the number of pairs of chromosomes in humans.

Through the Internet, 23andMe goes directly to the consumer and markets a genetic test related to health and genealogy. It is available over the counter on the Internet without a medical prescription. This test can provide genetic risk factors for diseases such as cancer. In 2008, *Time* magazine awarded the Invention of the Year prize to the saliva-based genetic test developed by 23andMe. Anne Wojcicki challenges the concept of health industries; she wants to give consumers access to personal health information by putting genetic tests in their hands. However, without any guidance or interpretation of the results by a doctor, 23andMe's offer remains uninformative for the consumer.

At times, access to genotyping has led to irreversible consequences, such as self-medication treatment changes without the advice of a physician, or unnecessary surgeries, such as mastectomy, based on a false-positive result. In 2013, the FDA questioned the quality of health-related genetic testing kits and called for proof of the reliability of results and explanation of results to consumers. It banned the tests from marketing pending compliance with regulations on predictive tests for medical risk factors, subject to medical device marketing and quality rules. From 2013 to 2015, the FDA exercised regulatory control over health tests, and only ancestry-related tests were allowed. Internal company changes resulted in FDA approval for marketing towards approximately 30 autosomal recessive diseases, such as cystic fibrosis, sickle cell disease and Tay–Sachs disease.

Along with MyHeritage, LabCorp and Myriad Genetics, 23andMe is one of the major players in the global DNA testing market. Stoeklé et al. analyzed the *new business model* of these biotech companies, whose specialty is to produce, process and analyze genetic data on a large scale. Stoeklé's study divides their business model into two segments and

describes 23andMe's model as a "two-sided data-banking market model" [STO 16].

This business model, which emerged with the explosion of the digital industry, manages data in two categories. On the one side, the company offers free services to the consumer and stores their data. On the other side, the consumer's data is sold to other companies. The content of the data differs from one company to another, but the example of Facebook illustrates the model of sharing personal information between users and also between companies.

In the case of 23andMe, the company collects biological samples through the offering of a direct sales genetic test at an attractive price. The protocol involves three simple steps (order–spit–discover). The customer orders a kit – health or genealogy – by creating an account on the Internet (order), where they are asked to fill out a form where information about their health, family, environment and lifestyle is declared. A few days later, they receive the kit containing a sampling tube labeled with a personal barcode into which they deposit a few milliliters of saliva (spit). The saliva sample is returned to the company by prepaid mail, analyzed and the results are sent via the Internet to a personal account within three to five weeks (discover).

The company 23andMe stores in its research division:

– biological samples;

– sequencing data;

– medical, family, personal and lifestyle information.

This is the definition of a biobank: 23andMe is one of the first biobanks to have integrated web tools. One of its first investors was Google, and the digital giants are now looking to invest in the market for genetic, genealogical and more broadly health data.

When announcing the collaboration with GSK in 2018, Anne Wojcicki explained that "I started 23andMe with the belief that by getting people interested in learning about themselves and participating in research, we would all benefit" in a post appearing on the 23andMe blog on July 25, 2018 [WOJ 18]. Anne Wojcicki holds the principle of sharing everyone's data for the health of all, which is what led to the sequencing of the human genome (Personal Genome Project (PGP)). This idea of sharing is implemented by

free access to PGP data (open access), but access to 23andMe data is forbidden, locked by the company. Different from "sharing", Stoeklé sees it as a "commercial exchange", giving biological data a market value. He recognizes one side of this business model as the collection of samples, where, for a relatively low introductory price, consumers gain access to their genealogical information and health risks. The second side compiles genetic data and personal information and allows clinical annotations to be added to the biobank samples. This makes them attractive to pharmaceutical companies who covet large-scale cohorts to detect disease exposure factors.

In France, open access genetic testing is prohibited. According to article 16-10 of the Civil Code, paragraph 1: "The examination of a person's genetic characteristics may only be undertaken for medical or scientific research purposes." However, many French people buy these kits online and enrich the French genetic data biobank. Companies have collected biological samples linked to personal data declared online such as family, medical and lifestyle information. Eighty-five percent of customers have agreed to share their anonymized results with biotechnology companies (Genentech), large pharmaceutical firms (Pfizer) or multinationals (Procter & Gamble).

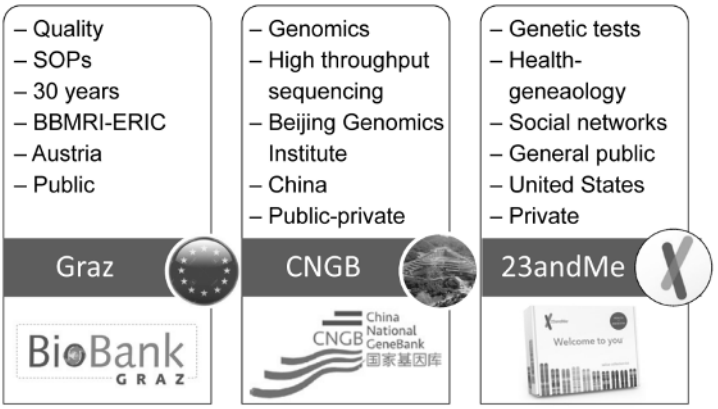

Figure 1.5. *Strengths of large international biobanks. SOPs: standard operating procedures; BBMRI-ERIC: European biobank research infrastructure; BGI: Beijing Genomics Institute; NIH: National Institutes of Health*

COMMENTARY ON FIGURE 1.5.– *a) The Graz Biobank, the oldest, is known for its quality standards; b) the China National Genebank (CNGB) has*

enormous high-throughput sequencing capabilities and collaborates with the Beijing Genomics Institute; c) the private biobank 23andMe markets genetic test kits via the Internet and has more than 1 million clients worldwide participating in collaborative studies with several pharmaceutical companies.

1.2.3. *The first results from the megacohorts*

Epidemiological cohorts are designed to study and follow up populations over long periods of time in order to monitor the development of various diseases. The goal is to identify genetic or environmental factors that lead to the development or progression of disease. Multifactorial studies are long term (10–15 years) and require the inclusion of hundreds of thousands of participants to detect the onset of disease. Two examples of "megacohorts" are presented: the French Constances cohort, which has been followed for nearly 20 years, and the EPIC cohort, which has been in place in Europe for about 30 years. "The data from the cohort allow us to analyze the health status of the population and to better understand what happens throughout life. After several years of recruitment, we have been able to analyze the data and provide initial estimates, in particular on the prevalence of being overweight and obesity", explains Marie Zins, epidemiologist and scientific and technical manager of the Constances cohort.

1.2.3.1. *Mapping obesity in France according to the Constances study*

"One out of two French people is overweight" was the headline of Pierre Le Hir's article in the newspaper *Le Monde* on October 25, 2016 [LEH 16]. It paints a portrait of the French man – and the French woman – based on the first results of the Constances cohort, published by Matta et al. in the weekly epidemiological bulletin of Santé publique France [MAT 16]. What is the Constances cohort? It is a major epidemiology and public health research project in France, which was launched in 2012. The cohort now has more than *200,000 participants*, and more people continue to join regularly. Conducted by Inserm and the *Caisse Nationale d'Assurance Maladie des Travailleurs Salariés* (CNAMTS), this longitudinal study tracks the health of French people, their diet, their environment and their working conditions. It brings together the scientific forces of several countries; about 80 research projects are underway and 11 French and international research consortia are participating. The number of scientific publications is accelerating, from

three articles in 2017, to 18 in 2018, then 28 in 2019, 23 in 2020 and 25 in December 2021, according to the PubMed NCBI bibliographic database.

Every year, participants complete a health questionnaire, and every five years they undergo a medical examination at a health insurance center. Since 2018, sample collection has been organized in partnership with the *Integrated Biobank of Luxembourg* (IBBL). It provides the necessary kits for the collection, and then takes care of the transport, processing and storage of the samples, guaranteeing their quality and traceability. Ninety-five percent of the participants have given their consent, and 250,000 samples – plasma, serum, buffy coat and urine – from 9,500 volunteers are stored at the IBBL. These samples are available to researchers who can request them by submitting a scientific project to Constances, which will be validated by the International Scientific Council and the Industrial Steering Committee. As for any project involving human samples, authorizations from the Institutional Review Board (*Comité de Protection des Personnes* (CPP)) and the *Commission Nationale Informatique et Libertés* (CNIL) are regulatory and essential.

"The scientific interest of the Constances biobank is to redistribute all these samples to researchers. These samples will help to elucidate the mechanisms of complex diseases, to understand the role of interactions between genetic, epigenetic and environmental factors in the development of diseases and to discover new biomarkers", explains Marie Zins, coordinator for Inserm of the Constances cohort.

The prevalence of being overweight and obese was studied in 29,000 participants aged 30–69 years, with as many men as women distributed in 16 French departments [MAT 16]. Two indicators were collected simultaneously: body mass index (BMI) and waist circumference. BMI is calculated from weight and height (weight in kg divided by the square of height in m), and usual marker of obesity is if it exceeds 30 kg/m^2, and of overweight above 25 kg/m^2. The results show that *one French* person *in two is overweight*: 56% of men and 40% of women are overweight, combining excess weight and obesity. Similarly, abdominal obesity, measured by waist circumference exceeding 94 cm for men and 80 cm for women, affects 41% of men and 48% of women.

Matta et al. showed a higher prevalence with age and a strong correlation with income level. Regional disparities were mapped by department: Paris

was the least affected department (10%), whereas Nord (25%) and Meurthe-et-Moselle (22%) had high prevalences. At the individual level, the study found that 30% of women with low monthly income (less than 450 euros) were obese, compared to 7% of women with a much higher income (more than 4,200 euros).

The results of Constances concern many research themes around aging, social and professional health determinants, women's health, genetic factors and interactions with the environment. One of the themes, stress at work, has demonstrated the chronic use of benzodiazepines, anxiolytic drugs with a high risk of dependence. Using the pharmacy-dispensed treatments to more than 9,000 participants in 2015, Airagnes et al. calculated prevalences of benzodiazepine use of 2.8% in men and 3.8% in women [AIR 19]. Chronic benzodiazepine use is thus particularly common in the general French population, and work stress is associated with an increased risk of chronic benzodiazepine use.

1.2.3.2. *The European Prospective Investigation into Cancer and Nutrition (EPIC)*

EPIC is a large European prospective investigation whose aim is to study the links between diet, lifestyle and cancer incidence.

The EPIC cohort was initiated by the World Health Organization (WHO) and launched under the coordination of the Lyon-based International Agency for Research on Cancer (IARC) with the support of the European Community's "Europe Against Cancer" program. Recruitment of participants took place from 1993 to 1999, involving more than *520,000 participants aged 35–70 years living in 10 European countries* (Denmark, France, Germany, Greece, Italy, the Netherlands, Norway, Spain, Sweden and the United Kingdom). The participants (153,427 men and 367,903 women) completed questionnaires about their diet, physical activity, alcohol and tobacco consumption, family history and lifestyle. They also gave blood samples, which constitute the International Agency for Research on Cancer (IARC) biobank, also called the IBB. For each participant, samples – plasma, serum, white blood cells and red blood cells – were collected. Each sample was aliquoted into plastic straws and divided into two batches, one stored locally and the other sent to the IARC biobank.

Initial results from the EPIC study show the role of fruits and vegetables in preventing digestive and respiratory cancers. The study by Bradbury et al. showed that the regular consumption of fruits and vegetables is associated with a decreased risk of developing cancers of the mouth, larynx, pharynx, esophagus, colorectum and lung [BRA 14]. The other cancers studied (stomach, biliary tract, pancreas, cervix, endometrium, prostate, kidney, bladder and lymphoma) did not show an association. These results reinforce the idea of cancer prevention through a balanced diet containing a daily intake of fruits, vegetables and whole grains. "This, however, is only likely to lead to significant cancer prevention in France if the nutritional message is accompanied by a general recommendation for lifestyle modification: not smoking, limiting alcohol consumption, increasing physical activity and avoiding obesity", comments Professor Elio Riboli, coordinator of the EPIC study and Director of the School of Public Health at Imperial College London.

From epidemic to prevention is the pathway traced by this study. The work of Trichopoulou et al. examined the effect of the *Mediterranean diet* on health and longevity [TRI 14]. By substituting olive oil (not widely consumed in northern European countries) for other unsaturated fats, they demonstrate that the "modified" Mediterranean diet is associated with increased longevity in older individuals. In all, more than 100,000 participants in the EPIC cohort over the age of 60 were recruited. Of the total, some participants were excluded, either because they had cardiovascular disease (15,362 participants) or because detailed dietary data were missing (10,340 participants), bringing the final number to 74,607 participants across nine European countries (France, Italy, Spain, the United Kingdom, the Netherlands, Greece, Germany, Sweden and Denmark). This observational study is a public health indicator in an increasingly elderly Europe.

The French part of the European EPIC study is represented by the epidemiological study of women in the French education system, called the *E3N cohort. The participants are female teachers* affiliated with the *Mutuelle Générale de l'Education Nationale* (MGEN), which supports the project alongside the *Ligue contre le Cancer*. The cohort includes approximately 100,000 women aged 40–65 years. They described their lifestyle and health status (cancer, cardiovascular disease, diabetes, depression, fracture, asthma, among others) by answering a questionnaire every two or three years. Biological samples supplemented these investigations with blood samples (whole blood, serum, plasma, erythrocytes and lymphocytes isolated in the buffy coat) and salivary and breast cancer tumor tissues.

The E3N cohort has produced broad results considering the complex role of diet, hormonal factors, physical activity and lifestyle in relation to health. According to Francoise Clavel-Chapelon, the major results concern hormonal treatments, being overweight and eating habits [CLA 15]:

– the benefit–risk balance of hormone replacement therapy during menopause, which has profoundly changed prescription habits in France;

– the association between chronic disease risk and anatomical characteristics;

– the influence of dietary habits on the risk of developing cancer, diabetes and asthma in middle-aged women.

Since 1990, the E3N cohort has been coordinated by Giluca Sevri and Marie-Christine Boutron-Ruault, both Inserm researchers. As an extension of this study, the *E3N–E4N family cohort* is composed of three generations: the women of the E3N study and their spouses (18,000 volunteers), their children and their grandchildren, whose recruitment began in 2021. The E3N–E4N study is interested in the influence of contemporary lifestyle on health in people of the same family over three generations.

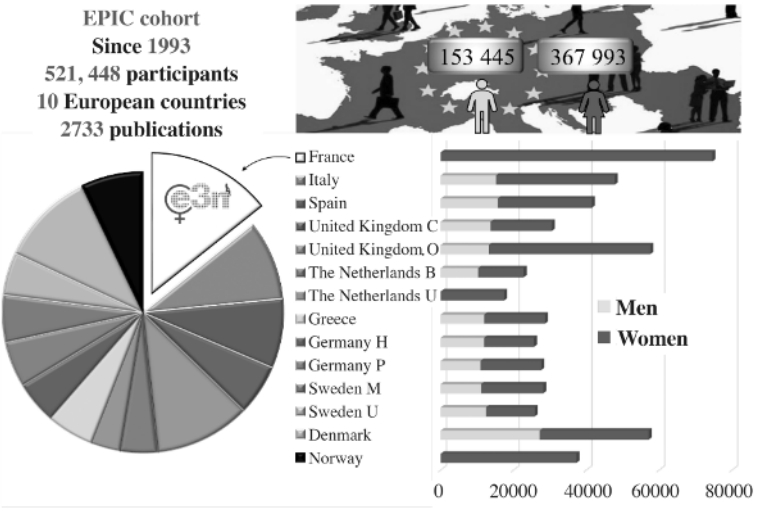

Figure 1.6. *The European Prospective on Cancer and Nutrition (EPIC). For a color version of this figure, see www.iste.co.uk/arrighi/biobanks.zip. UK C: Cambridge; UK O: Oxford; Holland B: Bilthoven; Holland U: Utrecht; Germany H: Heidelberg; Germany P: Potsdam; Sweden M: Malmo; Sweden U: Umea*

COMMENTARY ON FIGURE 1.6.– *Sometimes, two sites per country collected information and samples from the 153,445 male and 367,993 female volunteers. Like Norway and the Netherlands (Utrecht), France is represented by women only; this sub-cohort is named EPIC-E3N. The number of publications with the terms "EPIC" and "cancer" is given according to PubMed in December 2021*[3].

1.2.3.3. *The American precision medicine program: All of Us*

The *All of Us* research program is funded by the US government and led by the National Institutes of Health (NIH). Conceived under President Obama in 2015, the project was launched in 2018 along with the *precision medicine* initiative (PMI). Its goal is to recruit 1 million people in the United States to accelerate biomedical research and improve health. The program's challenges revolve around four major areas:

– identifying risk factors for certain diseases and discovering new biomarkers;

– understanding why treatments are effective in some people but fail in others;

– including all profiles in clinical studies to meet the needs of all;

– integrating new technologies into tomorrow's healthcare.

In short, this cohort will be a huge database for all prospective, retrospective and cross-sectional studies.

The recruitment of the *All of Us* cohort is the cornerstone of the project, hence the priority of raising awareness of the concept of medicine adapted to each person and convincing participants to enroll in this research program. In order to reach as many people as possible, a communication campaign has been launched on social networks and in the field, with a huge bus criss-crossing the country to meet the volunteers. The goal is to recruit 1 million adult participants by 2024 and to sequence 1 *million genomes*. Children will be included in a second phase. For 15 years, researchers will take dozens of blood and urine samples from each participant. At the same time, their health data, lifestyle and environment will be regularly monitored.

3 The information is taken from www.e3n.fr/epic.

The *Mayo Clinic* Biobank is responsible for the development of standards and protocols for the collection of specimens that will be stored and made available to researchers. Recognized as one of the world's leading health centers, the Mayo Clinic is a US research and teaching hospital located in Rochester, Minnesota. Its Center for Personalized Medicine is dedicated to promoting translational research and leveraging genomic data [LAZ 14].

The terms "personalized medicine", "individualized medicine" and "precision medicine" sometimes overlap, as the concept of personalized medicine remains unclear, as Bateman explains [BAT 14]. The idea of precision medicine entails a comprehensive view of the molecular and metabolic biology data, known as omics, that can characterize a individual:

– genomics: the sequencing of the individual's DNA (genome);

– transcriptomics: the set of transcribed messenger RNAs;

– proteomics: analysis of sample proteins;

– metabolomics: the set of small molecules (metabolites) present in the sample.

These data partially characterize the individual, as two other factors, intrinsic and extrinsic to the individual, are missing:

– the microbiome: the set of microorganisms present in the body;

– the exposome: the set of relations between the individual and the environment.

Finally, this complete profile lists all the factors that define the original and unique signature of each individual. This complex dataset is a haven of study for precision medicine to mimic, on the one hand, the individual, and, on the other hand, their belonging to a population and one of its subgroups.

Precision medicine needs a large number of participants who mirror the population. Very often, the scope of genetic studies is limited by small sample sizes, lack of diversity among participants or partial access to data, and thus the inability to link genotype and phenotype. Population-based genetic research requires very fine targeting of phenotypic data, which translates into access to participants of varied ancestry. The uniqueness of the *All of Us* cohort is its inclusion of minorities underrepresented in medical

research, ethnicities and communities. The goal is to create the most diverse biobank in the world.

This program of sequencing 1 million genomes is only of value if the associated data are added to the genetic information. In this large-scale study, the project wants to develop the use of *new information technologies*, cell phones and smart biosensors to continuously monitor health status, physical activity and different health parameters. Participants will share their electronic medical records and answer questionnaires on their diet, sleep and environment, in short their lifestyle.

Ethical, legal and social issues (ELSIs) highlight the three objectives of the study:

– cohort diversity;

– participant commitment;

– data confidentiality and security [SAN 17].

Diversity and health disparities are expressed by socioeconomic status, age, geographic location, health status, literacy level and personal information technology skills. Inclusion of disadvantaged populations, people with intellectual disabilities and families with children is requested. Priority is given to ethnically or racially diverse populations, as genetic backgrounds may reveal genetic variants. Inclusion of indigenous nations such as American Indian and Alaska Native populations is reinforced.

Security and privacy are essential. However, there are several difficulties: firstly, the need to protect the patient – in particular their electronic medical record, genetic information and health data posted on the digital platform – and secondly, the need to open up access to the data to research organizations that would like to include them in their work. A third constraint is ensuring the patient's follow-up and providing individual feedback. This is not a simple task given the quantity and complexity of the information. As for the interpretation of the data, particularly the genetic mutations, this will be handled by Color Genomics, who will support the patient when they receive the results.

The *New England Journal of Medicine* Special Report provided an update on the progress of the All of Us program as of the end of 2019 [DEN 19]. After a testing phase in 2017, the recruitment process was validated, as

well as data management and security. During this first phase, more than 27,000 participants were included. They committed to sharing their personal information, biospecimens and electronic medical records. After deploying 340 sites across the country, the recruitment rate was approximately 3,000 participants per week. The total cohort size was 230,000 participants. Of these, 51% were non-White and 80% were minorities who are typically underrepresented in medical research.

Access to the data will be available to the researchers from the beginning of the study and not after its closure, as is often the case. The data will be stored in a virtual space (Cloud) to which the researchers will have access through an access code, assigned once they have demonstrated the usefulness of these data for the advancement of their research.

Data access for participants is planned for the electronic medical record, routine biological analyses and genetic tests. It will be particularly monitored when results of genetic variants known to be risk factors for certain diseases are returned. They will then benefit from medical and therapeutic follow-ups.

In summary, this precision medicine program takes into account individual genomic differences, environmental factors and the American lifestyle. It creates a participatory health model, where the volunteers included in the All of Us cohort become the actors in looking after their own health. Francis Collins, Director of the NIH, sees a "coalition of partners" in "this largest and most ambitious health research project" on the scale of the American population [COL 16].

1.3. Process management in biobanks

Upon receipt of biospecimens in a biobank, patient consent is verified before the sample is registered. The flow of received samples is then monitored throughout the biobank.

The management of the whole process is divided into two parts: the management of the technical procedures that guarantee a high-quality sample and the management of the complex personal data that add value to the clinically annotated sample.

At each stage, the integrity of the sample is checked and maintained as it was when it arrived in the biobank. Maintaining the sample's high quality ensures high-quality results in biomedical research. *Technical factors* related to the sample such as transport (temperature, duration and mode of transport), processing (manual or automatic, duration and temperature) and storage (duration, temperature and freezing and thawing cycles) play an important role in the reproducibility of experiments. In the face of heterogeneous techniques, harmonization of operating protocols is required between the various methods and biobanks located around the world. Quality control and harmonization of procedures are the technical criteria for maintaining sample quality, accompanied by the automation of processing steps. These technical factors are essential for the collaboration of several research teams in a network.

In a human biospecimen biobank, *individual factors* are distinguished from technical factors. These individual factors depend on the clinical, genealogical, environmental and personal data related to the biospecimen donor. In medical research, these individual factors are as important as the others, so the biobank must be able to collect the personal and clinical data and then share it by clinically annotating each sample. The security of personal data and confidentiality of all information collected are the indicators of the trust between the physician, researcher and biobanker.

1.3.1. *Sample quality, the priority of biobanks*

Every year, the pharmaceutical industry spends billions of dollars on biomarker research in precision medicine. This research sets up collaborations between private and public partners to discover new biomarkers. The objective of partner biobanks is to rapidly build large-scale collections to accelerate the conduct of clinical and translational studies. Thousands of biological samples, such as tissues and blood samples, are cross-referenced with multiple analytical techniques such as immunohistology, genomics or, more broadly, omics. In order to ensure confidence in the analytical data generated by these various technologies, the quality of the biological materials and the associated data are the major challenges for biobanks. Throughout the sample's journey in the biobank, its quality is controlled, from reception (tissue or fluid), to preparation (extraction, centrifugation and aliquoting), then to conservation (storage at

–80°C or in nitrogen at –196°C) and finally release (transfer). Not only are biobanks confronted with the various types of samples, sometimes collected according to procedures that vary from one site to another, but they are also faced with the challenge of harmonizing the practices that they must put in place in order to guarantee research teams access to quality biological data and perfectly clinically annotated samples.

1.3.1.1. *Specific standards for biobanks: certification of professional quality*

In 2018, the International Standard Organization (ISO) published an international standard for biobanking, ISO 20387:2018 biotechnology – biobanking – general requirements for biobanking. Biobanks that have this label give users confidence in the overall functioning of the biobank and therefore in the samples transmitted. Moreover, this label increases the international visibility of the biobank. This ISO standard is defined by four points:

– it specifies general requirements for the competence, impartiality and consistent operation of biobanks, including quality control requirements, to ensure the appropriate quality of collections of biological materials and associated data;

– it applies to all organizations that conduct biobanking for research and development, including biobanking of biological materials from multicellular organisms (e.g. humans, animals, fungi and plants) and microorganisms;

– it recognizes and confirms biobanking activity with biobank users, regulatory authorities, peer review organizations and systems, accreditation bodies and others;

– it does not apply to biological materials intended for food/feed production or laboratories performing analyses for food/feed production and/or therapeutic use.

In France, the S96-900 standard for biobanks, implemented in 2008, combines quality management with technical skills in the field of biobanking.

To meet the criteria of these standards, biobanks can rely on a *quality management system* (QMS), which is based on the implementation of

standardized procedures and audits. Thanks to a high level of standardization and standardized procedures shared between the different centers, the European BBMRI-ERIC network of biobanks has set up its own validation routines. This allows the creation of expert biobanking centers, in sync with the advances in pharmaceutical or medical research and analytical techniques.

1.3.1.2. *The impact of the pre-analytical phase on sample quality*

Faced with the lack of reproducibility of experimental results, estimated to result in costs of more than 56 billion dollars according to the study by Freedman et al., the challenge for biobanks is to guarantee the integrity of the samples and therefore to control the pre- and post-collection process up to storage in the biobank [FRE 15]. The pre-analytical phase begins as soon as the bloodstream no longer supplies blood to the sample. During a surgical procedure, the warm ischemia ends when the sample is removed from the body. The time it takes to get the sample to the laboratory extends the duration of this ischemia, which is then called "cold ischemia". The recommendations of the HAS suggest transport "without breaking the cold chain, in a cooled container saturated with nitrogen vapor or in dry ice". When the sample is received at the biobank, depending on whether it is solid or liquid, the conservation technique differs:

– a tissue sample is fixed in formalin and embedded in paraffin blocks at a controlled temperature of 55–58°C to avoid tissue degradation;

– a fluid sample is frozen in the presence of a cryoprotectant (glycerol or dimethylsulfoxide).

In addition to maintaining quality during transport, a second factor depending on the nature of the sample itself can affect the robustness of the results. The sample, of small volume or size, contains little material; however, it must be a representative of the sampling site (organ) and of the patient's physiopathological state (healthy or suffering from a pathology). If the sample contains a majority of necrotic cells and few healthy or tumorous cells, its analysis will not reflect the whole organ, healthy or affected. Hence, it is important to control *pre-analytical variables* such as the nature of the sample, the storage technique and its duration.

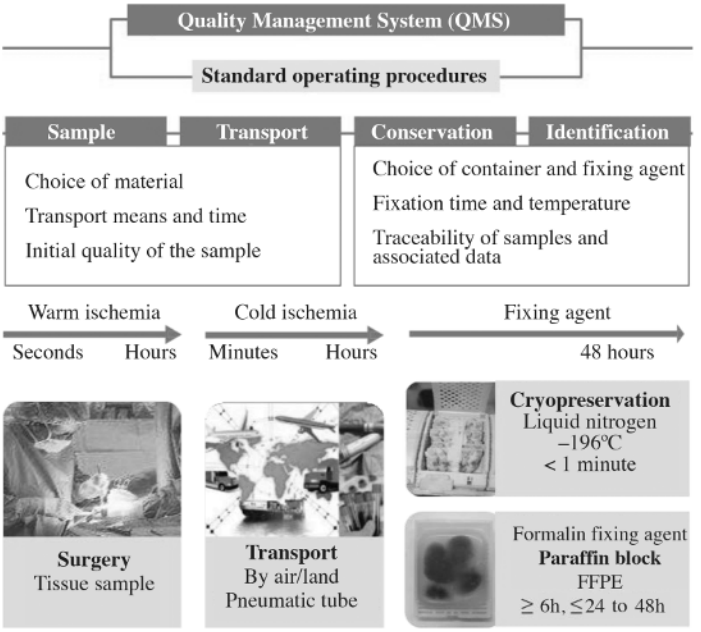

Figure 1.7. *Quality management system (QMS) activities. FFPE: formalin-fixed paraffin-embedded*

COMMENTARY ON FIGURE 1.7.– *Standard operating procedures are implemented at each step of the pre-analytical chain under the pathologist's supervision, starting with the surgical procedure, then the transport of the sample to the biobank and finally its storage. The data are taken from [AMA 20] with the authors' kind permission*.

FFPE tissues are stored at room temperature for decades without altering tissue integrity. Alternatively, after nucleic acid extraction, sample derivatives are stored at –80°C (freezer) or –196°C (liquid nitrogen container). At these very low temperatures, degradation enzymes such as ribonucleases are inactivated. The HAS recommend that these freezers be placed "in locked premises with air conditioning and effective alarm systems". These two preservation methods are the reference methods for genetic and histological analyses.

1.3.1.3. *Harmonized operating protocols*

Different types of processing and storage can alter the molecular composition of samples. To improve the maintenance of sample quality and consequently the quality of research resulting from these biological samples, it is important to accurately record the *pre-analytical history*. In a standardized way, the pre-analytical conditions are recorded in the form of a report (the pre-analytical coding for biospecimens (SPREC)) or more broadly in a biospecimen quality report (Biospecimen Reporting for Improved Study Quality (BRISQ)).

The BRISQ report is a data reporting tool that centralizes sample information. Biobank users enter all sample data and generate a report for researchers that can be included in publication appendices. The information describes the type of sample (tissue or fluid) and factors that may influence the integrity, quality and molecular composition of the biospecimen. These data do not guarantee quality or preclude quality control, but they do provide standardized information to researchers and regulatory agencies to improve the evaluation, interpretation and reproducibility of experimental results.

BRISQ recommendations are defined throughout the sample's journey through the biobank. Moore et al. constructed a hierarchy of information according to the four key stages of the process: sample preparation, collection, processing and storage [MOO 11]. The priority data consist of about 15 elements applicable to any type of sample, including the operative details of collection, stabilization and storage, as well as the clinical data associated with the patient. This first level of essential information is compulsorily documented in scientific publications. The second level includes additional data, i.e. 19 elements including the patient's personal information, stabilization times and temperatures. The last and third level of data are optional data that belong to the context and environment of the collection, such as parameters related to ischemia, therapy, health status, storage containers and transport. The biobank user must check the coded item and qualify each item with a short answer:

– biospecimen type;

– specimen's anatomical region;

– patient's health status;

– patient's clinical characteristics;

– collection method;

– type of stabilization after sampling;

– type of long-term storage of the sample;

– addition of a preservative;

– temperature and time for the transport and storage.

The first item defines the type of biospecimen, coded I.a., and is completed from the list of possible choices: solid tissue, whole blood, serum/plasma, isolated cells, urine, secretions or other human body derivatives.

The items in the BRISQ report are the result of discussions between researchers at the National Cancer Institute (NCI) and biobank users, including the International Society for Biological and Environmental Repositories (ISBER). After searching the biospecimen publications in the NCI bibliographic database, the BRISQ committee sorted the information according to its usefulness in the research project. The analysis of the information revealed three classes of information: parameters that were a priority to the research, those that were less of a priority and those that provided general information but did not advance the topic. Based on the terminology used in the bibliographic archive, the recommendations were prioritized, resulting in the BRISQ report. With this common terminology, information can be categorized in order of importance, including pre-analytical variables. The next step is to implement them in a standardized format.

The list of pre-analytical factors should cover the entire sample pathway in the biobank. All users of the biobank will use it, so it must be explicit and clear. In response to this demand, the ISBER developed the Standard Pre-Analytical Code (SPREC), *a way of coding current and future protocols*. The first objective is to annotate the sample so that researchers know the history of protocols that may influence the analytical results. Then, the second objective is to adopt these annotations routinely and include them in the biobank's quality management system. Lehmann et al. have identified the main pre-analytical factors that can impact the quality of a *clinical fluid or solid biospecimen* [LEH 12]. Let us take the example of a liquid fluid and see how the collection, processing and storage data are recorded according to the SPREC.

The first two parameters identify the type of sample and the initial collection tube. For example, whole blood is coded "BLD" (blood), plasma "PL1" after a single centrifugation or "PL2" after a second centrifugation and serum "SER". The container that receives the sample specifies the presence of additives or anticoagulants, for example, the tube containing EDTA is coded "SED" for sodium EDTA.

After identification of the sample and its collection tube, the centrifugation protocol is described by centrifugal force, time, temperature and presence (or absence) of a brake. The code contains a single letter. For example, "A" means a centrifugal force of less than 3,000 g, for 10–15 minutes, at room temperature and without a brake. If centrifugation is not immediate, the time before – and possibly after – centrifugation is also specified. For example, in post-centrifugation, "A" corresponds to a time of less than 1 hour at a temperature between 2°C and 10°C.

Finally, the storage code specifies the type of container used, its volume and the storage temperature. For example, "A" corresponds to a 0.5–2 mL polypropylene tube stored between –85°C and –60°C.

Launched in 2009 and regularly updated by the ISBER, the SPREC has been enhanced with new methods. Version 3.0, released in 2018, details the tree of choices for the seven elements comprising the SPREC code, as listed in the work of Betsou et al. [BET 18]. According to a recent survey on quality in European biobanks undertaken by the BBMRI, the SPREC was implemented in half of biobanks and was planned to be in the remaining biobanks surveyed [LIN 19]. However, this tool will be less and less often used with the implementation of the new ISO 20387 standard, which allows biobanks to opt for accreditation or certification of their activities as a way to formalize their skills. This survey also provides another piece of information about biobanks' quality intentions: if they were previously preparing to implement both ISO 9001 and ISO 20387, the majority are now aiming for ISO 20387. This biobank-specific standard aims at standardizing the entire process followed by the biospecimen, from collection to analysis, during the crucial steps of the pre-analytical phase.

1.3.2. *Protection of the human person and personal information*

Technical harmonization paves the way for standardized operating procedures in biobanks. This improves the robustness of results and

accelerates the development of medical research. Beyond the scientific interest, it should be kept in mind that biobanks are *collections of human material* (fluid, tissues, cells or DNA). The whole activity of the biobank is based on this "human" component, which is one of the biobank's priorities, and it must guarantee biospecimen donors their individual rights and protect them. In contrast to technical standardization, the ethical and legal framework of biobanks is far from being standardized with regard to the different actors and their varied interests.

Some Northern European countries have adopted a specific legal framework for biobanks. The Icelandic Act on Biobanks no. 110/2000 was the first European text to regulate the practices of Icelandic biobanks in 2000. Then, in Estonia, ethical, legal and societal issues were addressed in a special law: the Human Genes Research Act of Estonia. Sweden followed with the Swedish Act Biobanks in Medical Care (SFS 2002:297) and Finland biobank legislation (Biobank Act 688/2012) came into force in 2013. In the absence not only of a local law specific to the activity of biobanks, but also European regulation, biobanks must refer their activities to the general law. These activities raise a number of ethical questions about the patient and the biospecimen. How can the patient's wishes be respected? How can their agreement be obtained? Who owns the samples? How can they be distributed? We try to answer these questions by outlining the legal framework of biobanks in France. As there is no specific law, a complex system composed of a set of texts, laws and decrees of the French Public Health Code (*Code de la Santé Publique* (CSP)) takes up the international principles of ethics and research.

1.3.2.1. *The complex legal framework of French biobanks*

Promulgated in 1988, the *Huriet–Sérusclat law* is the founding text for the protection of persons who take part in biomedical research. Modified and then simplified in 2004, law no. 2004-806 relating to public health remains an essential turning point in the history of the development of clinical trials in France and in Europe at the end of the 20th century [FRE 04a]. In order to respond to the protection of participants and the requirement for transparency of clinical data, the law sets out the essential points consisting of the clinical trial sponsor, the personal protection committee and the patient's consent:

– the "*sponsor*" concept identifies the individual or legal entity that takes the initiative for medical research. They manage the implementation, the

management and the financing of the study. The sponsor entrusts the study to an experienced "investigator", a natural person who supervises the project. For multicenter trials, several investigators are recruited;

– clinical research protocols are then submitted to an advisory committee for the protection of individuals. The CPP, an independent body with a legal personality, is composed of representatives of patient associations or users of the health system (approved under article L. 1114-1 of the CSP). Its members are appointed by the legal representative of the state in the region where the CPP is located, which may also have several CPPs. The favorable opinion of the CPP is mandatory to start a research project;

– the individual included in the research project gives *free, written and informed consent*. Persons participating in the research may be healthy volunteers or sick persons. They must be affiliated with a social security system and must undergo a medical examination prior to the research, the results of which will be transmitted to them;

– non-respect of these rules can result in sanctioning by the Penal Code (articles 223-8 and 223-9).

In Europe, the ethical and regulatory framework of clinical research is governed by *Directive 2001/20/EC*, which harmonizes the rules on the safety and vigilance of therapeutic trials between member states. This European directive strengthened the rights of participants and created the database of serious adverse events (Eudravigilance). Directive no. 2001/20/EC was transposed into the French Public Health Code in *law no. 2004-806 on public health policy*. From then on, the authorization of clinical trials has been given by the ANSM after the mandatory favorable opinion of the CPP. Internationally recognized good clinical practices (E6: good clinical practice consolidated guideline) are published by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). Ensuring quality, reliability and harmonization of all documents essential to clinical research are the main objectives of this text, which defines the chain of responsibilities from the sponsor to the investigator and the coordinator designated by the clinical research investigator.

In 2004, the Huriet–Sérustat law was completed by the revisions of the law relating to bioethics and the so-called "CNIL 2004" law [FRE 04b]. *Law no. 2004-800 of August 6, 2004, relating to bioethics* establishes the legal rules organizing the obtaining of biological samples of human origin or

health information for medical or scientific purposes. It reinforces the rights of the individual regarding genetic studies and the use of elements and products of the human body. The State created the *Agence de biomédecine*, whose mission is to control research in the fields of transplantation, reproduction, embryology and human genetics. The same year, the law on the protection of individuals with regard to the processing of personal data (law no. 2004-801 of August 6, 2004) revised the framework for research on "human beings". Since 2004, the CNIL law and the bioethics law have been revised.

The law no. 2012-300 of March 12, 2012, relating to research involving the human person, known as the *Jardé law*, was implemented in 2016 by decree no. 2016-1537 of November 2016 [FRE 12]. This decree specifies the modalities for conducting research involving the human person. It defines two types of research – interventional and non-interventional – and replaces the notion of "biomedical research" with "research involving the human person". It determines the missions of the French National Commission for Research (*Commission nationale des recherches*) involving the human person, including the coordination of CPPs in conjunction with sponsors, the ANSM and the CNIL. In 2017, a new decree was published in the *Journal officiel* (decree no. 2017-1549 of November 8, 2017) relating to the conservation and preparation for scientific purposes of human body elements. It applies to the activities of conservation and preparation for scientific purposes of tissues and cells of the human body:

– for the specific needs of an organization's research programs, which requires a declaration to the Minister of Research (see article L. 1243-3 of the French Public Health Code);

– with a view to their transfer for scientific use, which requires authorization issued by the Minister of Research (see article L. 1243-4 of the French Public Health Code).

These declaration procedures applicable to the exercise of activities of conservation and preparation of elements of the human body for scientific purposes consequently modified the French Public Health Code. The modifications made by the decree were updated by the application for the management of the conservation of human body parts (Codecoh) of the Ministry of Higher Education, Research and Innovation. The declaration and authorization request forms are available online. In both cases, the

competent territorial CPP must also be notified. For a health establishment, the declaration – or the request for authorization – is also sent to the director of the regional health agency (*Agence Régionale de Santé* (ARS)) with territorial jurisdiction. For an army hospital, the declaration – or the request for authorization – is also addressed to the Minister of Defense.

Research activities on embryos and embryonic stem cells are subject to a specific legal framework. The law on bioethics (law no. 2004-800 of August 6, 2004), reformed with law no. 2011-814 of July 7, 2011, was modified by law no. 2013-715 of August 6, 2013, "authorizing under certain conditions research on embryos and embryonic stem cells". The 2019 bioethics bill aims to facilitate strictly supervised research on embryonic stem cells. It proposes the distinction of the legal regime for embryos and embryonic stem cells (article 14 voted on October 4, 2019).

Currently, the conservation of embryos and embryonic stem cells for research purposes requires authorization issued by the biomedicine agency (CSP, art. R. 2151-20). Any organization conserving embryos or embryonic stem cells for research purposes must be able to justify the collection of consent given by the embryos' parents. When embryonic stem cells have been imported, the organization must be able to justify that they have been obtained in compliance with the prescriptions of the Civil Code, with the prior consent of the couple who generated the embryo conceived in the context of medically assisted procreation, and that this embryo is no longer the subject of a parental project, without any payment, in whatever form, having been allocated to the couple (CSP, art. R. 2151-18).

1.3.2.2. *Protection of the biospecimen donor*

The biobank is at the center of a network of rights and of obligations towards the biological sample donor. Upstream, it must manage the rights of individuals by obtaining consent forms, and downstream, it must manage the conditions of access of donors and researchers to the data associated with biological samples. A body of laws and decrees indicates the necessity of informed consent in research involving the human person:

– the law on information technology and freedom, no. 78-17, modified by law no. 2004-801 of August 6, 2004, relating to the processing of personal data;

– law no. 2002-303 of March 4, 2002, relating to the rights of patients;

– law no. 2004-800 of August 6, 2004, relating to bioethics;

– the Jardé law on research involving the human person, decree no. 2016-1537 of November 16, 2016.

Free and informed consent is mandatory for the use of a human sample. According to article L. 1122-1-1 of the French Public Health Code, the rights of individuals must be respected: "No research may be carried out on a person without his or her free and informed consent, obtained in writing, after he or she has been given the information provided for in article L. 1122-1."

By signing this document, the individual expresses their willingness to participate in the research project, which will have been explained in a notice attached to the consent form. With this information, the investigator, or the authorized person, explains why the sample provided is essential to the conduct of the project. The investigator answers the donor's questions and both sign the consent form. The donor may withdraw from the study at any time. This consent is mandatory for the use of biological samples and personal data or for genetic research.

Consent to participate in the research project is based on:

– the participant's written consent (inclusion), which is required;

– the information about the research project provided by the investigator, such as the objectives, methods, benefits and risks associated with the study. It is also specified that the participant is free to accept or refuse to participate in the study in question;

– the investigator's prior declaration to the CPP, which is authorized to obtain the participant's consent.

Before the study begins, the consent procedure is explained orally to the patient by the investigator. The consent form mentions the patient's rights; it is dated and signed and kept by the investigator. The patient receives a copy which they also keep, as well as the information leaflet.

Authorization is granted in the context of a particular study. However, the participant's sample or associated data may have a different purpose. If the participant does not object to the reuse of their sample, it will be used in a future project (article L. 1211-2, paragraph 2 of the French Public Health

Code). In France, the secondary use of samples is based on the principle of non-opposition. More generally, if the reuse concerns several samples, the scientific manager may request a *global re-qualification of a collection* to the CPP, explaining the objectives of the new research and especially the need to include the samples of unreachable participants. The CPP can then accept the requalification, except in the case of genetic studies, for which individual consent is mandatory, and for the use of germ cells.

When a biobank *transfers a collection* for a new research project, the scientific director must make a declaration to the Ministry of Research, which takes into account the opinion of the CPP. The biobank may transfer its samples by means of a material transfer agreement (MTA) in accordance with articles L. 1243-3, L. 1243-4 and R. 1243-49 of the CSP. This document contains the references and the purpose of the applicant (the identity of the applicant, the structure of the institution, the context of the request, the use of the biological materials, the charges and indemnities, respect of confidentiality).

The principles of the non-opposition of the patient and of requalification, or even transfer, of a collection of biological samples show the limits of the traceability of consent. The future of consent is projected in the form of an electronic version of informed consent. Already regulated by FDA guidelines in the United States but still non-existent in France, *electronic informed consent (e-consent or eIC)* refers to "the use of electronic systems and processes that may employ multiple electronic media, including text, graphics, audio, video, podcasts, passive and interactive Web sites, biological recognition devices, and card readers, to convey information related to the study and to obtain and document informed consent. The information presented to the subject, processes used for obtaining informed consent, and documentation of the electronic informed consent (eIC) must meet the requirements of these and other applicable regulations" [FDA 16]. The 2016 FDA guidance for scientific leaders emphasizes the dynamics of patient interaction and information enrichment. It offers traceability of the sample in real time: the clinician controls the receipt of documents and evaluates the correct understanding of the information by the patient. For their part, manufacturers monitor the use of samples and data.

As well as requests for authorization and declarations to the competent authorities, the exchange of information between research teams is mainly

carried out electronically. These new computer technologies are being integrated into the lives of patients with the implementation of *electronic health records (EHRs)* in most US hospitals and shared medical records in France (*dossier médical partagé*). In clinical research, data collection is recorded in an *electronic case report form (eCRF).* This contains all the individual information collected during the study: patient identification, study inclusion criteria, informed consent and all medical examination results. It centralizes the individual management of the clinical study, which can potentially be accumulated for all patients in order to conduct a clinical trial that is largely numerical and respects the ethical and regulatory bases. Obviously, biological analysis samples are obtained in laboratories close to the patient's home and medical examinations are carried out by the attending physician. Over the past 20 years, digital transformation has made its entrance into the hospital and more broadly into healthcare with the rise of e-health.

1.3.2.3. *Protection of the data associated with the sample*

French law provides a protective framework for personal data with the law on data processing, files and freedoms (law of January 6, 1978, as amended), which the CNIL is responsible for enforcing. Article 8 of the law stipulates:

> It is prohibited to collect or process personal data that reveal, directly or indirectly, the racial or ethnic origins, political, philosophical or religious opinions or trade union membership of persons, or that relate to the health or sex life of such persons.

Computerized systems centralizing a vast amount of clinical, family, environmental and personal information must be able to ensure the *confidentiality* of all data associated with the donation of biological samples. The constitution of a collection of human biological samples for genetic or non-genetic research purposes is governed by articles L. 1131-4, L. 1243-3 and L. 1243-4 of the French Public Health Code. Once the collection project has been declared to the competent administrative authority, the latter ensures that the conditions of constitution, conservation and exploitation of the collection present sufficient guarantees to ensure the proper use, security and confidentiality of the data collected.

Upon receipt, the biospecimen starts its journey in the biobank with a coded identification. The management of the samples requires *the anonymization of data*. Absolute anonymization is not consistent with the principle of returning clinical information to the patient. Pseudonymization is reversible and enables re-identification of the patient. The person's identity information (surname, first names and social security number) is coded by a key or encryption code assigned to all samples from the same individual. This coding facilitates the longitudinal follow-up of the patient, which can extend over several years during an epidemiological study.

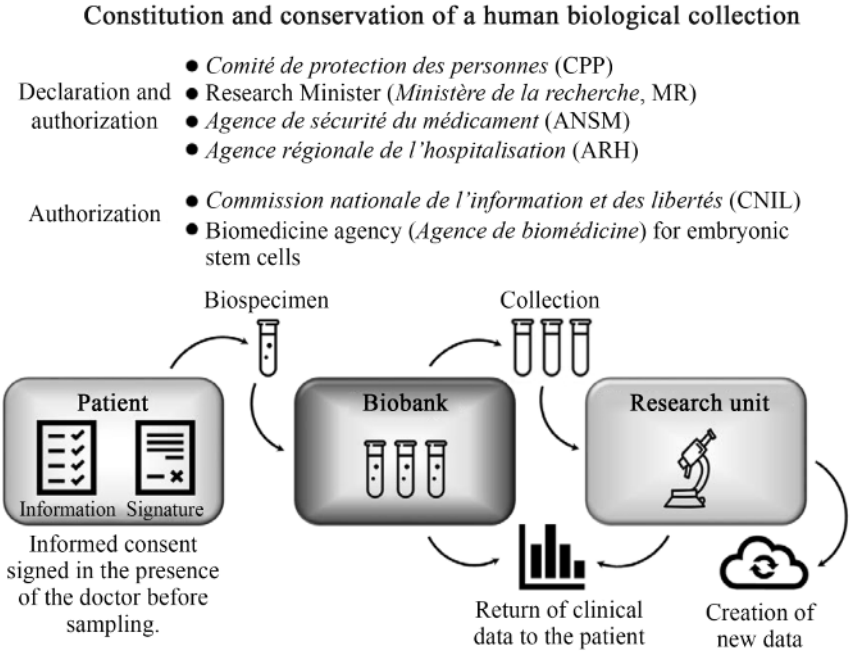

Figure 1.8. *The legal framework for biobanks for the conservation of a collection of human samples*

COMMENTARY ON FIGURE 1.8.– *After the favorable opinion of the CPP, the person in charge of the biobank declares the conservation of the collection to the Ministry of Higher Education, Research and Innovation and to the ANSM, as well as to the French Regional Agency for Hospitalization (Agence Régionale de l'Hospitalisation (ARH)). The creation of new data from research is submitted for authorization to the CNIL. In the case of*

embryonic stem cells, authorization to store the collection is requested from the Biomedicine Agency (Agence de Biomédecine).

At the European level, the new *General Data Protection Regulation* (GDPR) reinforces the principles of data protection and the rights of citizens, repealing Directive 95/46/EC (Data Protection Directive) [EUR 16]. This regulation applies to activities of the biobank both upstream, the missions related to the collection of personal data, and downstream, the creation of new data from and associated with the samples [CHA 16]. The processing of personal data must respect in particular law no. 78-17 of January 6, 1978. relating to "information technology, files and freedoms", consolidated version from July 23, 2019. The CNIL must be consulted before any data processing to ensure that security and confidentiality are respected, as well as the proper use of personal data, especially in the case of genetic data.

The third pillar of medical research, after the management of biological samples and complex data, is the regulatory and legal framework that ensures respect for the rights of individuals. Guaranteeing the protection of patients is a reinforced priority for biobanks, which are subject to the accelerated pace of scientific and technological innovations.