How to optimize the use of biobanks from population-based cohorts in aging research

C Berr[1,2,3], A Gabelle[2.3.4], N Fievet[5], M Goldberg[6,7], M Zins[6,7], I Carriere[1.2].

(1) Inserm U1061, La Colombière Hospital, F-34093 Montpellier, France.

(2) Montpellier University, F-34000 Montpellier, France

(3) Memory Research and Resources Center, Department of Neurology, Montpellier University Hospital Gui de Chauliac, F-34295 Montpellier, France.

(4) Inserm U1183, Saint Eloi Hospital, F-34295 Montpellier, France

(5) Institut Pasteur de Lille Laboratoire d'Analyse Génomique Inserm-U1167 F-59019 Lille

(6) Inserm, Population-based Epidemiologic Cohorts Unit, UMS 011, F-94807 Villejuif, France

(7) Versailles St-Quentin University, UMS 011, F-94807, Villejuif, France

Corresponding author and reprint requests to Dr Claudine BERR, Inserm U1061 Hôpital La Colombière, F-34093 Montpellier Cedex 5, France.

Phone: 33 (0)4 99 61 45 66; Fax : 33 (0)4 99 61 45 79.

@address: claudine.berr@inserm.fr

Abstract

In epidemiological cohorts, there is an increased interest for the implementation of biobanks. The potential role of biological determinants of diseases needs to be investigated before the onset of the event of interest in order to limit the problems encountered when examining biological determinants in classical case-control studies. Biobank is now a very sophisticated system that consists of a programmed storage of biological material and related data. Our aim in this paper is to document how biobank constitution is useful for studying biological determinants of aging and to give some indications on methodological issues that can be helpful to optimize the constitution and use of biobanks in aging cohorts. Optimization of sampling through two-phase designs (nested case control or case-cohort studies) allows better efficiency. These elements are, for most of them, not specific to aging populations but are useful more generally for the epidemiology of chronic diseases. Our purpose will be illustrated with some examples and results obtained in an ongoing aging cohort, the Three–City study.

Keywords: cohort, aging, biobank, sampling

Introduction

We now live in a global aging society; this is true for industrialized countries but it is also emerging in less developed countries. The World Health Organization (WHO) estimated that there were 841 million people aged 60 and over in the year 2013 and this will increase to 2 billion by 2050 (Chatterji et al.). Against the background of a steady increase in life expectancy, the question of how people will age has become increasingly important.

Since the 1990s, many epidemiological studies, mostly cohorts with large sample sizes have been initiated in different countries in order to identify factors associated with health events in aging or consequently those related to a successful aging. Cohort studies (Doll 2001) - longitudinal follow-ups of populations over time with respect to exposures (risk factors) and health event occurrence - are popular because they allow causal interpretations. But they require long-term investments, are expensive and, as all observational studies, subject to biases. There are fundamental criteria to allow investigation of the causative role of biological determinants in diseases(Hill 1965) . Dosages have to be performed before the onset of the event of interest in order to limit the problems encountered when looking at biological determinants in classical case-control studies (which compare cases recruited when disease has occurred to subjects not affected by the event). Potential effects of disease on biological status are controlled when biological samples have been collected before disease onset. This is the main reason which justifies the increasing number of cohorts with biological samples collected at baseline. Furthermore in aging epidemiological studies, most health events are chronic disorders with insidious onset and progression for some of them towards neurodegenerative disorders or chronic kidney failure and it is necessary to identify risk factors before the non-observed real onset.

Research programs utilize human biological samples; Biobank is a type of biorepository that stores biological samples for use in subsequent research. Since the late 1990s biobanks have become an important resource in medical research, supporting many types of contemporary research such as genomics and personalized medicine, set up either for population-based or disease-specific research purposes (Zika et al. 2011). Samples in biobanks and the data derived from these samples can often be used by multiple researchers for multiple purposes, which are not always defined when cohorts or research programs are planned. Advances in epidemiology and omics science have led to an increased interest in infrastructure development and data sharing facilitated by biobanks of specimens and linked health information (Artene et al. 2013). Biobank is now a very sophisticated system that consists of a programmed storage of biological material and related data.

Our aim in this paper is to document how biobank constitution is useful for studying biological determinants of aging and to give some indications on methodological issues that can be helpful to optimize the constitution and use of biobank in aging cohorts. These elements are, for most of them, not specific to aging populations but are useful more generally for the epidemiology of chronic diseases. Our purpose will be illustrated with some examples and results obtained from an ongoing aging cohort, the Three–City study (Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population 2003).

1- How to deal with biosamples, general considerations

When planning a cohort design, the choice of biological specimens is crucial and is guided by current but also future (and unknown) research questions. Ideally, the research staff must define the potential parameters to be studied so that appropriate collection and processing protocols can be designed. But for future research questions, it is clearly required to anticipate the future. In epidemiological studies on blood biomarkers, some dosages require serum samples, other plasma but plasma can be collected with different collection tube additives (heparin, EDTA, lithium…) allowing dosages of different biomarkers. Choice for storage, at liquid nitrogen temperature or in -80 degree centigrade freezers, must also be considered. For instance in microbiology, biological samples can be stored for up to 30 years but specific protocols are required to reduce the damage induced by preservation techniques (De Paoli 2005). Pre-acquisition phase and pre-storage phase can be crucial and security of storage itself frequently implies that it is carried out on at least two independent sites. Number and adequate volumes of aliquots are always questionable when planning a biobank. Experience of professionals involved in the biobank constitution is crucial for optimal choices.

In order to guarantee and to preserve the quality of the collections, biologists and epidemiologists need to have a comprehensive view of analytical and storage techniques to ensure long term integrity of the stored specimens (Henny J 2012). Sample collection and processing must be optimized to avoid approaches that may preclude future analyses. The quality of biospecimens and associated data must be consistent and collected according to standardized methods in order to prevent spurious analytical results that can lead to artifacts being interpreted as valid findings (Vaught and Lockhart 2012). All steps for the constitution of a biobank require standard operating procedures including traceability of sampling, pre-analytic phases, and storage. Guidelines should also include ethical, legal and social issues which depend on the different laws and regulations in various countries (Cambon-Thomsen et al. 2003; Noiville 2012). Published guidelines on specific biological processes should also be considered.

Alongside the biobank constitution a database must be set up to track the samples (traceability). Each aliquot must be identified in the database (protocol, individual ID and research consent, study site, collection time, aliquot number, storage conditions) and its location described (freezer, rack, slot, box, row, and column). It must be possible to track in this database location changes, aliquot consumption and trans-shipment in order to always provide an updated state of the biobank. In France, the necessity to ensure quality of samplings and associated data has led to the creation of Biological Resources Centers (CRB= Centre de Ressources Biologiques) which aim at harmonizing collection and storage of biological sampling. They must respond to quality norms.

When the repository is constituted, the rationale for its use must be considered carefully. Even if a large number of aliquots are collected, each one can become precious and rare (Clement et al. 2014), particularly aliquots from individuals who have become incident cases.

2- How to optimize sampling for biological studies

Because disease incidence rates are usually low, cohort studies need to follow large numbers of subjects during many years to obtain optimal numbers of cases and sufficient statistical power (Doll 2001). Precision for estimating risk associated with an exposure being mainly limited by the number of cases, it is not essential to collect complete information for all the controls. It is thus possible to conduct sub-studies with selected non-case participants so that expensive data such as biological dosages do not need to be ascertained for everyone. Investigators have to find the right strategy and decide on which sample the dosage will be performed in order to optimize the study cost efficiency. All designs constructed to include subjects in the sub-cohort on which the biobank is established must be based on the event (disease) occurrence and at risk status during the follow-up. The designs differ with respect to **how controls are sampled** and to how data will be analyzed. Thus, case-cohort studies and nested case-control studies enable cost reduction with a minimal loss of efficiency (Langholz and Thomas 1990).

Two-phase studies (see figure 1) were introduced into epidemiology more than twenty years ago but they have been relatively rarely applied in practice. In these studies, data are collected in two phases. First, the cohort participants are selected from a general population with or without randomization or stratification. During this phase, information is collected on all the subjects. At this phase information on disease status and (minimal) information on exposure are collected. Phase two uses this information to recruit a subsample, stratified on disease status and exposure. Further data are then collected to obtain more detailed exposure/confounder information. The final analysis uses data from both phases.

Two main types of designs can be proposed for the phase 2:

1. Nested case-control design: A nested case–control study (Langholz 2005) comprises subjects sampled from an assembled epidemiological cohort study in which the sampling depends on disease status. Nested case–control studies are generally used when disease is rare and, at the minimum, disease outcome has been obtained for all cohort subjects, but it is too expensive to collect and/or process information on covariates of interest for the entire cohort. After identification of cases, confounder information available in the cohort data is often used to select controls that closely match cases for age, sex or other characteristics. Most of the first studies conducted with this type of design were performed in the field of occupational epidemiology. In a lung cancer mortality study on an historical cohort of hard-metal workers (Moulin et al. 1998), the cases were the cohort workers who had died of lung cancer. Three controls per case were sampled from the set of those at risk, i.e. from all subjects under follow-up at the date of the death of the case and known to be alive on that date, with matching on sex and date of birth. Collection of data on job history was limited to cases and their controls and analyses were conducted with conditional logistic regression models.

2. Case-cohort design (Prentice and Self 1988): In this design the entire cohort is followed in order to determine the time of all the events considered in the study. Controls are randomly sampled without considering events or their timing as part of a "sub-cohort" (cohort random sample). Phase-2 expensive information is collected for this sub-cohort and for all the cases, belonging or not to the sub-cohort. Case cohort design is presented as more efficient than nested case-control since each case may be compared with the overall sub-cohort. It can be used for more than one outcome, e.g. dementia and stroke in an aged population. The same sub-cohort (same controls) is used to compare with cases for all the outcome types. It is possible to start analyzing biological material on the sub-cohort individuals from the beginning of follow-up, adding incident cases of one or more than one type as they occur. Even if not considered at the beginning of case-cohort design, use of cases not in the sub-cohort is possible. Most analyses utilize the "robust" approach of Barlow (Barlow et al. 1999) which involves Cox regression with case and control observations weighted by their inverse sampling probabilities.

3- Example of a case-cohort study in an aging study.

The 3C study is an ongoing population-based, prospective study of the relationship between vascular factors and dementia, carried out in three French cities: Bordeaux (South-West France), Montpellier (Southern France) and Dijon (Central Eastern France). Between January 1999 and March 2001, 9,294 non-institutionalized subjects aged $\geq$ 65 (selected from electoral rolls) agreed to participate in the study. The baseline data collection included socio-demographic and lifestyle

characteristics, symptoms and complaints, main chronic conditions, medication use, neuropsychological testing. Data have been updated at each follow up exam every 2 or 3 years until 2012. Incident cases for the three main outcomes (dementia, coronary heart diseases (CHD) and stroke) were considered, documented and validated. The study protocol has been described in detailed elsewhere (Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population 2003). It was approved by the Institutional Review Board at Kremlin-Bicêtre University Medical Center. Each participant gave their written, informed consent to participation.

Ninety-five percent of the participants accepted to have blood sampling for the measurement of biological parameters at baseline and the constitution of a blood specimen bank (serum, plasma, and DNA) stored at –80° C. Measurements of biological parameters [glycemia, cholesterol (total, high-density lipoprotein, and low-density lipoprotein), triglycerides, and creatinemia] were centralized and performed by the Biochemistry Laboratory of the University Hospital of Dijon. At baseline, we stored biological specimens for 8860 subjects, leading to 53160 blood tubes (6 per subject) and to more than 221500 aliquots located both in a regional and in a central center (Biological Resources Center, LAB-CRB of the Pasteur Institute of Lille) for security reasons.

A case-cohort study was set up at the end of the 4-year follow-up period, to investigate potential novel biological risk markers for the 3 events of interest. The case-cohort design involved a randomly selected sub-cohort from the initial cohort as well as all incident cases studied (Carcaillon et al. 2009). In practice, a random selection stratified by center, sex and age with a sampling ratio of 15% led to a sample of 1,254 subjects from the initial cohort. Within this sub-cohort, 84 incident events occurred during the 4 years of follow-up (33 CHD, 12 strokes, and 40 cases of dementia). For the study of dementia, we added all additional dementia cases identified from outside the sub-cohort (n=218) and removed prevalent cases (n=29) and missing values regarding incident dementia status (n=121) giving a total of 257 cases of incident dementia as illustrated in **figure 2**. For other events, similar additions and exclusions were proposed. For studying CHD, we included all additional CHD cases (n=166) and removed prevalent cases (n=131) and missing values regarding incident CHD status (n=37). The sub-cohort corresponding to the study of CHD (n=199) included 1086 subjects. Similarly, for stroke, we added all additional stroke cases (n=103) and removed prevalent cases (n=53) and missing values regarding incident stroke status (n=45). The sub-cohort corresponding to the study of stroke (n=111) included 1139 subjects.

Some of the analyses performed with this design have led to interesting results not limited to one of the events when studying inflammation (C-reactive protein, fibrinogen) and

hypercoagulability (fibrin D-dimers, thrombin generation). Carcaillon et al (Carcaillon et al. 2009) showed that elevated fibrinogen and D-dimer levels were associated with incident arterial disease (CHD). But in addition, they also found that high D-dimer level could represent a new risk factor for a subtype of dementia. They also used the case-cohort design limiting analysis to a random sub-cohort of 562 women not using hormone therapy and 132 incident dementia cases (Carcaillon et al. 2014).

Focusing on dementia risk factors, Lambert et al (Lambert et al. 2009) examined the association between plasma Aβ peptide levels and dementia based on dosages in the sub-cohort and on incident dementia cases and demonstrated that they may be useful markers to indicate individuals susceptible to short term–risk of dementia. Even if these dosages were performed only in 15% of the total sample representing 1254 individuals, it was also possible to use the sub-cohort itself as a sample to study relationships between these plasma Aβ peptide levels and mortality (Gabelle et al. 2014). Moreover, it was also possible to analyze the role of these peptides as potential prognosis markers in dementia (Gabelle et al. 2013), using dosages in incident cases diagnosed in the first years of follow-up. This example illustrates how dosages in this frame can be extensively beneficial. A number of papers have been published in the frame of the different ancillary studies which were proposed years after the constitution of the cohort.

Perspectives

Created in 2013 at the European level, the BBMRIERIC (Biobanking and Biomolecular Resources Research Infrastructure-European Research Infrastructure Consortium) (http://bbmri-eric.eu/) (Reichel et al. 2014) aimed at increasing efficacy and excellence of European bio-medical research by facilitating access to quality-defined human health/disease-relevant biological resources. Since the first biospecimen collections in cohorts and what are now biobank, there have been great changes. The scale of cohorts has changed and a number of ‘‘MEGA’’ cohorts often at a national scale are now conducted with very large numbers of subjects. In France, the Constances study on 200,000 subjects aged from 18 to 70 includes a specific program on aging (Zins et al. 2010) and plans to store 8,000,000 aliquots. In England, the UK Biobank (Elliott and Peakman 2008) aimed at investigating the role of genetic factors, environmental exposures and lifestyle in the causes of major diseases of late and middle age in a population of 500,000 individuals aged 40–69. By the end of the recruitment phase, they stored about 15 million sample aliquots with a highly automated approach for the processing and storage of the samples. In the Netherlands, the Lifelines cohort (Scholtens et al. 2014) is a three generation cohort study with a biobank with baseline data collected for more than 160,000 participants aged from 6 months to 93 years set up for studying the development of chronic diseases and healthy aging.

The huge size of the biobanks associated with very large cohorts generates new challenges in different areas. Regarding costs, all biobanks face the need to establish new economic models to take into account not only financing the initial investment of expansive facilities and sophisticated equipment but also to consider the cost of maintenance of the biobank and the rapid renewal of equipment due to the fast changes in technology. The cost of collecting and storing biosamples is thus very high and the price for accessing samples may become unaffordable for researchers funded through public grants (Clement et al. 2014). Ethics issues also pose new challenges. Even if consent was given at the time of biomaterial collection, it is simply not possible to forecast research questions that will be asked decades later. This is a common problem for all biobanks, but it is especially crucial for population-based cohorts where the number of deceased or lost to follow-up subjects is increasing according to the length of follow-up, making it almost impossible to retrieve the participants for asking them specific consents. In some countries the regulations may be a serious obstacle and this is amplified by the rapid development of international research consortia aiming at sharing data from different countries with different ethics regulations.

Faced with the inevitable multiplication of mega cohorts, an international effort of the scientific community and of the funding and administration of research bodies is therefore essential to find realistic solutions to these challenges.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Figure 1: Sheme for 2-phase sampling

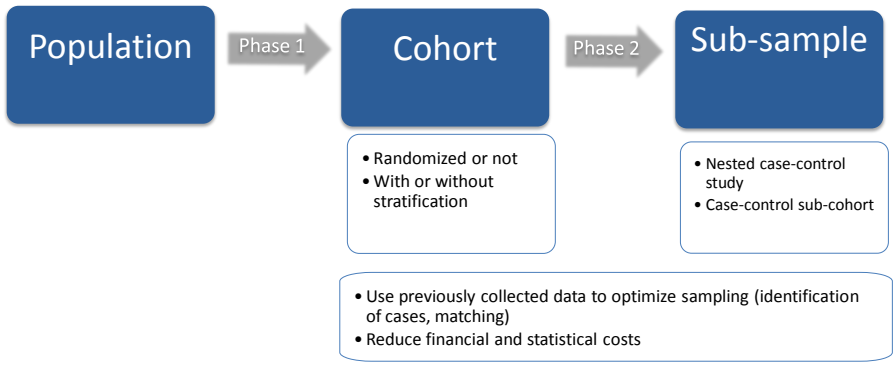
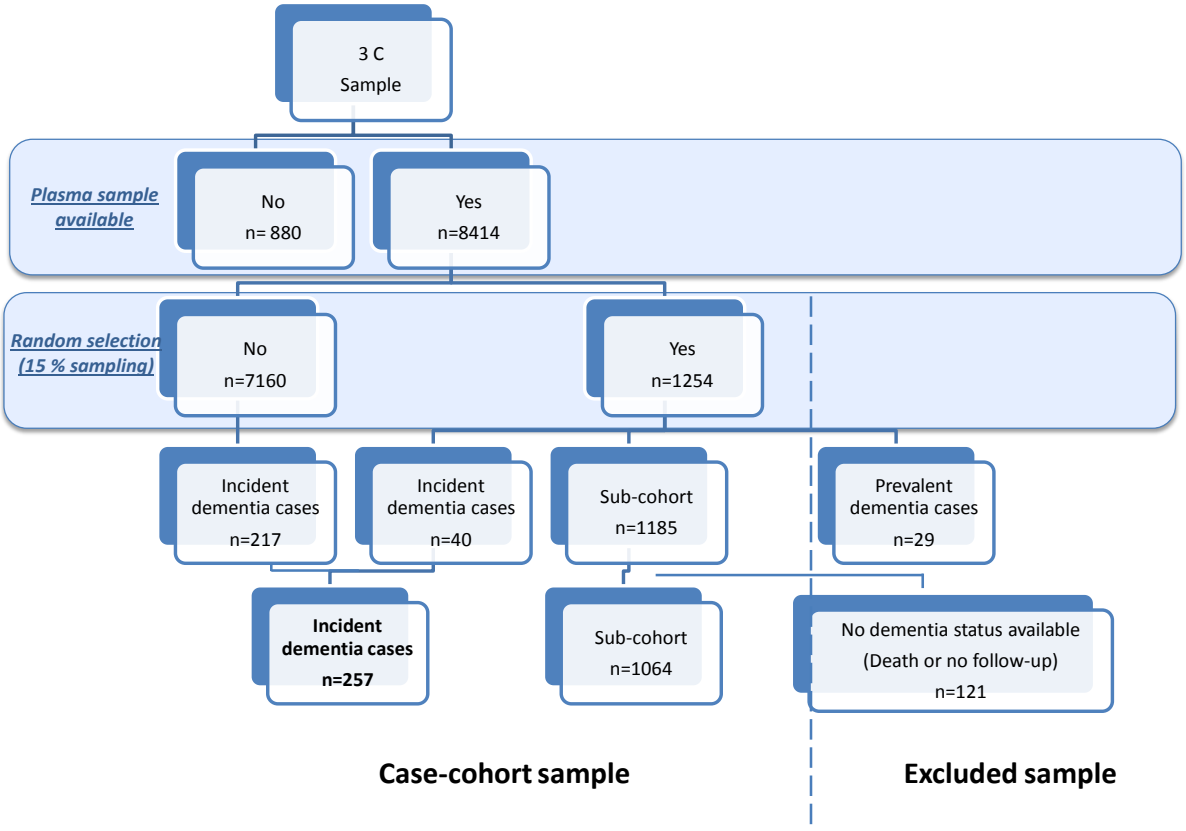


Figure 2: Design of the case-control study in the 3C Cohort, example of dementia cases at 4 –year follow up (adapted from Lambert, J. C., S. Schraen-Maschke, et al. (2009))



Conflict of Interest: The authors declare that they have no conflict of interest.

References

Artene SA, Ciurea ME, Purcaru SO, Tache DE, Tataranu LG, Lupu M, Dricu A (2013) Biobanking in a constantly developing medical world ScientificWorldJournal 2013:343275 doi:10.1155/2013/343275

Barlow WE, Ichikawa L, Rosner D, Izumi S (1999) Analysis of case-cohort designs J Clin Epidemiol 52:1165-1172 doi:S089543569900102X [pii]

Cambon-Thomsen A, Ducournau P, Gourraud P-A, Pontille D (2003) Biobanks for Genomics and Genomics for Biobanks Comparative and Functional Genomics 4:628-634 doi:10.1002/cfg.333

Carcaillon L, Brailly-Tabard S, Ancelin ML, Rouaud O, Dartigues JF, Guiochon-Mantel A, Scarabin PY (2014) High plasma estradiol interacts with diabetes on risk of dementia in older postmenopausal women Neurology 82:504-511 doi:WNL.0000000000000107 [pii]

10.1212/WNL.0000000000000107

Carcaillon L, Gaussem P, Ducimetiere P, Giroud M, Ritchie K, Dartigues JF, Scarabin PY (2009) Elevated plasma fibrin D-dimer as a risk factor for vascular dementia: the Three-City cohort study J Thromb Haemost 7:1972-1978 doi:JTH3603 [pii]

10.1111/j.1538-7836.2009.03603.x

Chatterji S, Byles J, Cutler D, Seeman T, Verdes E Health, functioning, and disability in older adults ”present status and future implications The Lancet doi:http://dx.doi.org/10.1016/S0140-6736(14)61462-8

Clement B et al. (2014) Public biobanks: calculation and recovery of costs Sci Transl Med 6:261fs245 doi:6/261/261fs45 [pii]

10.1126/scitranslmed.3010444

De Paoli P (2005) Biobanking in microbiology: From sample collection to epidemiology, diagnosis and research FEMS Microbiology Reviews 29:897-910 doi:10.1016/j.femsre.2005.01.005

Doll R (2001) Cohort studies: history of the method. I. Prospective cohort studies Soz Praventivmed 46:75-86

Gabelle A et al. (2013) Plasma amyloid-beta levels and prognosis in incident dementia cases of the 3-City Study J Alzheimers Dis 33:381-391 doi:Y51653341576G371 [pii]

10.3233/JAD-2012-121147

Gabelle A et al. (2014) Plasma beta-amyloid 40 levels are positively associated with mortality risks in the elderly Alzheimers Dement doi:S1552-5260(14)00643-8 [pii]

10.1016/j.jalz.2014.04.515

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

Hill AB (1965) The Environment and Disease: Association or Causation? Proc R Soc Med 58:295-300

Lambert JC et al. (2009) Association of plasma amyloid beta with risk of dementia: the prospective Three-City Study Neurology 73:847-853 doi:73/11/847 [pii]

10.1212/WNL.0b013e3181b78448

Langholz B (2005) Case-Control Study Nested. In: Colton PAT (ed) Encyclopedia of Biostatistics, vol 1. Second Edition edn. John Wiley & Sons, Ltd, , Chichester, pp 646-655

Langholz B, Thomas DC (1990) Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison Am J Epidemiol 131:169-176

Moulin JJ et al. (1998) Lung cancer risk in hard-metal workers Am J Epidemiol 148:241-248

Noiville C (2012) Biobanks for research. Ethical and legal aspects in human biological samples collections in France J Int Bioethique 23:165-172, 183-164

Prentice RL, Self SG (1988) Aspects of the use of relative risk models in the design and analysis of cohort studies and prevention trials Stat Med 7:275-287

Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population (2003) Neuroepidemiology 22:316-325

Vaught J, Lockhart N (2012) The Evolution of Biobanking Best Practices Clinica chimica acta; international journal of clinical chemistry 413:1569-1575 doi:10.1016/j.cca.2012.04.030

Zika E et al. (2011) A European Survey on Biobanks: Trends and Issues Public Health Genomics 14:96-103