cohorte CONSTANCES

UMS 11
Cohortes épidémiologiques en population

THE CONSTANCES COHORT

MARIE ZINS & MARCEL GOLDBERG

Janvier 2015

l'Assurance Maladie

l'Assurance Retraite

UNIVERSITÉ DE VERSAILLES
ST-QUENTIN-EN-YVELINES

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE DES
AFFAIRES SOCIALES ET
DE LA SANTÉ

Inserm

InsermTransfert

www.constances.fr
UMS 011 « Cohortes épidémiologiques en population »
Inserm - Université Versailles St-Quentin

CONTENT

FOREWORD

This report presents the main features of the protocol of the CONSTANCES cohort. Additional documents (mostly in French) provide details on specific aspects of the project.

1. Biobank technical report
2. Biology protocol
3. Charter, Call for research projects, project application form – French and English versions
4. Data Catalogue
5. Data flow
6. Inclusion and follow-up questionnaires
7. Information documents for the participants
8. Journal for the participants
9. Legal authorizations
10. Power calculations
11. Procedures for diagnosis ascertainment
12. Quality control for biology
13. Standardized operational procedures (SOPs)
14. Video presentation of the SOPs

To download documents, use the following link: www.constances.fr

SUMMARY

Large-scale, prospective observational cohorts have become essential resources for investigation into the causes of many diseases, especially common multifactorial diseases with multiple environmental and genetic determinants. When they are based on samples representative of the general population, prospective cohorts may also be used for descriptive and epidemiological surveillance purposes. "General-purpose cohorts" in epidemiology and public health are designed to cover a broad scope of determinants and outcomes, in order to answer several research questions, even questions not defined at study inception. Such cohorts constitute research platforms that can be set up to be open to the research community for developing multiple nested projects.

General objective of CONSTANCES

The general objective of the CONSTANCES ("*Cohorte des consultants des Centres d'examens de santé*") project is to set up a large population-based cohort designed and managed as a national infrastructure that will contribute to the development of epidemiologic research by hosting ancillary nested projects on a wide range of scientific domains, and to provide public health information.

Global description of CONSTANCES

CONSTANCES is designed as a randomly selected representative sample of French adults aged 18-69 years at study inception; 200,000 subjects will be included over a five-year period. **At inclusion**, the selected subjects are invited to attend one of the 17 participating Health Screening Centers (HSCs) for a comprehensive health examination based on standardized operational procedures (SOPs): weight, height, blood pressure, electrocardiogram, vision, auditory, spirometry, and biological parameters; for those aged 45 years and older, a specific work-up of functional, physical, and cognitive capacities are performed. A biobank, including blood and urine samples for each participant will be set up. Quality control procedures, including regular on-site visits of research assistants, are delegated to independent organizations. **The follow-up** includes a yearly self-administered questionnaire, and a periodic visit to an HSC. Data is regularly extracted from the French retirement (CNAV), health (CNAMTS-SNIIRAM and PMSI) and death register (CépiDc-INSERM) national databases. The data collected include social and demographic characteristics, socioeconomic status, life events, lifestyle and occupational factors. The health data cover a wide spectrum. Extensive procedures have been developed to use the national healthcare databases to allow identification and validation of diseases over the follow-up.

Other institutions involved

CONSTANCES is conducted in partnership with several institutions in charge of the main data sources of the cohort, each contributing to specific aspects. The *Caisse nationale d'assurance maladie des travailleurs salariés*-CNAMTS gives access to the HSCs for health examinations and linkage to the *Système national inter-régimes d'information de l'assurance maladie*-SNIIRAM, the National Health Insurance Information System. The *Caisse Nationale d'Assurance Vieillesse*-CNAV performs the sampling of the CONSTANCES eligible population and also allows linkage of the cohort to its professional, social and vital status databases, and the *Centre d'Épidémiologie des Causes de Décès*-CépiDc/INSERM provides the causes of

deaths. The National Institute for Public Health Surveillance (*Institut de Veille Sanitaire*-InVS) is developing two companion cohorts of CONSTANCES, one of agricultural workers and one of self-employed workers. CONSTANCES has numerous collaborations in France and abroad and is included in several international consortia.

Resources - Access to the infrastructure

The CONSTANCES Cohort project is conducted by the *Population-Based Epidemiological Cohorts Unit*-UMS 011 (UMS 011), recently created by INSERM and the Versailles-Saint Quentin en Yvelines University (UVSQ). UMS 011 is experienced in developing and managing large population-based cohorts. CONSTANCES benefits from the facilities already in use to safely store the database and ensure its confidentiality. All confidentiality, safety and security procedures were approved by the French legal authorities.

Every research group in France or in other countries, public or private, is entitled to apply for developing research projects within CONSTANCES after approval by the governing bodies of CONSTANCES: the Steering Committee (composed of representatives of INSERM, UVSQ, CNAMTS, CNAV and the Ministry of Health), the International Scientific Committee and the INSERM Ethics Committee.

1 CONTEXT

1.1 POPULATION-BASED GENERAL-PURPOSE COHORTS

Large-scale, prospective observational cohort studies have become essential resources for research on the causes of many diseases, especially common multifactorial diseases with multiple environmental and genetic determinants. These cohort studies include several hundreds of thousands of individuals and incorporate personal, social, lifestyle, occupational and environmental data as well as biobanks of blood and other biological materials. Population ageing implies that the burden of these diseases, often clustering in the same individuals, will further increase in the near future. Prospective cohort studies are indeed the optimal design for establishing causal pathways between exposure to specific agents and the onset of diseases, since they can take advantage of the longitudinal follow-up of individual participants to model the effects of multiple health, genetic and biologic, lifestyle or environmental factors and their interactions.

When based on population representative samples, prospective cohorts may also be used for descriptive and epidemiological surveillance purposes, allowing the study of the temporal evolution of the incidence of health outcomes and of the distribution of risk factors. Finally, cohort studies are also efficient tools for the evaluation of public health interventions either at an individual or a collective scale.

One can make a distinction between "population-based" and "patients" cohorts. The latter focus on specific diseases and include a relatively small number of patients who are followed for studying the course of the disease. Population-based cohorts include subjects in the population and are usually aimed at analyzing the causes of and risk factors for different health outcomes. They must be large, in order to include in sufficient numbers of persons exposed to various risk factors and to observe incident diseases. Some population-based cohorts focus on specific research areas regarding either risk factors (diet, smoking…), population sub-groups (children, women, ethnic groups…), or specific endpoints (cancer, type 2 diabetes, cardiovascular or kidney diseases, dementia…). Others are "general-purpose cohorts" designed to cover a broad scope of determinants and outcomes; they are designed to answer several research questions, even questions not defined at study inception. Often characterised as research infrastructures rather than as research projects, such cohorts are usually designed to be used by many investigators for multiple research projects and to operate for a fairly long period of time (up to 50 years in certain cases) [Burton et al. 2009a]. **The CONSTANCES cohort ("*Cohorte des consultants des Centres d'examens de santé*") is designed as a "population-based general-purpose cohort**".

1.2 THE NEED FOR LARGE-SIZED PROSPECTIVE COHORTS AND FOR DATA SHARING

Research on the causes of diseases in the field of environmental, occupational, social, genetic or pharmacoepidemiology often reveals small relative risks for individual risk factors. Very large-scale cohorts, providing high quality phenotyping and long-term follow-up, are required to ensure sufficient statistical power to better understand the role of various personal and environmental factors and their interaction with complex genetic traits. For instance, known associations between genetic variants and chronic diseases show typical allelic odds ratios in the range 1.1–1.4 (Burton al. 2009b]. The reliable identification of such effects demands vast data sets [Collins 2004]. Case-control studies show that thousands of cases are required even when interest focuses on the simplest situations, and when the research question focuses on the study of gene-environment and gene-gene interactions and the comprehensive exploration of causal pathways, tens of thousands of cases will often be required. Tens of thousands of subjects may also be required to study a quantitative phenotype (e.g. measured blood pressure), because allelic effect sizes may be as small as one-tenth of a standard deviation, or even less [Newton-Cheh et al. 2009]. Beginning with the Framingham Study, which follows-up from 1948 on a few thousand volunteers [Oppenheimer 2005], much larger prospective cohorts including hundreds of thousands of subjects were launched in different countries, such as the Nurses' Health Study [Egan et al. 2002], the One Million Women Study [Darling et al. 1998], the UK Biobank [Collins 2007], the Kadoorie Study of Chronic Disease in China [Chen et al. 2005], the Norwegian CONOR Consortium [Naess et al. 2008], or the EPIC European Prospective Investigation into Cancer and Nutrition [Riboli et al. 2002]. Other very large population-based cohorts with hundreds of thousands of participants are currently being implemented in different countries, such as the German National Cohort, LifeGene in Sweden, the Dutch LifeLines Study and Biobank or the Cartagene Cohort in Québec, Canada[1].

Nevertheless, as scientific progress demands access to databases providing comprehensive, accurate and precise information on relevant factors and disease traits and to very large sample sizes, even the largest studies only generate enough cases to investigate the commonest of complex diseases and relatively simple etiological patterns or have to focus on quantitative disease-related traits. It is thus often necessary to pool data across multiple studies [Thompson 2009], and large collaborative consortia have been responsible for much of the recent progress in human population genomics [Hindorff et al. 2009]. Large-scale data pooling is equally important in mainstream epidemiology, when power is a concern [Friedenreich 1993], as well as in public health and health-services research, and in comparative international analysis of social determinants of health [Backlund et al. 2007]. Such pooling not only supports the attainment of large sample sizes but is also useful to reduce potential biases arising from access to restricted sets of data.

The important investments necessary to build these large studies argue for their optimal utilization and have made the issue of access to data at center stage internationally. Numerous health research funding institutions have recently expressed their strong desire to promote data sharing [Walport &. Brest 2011]. The scientific community involved in developing and maintaining large population-based cohorts is trying to organize international collaborations taking into account the scientific, ethical and legal aspects of the use of collected materials [Doiron et al. 2013]. A major obstacle for sharing cohort epidemiologic and biological data, besides the technical and ethical aspects, is the lack of recognition of the efforts behind establishing and maintaining such resources and the absence of a clear link between their initiators/implementers and the impact of scientific research using them. The scientific community is currently trying to propose solutions to this problem, such as the creation of a "*Bioresource Research Impact Factor*" that would make it possible to trace the quantitative use of a bioresource, the kind of research utilizing it, and

[1] www.nationale-kohorte.de/index_en.html; lifegene.ki.se; www.lifelines.net; http://cartagene.qc.ca/en

the efforts of people and institutions that construct it and make it available [The BRIF Workshop Group 2011].

1.3 Population-based prospective cohorts in France and at the international level

There is indeed a wealth of **population-based epidemiological cohort studies in France**. The oldest, still in progress, is the *Paris Prospective Study*, following about 8,000 male Paris policemen since 1967 [Ducimetière et al. 1981]. Since the end of the 80's several other population-based cohorts were set up, such as the *PAQUID Cohort* [Dartigues et al. 1991] and the *EVA* [Berr et al. 2000] and *3C Cohorts* [Three-Cities Study Group 2003] devoted to the study of ageing, the *E3N-EPIC Cohort* [Clavel-Chapelon 2002] and the *SUVIMAX Cohort* [Hercberg et al. 2004] mainly for studying the effects of nutrition on specific outcomes, the *Epipage* [Ancel et al. 2008], *Eden* [Yazbeck et al. 2009] and *Elfe Cohorts*[2] which followed-up newborns. The GAZEL Cohort Study[3] designed and managed by our group for over 25 years was certainly the first attempt in France to develop a population-based general-purpose cohort [Zins et al. 2009]. The GAZEL Cohort Study was set up in 1989 among workers of Électricité de France-Gaz de France (EDF-GDF), the French national utility company. It was designed as an "open epidemiologic laboratory" characterized by a broad coverage of health problems and determinants and accessible to the community of researchers. At study inception in 1989, the GAZEL Cohort Study included 20,625 volunteers working at EDF-GDF (15,011 men and 5,614 women) then aged from 35 to 50 years. The data routinely collected cover diverse dimensions and come from different sources: annual self-administered questionnaire (morbidity, lifestyles, life events, etc.); personnel department of EDF-GDF for social, demographic, and occupational characteristics; EDF-GDF special social insurance fund (for sickness absences and cancer and ischemic heart disease registries), occupational medicine (occupational exposure and working conditions), Social Action Fund (healthcare utilization), Health Screening Centres for standardized health examination and the National Death Register (causes of death); a biobank including more than 8,000 blood samples was also set up. Today, more than 60 ancillary projects on diversified themes have been set up in the GAZEL Cohort Study by some 30 French and foreign teams from Belgium, Canada, Denmark, Finland, Germany, Great Britain, Sweden, and USA. Health problems as diverse as migraine, postmenopausal osteoporosis, ischemic heart disease, depression, musculoskeletal diseases, or traffic accidents have been the object of research projects in this cohort. They take into account risk factors that are lifestyle, social, psychological, occupational, and medical. A substantial proportion of the research work has focused on the problem of social inequalities in health and their occupational, personal, and social determinants using a life course perspective [Zins et al. 2009]. GAZEL is also involved in the IDEAR (*Integrated Datasets across Europe for Ageing Research*)[4], the IPD-Work[5], the P3G[6] and the BBMRI-LPC[7] consortia.

[2] www.elfe-france.fr/index.php/en/

[3] www.gazel.inserm.fr/en/

[4] www.idear-net.net/

[5] https://osha.europa.eu/sub/newoshera/en/noe-research-programme/project-abstracts/the-ipd-work-individual-participant-data-meta-analysis-of-working-populations-consortium

GAZEL has been quite productive and will continue to host research projects, especially on ageing and the study of long-term effects of various risk factors. It has nevertheless some important limitations. Albeit one of the largest French prospective cohorts, its size is limited and many uncommon conditions cannot be studied with a sample of 20,000 subjects. The age structure of the cohort was restricted to persons aged from 35 to 50 years at study inception; they are now aged 57 to 72 years, not allowing research on younger subjects. The study population is made of persons who were employed in a public company; this had some methodological and logistic advantages, but excluded certain categories of the population (e.g. unemployed, those with unstable jobs or foreigners).

To summarize the French situation, the existing prospective cohorts are limited in size (a few thousands or tens of thousands participants), focused on specific outcomes and/or risk factors and/or specific subgroups of the population, and (with few exceptions) not included in international cohort data pooling. Finally, there is a lack of very large general-purpose population-based cohorts, as highlighted by a report on the conditions of the development of epidemiology of the French Academy of Sciences [Valleron 2006]. It is widely recognized that we need very large cohorts for scientific as well as for public health purposes: in recent years, there were several official reports from different bodies that called for the constitution of new population-based general-purpose cohorts. Thus there is a need in France for a very large-sized prospective general-purpose population-based cohort such as CONSTANCES.

When we look at the international picture, CONSTANCES has some strengths compared to most of the existing cohorts and those currently under development. First, we collect very detailed data on personal, environmental, behavioral, occupational and social factors. There are over 100 prospective population-based cohorts in various stages of development and realization around the world [Doiron et al. 2013]. The vast majority of these have collected biological material (e.g. blood) with a view to correlating genetic characteristics with subsequent disease onset. Some of the cohorts have collected some environmental and lifestyle data, such as smoking history and maybe a food frequency questionnaire. But very few have collected extensive environmental, lifestyle and social data in a detailed manner, making it difficult to assess gene-environment interactions. For CONSTANCES, we are planning to collect genetic material, which is of course essential, but more uniquely, we also collect an extensive array of lifestyle and environmental data using advanced state-of-the-art methods.

1.4 CONSTANCES PARTNER INSTITUTIONS

The Caisse Nationale d'Assurance Maladie des travailleurs salariés-CNAMTS (www.ameli.fr)

The CONSTANCES cohort project is conducted in close partnership with the *Caisse nationale d'assurance maladie des travailleurs salariés*-CNAMTS. In France the General Health Insurance Fund administered by CNAMTS is compulsory for salaried workers and their family

[6] http://p3g.org/

[7] http://www.bbmri-lpc.org/

(including unemployed and foreigners); it covers about 85% of the French population. The partnership with CNAMTS involves two major contributions to CONSTANCES: (i) access to the CNAMTS' Health Screening Centers (HSCs): everyone with health insurance from CNAMTS, as well as their dependents, is entitled to receive health examinations entirely paid by CNAMTS, that include extensive work-ups conducted in some 110 HSCs located throughout France; the cohort participants will benefit from comprehensive health examinations in the participating HSCs, including the collection of blood and urine samples for the biobank; (ii) access to the *Système national d'information inter-régimes de l'Assurance maladie*-SNIIRAM, the National Health Insurance Information System managed by CNAMTS, which permanently collects health-care utilization claims, where individual health data, including hospital discharge records, are regularly extracted for each CONSTANCES participant at inclusion and during the follow-up of the cohort.

The Caisse Nationale d'Assurance Vieillesse-CNAV (www.lassuranceretraite.fr)

The National Retirement Insurance Fund administered by the *Caisse nationale d'assurance vieillesse*-CNAV includes all persons in France affiliated with the CNAMTS. CNAV permanently collects from different sources professional and social data compulsory for establishing the retirement benefits of all individuals in France. The initial sampling of the CONSTANCES eligible population is done within the CNAV database, and professional, social data and vital status are regularly extracted for each CONSTANCES participant at inclusion and during the follow-up.

The Centre d'épidémiologie des causes de décès-CépiDc/INSERM (www.cepidc.vesinet.inserm.fr)

CépiDc is part of INSERM; it is in charge of the National Death Registry and will be the source of information on causes of death during follow-up.

The National Institute for Health Surveillance (Institut de Veille Sanitaire-InVS) (www.invs.sante.fr)

A close collaboration has been established with the National Institute for Public Health Surveillance (*Institut de Veille Sanitaire*-InVS), the national agency in charge of the epidemiological surveillance of health in France. A major axis of the cooperation with InVS is the COSET Program, aimed at developing a surveillance tool for the whole working population. The general principle of our collaboration is based on the fact that CONSTANCES will include only salaried workers affiliated to CNAMTS, but not agricultural and self-employed workers for reasons which will be detailed below (see 1.3.13). Thus, the Occupational Health Department of InVS is in charge of developing two companion cohorts, one based on agricultural workers and the other on self-employed workers. Additionally, due to its expertise, the Occupational Health Department will be in charge of the coding of occupations according to French and international classifications for all three cohorts. While each partner is responsible for its cohorts, subsamples of data will be shared and included in the other's database: some CONSTANCES' data will be transferred to the COSET database, and reciprocally some data from COSET will be transferred to CONSTANCES' database; the utilization of these shared data will be under the control of the governing bodies of each project according to their own rules. All exchange of data will follow strict procedures in order to guarantee anonymity of the participants.

Other institutions

The above listed institutions are those that contribute directly to the functioning of the CONSTANCES cohort. There are also other institutions who are mainly interested in using the epidemiological and biological data from CONSTANCES, either for research or for public health purposes.

INSERM (www.inserm.fr) is the main medical research institution in France. Within INSERM, the Institute of Public Health (ISP) is in charge of coordinating research in epidemiology and public health, and is also in charge of the coordination of the epidemiological cohorts funded through the National Research Agency (ANR). The BIOBANQUES national infrastructure coordinated by INSERM is also a partner of the biobank component of CONSTANCES.

More generally, we expect numerous research projects from the French and international health and epidemiological research communities, as detailed in Section 4.

The main French Public health institutions are interested in the infrastructure, especially as CONSTANCES is designed to be a large representative and longitudinal sample of the adult French population with individual level data on social and demographic characteristics, socioeconomic status, life events, lifestyle, and occupational factors as well as biological data. The Ministry of Health expects CONSTANCES to regularly provide a series of health indicators; it has supported and partly funded the project since its beginning. InVS intends to use the database for its surveillance programs in the fields of chronic and infectious diseases and occupational and environmental health. Other public health institutions such as the National Cancer Institute-INCa (www.e-cancer.fr), the National Agency for Drug Safety-ANSM (www.ansm.sante.fr) expressed their interest and supported the project.

Finally, with the help of INSERM-Transfert, a subsidiary of INSERM devoted to collaboration with the private sector (www.inserm-transfert.fr), partnerships with major private pharmaceutical companies were established.

1.5 CONSTANCES ACCREDITATIONS

In addition to the above mentioned institutions that support the project in different ways, CONSTANCES is accredited as a "*National Research Infrastructure in Biology and Health*" by the National Research Agency-ANR and receives as such funding from the French governmental "*Investissements d'avenir"* program. CONSTANCES received also the "*Label of general interest and statistical quality*" from the French National Council for Statistical Information-CNIS.

2 OVERVIEW OF THE PROTOCOL

2.1 GENERAL OBJECTIVES

The general objective of the CONSTANCES project is to set up a large population-based general-purpose observational prospective cohort designed as a research infrastructure to contribute to the development of epidemiologic research and to provide useful public health information. CONSTANCES offers to the scientific community the possibility of developing large-scale projects at a minimal cost and effort by alleviating the financial, technical, and time burdens related to developing and maintaining large cohorts. Repeated collection of extensive information from personal questionnaires, medical examinations and health and socioeconomic databases covering many diverse domains such as lifestyle, environmental, occupational, social factors and of biologic and genetic markers allow the examination of pathways leading to disease development, establishing models for identifying individuals at increased risk of developing major diseases, evaluating markers for early detection of disease, and for assessing socioeconomic disparities in health in France and studying their determinants. Apart from its contribution to scientific knowledge, CONSTANCES provides new information on the impact of major determinants of health in the French population, providing a sound base for targeted prevention.

The CONSTANCES cohort is designed as a large sample of 200,000 persons aged 18-69 at inception, representative of the general French adult population affiliated to the General Health Insurance Fund (about 85% of the general population), and characterized by a broad coverage of health conditions and health determinants. **This cohort constitutes a national infrastructure to serve as an open epidemiologic platform widely accessible to the research community**, making it possible to conduct projects on a variety of scientific questions. It serves as a large scientific instrument, in a similar manner to a telescope or a particle accelerator, for example, or a genotyping laboratory — not built to answer a specific question, but rather to help analyze a wide range of scientific questions. In accordance with this general objective, its duration is not defined: the cohort is intended to serve as an object of longitudinal follow-up without a time limit, in order to study the effects of health determinants over the very long term and to be able to take into account progress in knowledge and techniques that continuously raises new scientific questions that it can help elucidate.

CONSTANCES is also designed as a tool to provide descriptive information in support to the objectives of the public health authorities in different domains. By its thorough system of follow-up and collection of very diverse information through a variety of methods and data sources on a large representative sample of the adult population, CONSTANCES contributes to better knowledge of the health of the French population. CONSTANCES also serves as a tool for the epidemiological surveillance of health in different domains.

A general-purpose cohort allowing for specific "nested" ancillary research studies

The usual way of using the CONSTANCES infrastructure is twofold: (i) by producing regularly health indicators for the public health authorities and health agencies; (ii) through specific "nested" ancillary studies conducted by teams of the French and international research community.

A nested study is an ancillary project designed to answer a specific research question; it uses a subset of the CONSTANCES database (epidemiological or/and biological data) selected according to the research hypotheses, including either the whole cohort or on subsample of subjects selected on the basis of personal characteristics, health or risk factor criteria. While based on the CONSTANCES cohort, they are conducted by independent investigators and independently funded. We describe below the procedure for eliciting and selecting ancillary projects based on the CONSTANCES infrastructure and we give some examples of such projects covering a wide range of scientific and public health topics already in preparation in collaboration with several French and international research groups.

Designing a population-based general-purpose cohort open to the scientific national and international communities, aimed at allowing the development of projects on very different specific research questions as well as providing descriptive public health information presents certain problems. A major challenge for a general-purpose epidemiological survey is to define which core set of basic data on various topics should be prospectively collected for the whole cohort that could be used either as variables of main interest or as covariates for a wide range of scientific objectives. This is an especially complex task in view of a long-term perspective when specific research questions that could be investigated in the future are not known at the time of the cohort is designed. We relied on our more than 25 years' experience with the GAZEL Cohort Study (www.gazel.inserm.fr/en/) to develop **a two-step research strategy**.

1 - First, an extensive set of data on biological, genetic, physiologic, personal, lifestyle, environmental, occupational and social determinants and many diverse health outcomes is collected prospectively for the whole cohort with annual repeated measurements and permanent linkage to the French health and social databases. This core set of data was defined after thorough discussions with researchers working in many different fields (cancer, diabetes, cardiovascular diseases, mental health, neurology, aging, social, occupational or environmental epidemiology, nutrition, etc.) interested in developing future research projects within CONSTANCES. The final list of information to be collected routinely was decided balancing scientific considerations and practical aspects: size of the questionnaires, capacity of the HSCs to perform the examinations at a large scale with a high level of quality and standardization, availability of data in the health and social databases, cost. Additionally, as we send a follow-up questionnaire each year (see below for the rationale for this procedure), supplementary data can also be prospectively collected when new research topics arise. The wealth of information that will thus be routinely available for the whole cohort will enable the health indicators needed by the public health authorities to be produced. Moreover, our experience with the GAZEL Cohort Study makes us confident that many ancillary research studies will be conducted using this "basic" CONSTANCES database.

2- However, it is of course not possible to systematically collect all the data that could be needed for any research purpose. Thus, in a second step some additional information has to be collected when needed for nested projects with specific research questions. In such cases, the management procedures allow for supplementary data collection, either on the whole cohort or on subsets of subjects selected from the basic CONSTANCES database on the basis of personal characteristics, health or risk factor criteria. The Charter of CONSTANCES details the rules to follow to collect such additional data[§].

2.2 A FOCUS ON SOME MAIN TOPICS

Apart from aiming to build a wide-spectrum infrastructure for epidemiology, CONSTANCES was also specifically shaped for the study of social, occupational and environmental determinants of health and aging, of the effects of diet on health, and of the functioning of the French health care system.

Social, lifestyle, occupational and environmental determinants of health and aging
We are especially interested in occupational, environmental, and lifestyle factors in the etiology of cancer and in exploring the genetic polymorphisms that make individuals susceptible to these factors [Siemiatycki et al. 2004]. Musculoskeletal disorders in relation to working conditions and biomechanical and psychosocial factors at work are also a key topic of interest, focused on the short and long-term medical, social and professional consequences of musculoskeletal disorders with a longitudinal perspective [Kilbom et al. 2008]. The effects of exposure to occupational chemicals on respiratory diseases (COPD and asthma) [Roche et al. 2008] and on neurodegenerative diseases (Parkinson disease, dementia) [Brayne 2007] and cognitive functioning is an important concern. Psychosocial factors at work in relation to coronary heart disease, depression and mental health and other outcomes; due to the economic context in industrial countries, there is also a major interest in workability and other determinants of early exit from the labor force, as well as on determinants and consequences of working beyond retirement age [Banks et al. 2006]. Within the associated COSET program (see below), it is planned to regularly produce national indicators of occupational health (distribution and time trends of occupational risk factors and diseases according to professions and economic sectors); for large occupational groups that are included in sufficient numbers (from 10,000 to 35,000), such as teachers, hospital workers, university students and agricultural or self-employed workers, more detailed analyses are possible.

Social determinants and health inequalities are another major area of interest for CONSTANCES. This covers social inequalities in the occurrence, treatment and socioeconomic consequences of common conditions such as diabetes, cancer, depression and other psychiatric problems or cardiovascular diseases [Marmot et al. 2008]. Childhood adversities and major life events and subsequent risks for physical and mental health can be analyzed in a life course perspective [Kuh & Ben-Shlomo 1997]. Work stress and work-family conflict in relation to subsequent risk for cardiovascular disease and cognitive decline taking gender differences into account can be studied [Cohidon et al. 2004]. The health impact of control and reward in core occupational and social roles is also a topic of interest [Marmot 2004].

Regarding **aging**, as CONSTANCES collects detailed data on cognitive and physical performance from the age of 45, which is earlier in life than most of the available cohorts [Finch 2009], many specific questions about aging can be addressed, such as the study of the occupational, personal and genetic determinants of cognitive decline, the effects of retirement on cognition, or factors that may lead to inactivity and isolation, factors and mechanisms that contribute to successful aging, and conversely those that contribute to disabilities and/or frailty [Christensen et al 2009]. Efforts are also be made to understand the causes of individual and social heterogeneity in aging by investigating the nature of the association between risk factors and cognitive aging in terms of cumulative risk, risk trajectories or critical period models [Gill et al. 2010]. Research on consequences of aging is focused on the impact of poor functional status on survival and functioning and the potential causes of its variation by socioeconomic position [Jagger et al. 2008]. As available cognitive tests used in common practice were developed for elderly populations, there are no reference values for younger people; it is thus of particular interest for both clinical practice and epidemiologic research to establish such normative values for the 45–65 age range.

The relationships between **health and environment** are also of particular interest. This is an essential area of research in public health and epidemiology today, involving numerous health problems and diverse populations. Health risk factors with origins in environmental exposures are numerous: outdoor and indoor air pollution, water pollution, noise, temperature, etc. These factors can induce various acute and chronic diseases, such as allergy, respiratory diseases, asthma, reproduction, cancer, cardiovascular diseases, or infectious diseases [WHO 2006]. Contextual studies have also shown that neighborhood characteristics, such as socioeconomic level, collective equipment, or density of food or alcohol outlets, can influence obesity, coronary heart disease, mortality, or smoking and physical activity independently of the effects of individual factors [Diez-Roux 2001]. CONSTANCES was designed to support research projects focused on environmental health through the systematic prospective collection and X/Y geocoding of residential addresses for all cohort participants. Several spatially resolved maps exist in France regarding different aspects of environmental exposure: noise, weather, air and water pollution, natural radiation, road traffic, etc.; through the linkage of residential addresses with these databases, individual exposures can be assessed for participants living in the areas covered by the available environmental surveillance systems[8]. Administrative databases containing numerous socioeconomic data (demography, economic activity, housing, or migration) are also available at different geographical levels, including small-scale units such as the IRIS ("*Ilots Regroupés pour des Indicateurs Statistiques*")[9], allowing for the use of recently developed French small area deprivation indices. It is thus possible to conduct many different studies on various exposures and outcomes, as well as contextual analyses of the effects of neighborhood characteristics.

Since decades, **diet** has been the subject of much research regarding its possible role in the occurrence of some major frequent chronic diseases, such as cancer, hypertension and cardiovascular diseases, diabetes, depression or osteoporosis. There is well-established evidence on the relationships of diet with the risk for some diseases. For instance, high levels of consumption of red and processed meat or alcohol is associated with increased risks of certain cancer types, while a reduction in some cancer risks for high intake levels of dietary fiber and whole-grain foods is likely; the risk of cardiovascular diseases decreases for high consumption levels of fruits, vegetables and n-3 polyunsaturated fatty acids and on the contrary increases in relation to consumption of saturated fatty acids and trans-fatty acids. However, there are still major inconsistencies in nutritional epidemiologic studies of chronic diseases, and many hypotheses on the etiological role of diet on disease incidence remain to be clarified [Schatzkin et al. 2009]. We prospectively collect basic data on dietary habits through the CONSTANCES annual questionnaires; more in-depth data could be later

[8] For details on available environmental databases, see: www.statistiques.developpement-durable.gouv.fr/theme/environnement/1097.html

[9] www.insee.fr/fr/methodes/default.asp?page=definitions/ilots-regr-pour-inf-stat.htm

collected from additional questionnaires and biological samples within the framework of ancillary projects focusing on some aspects diet and health.

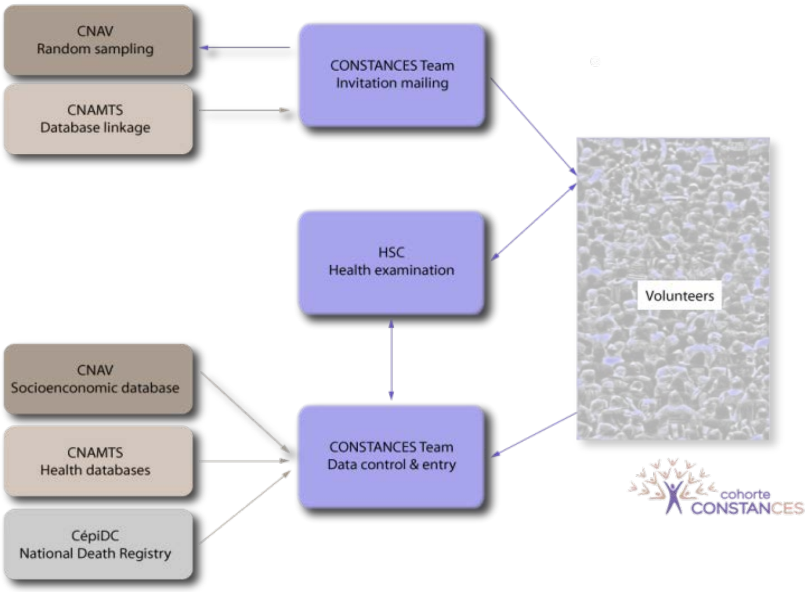
Functioning of the French health care system

Finally, as the main funder of the health care expenditures, CNAMTS is interested in many aspects of the functioning of the health system that can be studied taking into account the main personal and socioeconomic characteristics of the participants: behaviors of the patients and health care professionals and impact of the recommendations for good practice, health care and professional trajectories of the patients suffering from chronic diseases, inequalities in access to health care resources, etc.

2.3 GENERAL DESIGN OF THE COHORT

In this section, we summarize the main methodological and practical features of the protocol. The figure 1 presents a general overview of the design.

Figure 1. General overview of the design



The CONSTANCES cohort was designed as a randomly selected representative sample of 200,000 French adults aged 18-69 years. At study inception, the CNAV draws a random sample of the eligible population using our sampling scheme, and exchange with the CNAMTS database to complete the data needed to send an mail invitation to the selected persons and to set up the database of the future data transmission operations (study numbers, postal addresses, etc.; see below for a more detailed description of the data flow). The persons thus selected receive an invitation to participate and they have to send a completed form to their HSC and receive in return a questionnaire to fill out and an invitation for a comprehensive health examination based on SOPs specifically developed for CONSTANCES and biological samples collection. Data collected in the HSC are then sent to the CONSTANCES team and completed with data extracted from the national databases managed by CNAMTS and CNAV. Follow-up includes mailed questionnaires, linkage with the national social and health databases, and regular invitations to the HSCs. Quality control procedures, including regular on-site visits by research assistants, are delegated to independent organizations. To take into account non-participation at inclusion and attrition throughout the longitudinal follow-up, a cohort of non-participants was set up and followed through the same national databases as participants.

2.4 COMPOSITION, SIZE AND LENGTH OF FOLLOW-UP OF THE COHORT

As one of the main study objectives is to provide information on the health status and disease burden of the French adult population, the CONSTANCES cohort is a random sample representative of the general French population insured by CNAMTS in terms of age (18 to 69 years at inception), sex, and socioeconomic status; the collaboration with the COSET Program allows to cover also the fraction of the population which is insured by other insurance funds (see below).

Due to its general-purpose objectives, the size of the cohort was not defined by classical power calculations. However, it clear that to be able to answer the many questions raised in varied domains, CONSTANCES must be based on a large sample. Keeping this in mind, as well as costs and different practical constraints for the HSCs, which are basically funded by CNAMTS for delivering free extensive clinical screening for persons affiliated to its health insurance and their dependents, we decided that the optimal size would be 200,000. In order to assess the potential of CONSTANCES in terms of its capacity to conduct epidemiologic studies likely to have good statistical power, we estimated the number of major health outcomes expected over different periods of follow-up in a 200,000 persons cohort with an age and sex structure identical to that of the French general population aged 18 to 69 years. Table 1 presents the number of expected events at the end of 5, 10, and 15 years of follow-up for outcomes for which we have reliable national reference data from different sources: deaths and incidence of cancer, ischemic heart disease, and Alzheimer disease.

Table 1 - Expected number of major health outcomes during follow-up of the CONSTANCES cohort

	5-year follow-up			10-year follow-up			15-year follow-up		
	Men	Women	Total	Men	Women	Total	Men	Women	Total
Death, all causes	4131	2133	6264	9727	5502	15 229	16 983	10 736	27 719
Incident cancers	3162	2220	5381	7036	4855	11 892	11 444	7823	19 267
Ischemic heart disease (35-64 years)	681	138	819	1418	290	1708	2178	452	2630
Alzheimer disease	265	240	505	793	1007	1800	1548	2469	4018

For these major outcomes, the number of these serious events is quite high. Power calculations show that for a number of cases ranging from 500 to 5,000 it will be possible to detect odds ratios ranging from 1.4 to 1.10 for comparison of top to bottom quartiles of a quantitative factor with a power of 0.80 and a significance level of 0.05. Thus, for frequent diseases, numerous studies are possible with satisfactory power, and reliable descriptive data can be produced. For more infrequent outcomes, power will however be sometimes insufficient[§]; this is one the reasons for the collaborations we are developing with other population-based cohorts in different countries in order to be able to pool data (see below).

Because the advantages of longitudinal follow-up increase with its duration and in view of the broad objectives for the cohort, the duration of follow-up should be as long as possible; CONSTANCES planned duration is therefore unlimited.

2.5 COHORT COMPOSITION, REPRESENTATIVENESS AND SELECTION EFFECTS

The source population is that of the people living in France affiliated to CNAMTS. One of the main study objectives being to provide information on the health status and disease burden of this large part of the French adult population, the CONSTANCES cohort was designed as a random sample representative for age (18 to 69 years at inception), sex, and social category (table 2).

Table 2. Expected age distribution of the sample at inception

Age group	N	N cumul	%	% cumul
18-19	7 245	7 245	3,6%	3,6%
20-24	22 916	30 161	11,5%	15,1%
25-29	22 934	53 094	11,5%	26,5%
30-34	22 480	75 574	11,2%	37,8%
35-39	22 123	97 697	11,1%	48,8%
40-44	21 192	118 889	10,6%	59,4%
45-49	19 811	138 700	9,9%	69,4%
50-54	18 823	157 523	9,4%	78,8%
55-59	18 666	176 189	9,3%	88,1%
60-64	13 385	189 574	6,7%	94,8%
65-69	10 426	200 000	5,2%	100,0%

Selection effects are one of the major sources of bias in epidemiologic surveys. They can bias estimates of disease prevalence or incidence (or of prevalence of exposure to a risk factor) and of associations between exposures and diseases of interest. In longitudinal cohorts, selection effects may occur at inclusion and throughout follow-up because of cohort attrition.

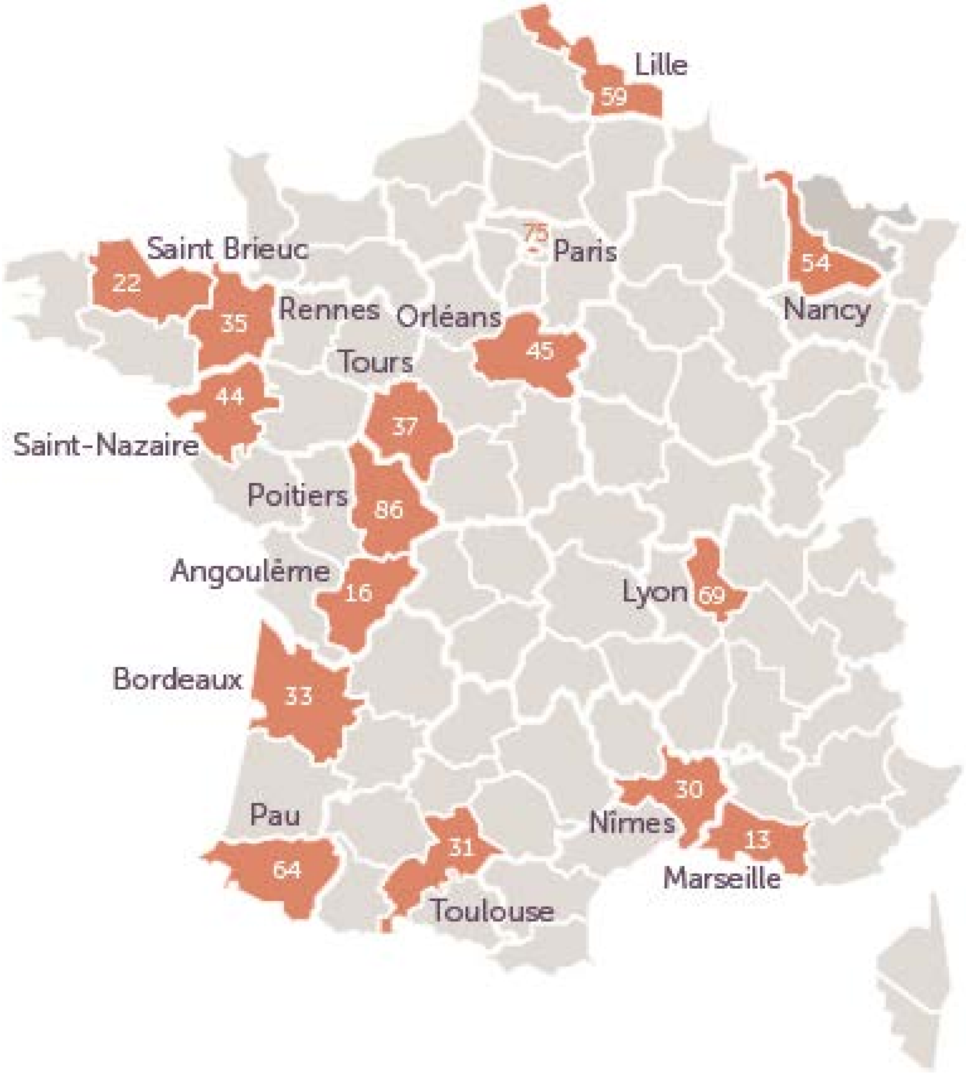
Regarding attrition, it is not possible at this stage to estimate precisely the number of subjects who will be lost to follow-up in the CONSTANCES cohort over the years. We can nonetheless make estimates based on our experience with the follow-up of GAZEL, which began in 1989 with more than 20,000 subjects. Active participation by self-administered questionnaires is high: after 25 years of follow-up, only 2.9% of the subjects who were part of the baseline study population never returned an annual questionnaire. The number of subjects truly lost to follow-up, that is, those we can no longer locate in the databases, is quite small: 107, or approximately 0.5%. It is reasonable to think that CONSTANCES, which will apply similar methods, will also have very high follow-up rates.

The problem of bias linked to selection effects is very different depending on whether the objectives are analytic or descriptive. In a cohort whose inclusion procedures are the same for all subjects (the case of CONSTANCES), in principle the exposure-disease relation does not differ between subjects who are included and those who are not. Therefore, the selection procedures at study inception for CONSTANCES participants should generate minimal bias, if any, in analytic studies. On the other hand, the problem of attrition during follow-up may cause substantial bias if the probability of continued follow-up is different in exposed and unexposed subjects or in those who do or do not become ill, which is often the case.

For descriptive studies of the frequency of health problems and exposures, the parameters of interest must be estimated in a representative sample of the target population. In this regard, the potential concerns for CONSTANCES are mainly incomplete geographical coverage of the districts of recruitment, and factors associated with voluntary participation.

As detailed below, the cohort participants are included in 17 Health Screening Centers (HSCs) located in 16 different districts (“*départements*”) in different regions of France (Figure 2).

Figure 2. Geographical recruitment of the CONSTANCES cohort



There are about 110 HSCs in France. In accordance with the CNAMTS representatives, **the participating HSCs were selected from them considering practical criteria**: their geographic distribution represents the principal regions of France, covering much of the heterogeneity regarding the distribution of risk factors and major diseases in the general population across the country; they have experience with the recruitment of large numbers of people and with participating in epidemiological studies; all are large, have a staff motivated to work in epidemiology, and use advanced medical equipment (including a biology laboratory), accepted that their personnel take a special training on CONSTANCES’ SOPs and that research assistants make on-site quality control monitoring visits. The number of CONSTANCES subjects is not equal for each HSC, and has been established accordingly to their size, considering that they have also to take in charge their regular patients (see below the sampling scheme).

We have verified that the structure of the population of the districts where the CONSTANCES HSCs are located is essentially identical to that for France as a whole for the principal demographic, social, economic and occupational characteristics; we should thus be able to generalize the CONSTANCES results to the French population as a whole (Table 3).

Table 3: Main demographic and socio-professional characteristics: comparison between all CONSTANCES districts and whole France (Source: Census)

	France	Constances
Age		
<20	24,57	23,96
20-29	13,48	14,97
30-39	14,66	14,78
40-49	14,47	14,19
50-59	11,49	11,42
60-69	9,37	9,01
70-79	7,96	7,74
80-89	3,14	3,09
90-99	0,86	0,83
Sex		
Men	48,56	48,16
Women	51,44	51,84
SES		
1- Agricultural worker	2,72	2,02
2-Self-employed.	6,62	6,44
3-Managers, teachers…	13,12	16,67
4- Professional & technical workers	23,07	24,07
5- Clerical workers	28,83	28,30
6-Blue collar workers	25,64	22,49
Economic sector		
1- Agriculture, hunting, forestry	4,00	3,19
2- Fishing	0,12	0,09
3- Mining and quarrying	0,20	0,16
4- Manufacturing	17,26	15,20
5- Electricity, gas and water	0,90	0,94
6-Construction	5,83	5,23
7- Wholesale and retail trade	13,24	13,07
8- Restaurants and hotels	3,54	3,57
9- Transport, storage and com	6,44	6,61
10- Financing, insurance	3,00	3,17
11- Real estate, business serv	11,20	12,93
12- Public administration	9,84	9,81
13-Education	7,30	7,98
14-Health and social action	11,56	11,91
15- Community and social serv	4,35	4,90
16- Personal services	1,12	1,10

Using volunteer subjects inevitably produces selection effects, even in studies that use random drawing from an appropriate sampling base, as it is the case of CONSTANCES. At inclusion, individuals may refuse participation (become non-participants), a potential source of bias. To compensate, researchers usually attempt to collect a minimum data set for the non-participants (mainly age, sex, and social category), to facilitate subsequent adjustments for estimating the relevant parameters. This approach nonetheless has some limitations. First, it is not always possible to collect these data for non-participating subjects. Nor is it always clear whether these data are sufficient to control for potential biases, because we know, for example, that within the same socioeconomic category there are many important differences in terms of health, lifestyles, or social networks. Finally, it is rarely possible to control completely for potential selection bias because it is rare to have the relevant data collected simultaneously for the participants and the non-participants. To obtain a representative sample of the target population and to minimize the bias associated with selection effects at inclusion and during follow-up in CONSTANCES, we took the following steps.

The sampling base at inclusion is composed of all persons aged 18 to 69 years and covered by CNAMTS in the catchment areas of the CONSTANCES HSCs. Sampling is done within the database of CNAV which includes exhaustively all the persons in France affiliated with the CNAMTS (including unemployed and foreigners). CNAV permanently collects professional and social data compulsory for establishing the retirement benefits of all individuals in France. The sampling scheme is stratified on districts where the 17 CONSTANCES HSCs are located in different regions of France and was designed to draw in each district a sample of the source population representative for age, sex, employment status and social category. To avoid an excessive burden for the participating HSCs, the size of the samples was defined so that in each district, the total number of CONSTANCES participants will not exceed 20% of the annual number of visits in the corresponding HSC. To take into account the fact that the participation rates to health surveys differ according to some socio-demographic characteristics, the random drawing is also stratified using unequal inclusion probabilities in relation to age, sex and social category, based on observed data from participation in previous surveys involving invitations to the HSC.

We are also setting up a “parallel cohort” from a random sample of 400,000 non-participants for whom we prospectively collect data on social and demographic characteristics (sex, age, employment status, social category, income), through the CNAV database, as well as information about health and health-care utilization from two other national databases covering the whole French population: SNIIRAM (the National Health Insurance Information System) and the National Death Registry managed by the *Centre d’épidémiologie des causes de décès* (CépiDc). As we have data from the SNIIRAM, CNAV and National Death Registry files for both the participants and the sample of non-participants, we are able to estimate the probabilities of participation in CONSTANCES with prediction models; the inverse of the probability of participation provide then an adjustment coefficient for each participant. This use of auxiliary data from administrative databases proved to effectively correct for nonresponse [Santin et al. 2014].

Regarding attrition, we can assume that almost none of the people included in CONSTANCES will be permanently lost to follow-up, since the participants will be followed “passively” (so-called because this follow-up will not require the subjects' participation) through the SNIIRAM, CNAV and National Death Registry files.

There will nonetheless be attrition due to the failure to return the annual questionnaire. Thus, adjustment for attrition is necessary if the analyses of the questionnaires variables are to be valid. Inclusion in CONSTANCES will take place over a 5-year period. For wave 1, we distinguish the participants who returned the self-administered questionnaire in year 2, that is, the year following their inclusion (which is year 1), from those who did not. We have the data collected at inclusion in CONSTANCES (year 1) for all participants as well as the SNIIRAM and CNAV data corresponding to year 1 or to year 0 (the year before inclusion). The coefficients of adjustment for attrition in year 2 can thus be calculated by a method similar to the one used to calculate the coefficient of adjustment for non-participation. For the following years, we again distinguish the participants who returned the self-administered questionnaire from those who did not. Thus, the longitudinal follow-up of participants in the SNIIR-AM and CNAV databases, whether or not they drop out of the cohort by not returning the annual questionnaire, makes it possible to update the coefficients of adjustment for attrition.

For descriptive purposes, each year we adjust the cohort data on the reference population. The first year, we randomly draw from the CNAV files a sample of CNAMTS members in the CONSTANCES districts aged 18 to 69 years. The second year, and respectively the third, fourth and fifth years, this sample of randomly drawn individuals will include people aged 18 to 70 years, then 18 to 71 years, 18 to 72 years and 18 to 73 years. Each of these samples is twice as large as that of the population of CONSTANCES' participants so that the reference population is significantly greater than the sample. After linking this file with the SNIIRAM files, we can calculate the relevant margins and thus, beyond socioeconomic and demographic characteristics, be able to integrate the variables relative to the health and healthcare utilization characteristics. The quality of weighting is therefore substantially improved by the calculation of margins specifically related to health, the focus of the CONSTANCES cohort.

2.6 PROCEDURES FOR INCLUSION

Randomly selected persons receive an invitation to come to their HSC for inclusion in the CONSTANCES cohort. The constitution of the cohort started in 2012, and we are proceeding gradually to include the entire cohort over a 5-year period.

At inclusion, several questionnaires are completed. Before coming to the HSC, participants have to complete two self-administered questionnaires: a "*Health and Lifestyle*" questionnaire, and a full "*Job history*" questionnaire; after coding of the occupations, the latter will be linked to the MATGENE job-exposure matrices developed by InVS [Févotte et al. 2011] to establish cumulative indices of exposure to various occupational hazards. In the HCS, the subjects complete an "*Occupational exposure*" questionnaire on past and current employment and working conditions; women complete also a specific "*Women's health*" questionnaire. The subjects undergo a medical examination, including the collection of biological samples for the biobank (details below) and the physician completes a "*Medical questionnaire*" on past and prevalent personal and familial diseases (questionnaires can be downloaded from the CONSTANCES website[§]). Finally, the participants have to sign an informed consent form to be included in the cohort.

We decided to use « classical » paper-and-pencil questionnaires for inclusion and not Internet-based questionnaires for several reasons. First, it is not possible to have the e-mail addresses of persons randomly selected from the CNAV database. Second, persons in France who use the internet for health purposes are distinctively different from the general population in regards to major cultural, social and behavioral characteristics [Renahy et al. 2010]. Additionally, participation rates are much lower when persons are initially asked to participate through an Internet questionnaire than by a letter with a paper questionnaire (a test during the field pilot of CONSTANCES gave very poor results). On the other hand, it may be that persons already included in the cohort respond well to Internet questionnaires during the follow-up [Touvier et al. 2010]: starting in 2015, we will thus offer to participants already included to complete their annual follow-up questionnaires through Internet. However, data collected through these different media may differ for several reasons: online control may indeed improve the quality of the information, but we may have problems in the comparability and standardization of data. It is also established that young persons are more familiar than older ones with the use of Internet, and that well-educated persons are better Internet users than less educated ones, and this could be also a source of selective non-response and undesired variability of the data obtained [Bowling 2005; Gmel 2000; Kwak & Radler 2002]; In addition, there are some practical issues and safety concerns involved in the application of Internet questionnaires in epidemiologic research [van Gelder et al. 2010]. We thus intend to test Internet follow-up questionnaires according to a specific protocol in order to assess the validity of such a medium and to compare Internet and paper-questionnaires respondents.

2.7 PROCEDURES FOR LONGITUDINAL FOLLOW-UP

An **annual self-administered follow-up questionnaire** (health conditions, lifestyle, life events, various health scales, working conditions, etc.) is sent to the subjects at home or by Internet according to the choice of the participants (see above). Post office procedures make it possible to obtain regular updates of participants' postal addresses. **Participants will also be invited to attend a medical examination every 5 years in their HSC**: based on the GAZEL experience, we expect that at least 50% will come again; participants having moved between two HSC visits will be invited to another HSC (there are about 110 HSCs spread all over France). Maximizing their personal participation rate is essential. Accordingly, regular contact with participants includes a website and a Cohort Newsletter, which presents results, ancillary studies, etc., sent every year to participants.

The subjects included in CONSTANCES are also **followed up passively for social and work-related events and health data by regular linkage with the national administrative databases.** The CNAV databases are essential for access to social and work-related data. CNAV regularly receives for its databases employers' annual reports, and information about periods of employment and unemployment from social welfare organizations (e.g., sick leave, maternity leave, unemployment, and diverse social benefits). Access to the SNIIRAM, which covers the entire French population, is an efficient method of obtaining information about health events. The SNIIRAM contains individual medical data from different sources, structured and coded in a standardized manner: reimbursement data (doctors and other health professionals visits, prescribed drugs and medical devices); so-called “long-term diseases” (serious diseases exempt from co-payments and user fees, coded according to the International Classification of Diseases 10th revision-ICD 10); hospital discharge records, including principal and associated diagnoses for each hospitalization, also coded according to ICD 10. Data on vital status are available through the CNAV database and causes of death through the CépiDc National Death Registry.

Procedures for the linkage to national databases

For the linkage of the cohort to the national databases we developed and tested complex data transfer procedures which have been approved by CNIL (the French legal authority for personal data privacy). We give here a short description of the French regulations regarding personal data privacy and of the procedures we accordingly designed.

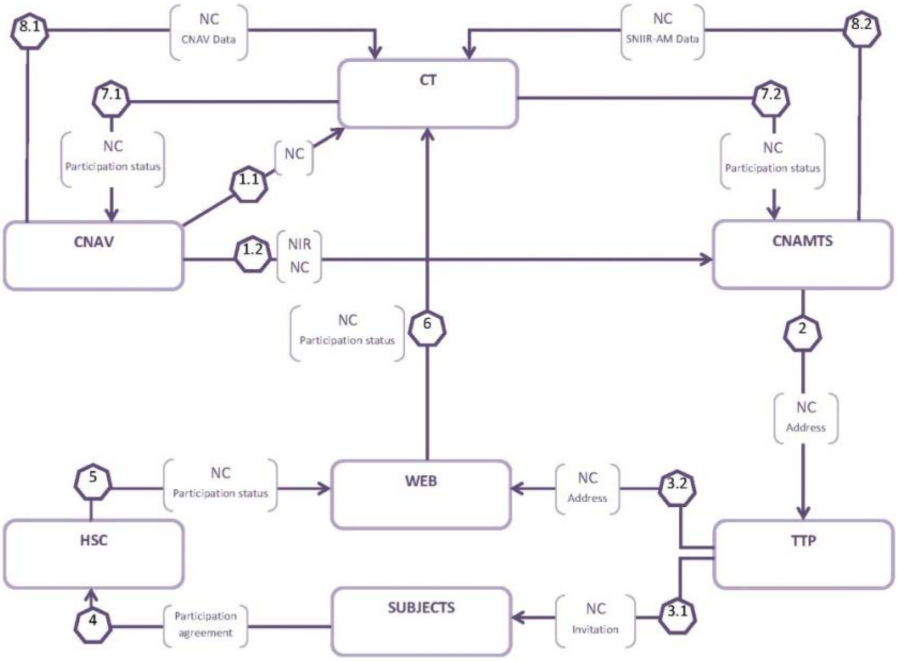
In France, every person has a unique national identification number given at birth or at immigration in France: the NIR ("*Numéro d'inscription au registre*"). This number is used by all administrations and employers to exchange personal information. NIR is also the personal identifier used in the national SNIIRAM and CNAV databases. To avoid the risk that data might be utilized without the agreement of the individuals the French law strictly regulates database linkage and particularly the use of the NIR; it is thus forbidden to directly collect the NIR from individuals in a survey. These legal restrictions concerning the use of the NIR for data access make usually impossible to link data from surveys with national health and socio-economic databases. Moreover, for security reasons, the SNIIRAM does not use directly the NIR as an identifier, but applies a hash technique using the "FOIN" (*Fonction d'Occultation des Informations Nominatives*) algorithm, which is based on the NIR, the sex and date of birth of the person yielding a non-reversible anonymous identifier, and making it impossible to go back to the nominative data from which it has been calculated.

For CONSTANCES it has been possible to overcome this problem, because the initial selection of the participants is made in the CNAV database: as CNAV and CNAMTS are allowed to use the NIR, they can exchange data between them using this identifier. We thus designed information transfer procedures which allow linking the cohort to the national databases thanks to the collaboration of CNAV and CNAMTS. Very schematically, these procedures follow several steps (figure 3)[10].

- CNAV draws a random sample according to the sampling scheme given by the CONSTANCES team (CT); a CONSTANCES study number (NC) is given to each selected person, and CNAV sends: (i) to CT a file where the subjects are identified by their NC (1.1); (ii) to CNAMTS a file with the correspondence between the NCs and the NIRs of the selected persons (1.2).
- CNAMTS updates the file with the name and postal address of the persons of the eligible sample and send it to a trusted third party (TTP) after suppressing the NIRs; TTP sends the invitations letters to the selected subjects (3.1) and generates a web application with a file containing names, addresses and NCs (3.2).
- Subjects who volunteer go to their HSC (4); the HSCs update the file of participants through the web application managed by TTP (5).
- Regularly, CT downloads participation state on the web application (6), draws a random sample of non-participants and sends to CNAV (7.1) and CNAMTS (7.2) the participation status of each person (identified by its NC).
- CNAV tags the participants and the subjects in the "parallel cohort" of non-participants, and erases all other non-participants from its files. Then CNAV extracts the needed data from its database using the NIR. CNAMTS generates from the NIR, plus sex and date of birth the "FOIN" anonymous identifier for the participants and the subjects in the "parallel cohort" of non-participants and erases all other non-participants from its files. Using the FOIN identifier CNAMTS extracts from the SNIIRAM the personal health data for both cohorts.
- Finally, CNAMTS and CNAV send the extracted data to CT using the NC identifier (8.1 and 8.2).

[10] We do not describe here the procedure for linking the cohort to the CépiDc database, which is regulated by a special decree.

Figure 3. Procedures for linking the cohort to the national databases



Indeed, the real data flow is much more complex as it has to manage situations such as withdrawals from the cohort, changes in address of people who move, geocoding of the addresses, numerous security procedures (data encryption, secured data transfer, traceability of the addresses, etc.)[§]. These procedures are fully operational and successful, with matching rates close to 100%.

While the linkage procedure itself works efficiently, there are some difficulties in using the data which are extracted, especially from the SNIIRAM database. These data are collected for administrative purposes and the database has a very complex architecture: it includes 7 dictionaries with 785,000 objects; the reimbursement claims are scattered over 12 tables including about 300 variables and there are 75 different lines on average for each person; the number of tables and variables of the hospital discharge records database changes every year (in 2007: about 25 tables and 370 variables; in 2008: about 30 tables / 450 variables). Numerous variables which are extracted form SNIIRAM are of no interest for our purposes, and selecting the useful information is a tedious process. Moreover, these unwieldy and very complex databases are permanently changing. We are working on a user-friendly computerized interface to allow an easy access to useful variables and we developed a series of tools such as an online data catalogue, or specific software for aggregating data in a usable form.

2.8 PRINCIPAL DATA COLLECTED FROM DIFFERENT SOURCES

Here we summarize the main data collected regularly for the whole cohort from different sources at each stage of the study; the full list of data is available in the Data Catalogue[§]. As already stated, the choice of the data was made after thorough discussions with researchers from many different fields, taking into consideration scientific criteria as well as practical and budget constraints: the HSCs are not equipped to perform some sophisticated investigations such as imaging or the collection of some complex biosamples, and their cost for a cohort of 200,000 subjects would be unrealistic. We thus restricted the parameters to be systematically assessed for the whole cohort to “basic” high-quality measurements in a wide range of domains, along the lines of a general-purpose cohort. Depending on separate funding opportunities, we intend at a later stage of the project to implement new measurements, such as collecting complex biological samples (RNA, faeces, etc.), inviting subsamples of the subjects to a hospital facility for magnetic resonance imaging, providing specialized devices such as accelerometers, or sending extra questionnaires, etc.

However, **additional data can be collected at any time for subsets of cohort members, according to specific research protocols**. For instance, some of the nested ancillary projects listed in the second part of this report need a complementary collection of data in addition to the data provided by the regular CONSTANCES database. In such cases, it is the responsibility of the investigators to find the funding and to manage for this collection of additional data through specific procedures.

For each CONSTANCES’ participant, data are collected from different sources: medical examination, national health and social databases, and questionnaires. We selected when possible variables already used in other surveys because they are validated measures; another reason is data sharing with other cohorts, which is much easier when the same variables are used in the different datasets [Fortier et al. 2011]. Whenever it was possible and pertinent, we also used scales already published in the literature, for which the psychometric properties are already established. Regarding occupational factors, we will in addition to questionnaire data, link the cohort with job-exposure matrices (JEMs) developed by InVS; currently available JEMS cover pesticides, chlorinated, petroleum and oxygenated solvents, asbestos, man-made mineral fibers, fuel, cement, leather, flour and wood dust, silica, formaldehyde; other JEMs are under development.

Health, personal and environmental data

Table 4 summarizes the main data collected for each cohort participant at different sources.

Table 4: Main data regularly collected for each cohort participants

DATA	SOURCES*
Social and demographic characteristics: social position, educational and income level, employment and marital status, geographic origin, household composition, socioeconomic status of parents and spouse, and material living conditions (type of housing, household income and wealth, etc.), including geocoding of successive residency addresses.	Q; CNAV
Health: personal and family history: cancer, cardiovascular, psychiatric; self-reported health scales: perceived health, quality of life (SF-12 [SF-36 Org]), mental health (CES-D [Fuhrer & Rouillon 1989], GHQ-12 [Goldberg 1979]**), and specific scales for cardiovascular, musculoskeletal, and respiratory diseases; incident and prevalent diseases: from self-reports, social security long-term diseases and hospital discharge (ICD-10 codes); sick leaves, handicaps, limitations, disabilities and injuries and healthcare utilization and management (visits to professionals, drugs and other prescriptions); date and cause of death; HSC examination (details below).	Q; CNAMTS; HSC; CépiDc
Lifestyle: smoking and alcohol consumption (past and present), dietary habits and physical activity, marijuana use, sexual orientation.	Q
Occupational factors: job history; current job title and employment status; lifelong and current occupational exposure to chemical, physical, and biological agents; postural, mechanical and organizational constraints; stress at work (job content questionnaire-JCQ [Karasek & Theorell 1990], effort-reward imbalance-ERI scales [Siegrist 2002])**	Q; MATGENE JEMs

Physical and cognitive functioning (45 years and older): evaluation of functional capacities: IADL (Instrumental Activities of Daily Living) scale [Katz et al. 1963], ability to use new technologies, and CASP (Control, Autonomy, Self-realization and Pleasure [Wiggins et al. 2004], a quality of life scale particularly appropriate for senior citizens); work-up of tests in HSCs (details below).	Q; HSC

*Q: questionnaire; HSC: HSC health examination; CNAV: CNAV socio-professional database; CNAMTS: SNIIRAM health database; MAGENE: InVS Job-exposure matrices; CépiDc: National Death Register
**These scales will be cyclically included in the follow-up questionnaires according to a planned calendar

Examination program in the HSCs: a full work-up of tests will be performed in the HSCs according to specific SOPs developed for CONSTANCES. The examinations are the following.

- Anthropometry: weight, height, waist-hip ratio.
- Visual acuity: Monoyer and Parinaud scales.
- Hearing: pure tone threshold audiometry for air conduction threshold rising in soundproof booth (500 – 1000 – 2000 – 4000 –-8000 Hertz).
- Spirometry: FEV and FVC (3 curves).
- Electrocardiogram (ECG): 12-lead resting ECG; recording of the curves.
- Blood pressure**:** one measure on each arm (2 mn spacing after a 5 mn rest) and one measure on the reference arm after a one mn rest; search of orthostatic hypotension for the subjects suffering from diabetes or/and aged 65 years and more.
- Biology: Blood: Glucose, Creatinine, Gamma GT, ALT, Total Cholesterol, HDL-Cholesterol, Triglycerides, White Blood Cell, Red Blood Cells, Hemoglobin, Mean Corpuscular Volume, Hematocrit, Platelets, Neutrophils, Eosinophils, Basophils, Lymphocytes, Monocytes, Leukocytes. Urine: Protein, Glucose, Nitrites, Microalbumin, Creatinine.

For subjects **45 years and older**, there are some additional tests.

Cognitive function: The cognitive test battery consists of tests for global cognitive status, the Mini Mental State Examination (MMSE) [Folstein et al. 1975] and the Digit Symbol Substitution Test (DSST), a subtest of the Wechsler Adult Intelligence Scale-Revised [Wechsler 1981], a timed paper- and pencil- task that measures psychomotor speed, sustained attention and logical reasoning, the Free and Cued Selective Reminding Test with Immediate Recall (FCSRT-IR), a measure of memory under conditions that control attention and cognitive processing in order to obtain an assessment of memory unconfounded by normal age-related changes in cognition [Grober et al. 2009]. Further tests of executive function include the Trail Making Test (TMT A& B) [Gaudino et al. 1995], and two tests of verbal fluency [Borkowski et al. 1967] where participants have 1 minute to recall as many animal words (semantic fluency) and then as many words as they know starting with the letter “f”.

Physical function: The battery of physical functioning administered consists of measures of upper and lower body function [Guralnik et al. 1994]. Lower body function is being assessed using measures of balance (Standing Balance Test), gait (Walking Speed) and upper body function using the Handgrip Strength Test and the Finger-Tapping Test (FTT) for psychomotor speed.

The total duration of the complete work-up is summarized in table 5. The median duration for the subjects under 45 is about 95 minutes (one and a half hour) and about 135 minutes (two and a quarter hour) for the 45 years and older who benefit from additional cognitive and physical tests.

Table 5 - Median duration of the HSC examination (minutes)

Administrative process	15
Biological samples collection	5
Occupational exposure interview	15
Doctor's examination	25
Tests (under 45 year-old)	35
Cognitive and physical (45 year and above)	40

2.9 BIOBANK

Designing the biobank associated with a long-term general-purpose large-size cohort is a complex challenge, as it must cover scientific needs in very different domains (cancer, cardiovascular diseases, psychiatry, aging, etc.), and be able to anticipate future research questions and the evolution of analytical and storage techniques. The biobank must also offer a high quality storage and long-term integrity. Another major challenge is the harmonization of the collection and storage methods in view of sharing of biological resources as large-scale data pooling will be crucial in the future.

Building on our experience with the biobank of the GAZEL Cohort Study [Zins et al. 2003], currently stored in the Dijon Centre of Biological Resources (CRB), we carefully prepared the design of the CONSTANCES biobank by following several steps. First, we consulted extensively with the scientific community potentially interested in using data from CONSTANCES to determine their future needs. We visited several large biobanks and we discussed the strengths and limits of different technical solutions based on their experience. We also consulted with companies manufacturing the various equipment needed for the biobank. Second, we studied the different available alternatives for each procedure and component of the biobank: containers, specimen identification, shipping of samples to the biorepository, aliquoting, storage technique, specimen handling and retrieval, quality control follow-up of the samples, biobank information system. For each option, we analyzed its technical features, feasibility, and cost in regard of the optimal functioning of the biobank; based on that work, we published guidelines for setting up a biobank associated to population-based surveys [Henny et al; 2012]. Finally, taking into account the large size of the cohort and the corresponding cost, we came to a compromise between an "ideal" biobank and a realistic project offering access to "basic" material for the whole cohort to a large diversified scientific community with the greatest possible choice of future laboratory analyses, while being able to take in charge more sophisticated needs for subsamples of the cohort, provided additional funding. Due to budget consideration, we will take samples on one half of the full cohort (100,000 participants), starting in 2016. Here, we summarize the main features of the CONSTANCES biobank.

Objectives

The conservation of biological samples is challenging. A biobank must achieve two main goals: ensure the integrity of biological samples and offer greater opportunities for prospective users.

1) **Ensuring the integrity of biological samples from sampling to final use**: artefacts due to cell lysis, cell metabolism and enzymatic degradation must be avoided (minimized) in the pre-storage phase. Artefacts during the pre-storage phase will be limited by separating the

serum and plasma in less than an hour after collection, biological fluids will be divided into aliquots of small volume (0.5 mL for serum and plasma, 1.0 mL urine) to avoid freeze-thaw cycles. The DNA will be extracted on demand from buffy-coat for economic reasons. The fragmentation of buffy coat will be automated if the technologies currently in development are recognized as reliable. A centralized aliquoting was selected for the basic program to improve efficiency, robotize all stages, ensure better traceability and reduce costs. Finally we will pay particular attention that sampling conditions, treatment for each sampling site, aliquoting and storage are fully standardized and identical for each specimen for the duration of the study; this will be supported by full traceability.

2) **Providing the most extensive opportunities to research teams**: as science and technology are constantly improving, it is impossible to predict all future uses of the bio-specimen. We however tried to cover broad scientific domains, and the basic program opens good prospects: we will collect **serum**, **plasma**, **whole blood, buffy coat** and **urine**, allowing for proteins, hormones, nutrients, genomics, epigenetics and biomarkers studies. This basic program can be completed according to the demands and needs, for specific research programs that require special care during the pre-storage phase, such as blood collection on ice, immediate pre-treatment, immediate freezing, isolation of mononuclear cells, collection of faeces or of other biological materials, etc.

Sample collection

Type of samples: We plan to collect biological samples (blood and urine) for 100,000 participants during a visit in a HSC. For the blood we shall store for every person 8 aliquots (0.5 mL): 2 EDTA plasma aliquots, 2 Heparinate Lithium plasma aliquots, 2 serum aliquots, 1 whole blood and/or RBC aliquot, 1 buffy-coat aliquot. For urine we shall store 2 aliquots (1.0 mL). In all, about 1,000,000 of aliquots will be stored. In complement to this basic storage program, we shall offer optional programs such as total RNA, proteins, mononuclear cells, saliva, hairs, and nails and possibly stool samples for specific research projects on subsets of population.

Standardization of samples collection: All steps of the process of the biological samples collection are standardized through specific SOPs: collection of blood and urine (fasting subjects), standardized preanalytical phase and pre-treatment of the samples in each recruitment center within 30 mn after the collection, transport from recruitment centers into the central laboratory of the biobank within 24 hours at a temperature of 4-8°C, in respect for the legal and regulatory rules. Data concerning the quality of every sample will be recorded in the biobank data base.

Biosamples storage

On receipt of the biological samples on the site of the biobank, the integrity of samples will be verified. Robotized liquid handling systems will aliquot every biological type of sample in different cryocontainers (0.5 mL, identified by a linear and/or 2D barcode system). The whole blood will be fragmented by a robotized device for extracting buffy-coat which will be aliquoted.

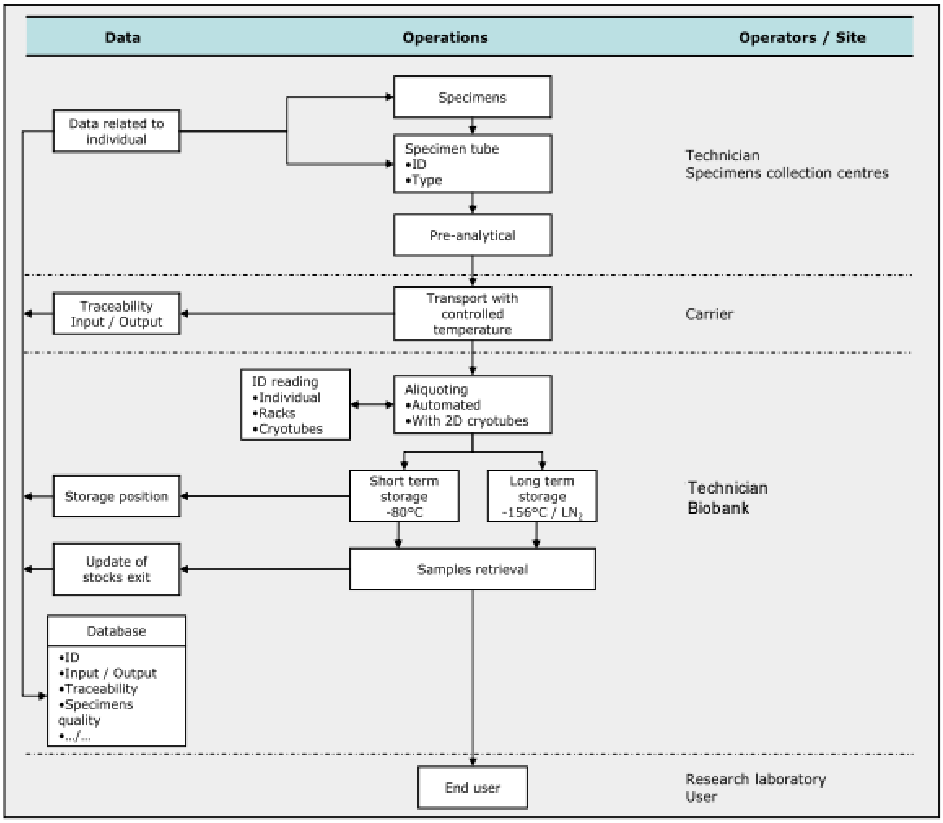
Then each aliquot will be transferred in the biorepository. A first half will be stored in –80°C deep freezers equipped of security systems (temperature registration and alarm, back-up freezers) for short and middle term preservation, and will constitute the “active” or “working” biobank). A second half will be stored in liquid nitrogen vessels in vapor phase, as an “archive” biobank for the long-term storage and as back-up. For safety reasons –80°C freezers and liquid nitrogen vessels will be located on two different sites.

Biobank Information System (BIS)

The BIS will work as a management system of laboratory (in- and output of specimen management, location and traceability of each sample, management and monitoring of the biobank and quality assurance scheme). A backup storage of the data and security systems insuring the protection of the data and of the power supply will be installed. The identification of the specimen within the biobank database will rely on ID numbers different from those used for the central database of the cohort; and a specific password-secured interface will allow for linkage of biological and other data for the same individuals. Finally, the interoperability of the CONSTANCES database will facilitate its integration in national, European and international biobank networks.

The general process of the CONSTANCES biobank is summarized in figure 4.

Figure 4. Flowchart of the biobank general process


Data
Operations
Operators / Site
Specimens
Data related to individual
Specimen tube •ID •Type
Technician Specimens collection centres
Pre-analytical
Traceability Input / Output
Transport with controlled temperature
Carrier
ID reading •Individual •Racks •Cryotubes
Aliquoting •Automated •With 2D cryotubes
Storage position
Short term storage -80°C
Long term storage -156°C / LN2
Technician Biobank
Update of stocks exit
Samples retrieval
Database
•ID •Input / Output •Traceability •Specimens quality •.../...
End user
Research laboratory User

Quality assurance

A quality assurance scheme will cover all the current functions and/or steps of the biobank from the collection of samples to the shipping of specimens to the end user, including: staff training, SOPs, quality control of the biological specimens, traceability of all the operations, maintenance of the infrastructure and equipment, safety of the building, certification and possibly accreditation. Regular training of the staff will be organized.

The CONSTANCES biobank will follow the “OECD Best Practices Guidelines for Biological Resources Centers" (OECD) and the French standard NF 96-600.

Official and professional bodies’ clearance

The CONSTANCES biobank will join the French BIOBANQUES Infrastructure and participate to the "Biobanking and Biomolecular Resources Research Infrastructure" (BBMRI) European network of biobanks.

Location of the biobank

At this stage of the preparation of the CONSTANCES project, a final decision is still to be made regarding the location of the biobank component of the cohort. For security and practical reasons (the storage of some 1,000,000 aliquots requires a large space with the required technical equipment, which is not readily available), we would like to split the biorepository in two parts: the "working" biobank and the "archive" part being located in different places; we plan to make in 2015 a call for proposals, in accordance with the BIOBANQUE Infrastructure, for selecting an appropriate CRB facility.

2.10 PERIODICITY OF FOLLOW-UP

The periodicity of follow-up will vary according to the sources.

The self-administered follow-up mail questionnaire is sent annually. This is an important feature of CONSTANCES, since the socioeconomic and health national databases do not contain any data on risk factors associated with lifestyle or major life events and do not allow exploration of changes in lifestyle and other risk factors over time. Personal questionnaires are also the only way to collect important data, such as quality of life, sleep, pain or mental health scales for instance. Previous experience with annual questionnaires in GAZEL showed the usefulness of this study design when the aim is to capture temporal changes in relation to an exposure. As an example, a series of articles published in high-impact journals showed that changes in perceived health, sleeping disturbances, depressive mood, fatigue, headache, driving behaviors or alcohol consumption associated with retirement are transient phenomena and would not have been fully captured with wider intervals between successive observations [Bhatti et al. 2008; Vahtera et al. 2009; Westerlund et al. 2009; Westerlund et al. 2010; Sjösten et al. 2011; Zins et al. 2011], and in fact, some of the ancillary research projects listed in the second part of this report basically rely on frequently repeated measures.

Another reason for having an annual follow-up questionnaire is allowing collecting numerous data without asking subjects for too much work which would discourage participation. At the same time, it will establish a sense of loyalty in the participants, as too long a delay between two questionnaires is a factor that promotes dropping out. Sending an annual questionnaire also facilitates rapid response for setting up new studies, as it makes it possible to investigate new specific topics each year. Finally an annual questionnaire is beneficial both in terms of high follow-up rates and excellent science.

Some questionnaire data are collected annually (health status and reported morbidity, life events and characteristics of place of residence, smoking, alcohol, etc.), while others are collected at longer intervals, according to a planned calendar (health scales and questionnaires for a specific health area or specific risk factors). The mailing of self-administered questionnaires takes place twice a year (May with a reminder in October; November with a reminder in April) to take seasonal variations into account, since these factors are important for some topics (in particular, morbidity, drug use, working and environmental conditions).

Because the **national databases** essentially record events continuously, the follow-up of the data they provide is permanent; for practical reasons, linkage to the national databases will take place once a year.

Participants are also asked to **visit the HSC** every 5 years for medical and laboratory examinations, and additional biological samples will be collected for the biobank.

2.11 QUALITY CONTROL AND VALIDATION OF HEALTH OUTCOMES

Data from questionnaires are computerized using an automated system including dedicated software that has numerous validity controls; any problem is addressed before it is integrated in the database. Percentage of non-response, missing data and delay in return are systematically monitored. Completion and coherence of data extracted from computerized data bases are systematically checked before their integration in the CONSTANCES database.

For the data collected during the inclusion visit to the HSC, we established detailed Standard Operational Procedures (SOPs) for each of the examinations, including biological and the physical and cognitive functioning tests[§]. Routine permanent quality control, based on regular on-site inspections by epidemiologic research assistants is performed, including monitoring of equipment used for the examinations, training of personnel, control of data completeness and validity from random samples of participants, etc. We also plan to perform reliability studies by repeating the test on subsamples of voluntary participants in order to assess within-subjects variability. These quality control procedures allow to assess the accuracy, reproducibility, concordance, and internal and external validity of the data collected and to study their factors of variability. Quality control procedures are taken in charge by independent organizations: ClinSearch for the data collected during the examination in the HSCs, and by Asqualab and EuroCell for the biological data[§].

One of the major challenges in following-up a population-based cohort is the identification of prevalent and incident diseases among the participants. In order to provide a high quality of phenotyping, particular attention is paid to ascertainment of health outcomes. We rely on different sources to track prevalent and incident diseases: self-report through questionnaires, medical examination in the HSCs, diagnoses extracted from the hospital discharge and health-related administrative national databases. However, while administrative data used for reimbursement of expenses (visits to health professionals, drug prescription, hospital stays…) are of high quality, the medical data (detailed diagnosis of diseases) are not formally validated before being recorded in SNIIRAM, and none of above cited source contains fully ascertained diagnoses. In order to provide a high quality of phenotyping, potential outcomes will be systematically verified. We obtained the authorization of CNIL to have a direct access to the hospitals or the general practitioners and other attending physicians, and each suspected outcome reported in the available sources will be routinely reviewed and validated by specialized adjudication committees. In a first phase, we particularly focus on some major chronic diseases: cardiovascular events (myocardial infarction, stroke, atrial fibrillation, heart failure, diseases of the heart valves, sudden death), cancers and neurodegenerative diseases. Each suspected outcome reported in the available sources is routinely verified at the hospitals or with the general practitioners and other attending physicians and validated by specialized adjudication committees. A "*Validation platform*" initially developed for the GAZEL Cohort Study was further enhanced in view of the large number of potential outcomes to ascertain (detailed procedures[§]).

2.12 LEGAL REQUIREMENTS

According to the French regulations, the CONSTANCES Cohort project has obtained the authorization of the National Data Protection Authority (*Commission Nationale de L'informatique et des Libertés*-CNIL). CNIL verified that before inclusion, clear information is provided to the eligible subjects (presentation of CONSTANCES, type of data to be collected, ability to refuse to participate, informed consent, etc.). Concrete procedures for setting up the two cohorts (participants and non-participants) ensure the confidentiality of the data at every point in its circulation as well as the anonymity of the cohort of non-participants. In addition, CONSTANCES was approved by the National Council for Statistical Information (*Conseil National de l'Information Statistique*-CNIS), the National Medical Council (*Conseil National de l'Ordre des Médecins*-CNOM), and the Institutional Review Board of the National Institute for Medical Research-INSERM[§].

2.13 PILOT

A field pilot of CONSTANCES took place in seven of the participating HSCs from May 2009 to May 2010 for a four- to five-month period in each center, including surveys on a random sample of non-participants and on a sample of participants who attended a HSC. About 3,700 subjects were included (women and men in almost equal numbers), and the preliminary analysis of the data showed that this sample was close to the general population of adults in France regarding sex, age and socioeconomic status. There was quite a diverse distribution of occupations, working conditions, and lifestyle factors; prevalence rates of various diseases and symptoms were close to those from other available French surveys, and globally, the quality of the health examination in the HSCs was satisfactory, in spite of some concern about the variation in the results as a function of the centers; in the post pilot phase several actions has been taken to rectify these issues: identifying the tests with problems and modifying the protocol to make the administration of the tests as standard as possible across the centers; a video has been prepared to allow easier understanding of the SOPs of the functioning tests, and the protocol of the Trail Making test A & B has also been modified due to variability between centers; we now follow the protocol set out in a recent publication in Nature Protocols [Bowie & Harvey 2006].

We also verified the **linkage with the CNAV and SNIIRAM databases**. Regarding the CNAV database, since the eligible sample was drawn from the CNAV database, the matching rate was by design of 100%; among the persons who were invited to participate, a full data record was retrieved for 96% of the subjects (participants and non-participants among which the parallel cohort of non-participants is randomly drawn). Regarding the SNIIRAM database, the matching rate for those who were randomly drawn from the CNAV database was 93.7% for the year before inclusion; one should notice that for a given year the SNIIRAM database records only operations that happened during this year. Thus it does not mean that the 6.3% of subjects who were not matched were lost, as they are likely to be found in the database in subsequent years.

Finally, the field pilot showed that the procedures for invitation of selected subjects, for data collection and medical examination and for linkage with the national databases were quite satisfactory; only some marginal adaptations of the protocol were necessary.

2.14 COMPANION COHORTS: THE COSET PROGRAM

The CONSTANCES cohort will include only persons insured by CNAMTS, *i.e.* salaried workers and their dependents, thus excluding agricultural workers, farmers and other self-employed workers from the eligible population. This is due to the French social security system. While health insurance is compulsory in France, persons are affiliated to specific funds depending on their employment status: the General Health Insurance Fund administered by CNAMTS for salaried workers and their family, the "*Mutualité sociale agricole*" (MSA) for agricultural workers and farmers, and the "*Régime social des indépendants*" (RSI) for self-employed workers; altogether, the three funds are covering almost all the French population. As the CONSTANCES cohort project is conducted in partnership with the General Health Insurance Fund, it cannot include those affiliated to the two other health insurance funds as they are not eligible for free medical examinations in the HSCs which belong to CNAMTS, and occupational and lifestyle factors specific to agricultural workers and farmers and to other self-employed workers will thus not be available.

This is why, in order to further enhance its coverage of specific populations and data, CONSTANCES is closely associated with another project currently in progress, the COSET Program (*"Cohortes santé et travail"*) developed by the Occupational Health Department of InVS (www.coset.fr). The COSET Program is composed of three cohorts: CONSTANCES and two other cohorts, the MSA and RSI Cohorts, which will include subjects affiliated to these funds and were designed to be closely coordinated with CONSTANCES. The MSA and RSI cohorts are currently in development under the responsibility of InVS. They will each include a nationally representative sample of 30,000 randomly selected participants: agricultural workers and farmers for the MSA Cohort, and other self-employed workers for the RSI Cohort. Data will be collected from the participants through mailed self-questionnaires at inception and during follow-up, and from linkage to the national databases to take into account non-participation at inclusion and attrition throughout the longitudinal follow-up. However, data do not include the CONSTANCES medical examination in the HSCs.

UMS 011 and InVS groups collaborated during the development of the protocols of the three cohorts for the design of the samples and the definition of the data to be collected on occupational factors. During follow-up, the same data on social and work-related events and health data than for CONSTANCES will be permanently collected from the MSA and RSI databases, and cohorts of non-participants will also be set up to take into account non-participation and attrition throughout the follow-up. Data collected for the MSA and RSI cohorts will be transmitted to the CONSTANCES database, allowing thus to cover all the French population and to increase the overall statistical power; data from CONSTANCES will be transmitted to InVS for occupational health surveillance purposes.

Finally, the collaboration with the COSET Program will allow for representative samples covering the whole French population. For some studies, pooling of data from the three cohorts will also increase the size of the study population to a total of 260,000 subjects.

2.15 STRENGTHS AND LIMITS

CONSTANCES has several strengths. It was designed both to help answer research questions in diverse areas and to provide public health information needed by the health authorities. To facilitate this goal, we designed for CONSTANCES and the associated MSA and RSI cohorts a complex sampling scheme including non-participants, and we developed original methods based on sophisticated statistical procedures to take into account selection effects at inception as well as during the follow-up.

When fully completed, the sample will be large (200,000 or 260,000 according to research objectives), including persons living and working in diverse settings, from large cities to small villages in different regions of France, with a broad range of socioeconomic statuses and trades, allowing the study of specific economic and professional categories, such as teachers, hospital workers, students or agricultural workers with a satisfactory power for many analyses.

The follow-up is very extensive, relying both on active participation of the volunteers through annual questionnaires and regular visits to the HSCs, and on passive methods through regular linkage to health and socioeconomic national exhaustive databases, one of the most innovative features of CONSTANCES. The latter methods are possible because we managed to use the national identification number of the participants and to develop secured procedures which were approved by the French national authority for personal data protection. Linkage to these databases also permits the study of comorbidities and ensures that there will be almost no lost to follow-up.

Numerous data are collected, including a comprehensive medical, physiological and biological examination and a large biobank. Compared to most of the existing or under development cohorts in other countries, CONSTANCES collects very detailed data on personal, lifestyle, environmental, lifestyle, occupational and social factors using state-of-the-art methods, making it possible to assess gene-environment interactions in an optimal way.

Specific efforts (validated SOPs, strict quality control procedures) are put into the quality of data collection and the validation of main outcomes in order to provide a highly phenotyped cohort; in this regard, the participation to CONSTANCES of the HSCs, which have a long experience in participating to various epidemiological surveys gives an additional strength.

A unique feature of CONSTANCES is also a comprehensive set of cognitive and physical tests starting as young as 45 years, which is earlier in the life course than most available studies on aging.

Of particular importance is the high frequency of measurements from many different sources, allowing for analyses of life course trajectories of health in relation to personal, social, occupational factors and major life events. In this regard, annual repeated measurements through self-administered questionnaires all along the longitudinal follow-up of the participants is a unique feature of the project, allowing for fruitful insight into temporal changes in health or lifestyle associated with various personal, social or health events, as demonstrated by the GAZEL Cohort Study experience.

Finally, the project is managed by an experienced team, with of more than 25 years of expertise in successfully designing, implementing and maintaining GAZEL, a large population-based prospective cohort, and in developing numerous fruitful collaborations with French and international groups.

The project also has some limitations. Due to the voluntary participation of cohort members, there will probably be an underrepresentation of hard-to-reach subjects, such as heavy drinkers or socially excluded persons. Comparisons between participants and non-participants at inclusion and during the follow-up through relying on the randomly selected "non-participants cohort" should allow assessment of potential biases due to selection effects and correction for non-participation.

However, lack of sufficient numbers in some categories might be a problem. Even more importantly, CONSTANCES and the MSA and RSI associated cohorts will not offer sufficient power to study rare outcomes or exposures despite their large size. Simulations[§] showed that in most of the situations where the relative risk is below 2, especially when interactions have to be taken into account, power will be satisfactory after at least 5 years of follow-up for situations when the incidence of the outcome is over 10/100,000 and the prevalence of exposure over 10%.

This limit is common to all longitudinal cohorts. We plan to pool data with the forthcoming *German National Cohort*. We participate to the BBMRI-LPC (Biobanking and Biomolecular Resources Research Infrastructure-Large Prospective Cohorts) Consortium[11], associating the main population-based cohorts in Europe aimed at promoting transnational access to large prospective cohorts. The objective is to increase the standardization and collaboration between these various prospective study resources to meet goals of increasing the sample size and statistical power for multi-factorial risk analyses of infrequent outcomes.

[11] http://www.bbmri-lpc.org/

3 DRIVING SCIENTIFIC PROJECTS

3.1 A WIDE RANGE OF SCIENTIFIC OBJECTIVES

CONSTANCES was designed as a general-purpose population-based cohort, open to the scientific national and international public health and research communities. In itself, it basically does not have specific scientific objectives: it is rather intended to be a research infrastructure supporting many diverse studies that have their own objectives.

3.2 PROCEDURES TO OPEN THE INFRASTRUCTURE FOR ANCILLARY NESTED PROJECTS

CONSTANCES Charter

Every group in France or in other countries, public or private, is entitled to apply to develop a nested project within CONSTANCES and to access to its database, including biological specimens. The selection of projects by the CONSTANCES Scientific Committee rely on scientific quality and feasibility criteria. Detailed rules have been established for using the CONSTANCES infrastructure, regarding legal aspects; data confidentiality and security; ethics; access to the database in the case where only available data are required or when the collection of supplementary data directly from the cohort participants is needed, as well as sharing of these supplementary data; access to the biological and genetic material; responsibilities of the CONSTANCES infrastructure and of external groups; dissemination of data and results; publications and authorship; acknowledgments; follow-up of the project and funding. In order to explicitly and clearly define these rules, we prepared a Charter[§] which was approved by the governance bodies of CONSTANCES.

Project proposals

An application form[§] has to be submitted, including: details on the PI and the teams involved in the project; state of the art, background and objectives; methods (including a detailed description of the data needed and procedures to insure data confidentiality); expected results and their dissemination; duration and schedule of the project; legal authorizations; budget. If the project is accepted by the governance bodies of CONSTANCES, the PI has to sign a commitment to respect the CONSTANCES Charter.

Tools for helping in the preparation and conduct of ancillary projects

The CONSTANCES database is quite complex and it could be difficult for external researchers to get familiar with it. In order to facilitate the preparation of nested ancillary projects, we developed several tools.

The CONSTANCES website gives a free open access to a full documentation useful for external researchers: protocols, questionnaires, SOPS, and the data catalogue[§] where it is possible to browse the full list of available data, including biological materials. We are also working on a highly secured interface that would permit direct access to the main database for those who are already conducting an ancillary study.

The CONSTANCES website will include a *Frequently Asked Questions* section and a hotline for technical support in case of difficulties. As proposals for ancillary projects have to be sent to the CONSTANCES team in the first place, previous informal exchanges can take place if necessary for helping in the preparation of the study. Once a year, an open CONSTANCES Scientific Symposium is organized, where the PIs of the ancillary projects present the advancement of their work, and share their experience and the problems they possibly met while conducting their study within CONSTANCES. This workshop is open to the scientific and public health communities, thus contributing to the dissemination of results and to the establishment of new scientific collaborations.

3.3 THE FIRST ANCILLARY STUDIES

A first call for proposals was launched in early 2014 among a restricted set of French investigators who collaborated in the preparation of the protocol of CONSTANCES. Almost 40 projects covering a wide range of topics were proposed and examined by the Scientific Committee.

Starting in 2015, a public call for ancillary projects proposals will be launched every year.

4 ORGANIZATION AND GOVERNANCE

4.1 THE CONSTANCES COORDINATING TEAM: THE POPULATION-BASED EPIDEMIOLOGICAL COHORTS UNIT-UMS 011

The CONSTANCES Cohort is intended to be a centralized infrastructure, under the scientific and technical responsibility of the “Population-Based Epidemiological Cohorts Unit” (UMS 011), which was recently created by INSERM and Versailles-Saint Quentin en Yvelines University (UVSQ) for conducting and managing CONSTANCES as well as other population-based cohorts open to the scientific French and international community.

Marie Zins, MD, PhD, is the Director of UMS 011, and is the PI of CONSTANCES project; Marcel Goldberg, MD, PhD act as a co-PI, with the help of Lisa Berkman, PhD. Both Marie Zins and Marcel Goldberg are epidemiologists, mainly working in occupational and social epidemiology; altogether, they have published some 300 peer-reviewed articles. They have a long experience in population-based epidemiological studies and in designing, implementing and maintaining large databases and prospective cohorts. UMS 011 has a good knowledge of the main French national health and socioeconomic SNIIRAM, CNAV and CépiDc databases which cover the entire French population. Marie Zins and Marcel Goldberg are also the co-PIs of the GAZEL Cohort Study designed and maintained by their group for more than 25 years (www.gazel.inserm.fr).

UMS 011 is composed of epidemiologists and medical doctors, biologists, biostatisticians, and database and telecommunication specialists, and it benefits from clerical and accounting support; currently, about 50 persons work in the Group, and there are plans to recruit additional personnel. In addition, UMS 011 subcontracts to private companies for specific aspects of the project: *ClinSearch* (www.clinsearch.net), a contract research organization (CRO) familiar with multicenter epidemiological studies, is in charge of the quality control of the clinical data collected in the HCS; *Asqualab* (www.asqualab.com)and *EuroCell* (www.eurocelldiag.com), both used to working with the HSCs, are in charge of the quality control of biological and hematological material; *Cemka-Eval* (www.cemka.fr), a CRO specialized in epidemiology is in charge of the relationships with physicians and hospitals in order to get the medical data for outcome ascertainment; *Imprimerie Nationale* (www.imprimerienationale.fr) the public company that manufactures secure documents such as passports takes care of the consent forms and some of the questionnaires.

The specific contribution of UMS 011 to the CONSTANCES infrastructure is to take in charge and coordinate all scientific and technical aspects of the infrastructure: general design of the cohort and the biobank, collecting individual data from the participants through questionnaires, designing and monitoring the examinations in the HSCs, applying quality control procedures, extracting individual data from the SNIIRAM and CNAV databases, managing the central database and insuring the protection and confidentiality of the data, organizing the access to the database for external researchers under the supervision of the governing bodies of CONSTANCES, obtaining legal authorizations and funding.

The PIs, along with the CONSTANCES team, are thus experienced in running large prospective population-based cohorts, in collaborating with the medical and scientific national and international community, the HSCs and the national organizations in charge of the health system and socioeconomic databases.

4.2 GOVERNING BODIES

The governance of CONSTANCES is provided by several committees.

The Institutional Steering Committee associates representatives of the main institutional partners of CONSTANCES: CNAMTS, CNAV, INSERM, UVSQ and the Ministry of Health. The Committee is in charge of defining the main scientific orientations of the cohort, supervising the organization and functioning, controlling the budget, ensuring the smooth functioning of the project, validating the reports elaborated by the PI. The COS is also in charge of designating the members of the Scientific Committee of CONSTANCES on proposal from the PI.

The Scientific Committee is in charge of advising on the orientations and on specific aspects it considers of importance. It also supervises the access of external groups to the infrastructure, and is in charge of the scientific evaluation of the proposals for ancillary projects. When it feels necessary, the Scientific Committee consults with the Ethics Committee; the current composition of the Scientific Committee is presented in the second part of this report.

Ethics Committee: ethical aspects are under the responsibility of the *Comité d'éthique pour la recherche médicale et en santé de l'Inserm*-ERMES, the Ethics Committee of INSERM. ERMES is in charge of all ethical aspects of biomedical research. On request of the CONSTANCES team, the Scientific Committee or the Institutional Steering Committee, any question relating to ethical problems may be transmitted to the Ethics Committee for its advice.

We also have established an **Advisory Committee,** associating recognized researchers in different fields related to the scientific objectives of CONSTANCES. The members of the committee were all involved in the preparation of the cohort’s protocol and are developing their own ancillary study nested within CONSTANCES. The Committee is intended to advise the core team in their domain of expertise. As its members are also in charge of personal nested projects, they can also advise the core team on difficulties they met in conducting their project, thus contributing to the improvement of the procedures, data or the functioning of the infrastructure.

Finally, we are considering setting up an **Industrials Committee** in a second step. Building on the discussions that are currently underway with private pharmaceutical and other companies about future collaborations, this advisory Committee will be the place to establish the rules of access to the CONSTANCES database by industry in accordance with the Institutional Steering, Scientific and Ethics Committees, and a place to exchange between researchers and industrials and develop future collaborations of topics of common interest.

5 REFERENCES

Ancel PY et al. Neurodevelopmental disabilities and special care of 5-year-old-children born before 33 weeks of gestation (the Epipage study) : a longitudinal cohort study. The Lancet 2008, 371: 813-820.

Backlund E et al.. Income inequality and mortality: a multi-level prospective study of 521, 248 individuals in 50 US States. Int J Epidemiol 2007;36:590–96.

Berr C, Balansard B et al. Cognitive decline is associated with systemic oxidative stress: the EVA study. Etude du Vieillissement Arteriel. J Am Geriatr Soc 2000; 48:1285-91.

Bhatti J et al. Impact of retirement on risky driving behaviors and attitudes toward road safety among a large cohort of French drivers (the GAZEL cohort). Scand J Work Envir and Health 2008; 22:307-315.

Borkowski JG, Benton AL, Spreen O. Word fluency and brain damage. Neuropsychologica 1967;5:135-140.

Bowie CR, Harvey PD. Administration and interpretation of the Trail Making Test. Nat Protoc 2006; 1:2277-2281.

Bowling A. Mode of questionnaire administration can have serious effects on data quality. J Pub Health 2005, 27(3), 281–291.

Brayne C. The elephant in the room - healthy brains in later life, epidemiology and public health.

Burton, PR et al. The global emergence of epidemiological biobanks: Opportunities and challenges, in Human Genome Epidemiology: Building the evidence for using genetic information to improve health and prevent disease. 2009a, Oxford University Press.

Burton PR et al. Size matters: just how big is BIG? Quantifying realistic sample size requirements for human genome epidemiology. Int J Epidemiol 2009b;38:263–73.

Chen Z et al. Cohort Profile: The Kadoorie study of chronic disease in China (KSCDC). Int J Epi 2005; 34: 1243-49.

Christensen K et al. Ageing populations: the challenges ahead. Lancet 2009;374:1196-1208.

Clavel-Chapelon F and the E3N-EPIC group. Secular trends of age at menarche and at onset of regular cycling in a large cohort of French women. Human Repr 2002;17: 228-32.

Cohidon C et al. Exposure to job-stress factors in a national survey in France. Scand J Work Environ Health. 2004;30:379-389.

Collins FS. The case for a US prospective cohort study of genes and environment. Nature 2004;429:475–77.

Collins, R. and UK Biobank Steering Committee. UK Biobank: Protocol for a large-scale prospective epidemiological resource. 2007, Manchester: UK Biobank Coordinating Centre.

Couris et al. French claims data as a source of information to describe cancer incidence: predictive values of two identification methods of incident prostate cancers. J Med Syst. 2006;30:459-63.

Couris et al. Breast cancer incidence using administrative data: correction with sensitivity and specificity. J Clin Epidemiol. 2009;62:660-6.

Darling GM et al. Hormone replacement therapy compared with simvastatin for postmenopausal women with hypercholesterolemia. N Eng J Med 1998 ; 338:64.

Dartigues JF, Gagnon M, Michel P, et al. Le programme de recherche paquid sur l'épidémiologie de la démence. Méthodes et résultats initiaux. Rev Neurol 1991; 147:225-230.

Diez-Roux AV. Investigating neighborhood and area effects on health. Am J Public Health. 2001;91:1783-9.

Doiron D, Burton P, Marcon Y, Gaye A, Wolffenbuttel BH, Perola M, Stolk RP, Foco L, Minelli C, Waldenberger M, Holle R, Kvaløy K, Hillege HL, Tassé AM, Ferretti V, Fortier I. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerg Themes Epidemiol. 2013 Nov 21;10(1):12. doi: 10.1186/1742-7622-10-12.

Ducimetière P, Richard J, Claude JR et al. Les cardiopathies ischémiques : incidence et facteurs de risque. L'Étude Prospective Parisienne. Paris, Éditions Inserm, 1981.

Egan KM et al. Active and passive smoking in breast cancer: Prospective results from the Nurses 'Health Study. Epidemiology 2002, 13, 138–145.

Févotte J et al. MATGENE: A program to develop job-exposure matrices in the general population in France. Ann Occup Hyg 2011 Sep 15. [Epub ahead of print].

Finch CE. The neurobiology of middle-age has arrived. Neurobiol Aging 2009;30:515-520.

Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J.Psychiatr.Res. 1975;12:189-98.

Fortier I et al, on behalf of the International Harmonization Initiative. Is Rigorous Retrospective Harmonization Possible? Application of the DataSHaPER Approach across 53 Large Bioclinical Studies. Int. J. Epidemiol 2011. doi: 10.1093/ije/dyr106.

Friedenreich CM. Methods for pooled analyses of epidemiologic studies. Epidemiology 1993;4:295–302.

Fuhrer R, Rouillon F. La version française de l'échelle CES-D (Center for Epidemiologic Studies-Depression scale). Description et traduction de l'échelle d'auto-évaluation. Psychiat Psychobiol 1989;4:163-6.

Gaudino EA, Geisler MW, Squires NK. Construct validity in the Trail Making Test: what makes Part B harder? J Clin Exp Neuropsychol 1995;17:529-535.

Gill TM et al. Trajectories of disability in the last year of life. N Engl J Med 2010;362:1173-1180.

Gmel G. The effect of mode of data collection and of non-response on reported alcohol consumption: a split-sample study in Switzerland, Addiction 2000, 95, 123–134.

Goldberg D. GHQ and psychiatric case. Br J Psychiatry. 1979;134:446-7.

Goldberg M. Administrative data bases: could they be useful for epidemiology? Rev Epidemiol Sante Publique, 2006, 54: 297-303. [French].

Goldberg M et al. Bases de données médico-administratives et épidémiologie : intérêts et limites. Courrier Stat; 2008, 124: 59-70. [French].

Goldberg M et al. The French Public Health Information System. Stat J Int Assoc Official Statistics 2011. 27: 1–11.

Grober E et al. Free and Cued Selective Reminding Test with Immediate Recall (FCSRT-IR). Psychol Sci Quart, 51, 2009: 266-282.

Guralnik JM et al. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. J Gerontol 1994;49:M85-M94.

HCSP. Les systèmes d'information pour la santé publique. Rapport du Groupe de travail du Haut Conseil de la Santé Publique. Paris, Haut Conseil de la Santé Publique, 2009.

Henny J, Goldberg M, and Zins M. Guide pour la constitution d'une Biobanque associée aux études épidémiologiques en population générale. Paris: Inserm-Lavoisier, Tec & Doc, 2012.

Hercberg S, Galan P, Preziosi P, Bertrais S, Mennen L, Malvy D, Roussel AM, Favier A, Briançon S. The SU.VI.MAX Study: a randomized, placebo-controlled trial of health effects of antioxidant vitamins and mineral. Arch Intern Med 2004; 164: 2335-42.

HID Survey: http://www.sante.gouv.fr/drees/serieetudes/serieetud16.htm

Hindorff LA et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci 2009;106:9362–67.

Iwatsubo Y et al. Prediction model of asthma using antiasthma drug claims for epidemiological surveillance of asthma in self-employed workers in France. EPICOH Conference, Oxford, 7-9 September 2011.

Jagger C et al. Inequalities in healthy life years in the 25 countries of the European Union in 2005: a cross-national meta-regression analysis. Lancet 2008;372:2124-2131.

Karasek R, Theorell T. Healthy Work: stress, productivity and the reconstruction of working life. New York: Basic Books, 1990.

Katz, S et al. Studies of illness in the aged. The index of ADL: a standardized measure of biological and psychosocial function. JAMA, 1963. 185: 914.

Kilbom S et al. Musculoskeletal Disorders: Work-related Risk Factors and Prevention. Int J Occup Environ Health 1996

Kuh D, Ben-Shlomo Y, eds. A lifecourse approach to chronic disease epidemiology. Oxford: Oxford University Press; 1997;3-14.

Kwak N, Radler B. A comparison between mail and web surveys: response pattern, respondent profile and data quality. J Official Stat 2002, 18: 257-273.

Marmot M (2004). Status Syndrome, Bloomsbury, London.

Marmot M et al. Closing the gap in a generation: health equity through action on the social determinants of health. Lancet 2008; 372: 1661–69.

Moisan F et al. Prediction model of Parkinson's disease based on antiparkinsonian drug claims. Am J Epidemiol 2011; 174:354-363.

Naess O et al. Cohort profile: cohort of Norway (CONOR). Int J Epidemiol. 2008 Jun;37(3):481-5.

Nat Rev Neurosci 2007;8:233-239 5. Banks J, Breeze E, Lessof C, Nazroo J (eds) (2006) Retirement, Health and Relationships of the Older Population in England. The Institute for Fiscal Studies, London.

Newton-Cheh C et al. Eight blood pressure loci identified by genomewide association study of 34,433 people of European ancestry. Nat Genet 2009;41: 666–76.

Oppenheimer GM: Becoming the Framingham Study. Am J Pub Health 2005; 95:602-610.

Renahy et al. Inform Health Soc Care. 2010 35:25-39.

Riboli E et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Pub Health Nutr. 2002; 5: 1113-24.

Roche N et al. Impact of chronic airflow obstruction in a working population. Eur Respir J. 2008;31:1227-33.

Santin G., Geoffroy B., Bénézet L., Delézire P., Chatelot J., Sitta R., Bouyer J, Guéguen A. SNIIRAM Cohorts Group. In an occupational health surveillance study, auxiliary data from administrative health and occupational databases effectively corrected for nonresponse. J Clin Epid. 2014; 67: 722-730.

Schatzkin A et al., Observational epidemiologic studies of nutrition and cancer: the next generation (with better observation). Cancer Epidemiology Biomarkers &Prevention, 2009; 18: 1026.

SF-36 Org: www.sf-36.org/tools/sf12.shtml

Siegrist J. Effort-reward Imbalance at Work and Health. In: Research in Occupational Stress and Well Being, Historical and Current Perspectives on Stress and Health. In: P. Perrewe (Eds).JAI Elsevier, London. Vol. 2, pp 261-291, 2002.

Siemiatycki J et al. Listing occupational carcinogens. Environ.Health Perspect., 2004. 112(15):1447–1459.

Sjösten N et al. Trajectories of headache in relation to retirement: a longitudinal modelling study. Cephalalgia 2011 Jan 10. [Epub ahead of print].

The BRIF Workshop Group. The role of a bioresource research impact factor as an incentive to share human bioresources. Nat Genet. 2011 43: 503-504.

Thompson A. Thinking big: large-scale collaborative research in observational epidemiology. Eur J Epidemiol, 2009. 24: 727-31.

Three-Cities study group. Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. Neuroepidemiology 2003; 22:316-325.

Touvier M et al. Eur J Epidemiol. 2010; 5:287-96.

Vahtera J et al. Effect of retirement on sleep disturbances: the GAZEL prospective cohort study. Sleep 2009;32:1459-1466.

Valleron AJ, Ed : Épidémiologie : conditions de son développement, et rôle des mathématiques. Rapport sur la Science et la Technologie n° 23, Comité RST de l'Académie des sciences, 2006.

van Gelder MHJ et al. Web-based Questionnaires: The Future in Epidemiology? Am J Epidemiol 2010;172:1292–1298.

Walport M. and Brest P. Sharing research data to improve public health. Lancet. 2011; 377: 2011– 2018.

Wechsler D. Manual for the Wechsler Adulte Intelligence Scale-Revised. New-York: Psychological Corporation, 1981.

Westerlund H et al. Self-rated health before and after retirement in France (GAZEL): a cohort study. The Lancet, 2009;374:1889-1896.

Westerlund H et al. Effect of retirement on major chronic conditions and fatigue: The French GAZEL occupational cohort study. BMJ 2010; 341:c6149.

Wiggins RD et al. Quality of life in the third age: key predictors of the CASP-19 measure. Ageing Soc 2004; 24:693-708.

World Health Organization. Preventing disease through healthy environments: towards an estimate of the environmental burden of disease. 2006, Geneva, Switzerland: World Health Organization.

Yazbeck C et al. Maternal Blood Lead Levels and the Risk of Pregnancy-Induced Hypertension: The EDEN Cohort Study. Environmental Health Perspectives 117, 2009: 1526-30.

Zins M et al. Mise en place d’une banque de matériel biologique associée à la cohorte GAZEL : aspects logistique et pratique. Rev Epidemiol Santé Publ, 2003,51:143-146.

Zins M, Leclerc A, Goldberg M: The French GAZEL Cohort Study: 20 years of epidemiologic research. Advances in Life Course Research 2009; 14: 135-146.

Zins M et al. The CONSTANCES Cohort: an Open Epidemiological Laboratory. BMC Public Health 2010; 10:479.

Zins M et al. Effects of retirement on alcohol consumption: longitudinal evidence from the French GAZEL Cohort study. PLoS ONE 2011, 6(10): e26531. doi:10.1371/journal.pone.0026531.