# Chapter 3

# Nordic Biological Specimen Bank Cohorts as Basis for Studies of Cancer Causes and Control: Quality Control Tools for Study Cohorts with More than Two Million Sample Donors and 130,000 Prospective Cancers

**Eero Pukkala**

## Abstract

The Nordic countries have a long tradition of large-scale biobanking and comprehensive, population-based health data registries linkable on unique personal identifiers, enabling follow-up studies spanning many decades. Joint Nordic biobank-based studies provide unique opportunities for longitudinal molecular epidemiological research. The Nordic Biological Specimen Banks working group on Cancer Causes and Control (NBSBCCC) has worked out very precise quality assurance principles for handling of the *samples*, based on the tradition in biobank culture. The aim of this paper is to demonstrate how high standards of quality assurance can also be developed for the *data* related to the subjects and samples in the biobanks. Some of the practices adopted from the strong Nordic cohort study experience evidently improve quality of nested case-control studies nested in biobank cohorts. The data quality requirements for the standardised incidence ratio calculation offer a good way to check and improve accuracy of person identifiers and completeness of follow-up for vital status, which are crucial in case-control studies for picking up right controls for the cases. The nested case-control design applying incidence-density sampling is recommended as an optimal design for most biobank-based studies. It is demonstrated how some types of biobanks have a period immediately after sampling, when the cancer risk is not comparable with the cancer risk in the base population, and how many of the biobanks never represent the normal average population of the region. The estimates on the population-representativeness of the biobanks assist in interpretation of generalisability of results of the studies based on these samples, and the systematic tabulations of numbers of cancer cases will serve in study power estimations. The well over 130,000 prospective cancer cases registered among subjects in the NBSBCCC biobank cohorts have already offered unique possibilities for tens of strong studies, but for rare exposure-outcome combinations predictions on future numbers of cases improve the chance to select the right moment when the study will have accurate statistical power.

**Key words:** Biobanks, Cancer incidence, Cohort study, Record linkage, Control selection, Selection bias, Inverse causality

Joakim Dillner (ed.), *Methods in Biobanking*, Methods in Molecular Biology, vol. 675,
DOI 10.1007/978-1-59745-423-0_3, © Springer Science+Business Media, LLC 2011

## 1. Introduction

In the Nordic countries, there exists a series of established biological specimen banks with many decades of follow-up that enable performing prospective epidemiological studies with adequate statistical power even for diseases and exposures that are not common. Between 1995 and 2006, more than 30 joint articles ((1–33); Table 1) were published by the Nordic

**Table 1**
**Use of serum samples in studies published by the Nordic Biological Specimen Banks for Cancer Causes and Control (NBSBCCC) network. Update 21 Nov 2006**

| NBSBCCC study number | Study (references) | Sera used in the study | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Finland | | | | Iceland | | Norway (Janus) | | Sweden | | | | | | | | | |
| | | | | | | | | | | | Sweden Health and Disease Study | | | | | Malmö Microbiology Biobank | | | |
| | | Maternity cohort (females) | Helsinki Heart Study (males) | ATBC (males) | Mobile clinic | Maternity cohort (females) | Heart preventive clinic | Health examinations | Blood donors | Northern Sweden Maternity Cohort (females) | VIP | MONICA | Mammography (females) | Diet and cancer | Preventive medicine | Maternity cohort (females) | Blood-borne viral screening | Other | Infectious disease control |
| 1 | Dillner et al. [1] | | | | X | | | | | | | | | | | | | | |
| 2 | Lehtinen et al. [2] | | .. | .. | X | | | | | | | | | | | | | | |
| 3 | Lehtinen et al. [3] | | | | X | | | | | | | | | | | | | | |
| 4 | Dillner et al. [4] | | .. | .. | | | | | | | | | | | | | | | X |
| 5 | Bjørge et al. [5] | X | | | | | | X | X | | | | | | | | | | |
| 6 | Dillner et al. [6] | X | .. | .. | | | | X | | | X | X | | | | | | | |
| 7 | Bjørge et al. [7] | | | | | | | X | X | | | | | | | | | | |
| 8 | Dillner et al. [8] | .. | | | X | .. | | | | .. | | | .. | | | .. | | | |
| 9 | af Geijersstam et al. [9] | X | .. | .. | | | | | | | | | | | | | | | |
| 10 | Luostarinen et al. [10] | X | .. | .. | | | | X | | | X | X | | | | | | | |
| 11 | Lehtinen et al. [11] | X | .. | .. | | | | | | | X | X | | | | | | | |
| 12 | Kibur et al. [12] | X | .. | .. | | | | | | | | | | | | | | | |

(continued)

## Table 1 (continued)

| NBSBCCC study number | Study (references) | Sera used in the study | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Finland | | | | Iceland | | Norway (Janus) | | Sweden | | | | | | | | | | |
| | | | | | | | | | | | Sweden Health and Disease Study | | | | | | Malmö Microbiology Biobank | | | |
| | | Maternity cohort (females) | Helsinki Heart Study (males) | ATBC (males) | Mobile clinic | Maternity cohort (females) | Heart preventive clinic | Health examinations | Blood donors | Northern Sweden Maternity Cohort (females) | VIP | MONICA | Mammography (females) | Diet and cancer | Preventive medicine | Maternity cohort (females) | Blood-borne viral screening | Other | Infectious disease control |
| 13 | Koskela et al. [13] | X | .. | .. | | | | X | | | X | X | | | | | | | | |
| 14 | Sigstad et al. [ 14] | X | .. | .. | | | | X | | | X | | | | | | | | | |
| 15 | Mork et al. [15] | X | X | | | | | X | | | X | | | | | | | | | |
| 16 | Anttila et al. [16] | X | .. | .. | | | | X | | | X | X | | | | | | | | |
| 17 | Stattin et al. [17] | .. | X | | | .. | | X | X | .. | X | X | .. | | | .. | | | | |
| 18 | Bjørge et al. [18] | X | | | | | | X | X | | | | | | | | | | | |
| 19 | Stattin et al. [19] | .. | X | | | .. | | X | X | .. | X | X | .. | | | .. | | | | |
| 20 | Lehtinen et al. [20] | X | .. | .. | | | | X | | | X | X | | | | | | | | |
| 21 | Lehtinen et al. [21] | X | .. | .. | | | | X | | | X | X | | | | | | | | |
| 22 | Youngman et al. [22] | X | .. | .. | | | | X | | | X | X | | | | | | | | |
| 23 | Paavonen et al. [23] | X | .. | .. | | | | X | | | X | X | | | | | | | | |
| 24 | Lehtinen et al. [24] | X | .. | .. | .. | X | .. | .. | .. | | .. | .. | .. | .. | .. | | .. | .. | .. | |
| 25 | Tuohimaa et al. [25] | .. | X | | | .. | | X | X | .. | X | X | .. | | | .. | | | | |
| 26 | Luostarinen et al. [26] | X | .. | .. | | | | X | | | X | X | | | | | | | | |
| 27 | Lehtinen et al. [27] | X | .. | .. | .. | X | .. | .. | .. | | .. | .. | .. | .. | .. | | .. | .. | .. | |
| 28 | Anttila et al. [28] | .. | X | | | .. | | X | X | .. | X | X | .. | | | .. | | | | |
| 29 | Stolt et al. [29] | X | .. | .. | .. | | .. | .. | .. | | .. | .. | .. | .. | .. | | .. | .. | .. | |
| 30 | Hakama et al. [30] | X | .. | .. | | | | X | | | X | X | | | | | | | | |
| 31 | Tedeschi et al. [31] | | | | | | | | | | X | | | | | | | | | |
| 32 | Tedeschi et al. [32] | X | X | X | | | | X | X | | X | | | | | | | | | |
| 33 | Korodi et al. [33] | .. | X | | | .. | | X | X | .. | X | X | .. | | | .. | | | | |

Participating serum banks marked with **X** and those which do not include sera from relevant persons with two dots (..). *Columns shaded with gray* indicate that sera from these serum banks have not been used in any NBSBCCC study so far. The table includes only studies that have got the internal NBSBCCC study number, i.e., officially accepted as network studies

Biological Specimen Banks working group on Cancer Causes and Control (NBSBCCC). The majority of studies so far were aimed at elucidating infections such as Human Papillomavirus (HPV) as causes of cancer. In addition to the joint Nordic studies, the biobanks operate independently with several hundred publications based on one or several of the biobanks described in this paper. Major subject areas for study have been hormones, nutrition, smoking, organochlorine compounds and genetic polymorphisms as causes of cancer in addition to a number of studies evaluating tumour markers. Still, the first systematic evaluation of characteristics and quality of the biobank cohorts or features of cancer risk pattern among the donors was done just recently and published in 2007 (34). This book chapter borrows much of the text of that publication, modified to give practical insight of thinking and methodology normally used in quality assurance (QA) of other types of epidemiological study cohorts than biobank cohorts.

This paper includes systematic descriptions of the participating biobanks: background, organisation, size, years of sample collection and administrative aspects. Numbers of cancer cases found among persons in the serum banks after serum drawing are given, advertising the unique possibilities of the national cancer registration systems in the Nordic countries. Population representativeness of the serum bank cohorts is estimated by comparing cancer incidence in the biobank cohorts with the respective national rates. Finally, issues to be taken into account in designing case-control studies nested in the Nordic biobanks are discussed.

In their classical assessment of the quantitative importance of avoidable causes of cancer, Doll and Peto estimated that a majority of human cancer was attributable to avoidable causes (35). They concluded that most of these avoidable causes remained unidentified. For risk factor identification and causality inference as well as for studies searching for mechanisms behind increases or decreases in cancer incidence, they recommended the use of prospective studies nested in cohorts of stored biological specimens.

This paper introduces the Nordic biobank network NBSBCCC to new potentially interested partners and serves as a general reference for specific studies based on these biobanks. NBSBCCC is a network of excellence that contains 17 independent biobank cohorts, five cancer registries and numerous expert user groups. The purpose of the network is to provide a concerted resource for etiologic studies of cancer, with a focus on longitudinal studies addressing unexplained causes and trends over time.

People who have donated samples to a biobank can also be considered as classical study cohort that is in most aspects technically comparable with, e.g., cohorts of occupationally exposed persons. Therefore, methods used in quality assurance and evaluation of accuracy of other types of cohorts can be adapted to biobank cohorts as well. Because that kind of approach has not been tradition

in biobank culture – most principles of quality assurance have their roots in laboratory sciences – this paper describes some of the practices typical to cohort studies that evidently also improve quality of nested case-control studies in biobank cohorts.

## 2. Participating Biobanks

The first crucial characteristic of any cohort study is to understand the *history* of the cohort collection:

- definition of the cohort; which type of persons were included (in the following I use the term "being exposed" to mean people who fulfil the inclusion criteria, although in biobank context the "exposure" simply means donation of sample),
- region of coverage; this can be a geographical area, or one institution (such as a factory in occupational studies, or a hospital in clinical cohorts),
- years of coverage; the cohort can be cross-sectional (including everyone under exposure at a point of time), or dynamic one (including everyone being exposed during a given period, no matter if the exposure started before that period or ended after that period),
- exclusions; there may be exclusion by purpose (defined in study protocol) or by accident (e.g., exclusion of deceased persons from the cohort because of lack of storage base; this would totally ruin the possibilities to use the cohort in any study on disease risk),
- other selection mechanisms; in biobank context one of the most important ones is selective participation that may decrease the population-representativeness of the cohort meant to be random sample of a population,
- variables collected, with full descriptions of principles in coding and input into the database; in the context of biobank samples such data may have existed but they have not always (e.g. if the samples are taken for clinical purposes) been collected systematically into databases in such format that they can be used afterwards in a scientific studies together with the sample,
- accuracy check-ups of the data in the context of storing them at the baseline; most importantly were the identification information of the person (in the Nordic countries: person ID codes) confirmed,
- collection of follow-up information; how were the data on possible deaths, migration out of the follow-up region and outcome events achieved; was the linkage procedure (key) fully complete, where there temporal or spatial holes in follow-up.

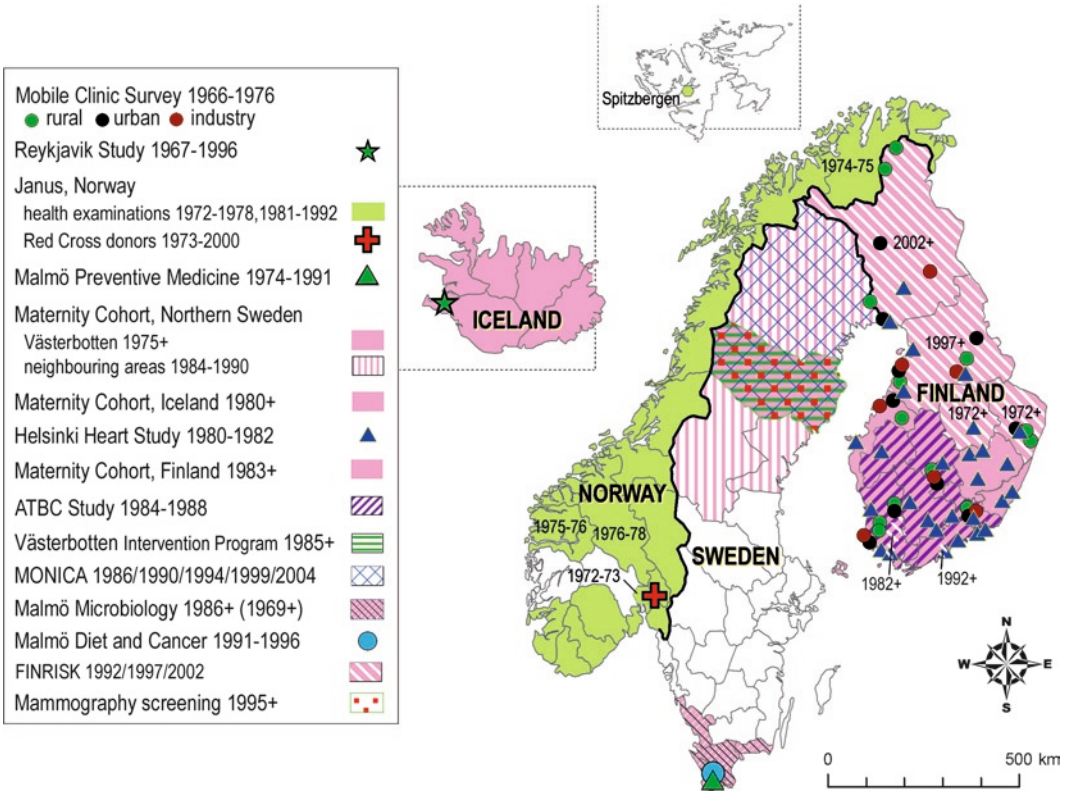


Fig. 1. Map of Nordic countries indicating the coverage areas of the serum banks.

In the following, some key characteristics of the background of the Nordic biobanks belonging to the NBSBCCC programme are described in a systematical way. The network so far consists of 12 biobanks in Finland, Iceland, Norway and Sweden, three of which are split into two to three independent subcohorts (Fig. 1, Table 2). Participating biobanks are independent entities that make their own decisions, but are committed to facilitate joint studies by working towards similar policies for quality assurance, logistics and study designs as well as for permission and terms of collaboration. NBSBCCC is funded by the Nordic Council of Ministries and as a European Union sixth framework programme Network of Excellence.

Research projects using the biobanks need appropriate permissions from the national Data Protection Authorities, National or Local Ethical Committees and from the boards of the biobanks. Informed consent is obtained from all persons donating samples, making it clear to the donors that the material will be used for future research purposes. Details of the permission procedure can be obtained via contact email addresses given in Table 2.

All samples have been stored at –20°C to –25°C except those of the Alpha-Tocopherol, Beta-Carotene Cancer (ATBC)

**Table 2**
**Characteristics of the serum banks included in the Nordic Biological Specimen Banks for Cancer Causes and Control (NBSBCCC) network. Status as of June 2005**

| Name, country (contact address) | Type | Years of first serum donation and subsequent samples of same individuals | Number of persons (+annual increase) | Number of sampling occasions (+annual increase) | Closing year in this study (complete cancer incidence & vital status) | Number of person-years |
|---|---|---|---|---|---|---|
| Finnish Maternity Cohort (helja-marja.surcel@ktl.fi) | S | 1983+ | 722,000 women (August 2005) (+30,000/year) | 1.47 million (+60,000/year) | 2005 | 9.71 million |
| Helsinki Heart Study, Finland (leena.tenkanen@ktl.fi) | S | 1980–1982 | 18,900 men | 117,000 | 2005 | 419,000 |
| Alpha-Tocopherol-Beta-Carotene (ATBC) Study, Finland (jarmo.virtamo@ktl.fi) | S | Baseline sera 1984–1988 | 29,200 male smokers | 55,000 {follow-up sera from all 1986–1993, annual sera from 800 men} | 2005 | 423,000 |
| Finnish Mobile Clinic Health Examination Survey (paul.knekt@ktl.fi) | R | 1966–1976 | 50,400 | 60,000 | 2005 | 1.42 million |
| FINRISK, Finland (pekka.jousilahti@ktl.fi) | R | 1992/1997/2002 (samples from years 1972/1977/1982/1987 incomplete) | 22,900 | 22,900 | 2005 | 185,000 |
| Icelandic Maternity Cohort (arthur@landspitali.is; Arthur Löve) | S | 1980+ | 53,000 women (+1,700/year) | 96,000 (+2,500/year) | 2005 | 768,000 |

(continued)

**Table 2 (continued)**

| Name, country (contact address) | Type | Years of first serum donation and subsequent samples of same individuals | Number of persons (+annual increase) | Number of sampling occasions (+annual increase) | Closing year in this study (complete cancer incidence & vital status) | Number of person-years |
|---|---|---|---|---|---|---|
| Reykjavik Study, Icelandic Heart Association (v.gudnason@hjartavernd.is; (Vilmundur Guđnason)) | R | 1967–1996 | 19,300 | 60,000 | 2005 | 457,000 |
| Janus, Norway (randi.elin.gislefoss@kreftregisteret.no) | | 1972–2005 | 332,000 | 493,000 | 2006 | |
| Health examinations | R | 1972–1978, 1980–1992 (+ few from Finmark and Troms counties in 2002) | 304,000 | 377,000 | 2001 | 6.96 million |
| Blood donors | S | 1973–1991, 1998–2000 | 31,900 | 117,000 (last samples 2005) | 2001 | 1.27 million |
| *Sweden* | | | | | | |
| Northern Sweden Health and Disease Study (goran.hallmans@nutrires.umu.se ) | R | 1985+ | 86,000 | 114,000 | 2003 | |
| VIP | R | 1985+ | 70,000 (+2000/year) | 83,000 | 2003 | 560,000 |
| MONICA | R | (1986)/1990/1994/1999/2004 | 9,000 | 14,000 | 2003 | 51,000 |
| Mammography | R | 1995+ | 27,500 women (+1,500/year) | 48,000 | 2003 | 158,000 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Northern Sweden Maternity Cohort (goran.wadell@climi.umu.se) | S | 1975+ | 86,000 women (+2,000/year) | 118,000 samples | 2003 | 1.24 million |
| Preventive Medicine in Malmö, Sweden (goran.berglund@medforsk.mas.lu.se) | R | 1974–1991 | 33,000 | 8,000 | 1999 | 560,000 |
| Malmö Diet and Cancer, Sweden (goran.berglund@medforsk.mas.lu.se) | R | 1991–1996 | 29,100 | – | 1999 | 159,000 |
| Malmö Microbiology, Sweden (joakim.dillner@med.lu.se) | | 1986+ (incomplete 1969+) | 454,000 (+40,000/year) | 1.24 million (+120,000/year) | 1999 | 1.84 million |
| Malmö Maternity Cohort | S | 1985, 1989+ (incomplete 1969+) | 70,000 women | 115,000 | 1999 | |
| Bloodborne virus screening | C | 1986+ | | | 1999 | |
| Other virus testing | C | 1990+ | | | 1999 | |
| Swedish Institute for Infectious Disease Control (joakim.dillner@med.lu.se) | | 1957+ (complete 1977+) | 358,000 in computerised files | >900,000 (629,000 computerised) | 2003 (test) | – |
| Population sample | R | 1968, 1977–1978, 1990–1991, 1997 | 12,000 | 12,000 | 2003 (test) | – |
| Diagnostic microbiological testing | C | 1990+ | 346,000 | 617,000 computerised | 2003 (test) | – |

Type: *R* random sample of population or other systematic invitation based on population register, *S* specific group with clearly defined enrolment criteria, *C* Clinical samples

Prevention Study, the Northern Sweden Health and Disease Study, the Malmö Diet and Cancer study and FINRISK Study (since 1997), which are stored at –70°C. Malmö Diet and Cancer biobank also has aliquots stored at –135°C.

Every resident of the Nordic countries has a unique personal identification code (PID) that is used in all main registers in these countries. The PID allows automatic and precise linkage of registers, without the need to use names. For meaningful research use, the PIDs have to be available for each person in the biobanks. Biobank cohorts are typically linked with the population-based cancer registries shortly before a new case-control set will be extracted for a specific study.

### 2.1. Finnish Maternity Cohort

Sera collected during the first trimester of pregnancy (two-thirds at 8–12 weeks) for screening of congenital infections and rubella immunity have been stored since late 1983 by the National Institute for Health and Welfare (THL). The biobank covers more than 98% of all pregnant women in Finland. So far, basic data for the sera up to 21 August 2005 have been transferred to Finnish Cancer Registry to be used in case-control studies.

Up to about 2005, The Finnish Cancer Registry took care of quality control of the data and also developed programs for precise random case-control selection within the FMC cohort. Record linkages for both incident cancers with cancer registry data and causes of death through Statistics Finland were administered from the Finnish Cancer Registry. This biobank has been used in more NBSBCCC studies than any other (Table 1). In the latest years, necessary quality assurance and record linkage routines have been developed at THL and no external consultancy is needed any more.

### 2.2. Helsinki Heart Study

The sera were collected during 1980–1982 for a trial to test the hypothesis that lowering serum LDL-cholesterol and triglyceride levels and elevating serum HDL-cholesterol levels with gemfibrozil (a fibric acid derivative) reduces the incidence of coronary heart disease (CHD) in middle-aged dyslipidaemic men (34). The volunteers for the trial were selected from men aged 40 through 55 years, employed by two government agencies and five industrial companies and living in different parts of Finland. Approximately 19,000 men participated in the first screening, and to be selected for the trial the participants ($N$=4,081) had to have their non-HDL cholesterol ≥5.2 mmol/l and no evidence of CHD or other major diseases.

Serum samples were collected from the participants at the first screening and from the participants in the trial at each follow-up visit during the trial. As the participants were followed up four times per year during the 5-year trial and twice a year during a subsequent extension of the trial, there are 28 serial samples from about 3,500 of those 4,081 who initially attended the trial. Also

from the last trial follow-up visit, blood samples were stored. This biobank has participated in numerous NBSBCCC studies since 1999 (Table 1).

### 2.3. Alpha-Tocopherol-Beta-Carotene Cancer Prevention Study

The Alpha-Tocopherol-Beta-Carotene (ATBC) study was a randomised, double-blind, placebo-controlled, primary prevention trial conducted in Finland by the National Institute for Health and Welfare in collaboration with the U.S. National Cancer Institute. The main aim of the study was to evaluate whether daily supplementation with alpha-tocopherol or beta-carotene would reduce the incidence of lung cancer and other cancers (35).

In 1985–1988, a questionnaire on current smoking and willingness to participate in the trial was sent to the total male population of 50–69 years living in south-western Finland ($n = 290{,}000$). Of them, 43,000 men smoked at least five cigarettes per day and were willing to participate. Men with prior cancer (except non-melanoma skin cancer and carcinoma-in-situ), severe angina pectoris, chronic renal insufficiency, alcoholism, or liver cirrhosis as well as those taking anticoagulants, beta-carotene, or vitamin A/E supplements in excess of defined doses were excluded. After exclusions and written informed consent, 29,133 eligible men were randomly assigned to receive either alpha-tocopherol 50 mg per day, or beta-carotene 20 mg per day, or both alpha-tocopherol and beta-carotene, or placebo.

At baseline, serum samples were collected. New serum samples were collected from all participants at the 3-year follow-up visit, and from about 800 randomly selected men a serum sample was collected annually throughout the trial. A whole blood sample was collected from the participants at the end of the trial between August 1992 and April 1993. This biobank was used for the first time in an NBSBCCC study just lately (33).

### 2.4. Finnish Mobile Clinic Health Examination Survey

The Mobile Clinic Health Examination Survey was carried out by the Social Insurance Institution during 1966–1972 in 34 rural, industrial or semiurban subpopulations (Fig. 1). Total populations aged 15 years or older or random samples of them were invited to participate in the study. On average 83% (57,400 men and women) participated in the health examination. Blood samples have been stored from 40,200 individuals in the baseline examination and from all 19,500 individuals in the re-examination survey of 12 subpopulations four to seven years later (1973–1976). This biobank participated particularly in early NBSBCCC studies (Table 1).

### 2.5. FINRISK

The National FINRISK Study has been conducted in Finland every 5 years since 1972. At the beginning, the Study was done only in eastern Finland as part of the North Karelia Project. The study area was expanded gradually. The serum samples are systematically available since 1992. In 1992, the Study was carried out in four areas: North Karelia and Kuopio Provinces in Eastern Finland,

Turku-Loimaa region in Southwest Finland, and cities of Helsinki and Vantaa in Southern Finland (Fig. 1). Oulu province in Northern Finland was included in 1997 and Lapland province in 2002. In each study year, a random sample of 2,000 individuals aged 25–64 years (stratified by sex and 10-year age group) has been taken in each study area according to the WHO MONICA protocol. Since 1997, a sub-sample of 1,500 men and women aged 65–74 years was included. Total cumulative sample size since 1992 is 33,000 and of them 22,900 (69%) have participated in the Study. DNA samples are available for most participants.

Study cohorts have been followed up through computerised register linkage of the National Causes of Death Register, the Hospital Discharge Register and the Finnish Cancer Register. The samples of the FINRISK Study have not been used in any NBSBCCC studies so far, but the general principle of the Study is that the collected samples can be utilised in large-scale collaborative studies that according to the FINRISK Steering Group are scientifically important.

### *2.6. Icelandic Maternity Cohort*

Sera generally collected at 12–14 weeks of pregnancy for rubella screening from all of Iceland have been stored since 1980 in the centralised Department of Medical Virology, Landspitali University Hospital. About 6% of the cohort members cannot be used in studies because they have moved out of the country, but the date of emigration is not registered. This biobank has participated in two NBSBCCC studies (Table 1).

### *2.7. Icelandic Heart Association, the Reykjavik Study*

The Reykjavik Study by the Heart Preventive Clinic and Research Institute of the Icelandic Heart Association is a prospective cardiovascular cohort study carried out in the Reykjavik capital area in 1967–1996. Selected birth cohorts of 14,923 men and 15,872 women in the Reykjavik area born in 1907–1935 were divided into six equally sized subgroups according to the date of birth and recruited systematically for collection of sera. The first subgroup was recruited in 1967–1969 and has attended altogether six times. The second one (first invited in 1970–1972) has attended twice. The later birth cohorts have been invited once (1974–1996) or never. Altogether 19,300 persons actually provided samples (annual participation rates between 71% and 76%), but about 200 of them cannot be used in analyses because of lacking dates of emigration. This biobank has not yet been used in any of the published NBSBCCC studies.

### *2.8. Janus Project (Norway)*

A project to collect and store blood samples from healthy persons for later scientific use was initiated in the 1960s and named Janus after the Roman god with two faces, one looking backward, and the other one looking forward (symbolising the retrospective and prospective directions of epidemiological research). The first collection, related to a survey of risk factors for cardiovascular disease

in ages of 35–49 years, covered four counties (Oslo 1972–1973, Finmark 1974–1975, Sogn og Fjordane 1975–1976 and Oppland 1976–1978; see Fig. 1). More subjects were added during 1985–1992 in the context of cardiovascular health examination of 40–42 years old Norwegians from all of the country except two counties (Hordaland and Buskerud; Fig. 1).

Red Cross blood donors in capital Oslo and surrounding areas were enrolled in 1973–1991 and 1999–2000. Every second year, these Janus donors donated 20 ml of extra blood to the biobank. Collection of later samples from these individuals ended in spring 2005.

The Janus bank consists of serum samples from 331,801 persons, 10% of them Red Cross donors. The average is two to three samples per donor, but some donors have given samples more than ten times. The Janus biobank is also collecting follow-up samples from cohort members who develop cancer. Before any treatment, a sample is collected when the donor is hospitalised at the Radium Hospital in Oslo (a nationally centralised cancer treatment hospital).

The Janus Project is funded by the Norwegian Cancer Registry which is also responsible for the data handling. This also allows frequent updates for incident cancer cases; several thousands of new prospective cancer cases have been registered after the closing date used in this study, and the addition in 2004 exceeded 3,000. Samples of the Janus health examination cohort have been used in 20 NBSBCCC publications, and the blood donors' sera in eight studies (Table 1).

### *2.9. The Northern Sweden Health and Disease Study Cohort*

The Northern Sweden Health and Disease Study (NSHDS) Cohort contains three subcohorts: the Västerbotten Intervention Program (VIP), the MONICA (Monitoring Trends and Determinants in Cardiovascular Disease) and the Mammography Screening in Västerbotten. The cohorts represent a population-based sample of the county of Västerbotten in Northern Sweden (254,000 inhabitants). The Monica study also contains a population-based sample from the adjacent county of Norrbotten.

The VIP is a long-term project intended for health promotion. Since 1985, all individuals of 40, 50 and 60 years of age are invited for screening. They are also asked to donate a blood sample for later research purposes. In June 2004, the cohort included 74,000 individuals, of whom 70,000 had donated blood. A second sample is taken after 10 years; this has produced 13,000 resampling occasions.

Samples taken in the context of the population-based mammography screening have been stored since 1995. Screening is done every second year among all women in the age group 50–69 years in the county. There have been 48,000 sampling occasions from 27,500 women. About 50% of the women in the mammography cohort have also attended VIP.

The Northern Sweden MONICA project contains material from population-based screenings for risk factors of cardiovascular diseases that were carried out in 1986, 1990, 1994, 1999 and 2004. There are 14,000 sampling occasions of 9,000 individuals, 50% of whom are also included in VIP. Samples from 1986 have not been used in NBSBCCC studies and they are not included in this standardised incidence ratio (SIR) analysis, either.

The VIP cohort has been used most frequently out of the numerous Swedish biobanks in NBSBCCC studies, and also MONICA cohort in 15 studies (Table 1).

### 2.10. Northern Sweden Maternity Cohort

Northern Sweden Maternity cohort consists of sera collected since 1975 from pregnant women screened for rubella immunity during week 14 of pregnancy in the Västerbotten county and especially in the 1980s also for some of the adjacent counties in Northern Sweden. So far, almost 120,000 samples from 86,000 women have been stored at the virus laboratory of Umeå University. This biobank has not yet been used in any of the published NBSBCCC studies.

### 2.11. Preventive Medicine in Malmö, Sweden

The prospective, population-based Preventive Medicine study, with main focus on cardiovascular disease, diabetes and cancer, includes sera from a population-based sample of 33,400 persons 40–60 years of age, resident in the city of Malmö. The samples were donated at baseline examination in 1974–1991. The biobank is owned by Lund University.

### 2.12. Malmö Diet and Cancer Study, Sweden

The prospective population-based Malmö Diet and Cancer study started with a baseline examination in 1991–1996. Main focus is on cancer and cardiovascular diseases. All men born between 1923 and 1945 and all women born between 1923 and 1950 living at the time in the city of Malmö were invited to participate. The participation rate was 40% (28,100 participants). Mean age at enrolment was 58.2 years. The biobank is owned by Lund University.

### 2.13. Malmö Microbiology Biobank, Sweden

The Malmö Microbiology Biobank is owned by the County Council of Skåne and contains samples submitted for clinical microbiological analyses to the University Hospital in Malmö that today serves the entire county of Skåne in southernmost Sweden. Samples have been saved for clinical diagnostic and documentation purposes, the majority of them taken for diagnosis of bloodborne viral infections, such as hepatitis viruses. The oldest samples are from 1969 and were submitted from the city of Malmö. The annual number of samples increased in 1986 when HIV testing started and the catchment area extended to cover most of the Skåne county (Fig. 1). Since 1990, also the samples submitted for virus serology (typically because of clinical suspicion of virus infection or desire to investigate viral immunity) have been stored.

In recent years, a large number of samples have been submitted from the microbiology laboratories of adjacent counties in southern Sweden (Blekinge and Halland), raising the annual number of samples added to the biobank to about 60,000.

The Malmö Microbiology Biobank also includes samples of the population-based serological screening for virus infections and rubella immunity during pregnancy scheduled to be taken during week 14 of pregnancy (Malmö Maternity Cohort). The maternity cohort contains all samples from 1986 and from 1989 onwards, altogether more than 100,000 samples from 74,000 mothers.

Malmö Microbiology Biobank was computerised in 1997. NBSBCCC studies with MMB participation have as yet not been published.

### *2.14. Swedish Institute for Infectious Disease Control Biobank*

The Swedish Institute for Infectious Disease Control (SIIDC) has performed a series of population-based, nationwide investigations on the immunity against infections in the Swedish population.

A small fraction of the biobank consists of randomly selected persons sampled in 1968 (3,000 subjects), 1977–1978 (1,845), 1990–1991 (4,800) and 1997 (2,400) and analysed to estimate age-specific population immunity rates of, e.g., polio, parotitis, measles, rubella, diphtheria and tetanus.

Most of the about 900,000 biological samples in the SIIDC biobank are diagnostic ones, submitted for microbiological analyses from all over Sweden. The oldest stored samples are from 1957, and complete series exist since 1977.

The information on the samples has been transferred from paper documentation to computerised files for about 629,000 samples. The biobank has recently been linked with the Swedish Cancer Registry, and the quality control of the result of the linkage is on-going. Samples of the Swedish Institute for Infectious Disease Control have been utilised in one NBSBCCC study (Table 1).

## 3. Some Quality Control Tools for Study Cohorts

The same methods that are used to check accuracy of any new study cohort, e.g., in occupational cancer epidemiology, can and should also be used for biobank cohorts. The following types of evaluations – presented below as a cookbook type list – were made for the NBCBCCC biobanks systematically in the context of a specific quality assurance study (34). For some of the biobank that kind was never done before and numerous gaps in the quality were revealed (and corrected).

### *3.1. Irregularities in the Entry?*

The first tabulation to control completeness of any study cohort is to count numbers of cohort members by gender, year of entry

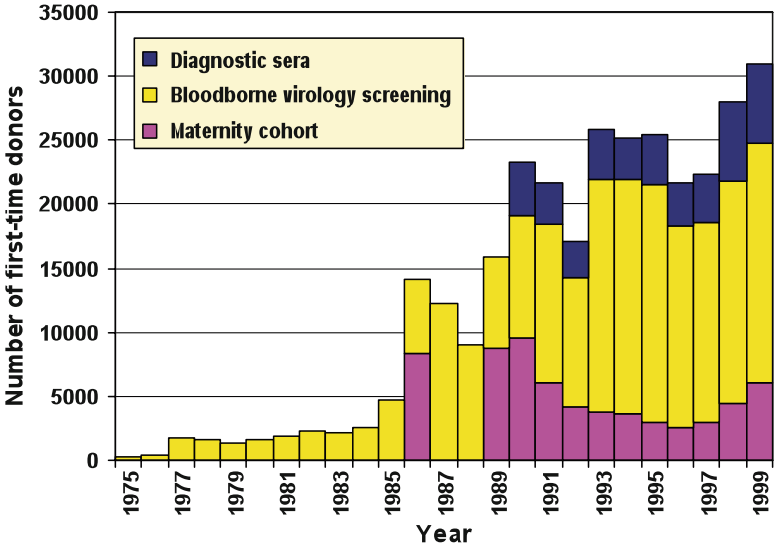


Fig. 2. Annual numbers of first-time donors in the Malmö Microbiology Biobank 1975–1999, by subcohort.

and similar simple classifiers. The numbers should correspond to the known facts about the size of the subjects. Any irregularity in the time series should be documented in the history of cohort formation.

For instance, the distribution of the number of pregnant women in Malmö Maternity Cohort (Fig. 2) reveals that the samples from 1987 and 1988 have been destroyed to save storage space. It has been quite common that such tabulations reveal differences between what has been believed to be the historical coverage of a biobank and what is the actual one. It is better to know this type of discrepancies *before* designing a biobank-based study than *after* the laboratory analyses have been done.

### 3.2. Are the End-of-Follow-Up Data Complete?

It is crucial to know for how long the cohort members are at risk. Therefore, information on vital status and emigration should also have been obtained for every cohort member. The simplest tabulation to control completeness of follow-up data is to count the annual numbers of deaths (and emigrations) of the cohort members by year of death (or emigration). Figure 3 demonstrates two such trends for real biobank cohorts and two artificial situations demonstrating problems in follow-up of vital status.

Annual numbers of deaths among the 722,500 women in the Finnish Maternity Cohort (FMC) are very small in the 1980s but increase heavily during the present millennium. This is expected because the women who were pregnant in the beginning of the biobank collection in the 1980s now gradually reach ages when the mortality among women starts to increase. The dynamic nature of the FMC (new women join the cohort every year) also increases the number of annual deaths.

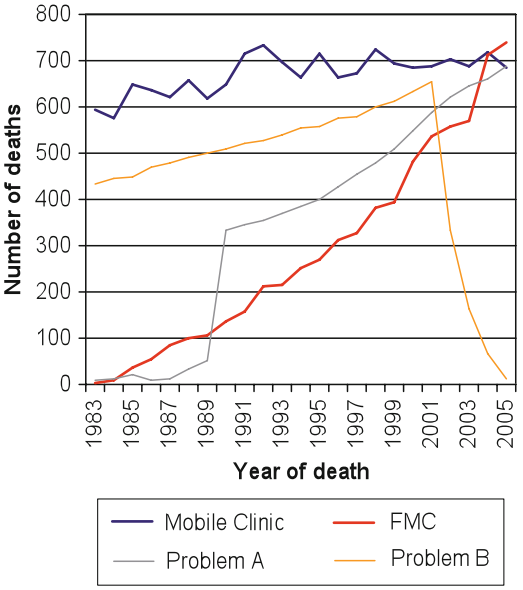


Fig. 3. Annual numbers of deaths among the 722,500 women in the Finnish Maternity Cohort (FMC), and the 50,400 persons in the Finnish Mobile Clinic Health Examination Survey, and two artificial situations demonstrating problems in follow-up of vital status (*see* text).

The trend of the annual numbers of deaths among the 50,400 persons who were 15+ years old when they participated in the Finnish Mobile Clinic Health Examination Survey in 1966–1972 is quite stable over the years. This is a correct trend in a cohort with large amount of old people. The number of persons alive in the cohort simply decreases so quickly that despite the strongly increasing relative mortality rate of the remaining cohort members the absolute numbers of annual deaths start to decrease.

In numerous newly collected study cohorts, we see a mortality trend demonstrated as "Problem A" in Fig. 3: there is a very small mortality in the beginning of follow-up. The reason for this problem is that some old samples (or their records) have been destroyed, if the person had died, to save space. Often this type of deletion is not documented and may (or may not) be revealed as too low mortality rates in the first years of follow-up.

Trend curve "Problem B" in Fig. 3 demonstrates another common problem in follow-up of vital status: the number of deaths decreases in the most recent years. This happens when the systematical follow-up for vital status via national death register files has not been done.

### *3.3. Are the Person Identification Data Accurate?*

Every resident of the Nordic countries has a unique personal identification (PID) code that is used in all main registers and makes computerised linkages accurate and effective (36). The identification data of each biobank cohort member should have been compared with the national Population Register data to check that

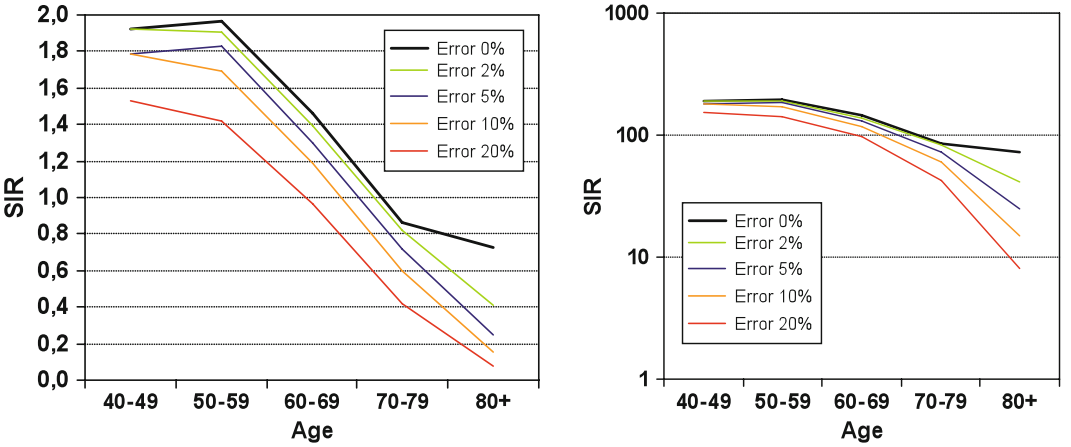


Fig. 4. Example of the effect of error in the identifier (*link key*) to the relative risk estimate. Standardised incidence ratio (SIR) for cancer (all sites combined) during 1953–2005 among 750 male workers of an anthophyllite asbestos mine, by age. The correct SIRs are indicated with "Error 0%" line; the other lines demonstrate situations when part of the cancer cases and deaths of the cohort members are missed because of failure in the person identifier data randomly produced in the cohort.

the personal identifiers are the correct ones and persons really exist in the population. In the same occasion, information on vital status and emigration can be obtained for every cohort member, but this information needs to be updated regularly (see Subheading 3.2).

The bias related to failures in record linkage with vital status and cancer very much increases along with increasing age at follow-up. Because the biobank cohorts are still quite young, the example is taken from another type of cohort, namely workers of an old asbestos mine in Finland (36). Figure 4 illustrates effect of error in the identifier (link key) to the SIR for cancer among male anthophyllite asbestos miners. The true SIR (all ages combined) during 1953–2005 is 1.35 (95% confidence interval 1.17–1.55). If there would be an error in 2% of the identifiers (at random), the observed number of the cases would decrease by 2% but the expected number would increase by 8% because of missing death information and subsequent addition of person-years at risk in the oldest age groups. The SIR related to 2% linkage error would be 1.23 (1.06–1.41), to 5% linkage error 1.06 (0.91–1.22), to 10% linkage error 0.87 (0.75–1.01) and to 20% linkage error 0.62 (0.52–0.72). Hence, a highly significant cancer risk related to asbestos mining would look like a significant protective effect if about 10% of the person IDs would be incorrect. In some of the older biobanks originally collected for non-scientific use, the proportion of incomplete IDs may have been several percentages. If these errors would not have been corrected, they would in long run have had serious effects on the results based on those samples.

The possibility of such linkage errors when the events of a *different* persons are linked together would be connected is not a major issue in the Nordic countries where all linkages nowadays are based on the unique PIDs. Although it is possible that a registered person may get a PID of another person, this event is so rare that it hardly affects conclusions of any study. There are examples of older times when the linkage was done manually based, e.g., on name, date and place of birth, place of residence. The linkage failures were much more common than in the later automatic linkages based on PIDs (37).

### 3.4. Standardised Incidence Ratios as Tools of Cohort Quality and Representativeness

The SIR is a useful tool to reveal occasions when the cohort members do not represent typical risk situation of the base populations. Although in nested case-control studies, it is not required that the cases and controls are selected from a population with average population risk level, understanding of the baseline risk level is important if the relative risk estimates are generalised to population attributable risk fractions. There are also situations when cohort members have temporally very special conditions that modify their risk level in a way that is not be easy to take into account in the RR analyses (see examples below).

The cancer cases among the serum donors included in the NBSBCCC serum banks have been traced through automatic record linkages with the national Cancer Registries.

In the person-year calculation needed for calculation of expected numbers of cancer cases, the follow-up starts at the date of first serum donation and ends at death, emigration or on the general closing date (depending on the lag of national cancer registration), whichever is first. Because the dates of emigration are not known in the Icelandic biobanks, about 4,000 emigrated persons of the Icelandic biobanks have to be excluded.

The numbers of observed cases and person-years at risk are counted for each calendar year, by gender and five-year age group. Sometimes it is useful to make further stratification according to the time elapsed since the sample donation. The expected numbers of cases for total cancer and for selected specific cancer types in the following examples were calculated by multiplying the number of person-years in each stratum by the corresponding cancer incidence rate in the national population, but sometimes regional cancer incidence rates may be a more informative reference.

The SIR was defined as the ratio of the observed to expected number of cases. The 95% confidence intervals (CI) for the SIR were based on the assumption that the number of observed cases followed a Poisson distribution.

#### 3.4.1. Selection Related to the Indication of the Sampling?

Sometimes factors related to the reason of the serum donation may make the cohort temporally quite different from average of the baseline population it presents. In screenings of random samples

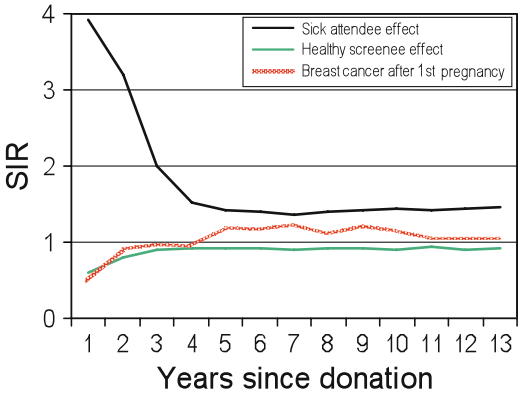


Fig. 5. Examples of biases related to the indication of serum donation that can be studied via trends of the standardszed incidence ratio (SIR) stratified by time elapsed since serum donation.

of population, incidence and mortality of chronic diseases tends to be decreased during the first months or years after the baseline study. This bias, illustrated with the "healthy screenee effect" curve in Fig. 5, is related to the selective participation: those who have severe early symptoms of a disease participate less frequently than the other people.

In the biobanks including samples of symptomatic persons, part of the symptoms may actually reveal to be symptoms of the outcome disease and therefore the SIR of that disease is very high soon after the serum donation. In the situation illustrated by the "sick attendee" curve in Fig. 5, the risk level is stabilised to the normal level of the base population of the cohort after about 5 years. If one would design a study within that cohort, it would be safest to exclude cases diagnosed during the five first years. If these cases would be included, there would be a risk of "reverse causality bias" (see Chapter 5): the hidden disease may have affected the values of biological parameters at baseline.

The third curve in Fig. 5 describes another atypical risk pattern, namely the "dual effect" of the pregnancy to the risk of breast cancer (38). The real-data example is taken from the Finnish Maternity Cohort. The SIR for breast cancer is first low but there is peak of increased risk some years after the pregnancy before the protective effect of the pregnancy starts to decrease the risk.

### *3.4.2. What Does Cancer Incidence Pattern Reveal of the Biobank Cohort?*

The following observed and expected numbers of cancers are based on altogether 1.95 million subjects under follow-up in the 17 biobank cohorts, which were ready to produce person-years at risk calculations. The accumulated number of person-years from the date of first donation until the closing date (1999–2006, depending on the biobank) was 29.3 million (Table 2). The mean length of follow-up of a person was 13.4 years and the

longest follow-times almost 40 years. The number of malignant cancer cases diagnosed between sampling and closing date exceeds 130,000.

The above numbers exclude the subjects from the Swedish Institute for Infectious Disease Control biobanks and those donors from other biobanks who donated their first sample after the closing date, altogether more than one million donors.

The specific cancer types selected a priori for the analysis (see Table 3) included cancer sites with known risk factors that reveal deviating risk behaviour among the cohort members, and other common cancer types selected to give a representative picture of the cancer situation among the cohorts.

#### 3.4.2.1. Biobanks Based on Invitation of the General Population

The observed number of malignant neoplasms among persons (6,219) in the prospective cardiovascular *Reykjavik Study* exceeded slightly the expected rates based on Icelandic national rate, similarly in both genders, yielding an SIR of 1.06 (95% CI 1.03–1.08) (Table 3). Men had significantly elevated incidence of cancers of the prostate (SIR 1.13; 1.06–1.19) and kidney (1.23; 1.04–1.43) and significantly low risk of lip cancer (SIR 0.59; 0.35–0.94). Women had significant excess risk of ductal carcinoma of the breasts (SIR 1.17; 1.06–1.27) and leukaemia (1.29; 1.02–1.61). Incidence of malignancies of unknown primary was lower than in the general population of Iceland (SIR 0.82; 0.68–0.98). From the malignancies not included above, basal cell carcinoma of the skin showed elevated incidence (SIR 1.26; 1.18–1.34). The low lip cancer rate indicates a low proportion of high-risk categories, namely farmers and fishermen (39) that was expected because the cohort represents city people. All elevated SIRs are in cancers that are most common among urban populations.

There were 7,754 malignant neoplasms diagnosed between serum donation (1966–1972) and 31 December 2005 among the 50,448 subjects of the *Finnish Mobile Clinic Health Examination Survey* for whom serum sample is available. The SIR for all cancers combined was 0.94 (95% CI 0.92–0.95), similarly in both genders. Incidence of cancers of the genital organs is significantly below the national average: prostate (0.93; 0.88–0.98), penis 0.50 (0.21–0.97), breast (0.90; 0.84–0.95), cervix uteri (0.75; 0.57–0.97) and Fallopian tube (0.33; 0.07–0.97). Only penile cancer showed an SIR above 1.0 (1.98; 1.17–3.12). From the other cancers low SIRs were seen in adenocarcinoma in lungs (0.80; 0.68–0.93), liver cancer (0.78; 0.62–0.95). In males but not in females, there was a low SIR in non-Hodgkin lymphoma (0.81; 0.68–0.95) and cancer with unknown primary site (0.74; 0.58–0.92).

The total number of cancer cases among the *FINRISK* study members who have donated serum in 1992, 1997 or 2002 was 1,104 (SIR 0.97; 0.92–1.03). There was an excess of prostate cancer of localised stage (SIR 1.26; 1.05–1.48) but no excess

**Table 3**
**Numbers of observed (O) and expected (E) cancer cases diagnosed between first serum donation (1,967+) and 31 December 2005 among the 19,257 participants of the cardiovascular Reykjavik Study. Expected numbers based on national population; standardised incidence ratios (SIR = O/E) given with 95% confidence intervals (CI). Statistically significant SIRs are in bold**

| ICD-7 | Cancer site | O | E | SIR | 95% CI |
|---|---|---|---|---|---|
| 140–207 | All malignant neoplasms | 6,219 | 5874.45 | **1.06** | **1.03–1.08** |
| 140 | Lip | 20 | 33.18 | **0.60** | **0.37–0.93** |
| 143–144 | Oral cavity | 14 | 13.13 | 1.07 | 0.58–1.78 |
| 145–148 | Pharynx | 31 | 34.03 | 0.91 | 0.62–1.29 |
| 150 | Oesophagus | 81 | 90.25 | 0.90 | 0.71–1.11 |
| 151 | Stomach | 347 | 349.52 | 0.99 | 0.89–1.10 |
| 153 | Colon | 523 | 494.65 | 1.06 | 0.97–1.14 |
| 154 | Rectum | 184 | 170.57 | 1.08 | 0.93–1.24 |
| 155 | Primary liver | 46 | 46.19 | 1.00 | 0.73–1.32 |
| 155.1 | Gall-bladder, biliary tract | 45 | 44.29 | 1.02 | 0.74–1.35 |
| 157 | Pancreas | 188 | 180.37 | 1.04 | 0.90–1.19 |
| 161 | Larynx | 34 | 41.55 | 0.82 | 0.57–1.14 |
| 162–163 | Lung | 790 | 757.58 | 1.04 | 0.97–1.11 |
| 170 | Breast | 714 | 635.16 | **1.12** | **1.04–1.20** |
| 171 | Cervix uteri | 41 | 51.46 | 0.80 | 0.57–1.08 |
| 172 | Corpus uteri | 149 | 139.12 | 1.07 | 0.91–1.24 |
| 175 | Ovary | 101 | 120.06 | 0.84 | 0.69–1.01 |
| 177 | Prostate | 1,013 | 900.18 | **1.13** | **1.06–1.19** |
| 180 | Kidney | 221 | 201.39 | 1.10 | 0.96–1.24 |
| 181 | Bladder | 353 | 315.92 | **1.12** | **1.00–1.23** |
| 190 | Melanoma of the skin | 83 | 79.16 | 1.05 | 0.84–1.29 |
| 191 | Non-melanoma skin | 206 | 202.07 | 1.02 | 0.88–1.16 |
| 193 | Brain and nervous system | 165 | 151.71 | 1.09 | 0.93–1.25 |
| 194 | Thyroid | 123 | 107.61 | 1.14 | 0.95–1.35 |
| 200,202 | Non-Hodgkin lymphoma | 141 | 139.51 | 1.01 | 0.85–1.18 |
| 204 | Leukaemia | 164 | 136.20 | **1.20** | **1.03–1.39** |
| 199 | Unknown site | 116 | 140.77 | **0.82** | **0.68–0.98** |
| *Not included above* | | | | | |
| Basal cell carcinoma of the skin | | 824 | 653.77 | **1.26** | **1.18–1.34** |

of non-localised prostate cancer. Also diagnosis of basal cell carcinoma of the skin was more common than in the general population (SIR 1.18; 1.05–1.32). Cancer of gallbladder was rare (SIR 0.33; 0.07–0.97).

Persons participating in health examinations in Norway (and allowing use of their sera for anonymous cancer research) in the *Janus biobank* cohort had less incidence of cancer than the general Norwegian population (21,889 cases observed by the end of 2001 vs. 24,086 expected). Significantly decreased SIRs were observed for cancers of the oral cavity (0.84; 95% CI 0.70–1.00), pharynx (0.77; 0.65–0.92), oesophagus (0.83; 0.70–0.97), primary liver (0.77; 0.61–0.95), lung (0.87; 0.83–0.90) and cervix uteri (0.82; 0.75–0.90), i.e., cancers related to high alcohol consumption and generally way of life not directed to healthy habits. None of the SIRs was significantly elevated.

In the *Malmö Diet and Cancer Study* cohort, there were 1,852 cancer cases, while the expected number based on incidence rates of the entire Swedish population was 1,568 (SIR 1.18; 1.13–1.24). This significant excess was mainly attributable to excesses in prostate cancer (84 excess cases, SIR 1.40; 1.25–1.57), breast cancer (59 excess cases, SIR 1.22; 1.09–1.38), skin melanoma (39 excess cases, SIR 1.72; 1.39–2.11) and bladder cancer (30 excess cases, SIR 1.42; 1.16–1.72). There were no significantly decreased SIRs in the cohort. This cancer incidence pattern is typical to a cohort representing population from southernmost Sweden with rates often more similar to the Danish cancer incidence rates than the Swedish average (Fig. 6).

The other invitational Southern Swedish cohort, that of the *Preventive Medicine in Malmö* project, produced more cancers (4,343), but the SIR was similar (1.17; 1.13–1.20). The pattern of cancer sites with increased incidence was partly similar to that of Malmö Diet and Cancer Study – breast cancer (SIR 1.24; 1.13–1.36), bladder cancer (1.46; 1.30–1.63), and skin melanoma (1.33; 1.16–1.53) – but some other cancers also had increased SIRs: lung cancer 1.48 (1.36–1.61), laryngeal cancer 1.41 (1.02–1.89), pharyngeal cancer 1.56 (1.14–2.08) and pancreatic cancer 1.23 (1.02–1.48). Despite the large numbers of cases, none of the 22 primary sites studied separately showed an SIR significantly below unity.

The *Northern Sweden Health and Disease Study* consists of three cohorts randomly selected from the population of given ages in that region. The largest number of cancer cases (2,426) was found among members of the Västerbotten Intervention Program (VIP). The expected number was slightly higher (2,531). The SIR was significantly decreased for lung cancer (0.82; 95% CI 0.68–0.98); otherwise, there were no major aberrations from 1.0.

There were 289 cancer cases in the smaller MONICA cohort as compared to 310 cases expected. The difference is not significant, and none of the site-specific SIRs was significantly different from unity.

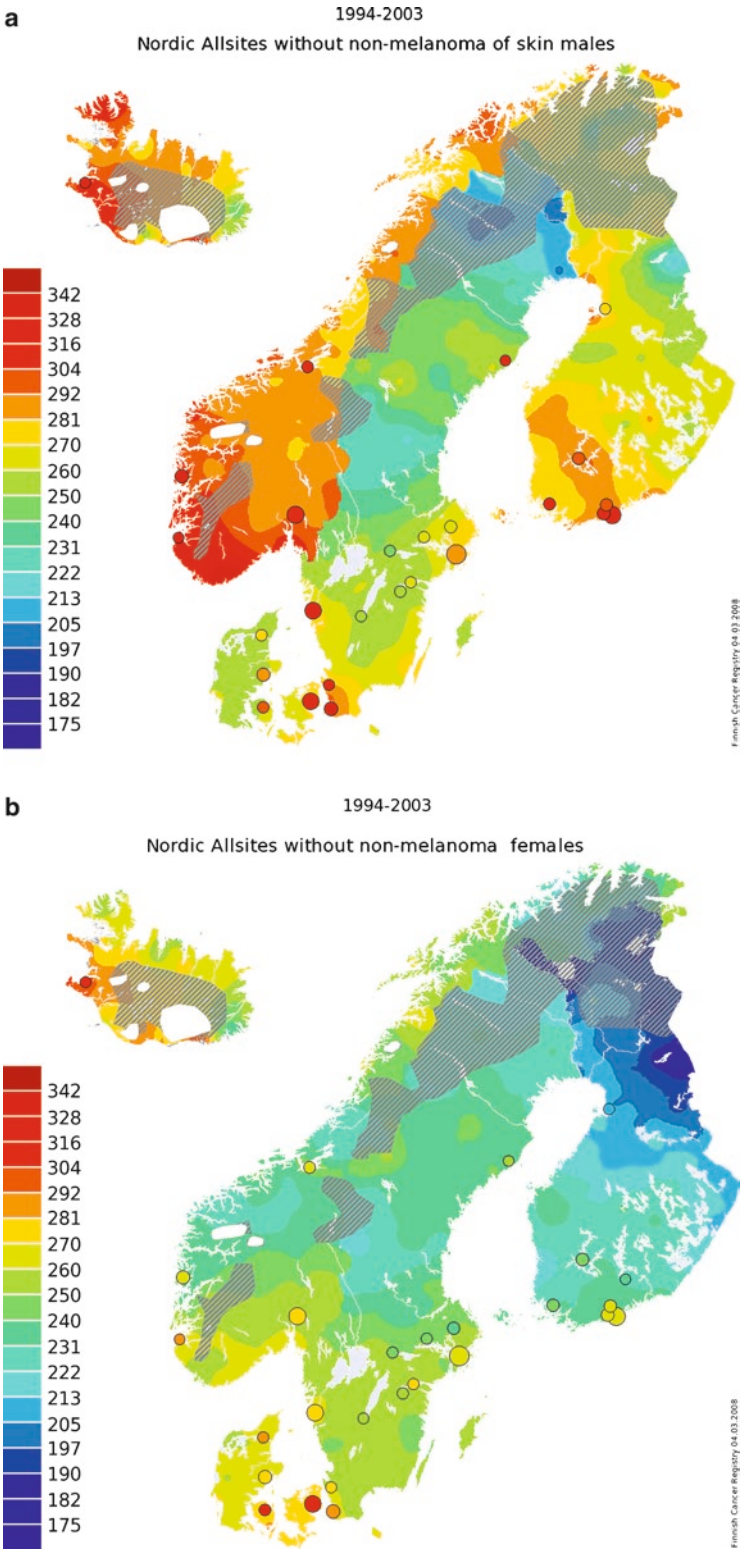


Fig. 6. Spatial variation of age-adjusted incidence rates per 100,000 of cancer (all sites, excluding non-melanoma skin cancer) in the Nordic countries, 1994–2003. For mapping method, *see* (57).

The mammography screening cohort (also part of Northern Sweden Health and Disease Study) showed an SIR of 0.99 (1,159 observed cases vs. 1,174 observed). These women – age range 50–69 years – had a significantly lowered SIR for lung cancer (0.75; 0.56–0.98), while none of the other sites showed an SIR significantly different from unity. Incidence of breast cancer was significantly increased during the first year after mammography and serum sampling date (SIR 1.89; 1.58–2.24), but this excess was compensated by a significantly decreased incidence in the later years (SIR 0.88; 0.78–0.98).

#### 3.4.2.2. Maternity Cohorts

There were 14,973 cancer cases observed after sampling (from 1983 until August 2005) and before 31 December 2005 in the *Finnish Maternity Cohort*, which is the biggest one of the four Nordic biobanks based on the population screening of pregnant women. The expected number based on average Finnish female population was 15,770 and the SIR was 0.95 (95% CI 0.93–0.96).

There were 6,861 cases of breast cancer, equal to the expected number. The SIR for lobular type of breast cancer (15% of breast cancers) was 1.12 (1.05–1.18). The incidence of breast cancer was above the national average after sera drawn in the context of the first pregnancy (SIR 1.08; 1.04–1.11) but the SIR gradually declined along with the subsequent pregnancies and was after fifth pregnancy 0.62 (0.38–0.94). The SIR for endometrial cancer among all pregnant women was 0.64 (0.57–0.70) and decreased after the third pregnancy to only 0.30 (0.15–0.56).

There was an excess of the rare placental choriocarcinoma during the first year after sampling (11 cases; SIR 6.06; 3.03–10.84), which is by definition related to pregnancy. Borderline tumours of the ovary were less frequent than in the population on average (SIR 0.85; 0.76–0.94) and invasive ovarian tumours even more rare (SIR 0.73; 0.67–0.79). The SIR for lung cancer was 0.79 (0.69–0.89), with the strongest decrease in adenocarcinoma (SIR 0.60; 0.48–0.74). The SIR for stomach cancer was 0.88 (0.77–0.99), for soft tissue sarcoma 0.84 (0.69–0.99) and for cancer with unknown primary site 0.82 (0.68–0.96).

In the *Icelandic Maternity Cohort*, there were 1,453 malignant neoplasms observed versus 1,466 expected (SIR 0.99; 0.94–1.04). The SIRs for single cancer sites were similar as those reported above for the Finnish Maternity cohort but none of them reached statistical significance in this ten times smaller data set

Women in the *Malmö Maternity Cohort* (part of Malmö Microbiology biobank) also had overall cancer incidence similar to the national population (493 observed cases vs. 498 expected, SIR 0.99; 0.91–1.08), but there was a tendency for higher lung cancer incidence than the reference population (SIR 1.28; 0.68–2.18). None of the other cancer sites deviated significantly from the expected incidence.

In the *Northern Sweden Maternity Cohort*, there were 1,625 cancer cases observed after sampling and before end of follow-up. The expected number was 1,717 and the SIR 0.95 (0.90–0.99). Significantly decreased SIRs were seen for lung cancer (0.59; 0.40–0.83) and endometrial cancer (0.69; 0.49–0.94).

3.4.2.3. Specific Cohorts with Clearly Defined Enrolment Criteria

Men in the *Helsinki Heart Study* had 3,638 cancer cases, less than expected (SIR 0.92, 95% CI 0.89–0.94). The SIRs were significantly decreased for cancers of the pharynx (SIR 0.55; 0.28–0.99), stomach (0.78; 0.65–0.91), pancreas (0.81; 0.66–0.96), nose (0.14; 0.00–0.77) and unspecified sites (0.69; 0.52–0.90). SIR for lung cancer was below the national average in all main histological types: in squamous cell carcinoma 0.72 (0.61–0.83), adenocarcinoma 0.71 (0.57–0.88), and small cell carcinoma 0.63 (0.49–0.78).

Incidence of non-melanoma skin cancer (SIR 1.37; 95% CI 1.15–1.59) and basal cell carcinoma of the skin (1.24; 1.16–1.31) was significantly above the national average. Also meningiomas of the brain were in excess (SIR 1.59; 95% CI 1.05–2.29).

The incidence pattern of the health-interested volunteers of Helsinki Heart Study is very different from that of the cohort of smoking men in the *Alpha-Tocopherol-Beta-Carotene (ATBC) study* (Fig. 7). The latter cohort has been utilised in studies aiming to confirm whether various diseases are related to smoking or not (40).

In the ATBC cohort, there is an excess risk of cancer in most sites. The observed number of cancers in the end of 2005 was

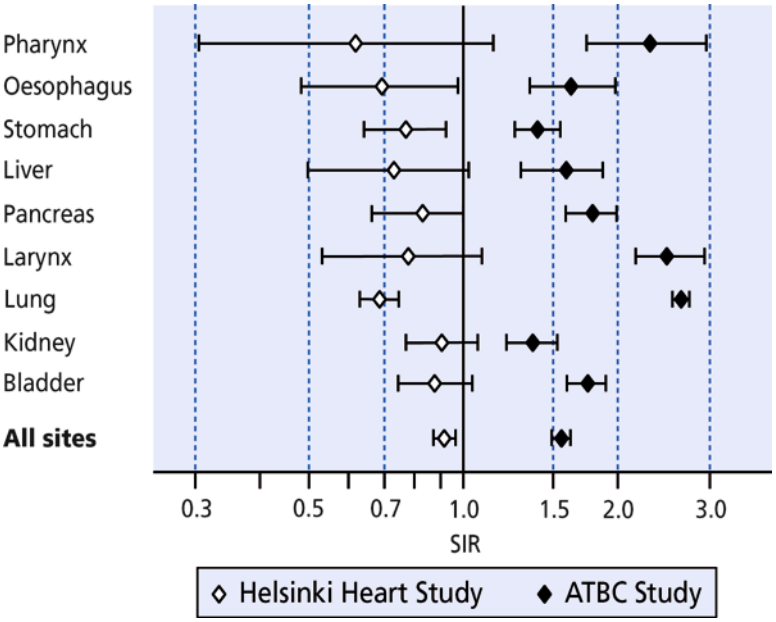


Fig. 7. Standardised incidence ratios (SIR) of selected cancers 1984–2003 among the 19,000 Finnish men in Helsinki Heart Study, and the 29,000 men in Alpha-Tocopherol-Beta-Carotene (ATBC) Study, with 95% confidence interval bars.

9,420, i.e., 3,202 more than the expected number calculated on the basis of incidence rate of average Finnish same-aged men (SIR 1.52; 1.48–1.54). The SIRs were significantly increased for cancers of the tongue (SIR 1.77; 95% CI 1.23–2.46), other oral cavity (1.86; 1.31–2.54), pharynx (2.22; 1.68–2.88), oesophagus (1.64; 1.36–1.95), stomach (1.36; 1.23–1.50), colon (1.14; 1.02–1.25), liver (1.61; 1.36–1.87), pancreas (1.76; 1.58–1.94), larynx (2.50; 2.13–2.89), lung (2.71; 2.62–2.80), prostate (1.11; 1.06–1.15; for non-localised prostate cancers 1.22; 1.10–1.34), kidney (1.36; 1.22–1.50; includes renal pelvis 2.41; 1.69–3.33), bladder (invasive 1.77; 1.64–1.91, and for papilloma 2.22; 1.32–3.51), and unspecified sites (1.79; 1.57–2.03). Excess risk was seen in acute myeloid leukaemia (SIR 1.37; 1.00–1.83) –according to literature that seems to be related to smoking (41) – but not in other types of leukaemia. The only sites with an SIR below 1.0 were the skin melanoma (0.69; 0.56–0.84) and the basal cell carcinoma of the skin (0.89; 0.84–0.94).

The *Red Cross blood donors* in capital Oslo and surrounding areas (the smaller part of Janus biobank) had lower than average overall cancer incidence (2,286 cases observed vs. 2,399 expected; SIR 0.95; 95% CI 0.91–0.99). The SIRs of cancers of the stomach, primary liver and larynx are as low as 0.36–0.46, all significantly decreased. The SIR for lung cancer was 0.76 (95% CI 0.65–0.88). The SIR for breast cancer was significantly elevated (1.29; 1.17–1.42), and so was the SIR for skin melanoma (1.24, 1.07–1.42).

#### 3.4.2.4. Viral Screening and Clinical Testing Biobanks

In the part of the *Malmö Microbiology cohort* including samples submitted for testing because of clinical suspicion of infection with blood-borne viruses (e.g. jaundice or impaired liver function, drug addicts, haemophiliacs and dialysis patients), there were 2,055 cancer cases more than the expected number 4,455 (SIR 1.46; 95% CI 1.43–1.50). All SIRs were above 1.0, except those for breast cancer and endometrial cancer. The highest SIRs were seen for primary liver cancer (5.58; 4.87–6.36), pancreatic cancer (3.28, 2.93–3.67) and gall-bladder cancer (2.52; 1.94–3.22).

The *Malmö Microbiology* subcohort consisting of sera submitted for other virus serology had even higher relative overall cancer risk (SIR 2.08; 95% CI 1.97–2.20; 1,328 cases observed vs. 638 expected). Very high SIRs were seen in primary liver cancer (4.15; 2.63–6.22), pancreatic cancer (2.71, 1.91–3.74), lung cancer (2.95; 2.46–3.53) and cancers of the brain and nervous system (3.05; 2.37–3.88).

### *3.5. Accuracy of Variables Associated to Persons and Samples*

Most samples of the biobanks include variables related to the sample itself, to the sampling occasion, or to the person who donated the sample. If these variables were in major role in the original setting of a study – such as the questionnaire data related to study persons' health habits (for example smoking, diet,

physical exercise, body mass index) – they are normally stored and documented systematically. If such data has been asked in a context of clinical practice, these data may be kept non-systematically, possibly on paper format only, or even lost.

A high-quality biobank database should include some variables directly related to the sample that are crucial for nested case-control studies based on the biobank:

1. date of sampling,
2. indication of sampling (in biobanks with mixed origin of sampling),
3. number of freeze–thaw cycles,
4. amount of sample left,
5. indicator of damaged sample.

All of these variables may be used as matching criteria in control selection. If they are missing, the quality of the study will not be as good as it could be. If these factors can only be confirmed after search the samples from the fridge, the logistics of any such study becomes clumsy and laborious.

## 4. Prospective Cancers: Basis for Nested Case Control Studies

### *4.1. Numbers of Prospective Cancer Cases*

Maybe the most important tool for quality assurance of a big biobank network is a simple tabulation of numbers of persons in the biobank cohorts and numbers of cancer cases. The NBSBCCC network has agreed to collect such data – stratified by biobank, year, gender, age and cancer type – in a centralised database, which will be automatically updated after each new linkage of the records of any biobank and cancer registry data.

Even though the number of new donors to the NBSBCCC has been decreasing in the latest years (Fig. 8), the number of prospective cancer cases increases year by year (Fig. 9). The annual number seems to drop in the very latest years in some of the regions. This is an artefact related to technical reasons. For instance, the biobanks in Malmö have had to wait for a new update because of slow progress in getting permission from privacy issue officials to link cancer data with their cohort; this problem is now solved and very soon the closing date for the Malmö cohorts will be moved from 1999 to 2005. In Norway, the Janus cohort has been linked with cancer data until about 2005, but there have been problems and principle issues to tabulate the numbers for the NBSBCCC quality assurance database.

The numbers of cancer cases diagnosed after serum donation among persons in each serum bank are given in Table 4, for all cancers combined and for 64 subcategories. These numbers are based on the routine linkages between the serum banks and cancer

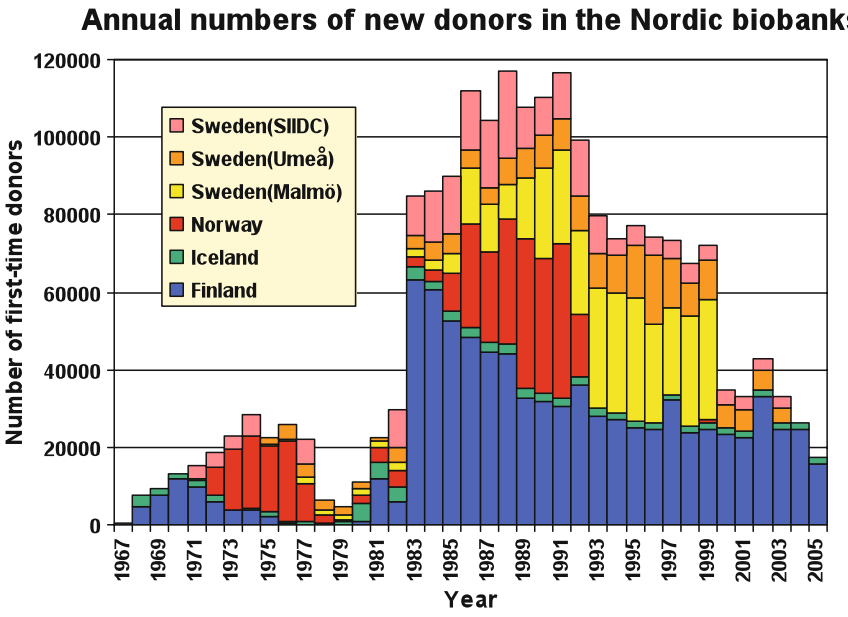


Fig. 8. Annual numbers of first-time donors of the Nordic biobanks as reported to the joint NBSBCCC quality assurance surveillance database by June 2007, by region.

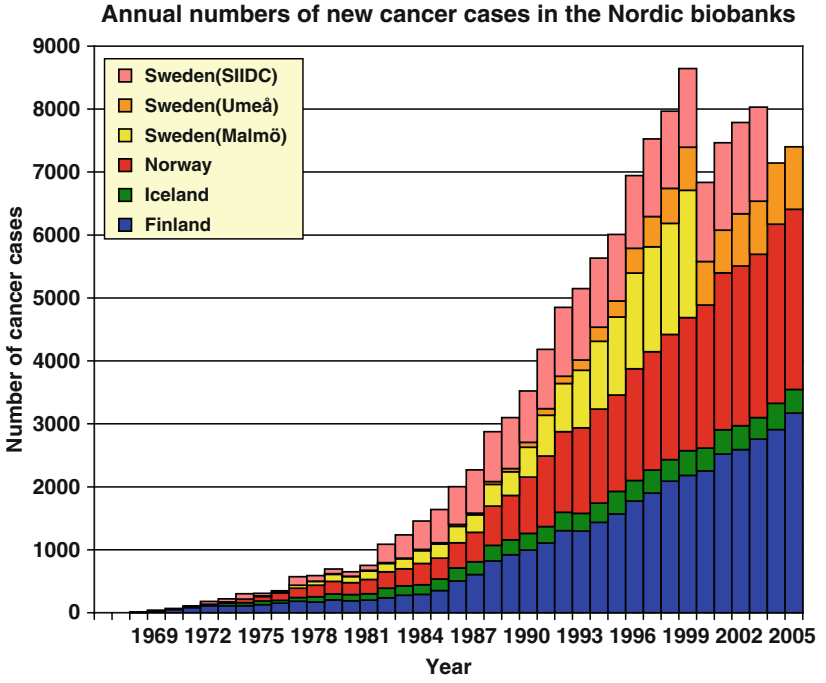


Fig. 9. Annual numbers of registered cancer cases among subjects in the Nordic biobanks diagnosed after serum donation and before 31 Dec 2005, as reported to the joint NBSBCCC quality assurance surveillance database by March 2008 by region.

**Table 4**

**Numbers of cancer cases registered among serum donors after donation to the Nordic Biological Specimen Banks for Cancer Causes and Control (NBSBCCC) according to the latest cancer registry linkages, by cancer site and serum bank. Table constructed February 28 2008**

| ICD-7 (or internal code) | Cancer site/type | Number of cancer cases | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Finland | | | | | Iceland | | Norway | | Sweden | | | | | | | | | | Total nordic |
| | | | | | | | | | Janus | | Northern Sweden Health and Disease Study | | | | | | | Malmö Microbiology Biobank | | | |
| | | Maternity cohort (females) | Mobile Clinic | FINRISK | Helsinki Heart Study (males) | ATBC (males) | Maternity cohort (females) | Heart Preventive Clinic | Health examinations | Blood donors | Northern Sweden Maternity Cohort (females) | VIP | MONICA | Mammography (females) | Institute for Infectious Disease Control | Diet and cancer (Malmö) | Preventive medicine (Malmö) | Maternity cohort (females) | Blood-borne virus screening | Clinical viral serology | |
| **Years covered** | | 1983–2005 | 1967–2005 | 1992–2005 | 1980–2005 | 1985–2005 | 1980–2005 | 1967–2005 | 1973–2006 | 1973–2006 | 1975–2005 | 1985–2005 | 1986–2005 | 1995–2005 | 1971–2003 | 1991–1999 | 1974–1999 | 1985–1999 | 1969–1999 | 1990–1999 | |
| 140–207 | All malignant neoplasms | 14,973 | 8,463 | 1104 | 3,638 | 9,420 | 1,453 | 6,219 | 36,371 | 2,406 | 2,192 | 3,786 | 476 | 1,864 | 21,802 | 1,852 | 4,343 | 493 | 6,511 | 1,328 | 128,694 |
| 140 | Lip | 4 | 78 | 4 | 31 | 70 | 1 | 20 | 131 | 7 | 2 | 9 | 1 | 1 | 76 | 7 | 18 | 0 | 32 | 10 | 502 |
| 141 | Tongue | 48 | 41 | 9 | 9 | 35 | 1 | 15 | 154 | 14 | 6 | 6 | 2 | 3 | 54 | 3 | 18 | 2 | 25 | 2 | 447 |
| 142 | Salivary glands | 61 | 28 | 2 | 11 | 8 | 6 | 10 | 67 | 4 | 4 | 7 | 1 | 5 | 39 | 3 | 7 | 0 | 18 | 6 | 287 |
| 143–144 | Oral cavity | 35 | 29 | 5 | 15 | 38 | 5 | 29 | 183 | 8 | 2 | 11 | 1 | 6 | 68 | 10 | 22 | 1 | 26 | 6 | 500 |
| 145–148 | Pharynx | 34 | 30 | 2 | 11 | 57 | 3 | 16 | 194 | 17 | 5 | 14 | 3 | 4 | 115 | 10 | 46 | 1 | 40 | 7 | 609 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 150 | Oesophagus | 14 | 107 | 12 | 35 | 119 | 1 | 81 | 263 | 20 | 2 | 24 | 2 | 10 | 108 | 18 | 53 | 1 | 51 | 10 | 931 |
| 151 | Stomach | 246 | 516 | 37 | 133 | 369 | 8 | 347 | 995 | 38 | 27 | 92 | 14 | 38 | 392 | 36 | 125 | 6 | 157 | 26 | 3,602 |
| 152 | Small intestine | 33 | 32 | 7 | 13 | 19 | 2 | 27 | 135 | 13 | 8 | 12 | 2 | 6 | 92 | 10 | 23 | 0 | 37 | 5 | 476 |
| 153 | Colon | 525 | 445 | 56 | 220 | 366 | 42 | 523 | 3,104 | 195 | 67 | 254 | 31 | 161 | 1,066 | 105 | 232 | 16 | 311 | 75 | 7,794 |
| 154 | Rectum, rectosigma | 275 | 351 | 38 | 160 | 264 | 25 | 184 | 1,697 | 118 | 38 | 155 | 23 | 75 | 545 | 85 | 172 | 6 | 205 | 28 | 4,444 |
| 155 | Primary liver | 42 | 89 | 19 | 48 | 148 | 2 | 46 | 141 | 4 | 7 | 26 | 5 | 13 | 359 | 15 | 39 | 0 | 223 | 23 | 1,249 |
| 155.1 | Gall-bladder, biliary tract | 55 | 109 | 3 | 23 | 57 | 1 | 45 | 172 | 8 | 6 | 9 | 0 | 7 | 132 | 9 | 16 | 0 | 65 | 6 | 723 |
| 157 | Pancreas | 135 | 354 | 38 | 111 | 368 | 7 | 188 | 828 | 49 | 22 | 100 | 11 | 47 | 476 | 42 | 119 | 3 | 316 | 37 | 3,251 |
| 160 | Nose | 9 | 14 | 5 | 1 | 12 | 1 | 10 | 63 | 5 | 4 | 9 | 0 | 3 | 17 | 1 | 6 | 0 | 5 | 4 | 169 |
| 161 | Larynx | 2 | 80 | 6 | 37 | 165 | 1 | 34 | 260 | 9 | 1 | 8 | 1 | 1 | 71 | 4 | 44 | 0 | 43 | 1 | 768 |
| 162–163 | Lung | 236 | 1,157 | 108 | 485 | 3,125 | 63 | 790 | 3,548 | 190 | 62 | 185 | 22 | 93 | 1,698 | 137 | 535 | 13 | 523 | 127 | 13,097 |
| (162A) | lung, adenoca | 85 | 146 | 18 | 86 | 416 | 32 | 269 | 1,075 | 60 | 26 | 65 | 7 | 34 | 466 | 45 | 149 | 2 | 160 | 40 | 3,181 |
| (162S) | lung, small cell | 41 | 167 | 16 | 76 | 519 | 10 | 154 | 685 | 28 | 7 | 28 | 7 | 17 | 231 | 27 | 93 | 2 | 80 | 20 | 2,208 |
| (162E) | lung, squamous cell | 32 | 354 | 29 | 155 | 945 | 8 | 172 | 801 | 74 | 9 | 17 | 2 | 13 | 406 | 25 | 125 | 2 | 115 | 21 | 3,305 |
| 170 | Breast | 6,861 | 998 | 173 | 4 | 10 | 512 | 714 | 5,215 | 413 | 903 | 696 | 75 | 614 | 1,265 | 324 | 466 | 165 | 635 | 112 | 20,155 |
| (170D) | breast, ductal | 5,389 | 703 | 129 | 3 | 9 | 406 | 485 | 3,925 | 192 | 560 | 478 | 53 | 424 | 1,047 | 138 | 117 | 67 | 243 | 57 | 14,425 |
| (170L) | breast, lobular | 1,031 | 100 | 33 | 1 | 0 | 50 | 48 | 507 | 52 | 99 | 99 | 7 | 100 | 194 | 66 | 50 | 13 | 75 | 12 | 2,537 |
| 171 | Cervix uteri, invasive | 559 | 57 | 4 | 0 | 0 | 134 | 41 | 539 | 54 | 174 | 36 | 5 | 17 | 1,795 | 13 | 27 | 56 | 101 | 10 | 3,622 |
| 172 | Corpus uteri | 316 | 237 | 39 | 0 | 0 | 33 | 149 | 1,176 | 59 | 54 | 152 | 17 | 161 | 329 | 48 | 75 | 2 | 76 | 18 | 2,941 |
| 173 | Choriocarci-noma | 28 | 1 | 0 | 0 | 0 | . | . | 2 | 2 | 5 | 0 | 1 | 0 | . | 0 | 0 | 0 | 2 | 0 | 41 |
| 175.0 | Ovary | 508 | 163 | 30 | 0 | 0 | 41 | 101 | 925 | 59 | 103 | 100 | 7 | 65 | 638 | 35 | 59 | 13 | 103 | 16 | 2,966 |

(continued)

**Table 4**
**(continued)**

| Cancer site/type | | Number of cancer cases* | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Finland | | | | | Iceland | Norway | Sweden | | | | | | | | | | | | | Total nordic |
| | | | | | | | | Janus | | | | Northern Sweden Health and Disease Study | | | | | | Malmö Microbiology Biobank | | | | |
| ICD-7 (or internal code) | | Maternity cohort (females) | Mobile Clinic | FINRISK | Helsinki Heart Study (males) | ATBC (males) | Maternity cohort (females) | Heart Preventive Clinic | Health examinations | Blood donors | Northern Sweden Maternity Cohort (females) | VIP | MONICA | Mammography (females) | Institute for Infectious Disease Control | Diet and cancer (Malmö) | Preventive medicine (Malmö) | Maternity cohort (females) | Blood-borne virus screening | Clinical viral serology | |
| 175.1 | Tuba | 18 | 3 | 1 | 0 | 0 | 2 | 3 | 43 | . | 2 | 2 | 0 | 3 | 11 | 4 | 4 | 0 | 7 | 0 | 103 |
| 176.0 | Vulva | 47 | 23 | 4 | 0 | 0 | 4 | 14 | 87 | 6 | 9 | 3 | 0 | 6 | 50 | 5 | 8 | 2 | 6 | 3 | 277 |
| 176.1 | Vagina | 11 | 8 | 2 | 0 | 0 | 2 | 3 | 30 | 1 | 6 | 2 | 1 | 2 | 11 | 1 | 0 | 1 | 2 | 0 | 83 |
| 177 | Prostate | 0 | 1,076 | 227 | 1,127 | 2,000 | 0 | 1,013 | 4,632 | 266 | 0 | 740 | 100 | 0 | 1,985 | 292 | 606 | 0 | 694 | 128 | 14,886 |
| (177L) | prostate, localised | 0 | 501 | 133 | 546 | 854 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 2,034 |
| (177N) | prostate, non-localised | 0 | 214 | 23 | 156 | 366 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 759 |
| 178 | Testis | 0 | 8 | 2 | 9 | 8 | 0 | 2 | 220 | 30 | 0 | 22 | 3 | 0 | 117 | 2 | 17 | 0 | 70 | 2 | 512 |
| (178S) | testis, seminoma | 0 | 5 | 2 | 7 | 6 | 0 | 1 | 150 | 20 | 0 | 16 | 2 | 0 | 117 | 1 | 12 | 0 | 36 | 1 | 376 |

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (178N) | testis, non-seminoma | 0 | 3 | 0 | 2 | 2 | 0 | 1 | 57 | 3 | 0 | 6 | 1 | 0 | 162 | 1 | 5 | 0 | 34 | 1 | 278 |
| 179.0 | Penis | 0 | 18 | 2 | 10 | 10 | 0 | 11 | 74 | . | 0 | 8 | 2 | 0 | 20 | 1 | 4 | 0 | 8 | 2 | 170 |
| 180 | Kidney | 211 | 296 | 37 | 156 | 331 | 19 | 221 | 1,009 | 71 | 22 | 79 | 10 | 45 | 589 | 43 | 141 | 5 | 174 | 41 | 3,500 |
| (180.1) | renal pelvis | 7 | 22 | 3 | 12 | 36 | 0 | 25 | 98 | 1 | 2 | 8 | 3 | 9 | 57 | 7 | 25 | 0 | 21 | 2 | 338 |
| 181 | Bladder, ureter, urethra | 84 | 327 | 25 | 180 | 623 | 19 | 353 | 1,757 | 60 | 20 | 149 | 25 | 51 | 728 | 102 | 310 | 5 | 292 | 28 | 5,138 |
| 190 | Melanoma of the skin | 778 | 209 | 28 | 107 | 95 | 149 | 83 | 2,036 | 200 | 138 | 123 | 9 | 43 | 623 | 93 | 208 | 65 | 256 | 33 | 5,276 |
| 191 | Non-melanoma skin cancer | 140 | 250 | 24 | 143 | 177 | 20 | 206 | 1,029 | 83 | 13 | 36 | 9 | 29 | 2,792 | 71 | 171 | 8 | 489 | 144 | 5,834 |
| 192 | Eye | 43 | 22 | 3 | 8 | 9 | 3 | 13 | 106 | 11 | 9 | 17 | 0 | 7 | 82 | 1 | 5 | 1 | 8 | 4 | 352 |
| 193 | Brain and nervous system | 1,149 | 238 | 39 | 114 | 129 | 65 | 165 | 1,280 | 95 | 143 | 154 | 23 | 72 | 817 | 58 | 162 | 40 | 242 | 67 | 5,052 |
| (193G) | glioma | 415 | 84 | 13 | 47 | 56 | 22 | 73 | 534 | 49 | 28 | 54 | 11 | 27 | 114 | 1 | 8 | 4 | 30 | 10 | 1,580 |
| (193M) | meningeoma | 384 | 65 | 10 | 28 | 29 | 36 | 65 | 430 | 23 | 49 | 52 | 7 | 31 | 222 | 19 | 48 | 8 | 70 | 12 | 1,588 |
| 194 | Thyroid | 1,108 | 116 | 14 | 17 | 25 | 127 | 123 | 415 | 27 | 49 | 18 | 1 | 13 | 94 | 10 | 28 | 17 | 60 | 11 | 2,273 |
| (194F) | follicular | 52 | 17 | 1 | 5 | 3 | 10 | 20 | 48 | 4 | 3 | 1 | 0 | 1 | 5 | 2 | 2 | 0 | 5 | 2 | 181 |
| (194P) | papillary | 1,017 | 77 | 9 | 10 | 14 | 113 | 90 | 311 | 11 | 24 | 11 | 1 | 8 | 73 | 4 | 6 | 9 | 23 | 4 | 1,815 |
| 195.0 | Glandula suprarenalis | 20 | 7 | 0 | 3 | 5 | 1 | 2 | 34 | 5 | 11 | 8 | 0 | 7 | 43 | 3 | 11 | 2 | 24 | 4 | 190 |
| 195.1 | Glandula parathyreiodea | 0 | 1 | 0 | 0 | 0 | 0 | 3 | . | . | 34 | 11 | 5 | 4 | 232 | 16 | 53 | 8 | 75 | 13 | 455 |
| 195.2 | Thymus | 10 | 2 | 2 | 0 | 1 | 0 | 2 | 14 | . | 2 | 4 | 0 | 2 | 17 | 1 | 4 | 2 | 10 | 5 | 78 |
| 195.3 | Hypophysis | 0 | 0 | 0 | 0 | 0 | 8 | 22 | 1 | . | 14 | 16 | 0 | 6 | 78 | 9 | 17 | 6 | 40 | 8 | 225 |
| 195.4 | Corpus pineale | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | . | . | . | . | . | . | 0 | 0 | 0 | 0 | 0 | 6 |

(continued)

## Table 4 (continued)

| Cancer site/type | | Number of cancer cases* | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Finland | | | | | Iceland | Norway | Sweden | | | | | | | | | | | | | Total nordic |
| | | | | | | | | Janus | Northern Sweden Health and Disease Study | | | | | | | Malmö Microbiology Biobank | | | | | | |
| ICD-7 (or internal code) | | Maternity cohort (females) | Mobile Clinic | FINRISK | Helsinki Heart Study (males) | ATBC (males) | Maternity cohort (females) | Heart Preventive Clinic | Health examinations | Blood donors | Northern Sweden Maternity Cohort (females) | VIP | MONICA | Mammography (females) | Institute for Infectious Disease Control | Diet and cancer (Malmö) | Preventive medicine (Malmö) | Maternity cohort (females) | Blood-borne virus screening | Clinical viral serology | |
| 196 | Bone | 54 | 14 | 3 | 4 | 4 | 10 | 14 | 50 | 4 | 6 | 4 | 0 | 1 | 43 | 3 | 5 | 1 | 15 | 3 | 238 |
| 197 | Soft tissue | 118 | 55 | 6 | 21 | 27 | 10 | 16 | 131 | 18 | 17 | 24 | 1 | 6 | 193 | 14 | 26 | 4 | 46 | 12 | 745 |
| 199 | Other/unknown site | 133 | 190 | 16 | 54 | 230 | 19 | 116 | 760 | 53 | 29 | 94 | 16 | 62 | 755 | 37 | 103 | 3 | 211 | 40 | 2,921 |
| 200.202 | Non-Hodgkin's lymphoma | 457 | 249 | 29 | 167 | 226 | 43 | 141 | 1,266 | 72 | 58 | 138 | 17 | 59 | 1,429 | 69 | 138 | 13 | 271 | 111 | 4,953 |
| 201 | Hodgkin's disease | 216 | 41 | 1 | 12 | 14 | 19 | 14 | 112 | 12 | 22 | 10 | 3 | 6 | 340 | 3 | 15 | 9 | 52 | 40 | 941 |
| 203 | Multiple myeloma | 68 | 106 | 15 | 47 | 60 | 8 | 91 | 464 | 12 | 11 | 80 | 8 | 30 | 308 | 28 | 57 | 1 | 83 | 17 | 1,494 |
| 204–207 | Leukaemia | 243 | 192 | 21 | 84 | 129 | 26 | 164 | 646 | 40 | 40 | 71 | 12 | 34 | 1,591 | 42 | 91 | 11 | 158 | 67 | 3,662 |
| (204CLL) | chronic lymphocytic | 35 | 76 | 4 | 45 | 52 | 1 | 61 | 223 | 4 | 7 | 32 | 7 | 15 | 253 | 14 | 32 | 1 | 41 | 17 | 920 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (204AML) | acute myeloid | 109 | 57 | 7 | 17 | 45 | 6 | 46 | 250 | 10 | 13 | 16 | 2 | 8 | 81 | 10 | 28 | 5 | 54 | 23 | 787 |
| *Not included above* | | | | | | | | | | | | | | | | | | | | | |
| 171C | Cervix, CIN3/ in situ/ dysplasia gravis | 5,335 | 88 | 23 | 0 | 0 | . | . | . | . | 2406 | 225 | 26 | 53 | 1,607 | 0 | 0 | 0 | 0 | 0 | 9,763 |
| 175B | Ovary, borderline tumour | 345 | 28 | 7 | 0 | 0 | 47 | 20 | 273 | 7 | 61 | 25 | 5 | 13 | 73 | 0 | 0 | 0 | 0 | 0 | 904 |
| 181P | Bladder, papilloma | 10 | 11 | 1 | 10 | 18 | 1 | 23 | . | 60 | 16 | 140 | 23 | 47 | 3 | 90 | 271 | 4 | 252 | 22 | 1,002 |
| 191B | Skin, basal cell carcinoma | 2,269 | 1677 | 270 | 920 | 1,038 | 254 | 824 | . | . | . | . | . | . | . | . | . | . | . | . | 7,252 |
| Total benign/ semima-lignant | 7,959 | 1,804 | 301 | 930 | 1,056 | 302 | 867 | 273 | 67 | 67 | 390 | 54 | 113 | | 90 | 271 | 4 | 252 | 22 | 17,238 | |

Not applicable
Not registered by the national cancer registry
Classification not available from the national cancer registry

registries (status in June 2007). There were altogether 110,217 cases traditionally counted as real cancers, and 15,428 basal cell carcinomas of the skin, precancerous cervical lesions, borderline tumours of the ovaries and cancers of low malignant potential of the bladder. The registration of the latter outcome categories varies over the Nordic cancer registries, and the same is true for in situ cancers and several other cancer-like lesions not included in Table 4.

After update of the biobanks in Malmö and taking into account the cancer cases for Janus biobank missing from the above tabulation, the numbers of subsequent cancer cases exceed 30,000 in Sweden, Finland and Norway (27,000), giving a balanced three-country setting in the future studies. The Icelandic number (7,700) is smaller but very large as compared to the small population size in Iceland (less than 300,000).

### 4.2. Simple Cross-Tabulations: A Powerful Quality Assurance Tool

There are several ways how simple tabulations of numbers of persons and prospective cancer cases may improve the quality of network activity.

When designing a case-control study nested in the biobank cohorts, it is good to know how many eligible cases there will be to be sure that the study power will be satisfactory. The number of cancer-free individuals makes it easy to select matching criteria in such a way that required number of eligible controls will be found but there will not be unnecessary large variation in matching criteria such as storage time of the sample or age of the individual.

Once the cases and controls have been selected in each participating biobank, it is always good to check whether the numbers match with those to be found from the NBSBCCC tabulation. If the numbers do not match, there are two possibilities:

1. There has been an error in the case-control selection. In this case, the error can be corrected before sending the samples to laboratory analyses.
2. There are good explanations for the drop of the case number, such as additional exclusion criteria whose prevalence was wrongly estimated. If this happens, the design may be modified to replace the missed cases from other biobanks or by extending the period of case recruitment.

This gives the principle investigator of any network-based study tools to control for accuracy of the study materials, and also eliminate attempts of fraud, such as fabrication of data. Although the scientific moral in the NBSBCCC network has been high and fraud would never been expected in this research society, there are examples from the latest years from other research groups that makes it important to be able to demonstrate that even such extreme possibilities can be controlled. Those research institutes whose studies have never been linked to any scientific miscarriage

are most eager to offer materials to any quality assurance operation that would improve possibilities to external evaluation of the accuracy of the data, such as the simple tabulations described above.

### 4.3. Future Numbers of Prospective Cancer Cases

The tabulations of numbers of prospective cancer cases indicate that the annual number of new prospective cancer cases in the NBSBCCC biobank cohorts will be about 10,000 cancer cases in the next few years. The basic data readily collected to the joint database allow predictions based on age-period-cohort analyses, and a project to predict numbers of cancer cases up to the year 2015 has been started. In short, this requires estimation of future incidence rates of selected cancers, and prediction of future person-years at risk.

Even when the total number of prospective cancer cases is huge, there are rare cancer types, special subcategories and rare exposures where the current study power may not be satisfactory. Therefore, in a situation when we anyway have to set priorities on what to do first, an optimally coordinated research cluster should postpone underpowered studies to the future. Predictions on the future numbers of case help in long-term planning of study schedules.

## 5. Discussion

### 5.1. Strengths of Biobank-Based Study Designs

The infrastructure described allows multi-national and multi-disciplinary networking for comprehensive prospective epidemiological studies nested in several biological specimen banks. There are several strengths in studies based on samples readily collected in biobanks to the alternative situation that there is no biobank, i.e., samples from cases and controls have to be collected after the disease of the case has been diagnosed:

(a) Use of biobank data offers proper time order of exposure data collection and outcome and decreases the possibility of "*reverse causality bias*", i.e. the mixing up of cause and effect. For instance, herpes viruses are frequently reactivated by severe diseases, such as cancer, and may indeed induce cellular genes related to cellular proliferation (4, 21). If the virus is measured from a sample taken at the time of cancer diagnosis, it is difficult to assess whether associations between reactivatable viruses and cancer are causal or mere secondary associations with opportunistic infectious agents. In the prospective design, we have been able to show that cancer reactivates herpes simplex virus type 2 and not vice versa (21, 42, 43).

(b) A related type of bias is the *differential measurement bias*, i.e., situations where the fact that the patient has disease

influences measurements. Even existent (pre)cancer may influence both antibody levels and cellular immunity because of the immune dysfunctions seen in cancer (11, 25, 27). Also, it may be easier to obtain cancer tissue than control tissue. When measurement biases are related to case status, their effect is particularly unpredictable. Studies using samples taken from individuals long time before the cancer diagnosis suffer only from misclassification bias that is non-differential with regard to case status, which may result in a conservative and readily quantifiable bias.

(c) Many exposures are associated with *non-attendance* in retrospective case-control studies, biasing results. In biobank-based studies, there may be baseline selection in the formulation of the study base (that makes the study base different as compared to the population from which it was originally drawn), but after that all samples from the study base are available for testing, and there is no selection related to later case-control status.

(d) Studies based on readily collected biosamples are *time-effective* and – if the biobank is used in many studies –*cost-effective*. The classical prospective cohort study, where samples are not stored but analysed immediately after sampling, requires very long follow-up, often decades. Study hypotheses and measurement assays may be outdated when the outcomes are finally obtained. The establishment and maintenance of population-based biological specimen banks is costly, but when such banks are established they can be used for a variety of prospective studies on the aetiology of several reasonably common diseases, e.g., association of HPV infections with various human cancers (2, 3, 6–9, 16, 19). The marginal cost for a prospective study can be reduced to the level where also rather unlikely, innovative hypotheses (that may result in breakthroughs) can be reliably evaluated., e.g., the role of *Chlamydia trachomatis* in cervical cancer causation (14, 17, 44). Since biological specimen banks are already established, the time required for completion of a reliable prospective study with decade-long follow-up of a recently emerged epidemiological problem is short.

(e) The Nordic biological sample banks contain a very high proportion of *serial samples*; the mean number of samples per person is two to three (Table 2). For instance, the maternity cohorts includes complete sets of serial samples related to pregnancies of majority of the parous women (13), and some specific research cohorts may include very tight set of samples, e.g. there are up to 28 samples from part of the Helsinki Heart Study subjects. For studies of chronic diseases, such as cancer, that develop over a very long time span, a considerably

more reliable and complete assessment of the importance of various exposures can be obtained by studying multiple serial measurements of the same person (10, 13). Furthermore, the unavoidable variability of measurements in a single sample will cause a systematic underestimation of the importance of a risk factor (regression dilution towards the mean) (45), which can be corrected for using serial measurements. Serial samples can also be used to pin point the time-point of exposure (13). There are many examples of disease causation by an exposure that occurs only if the exposure occurs at a certain time-point. Poliomyelitis as a result of delayed exposure to poliovirus is a well-known example. Only if samples taken at many different time-points preceding development of disease have been stored, one can attempt to study time-point of exposure by biochemical and molecular assays.

Biobank samples may sometimes offer *objective measurements* for variables that are hard to be accurately registered via questionnaire surveys. The extent of misclassification of self-reported exposures can be considerable, especially for sensitive questions, such as addictions. Even a very modest amount of misclassification may lead to very misleading conclusions. There now exist an increasing arsenal of biochemical measurements that can be used for objective measurement of exposures in stored biological specimens, e.g. serum cotinine measurements for assessment of smoking habits (46). The accuracy of extensive questionnaires on environmental exposures, diet and life style that most of the NBSBCCC banks contain may be validated with biochemical measurements in a relatively small amount of samples derived from the biobank.

### 5.2. Stability and Validity of Old Samples

A potential weakness of studies based on historical biobank samples is the stability and validity of the old samples. The oldest samples in the Nordic biobanks are more than 30 years old and many are stored at –25°C. Validations of the Janus biobank have shown that most of the substances commonly analysed in epidemiological studies, for instance proteins (in particular antibodies), organic acids, carbohydrates, trace metals, inorganic salts and polyunsaturated fatty acids are stabile when they are stored at –25°C. However, not all enzymes and vitamins are stabile under these conditions (46).

Genotyping from archival serum and plasma samples is, following the development of efficient whole genome amplification methods, a fairly routine method also from very old samples stored at –25°C (47). However, investigators contemplating amplification-based methods such as PCR should be aware that in the 1960s and 1970s disposable pipettes and tips may not always have been used in all biobanks.

Possible deterioration of the oldest sera is commonly outweighed by consideration of increased statistical power, reduced reverse causality biases with longer follow-up and possibility to detect causative exposures that occur many years before diagnosis of disease and may not be detectable in samples taken at or close to diagnosis.

### 5.3. Follow-Up Procedures

Initial calculations of SIRs in some of the biobanks did not include *follow-up for vital status*, which produced erroneous, markedly lowered SIRs in older ages. As demonstrated above for cohort analyses, the problem with missing data on vital status slowly becomes a serious problem also in case-control settings: a control subject that is registered as being alive may actually have had died before the respective case is diagnosed with cancer. For the quality assurance tabulations presented in this paper, all NBSBCCC biobanks were linked with national population registers to get dates of death up-to-date, and the procedure will from now on become a regular routine procedure.

*Follow-up for emigration* has not been considered very important because its magnitude has been rather small. However, in younger cohorts of modern Europeans emigration really has an effect. For instance, almost 4,000 women (6%) of the Icelandic maternity cohort had emigrated after serum sampling. Because the Icelandic registration system did not give the dates of emigrations, there was no information on how long the persons had been at reach of Icelandic follow-up possibilities, and all emigrated persons have to be excluded from all studies. In this type of situation, additional effort in seeking the missing dates of emigration would return several thousands of readily collected and carefully stored samples back to useful study materials and might be worth doing.

*Incorrect PIDs* is another source of errors on cancer risk estimates as demonstrated for cohort analyses in Fig. 4. The practice to check all PIDs against the population registries was not in routine use by all biobanks before the NBSBCCC quality assurance evaluation, but the procedure will from now on become a regular routine procedure.

The data quality requirement for the standardised incidence ratio calculation was a good way to improve accuracy of identifiers and completeness of follow-up for vital status, which is crucial in case-control studies for picking up controls that really are at risk of getting the cancer. Lack of follow-up for vital status and presence of some incorrect identifiers are likely to have caused minor errors in control selection in previous studies (Table 1): controls might have died or got cancer which was not known to the researchers. This type of errors would have reduced the risk estimates towards unity, i.e., any excess risks published so far are rather under- than overestimates of the true risk. Computerised record linkage procedures based on the unique PIDs are

unambiguous (48). Therefore, linking failures do not bias cancer risk estimates.

### 5.4. Follow-Up for Cancer Incidence

The nationwide Nordic cancer registries have been in operation since 1950s and have virtually complete coverage for cancer incidence (49). The tabulation of observed numbers of cancer cases given in Table 4 demonstrate that the cancer registries are able to produce data also by cancer classifications based on variables other than the topography alone (such as subtypes of leukaemia, histology and stage-specific categories) and tabulations of certain precancerous lesions. These specific categories are often useful for focused hypotheses testing. The data collection procedures prepared for the NBSBCC must be made to be able to design nested case-control studies, as knowledge of the number of cases is required to estimate statistical power. The predicted numbers for the years to come, further help in deciding the optimal time to start a given study.

The numbers of cancer cases diagnosed after sample donation and accumulated to Table 4 are based on the latest linkages between the biobanks and cancer registries: 10,000–20,000 of newly diagnosed cases are missing due to the normal delay of cancer registration and about 10,000 are missing because some biobanks are not linked with cancer registry very often. In some countries, each linkage for a specific research purpose requires a new ethical permission.

### 5.5. Cancer Incidence Rates in Cohorts in Relation to National Cancer Incidence Rates

None of the biobank cohorts had exactly the incidence pattern of the national general population. Some of them were known to deviate from the general population by enrolment design. For instance, the maternity cohorts included only pregnant women who are known to have lower risk of cancers of breast, corpus uteri and ovary than nulliparous women. Information on parity and age at first pregnancy is available from the databases and can be taken into account when designing studies on diseases related to reproductive parameters.

Studies on samples taken during pregnancy are not necessarily generalisable to non-pregnant women. On the other hand, these samples offer a unique possibility to study the effect of in utero exposures to the health of the children (25, 28). The large Nordic Maternity cohorts are the main source of prospective cancer cases diagnosed in ages before the age of 50 (Fig. 10).

The most extreme example of an a priori known selection was the ATBC cohort which included only smoking men, who have a more than twofold excess incidence of numerous cancer types than the average male population (Fig. 7). Clinical biobanks also deviated from population averages due to the clinical diagnostics selection process, the impact of which could not have been estimated in advance.

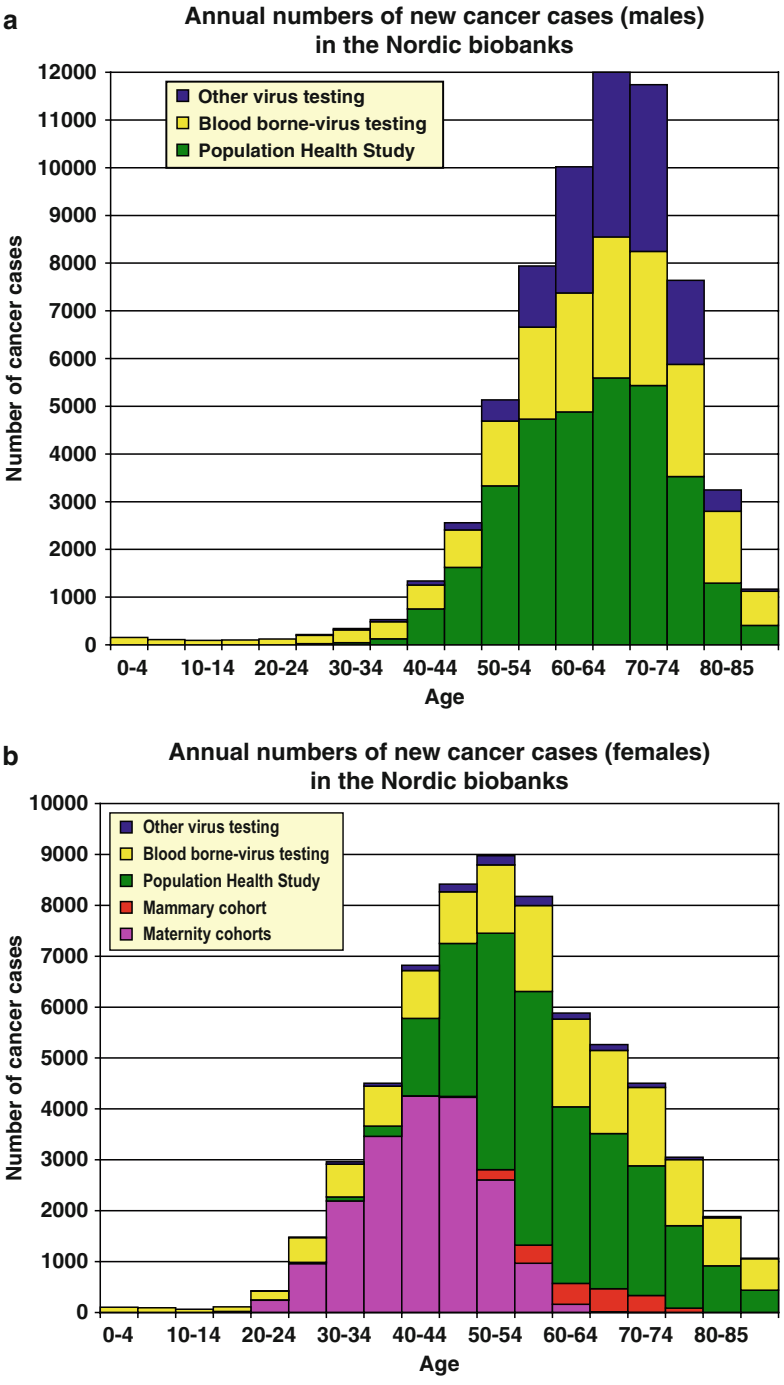


Fig. 10. Numbers of registered cancer cases among subjects in the Nordic biobanks diagnosed after serum donation, by sex, age and type of biobank. The numbers refer to cancer update status in March 2008, when coverage was complete only until 1999–2006.

The overall cancer incidence among men increases and among women decreases towards the lower socio-economic position (37, 39, 50). Typical cancers associated with low socio-economic

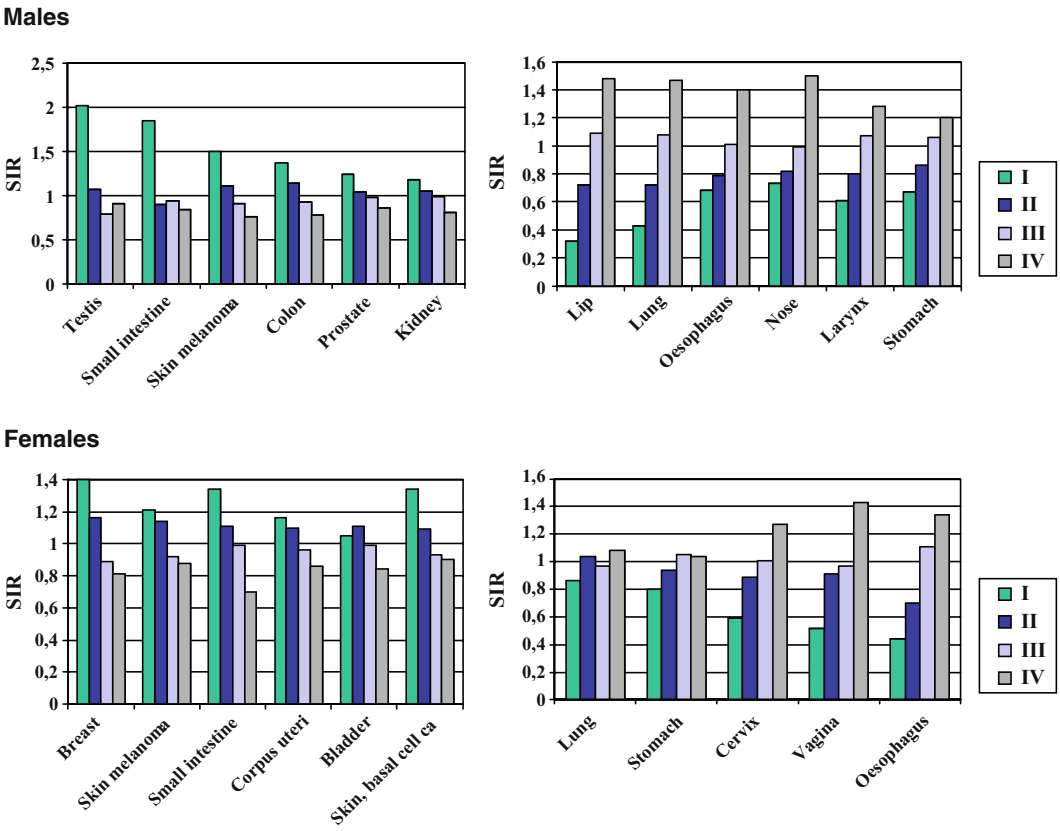


Fig. 11. Standardised incidence ratio (SIR) of selected cancers according to social class in Finland, 1971–1995. Social classes: I = managers, higher administrative; II = lower administrative/clerical; III = skilled/specialised blue-collar; IV = labourers (for details, *see* 51). Cancers of the *left* are related to high and on the *right* to low socioeconomic position.

status or educational level are cancers of the lip, oesophagus, stomach, larynx and nose, and multiple myeloma in both sexes, cancers of cervix uteri and vagina in women and lung cancer in men (Fig. 11). Cancers of the colon, breast, testis and soft tissue, and skin melanoma (especially in the trunk and limbs) are most common in high social strata. A person who knows the variation of cancer incidence over socio-economic or health habit strata can estimate from the cancer pattern whether a cohort is representative of the general population in terms of these factors.

Most biobank cohorts showed slightly lower than average cancer risk. The biobanks that were based on population registry-based invitations presumably contain a *representativity bias* related to better participation rate among health-conscious subjects. Participation rate seems not to be a especially strong indicator of this selection; e.g., the cancer pattern for the Malmö Diet and Cancer Study, with participation rate of only 40%, was rather typical for the entire population in Southern Sweden, and similar to the population samples with higher participation rates,

suggesting that selection is commonly related to a never-attending non-health-conscious population.

Some serum banks contain clearly *discernible subcohorts* with obviously different cancer incidence patterns. In nested case-control studies, it is therefore recommended to consistently match for such subcohorts. Malmö Microbiology Biobank is the best example of a biobank technically collected in same place by the same organisation, but that contains clearly discernible subcohorts enrolled for different reasons. As described in this paper, these subcohorts have clearly different background cancer risks. Matching for subcohort in case-control selection is important to maintain validity in the rate ratio estimation.

The fact that symptoms related to the outcome disease of the study may increase the likelihood for sampling will increase the likelihood to encounter reverse causality biases (mix-up of cause and effect). In Malmö Microbiology Biobank, the SIRs for liver, gallbladder and pancreatic cancer were extremely high during the first year after serum sampling. Symptoms from these cancers (such as jaundice) are likely to cause testing for hepatitis viruses. While the risk for gallbladder and pancreatic cancers were not elevated after the first year after sampling, the risk for liver cancer remained elevated, presumably reflecting a true etiologic link (such as infections with hepatitis B and C viruses being causes of liver cancer). When using clinical biobanks for prospective studies, we therefore suggest excluding samples that do have shorter follow-up between sampling and diagnosis of the endpoint disease than the length of the "sick attendee effect" as demonstrated in Fig. 4.

In the cohort collected in association of mammography screenings in Northern Sweden, there was an almost twofold incidence of breast cancer during the first year after sampling. Mammography screening is indeed expected to find non-symptomatic breast cancer cases that will have a diagnosis date shortly after the screening visit. The *cohort formation principle* therefore produces an atypical collection of breast cancers in terms of timing of diagnosis and stage distribution that must be considered if these cases are used, e.g., in studies on natural latency times.

While calculation of observed and expected rates is very helpful for characterising cohorts and estimating generalisability, it should be pointed out that the main focus of biobank-based studies is more on studies of new aetiologies than on generalising to total cancer occurrence in national populations. When cases and controls are selected from the same prospectively followed cohort (strictly defined using personal identifiers and enrolment date) representing relatively homogeneous baseline population there is internal validity and possibility to make valid aetiologic inferences regardless of the degree of population representativeness of the cohort.

### 5.6. Recommended Study Design for Biobank-Based Studies

The nested case-control design and the case-cohort design are commonly used in molecular epidemiological studies within cohorts. They are methods of sampling from an assembled cohort study (51).

In the nested case-control design, for each case controls are randomly sampled from those eligible to be controls. In the classical case-cohort design, a simple random sample of the cohort, a subcohort, is used as a comparison group for all cases in the cohort. In the stratified case-cohort design, the subcohort is selected applying stratified random sampling.

The controls for nested case-control studies are appropriately selected applying incidence-density sampling. For each case controls are randomly sampled from all control candidates alive and free of cancer at the time of case's diagnosis. A subject is eligible to be selected as a control for more than one case and a case can serve as a control for cases with earlier date of diagnosis (52). This is called sampling with replacement. The odds ratio will then be an estimate of the incidence rate ratio in the source population between those exposed and not exposed. This holds true regardless of a disease rarity assumption, provided that the control sampling is independent of the exposure given the factors used in matching (53). In sampling with replacement, there is a small probability of multiple use of the same sample, and therefore more complicated statistical pseudo-likelihood approach is usually not necessary.

In case-cohort studies, the subcohort is selected without regard to disease status. The subcohort provides information about the person-time experience in the random sample. The case-cohort design allows direct estimation of risk ratio.

Among the advantages of the nested case-control design is that there is no need to follow up the controls beyond case's diagnosis. Effects of analytic batch, storage time and freeze–thaw cycles can be removed by matching (54). The major advantage of the case-cohort design is that the subcohort can be used for several diseases and for extended follow-up.

Among the drawbacks of the nested case-control design is that the controls are not a representative sample of the cohort and thus cannot necessarily be used as controls for future cases. Control for batch and storage effects and freeze–thaw cycles is cumbersome in case-cohort design compared to nested case-control design. Batch effect will cause bias when subsequent case series are studied in case-cohort design (55).

The case-cohort design might be preferable if the biomarkers would not suffer from storage length, batch effects and freeze–thaw cycles. The nested case-control design provides tools for dealing with such issues in principle, and is therefore more appropriate design for the NBSBCCC studies. Hence, *the optimal design is the nested case-control design applying incidence-density sampling with replacement.*

*5.6.1. Matching in Nested Case-Control Design*

Matching is restriction on selection of control series. The goal of matching is to balance the ratio of cases to controls within matched sets, and to make controls' distributions of the potentially confounding matching variables more like those of cases'.

The network of Nordic biobanks has attempted to use *uniform control selection algorithms* in all biobanks participating in a given joint study. For each cancer case of interest, typically one to four control donors of same sex are randomly selected among persons who were alive at the time of case's diagnosis have donated a sample around the same time as the case and were born within two years of the case's date of birth. As pointed out above, in the case of heterogeneous biobanks, matching for subcohort (e.g. Malmö Maternity Cohort and Blood-borne virus screening within Malmö Microbiology Bank) is essential. Rather exact matching for sampling date has been considered important, because different length of storage time in the bank can have profound influence on some biological markers. For some markers, seasonal variation is so large that it is also therefore important to select the control samples from same time of the year as the sample of the case. In NBSBCCC studies typically only a difference of 1–2 months in sampling date is accepted. As freezing and thawing can affect a number of biomarkers, it is also highly recommended to match on the number of freeze-thawing cycles a sample has been subjected to. The biobanks have not necessarily recorded the numbers of freeze–thaw cycles. The effect freeze–thaw cycles should be in any case prevented by sufficient aliquoting or other suitable methods, for example the straws in the EPIC study (55). Samples of the matched set are typically pipeted in random order on same panel to minimise the effects of analytic batch.

While matching is a means of reducing bias due to confounders, matching on variables intermediate in the causal pathway between exposure and disease will bias estimates (56). This is also true for matching on variables affected by exposure and disease. Therefore, matching on other variables than those mentioned above is generally not allowed in NBSBCCC studies. Matching may increase the random error, e.g., matching on a non-confounder associated with exposure but not disease reduces efficiency. Hence, matching for only a limited number of variables, typically sex, age, storage time and subcohort, is preferable.

There are certain practices in control selection that are bound to specific features of the unique sample materials. First, because most biobank databases do not include variables indicating how many times a sample has been used as a control and how much serum is left, it is often necessary to pick up one or two *extra control candidates* that will be used if the actual controls are missing or do not contain enough materials. Second, persons who have been diagnosed with other cancers have in some studies in some biobanks not been accepted as controls (to save these valuable samples), although formally they would be eligible at least until the date of

cancer diagnosis of the respective case. This causes only a negligible theoretical error, because the pool of eligible controls for each case normally includes hundreds of subjects.

To assure protection of integrity and ensuring equal analyses of cases and controls, the samples must be blinded before they are sent to the analysing laboratory. After the laboratory analyses are ready, the researchers receive the code key that tells them which samples are cancer cases and which are controls.

### *5.7. Quality Assurance*

Since 2001, a system of Quality Assurance (QA) for Good Biobanking Practice is used on a routine basis at the Medical Biobank in Umeå. Quality control and auditing by an external expert or organisation is performed at regular intervals. QA is a process that aims at measuring, evaluating and continuously re-evaluating the quality and when required, improving the quality. The QA work should have a plan of activity and schedule for work, and all employees of a biobank should be involved in the QA system. The QA system is supposed to guarantee that the biological samples, questionnaires and data have the quality that corresponds to the intended use. Database systems that document historical storage conditions, aliquoting history, number of thawings/freezings and amounts available are highly recommended.

The QA system should include procedures for how the completeness and accuracy of the attached database (non-material part of the biobank) should be maintained, kept up-to-date and how the pitfalls of selection and follow-up biases should be traced. Many biobanks have no instruments to make basic person-year at risk calculation from their cohorts or other means to control the coverage and population representativeness of their data. We suggest that calculation of cancer incidences and SIRs should be included as a basic QA practice of essential importance in biobanking QA, which should be asked for in reviews of biobank-based studies.

Many clinical biobanks do not give high priority to such check-ups of registered data that are absolutely necessary for epidemiological follow-up studies. The system described in this paper, where the data management of clinical biobanks was entrusted to cancer registries or experienced epidemiological biobanks, is likely to be essential for valid use of clinical biobanks for epidemiological studies.

## 6. Conclusions

The high internal validity of internal comparisons within a defined biobank cohort make prospective biobank-based study designs preferable for aetiological studies. Limited population-representativeness implies that generalisation of results to entire national populations

should be made with caution. Because the described biobanks are committed to work towards joint Quality Assurance standards, including defined accessibility to external requests for samples and as the biobanks together contain a huge numbers of prospectively occurring cases of cancer, the Nordic biobank cohorts provide a solid basis for prospective studies on cancer causes and control.

In practical terms, each biobank cohort should at least once be checked using the best quality assurance methods traditionally used for many other types of study cohorts, including calculations of standardised incidence ratios and correction of any erroneous data. After that, regular simple cross-tabulations such as those described in this chapter may well be enough to keep the quality high. A real quality biobank also takes care of the future of its materials; future predictions of outcome events belong to this vision.

## Acknowledgement

The publisher of the main source article of this chapter (34), Acta Oncologica (http://www.informaworld.com) kindly granted the permission to copy parts of that article. I am also pleased to acknowledge my co-authors of that article, all of whom made a great contribution to that text: Aage Andersen and Randi Gislefoss from the Cancer Registry of Norway, Institute of Population-Based Cancer Research, Oslo, Norway; Göran Berglund from Malmö Diet and Cancer Study, Lund University, Malmö, Sweden; Joakim Dillner and Anders Widell from the Department of Medical Microbiology, Lund University, University Hospital at Malmö, Sweden; Vilmundur Guđnason from the Icelandic Heart Association, Kópavogur, Iceland; Göran Hallmans from the Department of Public Health and Clinical Medicine, Nutritional Research, Umeå University, Umeå, Sweden; Egil Jellum from the Institute of Clinical Biochemistry, Rikshospitalet University Hospital, Oslo, Norway; Pekka Jousilahti, Paul Knekt, Pentti Koskela, Matti Lehtinen, Leena Tenkanen and Jarmo Virtamo from the National Institute for Health and Welfare, Helsinki and Oulu, Finland; Pentti Kyyrönen, and Tapio Luostarinen from the Finnish Cancer Registry, Institute for Statistical and Epidemiological Cancer Research, Helsinki, Finland; Per Lenner from the Cancer Registry of Northern Sweden and Department of Radiation Sciences, Umeå University Hospital, Umeå, Sweden; Arthur Löve from the Department of Medical Virology, Landspitali-University Hospital, University of Iceland, Reykjavik, Iceland; Helga Ögmundsdóttir from the Molecular and Cell Biology Laboratory, Icelandic Cancer Society, Reykjavik, Iceland; Pär Stattin from the Department of Urology, Umeå University

Hospital, Umeå, Sweden; Laufey Tryggvadóttir from the Icelandic Cancer Registry, Reykjavik, Iceland; and Göran Wadell from the Department of Virology, University of Umeå, Sweden. Several network researchers – in addition to those listed above – gave valuable comments in topics related to this study discussed in numerous joint network meetings. Special thanks are due to Tapio Luostarinen, key statistician in the biobank-related analyses in the Finnish Cancer Registry, who drafted virtually the entire text for section 5.6.; and Timo Hakulinen (Finnish Cancer Registry) and Esa Läärä (University of Oulu, Finland) who participated actively in finalisation of the paragraphs related to the principles of theoretical statistics related to this chapter. Joakim Dillner and Matti Lehtinen as present and former coordinators of the network had an essential impact in pinpointing examples of the general strengths and weaknesses experienced in practical research projects during the years of network operation.

Jan Ivar Martinsen at the Cancer Registry of Norway, Björn Tavelin in Umeå, and Håkan Krzeszowski and Henrik Månsson in Malmö made a great effort in creating the O/E calculation procedures for biobank cohorts in Norway and Sweden. Guðríður Olafsdóttír took care of quality control of the data related to the Icelandic serum cohorts. Anna Törner kindly offered the numbers of cancer cases from the test linkage from the Swedish Institute of Infectious Disease Control. Kari Pasanen from the University of Kuopio prepared the informative map on biobanks' coverage areas and Toni Patama from the same institute the wonderful maps on cancer incidence in the Nordic countries.

The study was supported by the Nordic Council of Ministers longitudinal epidemiology programme, by the European Union fifth framework Concerted Action on Evaluation of the Role of Infections in Cancer and by the sixth framework Network of Excellence on Cancer Control using Population-based Registries and Biobanks. The map production was supported by the Nordic Cancer Union.

## References

1. Dillner, J., Knekt, P., Schiller, J.T. and Hakulinen, T.: Prospective seroepidemiological evidence that Human Papillomavirus type 16 infection is a risk factor for oesophageal squamous cell carcinoma. Br. Med. J. 311: 1346 (1995)
2. Lehtinen, M., Dillner, J., Knekt, P., Luostarinen, T., Aromaa, A., Kirnbauer, R., Koskela, P., Paavonen, J., Peto, R., Schiller, J.T. and Hakama, M.: Serologically diagnosed infection with human papillomavirus type 16 and risk for subsequent development of cervical carcinoma: nested case-control study. Br. Med. J. 312: 537–539 (1996)
3. Lehtinen, T., Luostarinen, T., Dillner, J., Aromaa, A., Hakama, M., Hakulinen, T., Knekt, P., Leinikki, P., Lumio, J., Lähdeaho, M.-L., Maatela, J., Teppo, L. and Lehtinen, M.: Serum p53 accumulation and altered antibody responses to Epstein-Barr virus proteins precede diagnosis of haemopoietic malignancies of lymphoid origin. Br. J. Haematol. 93: 104–110 (1996)
4. Dillner, J., Kallings, I., Brihmer, C., Sikström, B., Koskela, P., Lehtinen, M., Schiller, J.T.,

Sapp, M. and Mårdh, P.A.: Seropositivities to Human Papillomavirus type 16, 18, or 33 capsids and to *Chlamydia trachomatis* are markers of sexual behavior. J. Infect. Dis. 173: 1394–1398 (1996)

5. Bjørge, T., Dillner, J., Anttila, T., Engeland, A., Hakulinen, T., Jellum, E., Lehtinen, M., Luostarinen, T., Paavonen, J., Pukkala, E., Sapp, M., Schiller, J., Youngman, L. and Thoresen, S.: Prospective seroepidemiological study of role of human papillomavirus in non-cervical anogenital cancers. Br. Med. J. 315: 646–649 (1997)
6. Dillner, J., Lehtinen, M., Bjørge, T., Luostarinen, T., Youngman, L., Jellum, E., Koskela, P., Gislefoss, R.E., Hallmans, G., Paavonen, J., Sapp, M., Schiller, J.T., Hakulinen, T., Thoresen, S. and Hakama, M.: Prospective seroepidemiologic study of human papillomavirus infection as a risk factor for invasive cervical cancer. J. Natl. Cancer Inst. 89: 1293–1299 (1997)
7. Bjørge, T., Hakulinen, T., Engeland, A., Jellum, E., Koskela, P., Lehtinen, M., Luostarinen, T., Paavonen, J., Sapp, M., Schiller, J., Thoresen, S., Wang, Z., Youngman, L. and Dillner, J.: A prospective, seroepidemiological study of the role of Human Papillomavirus in esophageal cancer in Norway. Cancer Res. 57: 3989–3992 (1997)
8. Dillner, J., Knekt, P., Boman, J., Lehtinen, M., af Geijersstam, V., Sapp, M., Schiller, J., Maatela, J. and Aromaa, A.: Seroepidemiological association between human-papillomavirus infection and risk of prostate cancer. Int. J. Cancer. 75: 564–567 (1998)
9. af Geijersstam, V., Kibur, M., Wang, Z., Koskela, P., Pukkala, E., Schiller, J., Lehtinen, M. and Dillner, J.: Stability over time of serum antibody levels to Human Papillomavirus type 16. J. Infect. Dis. 177: 1710–1714 (1998)
10. Luostarinen, T., af Geijersstam, V., Bjørge, T., Eklund, C., Hakama, M., Hakulinen, T., Jellum, E., Koskela, P., Paavonen, J., Pukkala, E., Schiller, J.T., Thoresen, S., Youngman, L.D., Dillner, J. and Lehtinen, M.: No excess risk of cervical carcinoma in women seropositive for both HPV16 and HPV6/11. Int. J. Cancer. 80: 818–822 (1999)
11. Lehtinen, M., Luostarinen, T., Youngman, L.D., Anttila, T., Dillner, J., Hakulinen, T., Koskela, P., Lenner, P. and Hallmans, G.: Low levels of serum vitamins A and E in blood and subsequent risk for cervical cancer: interaction with HPV seropositivity. Nutr. Cancer. 34: 229–234 (1999)
12. Kibur, M., af Geijerstamm, V., Pukkala, E., Koskela, P., Luostarinen, T., Paavonen, J., Schiller, J., Wang, Z., Dillner, J. and Lehtinen, M.: Attack rates of Human Papillomavirus type 16 and cervical neoplasia in primiparous women and field trial designs for HPV16 vaccination. Sex. Transm. Infect. 76: 13–17 (2000)
13. Koskela, P., Anttila, T., Bjørge, T., Brunsvig, A., Dillner, J., Hakama, M., Hakulinen, T., Jellum, E., Lehtinen, M., Lenner, P., Luostarinen, T., Pukkala, E., Saikku, P., Thoresen, S., Youngman, L. and Paavonen, J.: *Chlamydia trachomatis* infection as a risk factor for invasive cervical cancer. Int. J. Cancer. 85: 35–39 (2000)
14. Sigstad, E., Lie, A.K., Luostarinen, T., Dillner, J., Jellum, E., Lehtinen, M., Thoresen, S. and Abeler, V.: A prospective study of the relationship between prediagnostic Human Papillomavirus seropositivity and HPV DNA in subsequent cervical carcinomas. Br. J. Cancer. 87: 175–180 (2002)
15. Mork, J., Lie, A. K., Glattre, E., Hallmans, G., Jellum, E., Koskela, P., Møller, B., Pukkala, E., Schiller, J.T., Youngman, L., Lehtinen, M. and Dillner, J.: Human Papillomavirus infection as a risk factor for squamous-cell carcinoma of the head and neck. N. Engl. J. Med. 344: 1125–1131 (2001)
16. Anttila, T., Saikku, P., Koskela, P., Bloigu, A., Dillner, J., Ikäheimo, I., Jellum, E., Lehtinen, M., Lenner, P., Hakulinen, T., Närvänen, A., Pukkala, E., Thoresen, S., Youngman, L. and Paavonen, J.: Serotypes of *Chlamydia trachomatis* and risk for development of cervical squamous cell carcinoma. JAMA. 285: 47–51 (2001)
17. Stattin, P., Adlercreutz, H., Tenkanen, L., Jellum, E., Lumme, S., Hallmans, G., Harvei, S., Teppo, L., Stumpf, K., Luostarinen, T., Lehtinen, M., Dillner, J. and Hakama, M.: Circulating enterolactone and prostate cancer risk: a Nordic nested case-control study. Int. J. Cancer. 99: 124–129 (2002)
18. Bjørge, T., Engeland, A., Luostarinen, T., Mork, J., Gislefoss, R.E., Jellum, E., Koskela, P., Lehtinen, M., Pukkala, E., Thoresen, S.Ø. and Dillner, J.: Human Papillomavirus infection as a risk factor for anal and perianal skin cancer in a prospective study. Br. J. Cancer. 87: 61–64 (2002)
19. Stattin, P., Lumme, S., Tenkanen, L., Alfthan, H., Jellum, E., Hallmans, G., Thoresen, S., Hakulinen, T., Luostarinen, T., Lehtinen, M., Dillner, J., Stenman, U-H. and Hakama, M.: High levels of circulating testosterone are not associated with in creased prostate cancer risk: a pooled prospective study. Int. J. Cancer. 108: 418–424 (2004)
20. Lehtinen, M., Koskela, P., Jellum, E., Bloigu, A., Anttila, T., Hallmans, G., Luukkaala, T., Thoresen, S., Youngman, L., Dillner, J. and Hakama, M.: Herpes simplex virus and risk of

cervical cancer: a longitudinal nested case-control study in the Nordic countries. Am. J. Epidemiol. 156: 687–692 (2002)

21. Lehtinen, M., Pawlita, M., Zumbach, K., Lie, K., Hakama, M., Jellum, E., Koskela, P., Luostarinen, T., Paavonen, J., Pukkala, E., Sigstad, E., Thoresen, S. and Dillner, J.: Evaluation of antibody response to human papillomavirus early proteins in women in whom cervical cancer developed 1 to 20 years later. Am. J. Obstet. Gynecol. 188: 49–55 (2003)
22. Kapeu, A.S., Luostarinen, T., Jellum, E., Dillner, J., Hakama, M., Koskela, P., Lenner, P., Löve, A., Mahlamaki, E., Thoresen, S., Tryggvadóttir, L., Wadell, G., Youngman, L., Lehtinen, M.: Is smoking an independent risk factor for invasive cervical cancer? A nested case-control study within Nordic biobanks. Am J Epidemiol. 169: 480–8 (2009)
23. Paavonen, J., Karunakaran, K.P., Noguchi, Y., Anttila, T., Bloigu, A., Dillner, J., Hallmans, G., Hakulinen, T., Jellum, E., Koskela, P., Lehtinen, M., Thoresen, S., Lam, H., Shen, C. and Brunham, R.C.: Serum antibody response to the heat shock protein 60 of *Chlamydia trachomatis* in women with developing cervical cancer. Am. J. Obstet. Gynecol. 189: 1287–1292 (2003)
24. Lehtinen, M., Koskela, P., Ögmundsdóttir, H.M., Bloigu, A., Dillner, J., Gudnadóttir, M., Hakulinen T., Kjartansdóttir, A., Kvarnung, M., Pukkala, E., Tulinius, H. and Lehtinen, T.: Maternal Herpesvirus infections and risk of acute lymphoblastic leukemia in the offspring. Am. J. Epidemiol. 158: 207–213 (2003)
25. Tuohimaa, P., Tenkanen, L., Ahonen, M., Lumme, S., Jellum, E., Hallmans, G., Stattin, P., Harvei, S., Hakulinen, T., Luostarinen, T., Dillner, J., Lehtinen, M. and Hakama, M.: Both high and low levels of blood vitamin D are associated with a higher prostate cancer risk: a longitudinal, nested case-control study in the Nordic countries. Int. J. Cancer. 108: 104–108 (2004)
26. Luostarinen, T., Lehtinen, M., Bjørge, T., Abeler, V., Hakama, M., Hallmans, G., Jellum, E., Koskela, P., Lenner, P., Lie, A. K., Paavonen, J., Pukkala, E., Saikku, P., Sigstad, E., Thoresen, S., Youngman, L.D., Dillner, J. and Hakulinen, T.: Joint effects of different Human Papillomaviruses and *Chlamydia trachomatis* infections on risk of squamous cell carcinoma of the cervix uteri. Eur. J. Cancer. 40: 1058–1065 (2004)
27. Lehtinen, M., Ögmundsdóttir, H.M., Bloigu, A., Gudnadóttir, M., Hakulinen, T., Hemminki, E., Kjartansdóttir, A., Paavonen, J., Pukkala, E., Tulinius, H., Lehtinen, T. and Koskela, P.: Associations between three types of maternal bacterial infection and risk of leukemia in the offspring. Am. J. Epidemiol. 162: 662–667 (2005)
28. Anttila, T., Tenkanen, L., Lumme, S., Leinonen, M., Gislefoss, R.E., Hallmans, G., Thoresen, S., Hakulinen, T., Luostarinen, T., Stattin, P., Saikku, P., Dillner, J., Lehtinen, M. and Hakama, M.: Chlamydial antibodies and risk of prostate cancer. Cancer Epidemiol. Biomarkers Prev. 14: 385–389 (2005)
29. Stolt, A., Kjellin, M., Sasnauskas, K., Luostarinen, T., Koskela, P., Lehtinen, M. and Dillner, J.: Maternal Human Polyomavirus infection and risk of neuroblastoma in the child. Int. J. Cancer. 113: 393–396 (2005)
30. Hakama, M., Luostarinen, T., Hallmans, G., Jellum, E., Koskela, P., Lehtinen, M., Thoresen, S., Youngman, L. and Hakulinen, T.: Joint effect on HPV16 with *Chlamydia trachomatis* and smoking on risk of cervical cancer: antagonism or misclassification (Nordic countries). Cancer Causes Control. 11: 783–790 (2000)
31. Tedeschi, R., Bidoli, E., Agren, Å., Wadell, G., De Paoli, P. and Dillner, J.: Epidemiology of Kaposi's sarcoma herpesvirus (HHV8) in Västerbotten county, Sweden. J. Med. Virol. 78: 372–378 (2006)
32. Tedeschi, R., Luostarinen, T., De Paoli, P. Gislefoss, R.E., Tenkanen, L., Virtamo, J., Koskela, P., Hallmans, G., Lehtinen, M. and Dillner, J.: Joint Nordic prospective study on human herpesvirus 8 and multiple myeloma risk. Br. J. Cancer 93: 834–837 (2005)
33. Korodi, Z., Dillner, J., Jellum, E., Lumme, S., Hallmans, G., Thoresen, S., Hakulinen, T., Stattin, P., Luostarinen, T., Lehtinen, M. and Hakama, M.: Human papillomavirus 16, 18, and 33 infections and risk of prostate cancer: a Nordic nested case-control study. Cancer Epidemiol. Biomarkers Prev. 14: 2952–2955 (2005)
34. Pukkala, E., Andersen, A., Berglund, G., Gislefoss, R., Gudnason, V., Hallmans, G., Jellum, E., Jousilahti, P., Knekt, P., Koskela, P., Kyyronen, P. P., Lenner, P., Luostarinen, T., Love, A., Ogmundsdottir, H., Stattin, P., Tenkanen, L., Tryggvadottir, L., Virtamo, J., Wadell, G., Widell, A., Lehtinen, M. and Dillner, J.: Nordic biological specimen banks as basis for studies of cancer causes and control – more than 2 million sample donors, 25 million person years and 100,000 prospective cancers. Acta Oncol. 46: 286–307 (2007)
35. ATBC (The Alpha-Tocopherol B-CCPSG): The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. N. Engl. J. Med. 330: 1029–1035 (1994)

36. Meurman, L.O., Pukkala, E. and Hakama, M.: Incidence of cancer among anthophyllite asbestos miners in Finland. Occup. Environ. Med. 51: 421–425 (1994)
37. Pukkala, E. (1995) Cancer risk by social class and occupation. A survey of 109,000 cancer cases among Finns of working age. Contributions to Epidemiology and Biostatistics. Basel: Karger; 7
38. Liu, Q., Wuu, J., Lambe, M., Hsieh, S.F., Ekbom, A. and Hsieh, C.C.: Transient increase in breast cancer risk after giving birth: postpartum period with the highest risk (Sweden). Cancer Causes Control. 13: 4, 299–305 (2002)
39. Andersen, A., Barlow, L., Engeland, A., Kjaerheim, K., Lynge, E. and Pukkala, E.: Work-related cancer in the Nordic countries. Scand. J. Work Environ. Health. 25, Suppl. 2 (1999)
40. Malila, N., Virtanen, M.J., Jarmo Virtamo, J., Albanes, D. and Pukkala, E.: Cancer incidence in a cohort of Finnish male smokers. Eur. J. Cancer Prev. 15: 103–107
41. Dreyer, L., Winther, J.F., Pukkala, E. and Andersen, A.: Tobacco smoking. APMIS, 105, Suppl. 76: 9–47 (1997)
42. Lehtinen, M., Leminen, A., Kuoppala, T., Tiikkainen, M., Lehtinen, T., Lehtovirta, P., Punnonen, R., Vesterinen, E. and Paavonen, J.: Pre- and post-treatment serum antibody responses to HPV16 E2 and HSV-2 ICP8 proteins in women with cervical carcinoma. J. Med. Virol. 37: 180–186 (1992)
43. Lehtinen, M., Hakama, M., Knekt, P., Heinonen, P.K., Lehtinen, T., Paavonen, J., Teppo, L. and Leinikki, P.: Serum antibodies to the HSV-2 specified major DNA-binding protein are elevated before the diagnosis of cervical cancer. J. Med. Virol. 27: 131–136 (1989)
44. Wallin, K.L., Wiklund, F., Luostarinen, T., Hallmans, G., Anttila, T., Koskela, P., Lehtinen, M., Paavonen, J., Stendahl, U., Wadell, G. and Dillner, J.: *Chlamydia trachomatis* infection: a risk factor in cervical cancer development – a population based prospective study. Int. J. Cancer. 101: 371–374 (2002)
45. Clarce, R., Shipley, M., Lewington, S., Youngman, L., Collins, R., Marmot, M. and Peto, R. Underestimation of risk associations due to regressios dilution in long-term follow-up of prospective studies. Am. J. Epidemiol. 150: 341–353 (1999)
46. Parish, S., Collins, R., Peto, R., Youngman, L., Barton, J., Jayne, K., Clarke, R., Appleby, P., Lyon, V., Cederholm-Williams, S., Marshall, J. and Sleight, P.: Cigarette smoking, tar yields, and non-fatal myocardial infarction: 14,000 cases and 32,000 controls in the United Kingdom. The International Studies of Infarct Survival (ISIS) Collaborators. Br. Med. J. 311: 471–477 (1995)
47. Jellum, E., Andersen, A., Lund-Larsen, P., Theodorsen, L. and Orjasaeter, H.: Experiences of the Janus Serum Bank in Norway. Environ. Health Perspect. 103, Suppl. 3: 85–88 (1995)
48. Pukkala, E. Use of record linkage in small-area studies. In: Elliot, P., Guzick, J., English, D., Stern, R., eds. (1992) *Geographical and environmental epidemiology*. Oxford: Oxford University Press: 125–131
49. Teppo, L., Pukkala, E. and Lehtonen, M.: Data quality and quality control of a population-based cancer registry. Acta Oncol. 33: 365–369 (1994)
50. Pukkala, E. and Weiderpass, E.: Time trends in socio-economic differences in incidence rates of cancers of the breast and female genital organs (Finland 1971–1995). Int. J. Cancer. 81: 56–61 (1999)
51. Langholz, B. Entries: case-cohort study and case-control sudy, nested. In: Armitage, P., Colton, T., eds. (1999) *Encyclopedia of biostatistics*. Chichester: John Wiley & Sons: 497–503 & 514–519
52. Greenland, R. and Thomas, D.C.: On the need for the rare disease assumption in case-control studies. Am. J. Epidemiol. 116: 547–553 (1982)
53. Rothman, K.J. and Greenland, S. (1998) *Modern Epidemiology 2nd ed*. Philadelphia: Lippincott-Raven: 95–96
54. Rundle, A.G., Vineis, P. and Ahsan, H.: Design options for molecular epidemiology research within cohort studies. Cancer Epidemiol. Biomark. Prev. 14: 1899–1907 (2005)
55. Riboli, E., Hunt, K.J., Slimani, N., Ferrari, P., Norat, T., Fahey, M., et al.: European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr. 5: 1113–1124 (2002)
56. Wacholder, S., Silverman, D.T., McLaughlin, J.K. and Mandel, J.S.: Selection of controls in case-control studies. III. Design options. Am. J. Epidemiol. 135: 1042–1050 (1992)
57. Pukkala, E., Söderman, B., Okeanov, A., Storm, H., Rahu, M., Hakulinen, T., Becker, N., Stabenow, R., Bjarnadottir, K., Stengrevics, A., Gurevicius, R., Glattre, E., Zatonski, W., Men, T. and Barlow, L. (2001) Cancer atlas of Northern Europe. *Cancer Society of Finland* Publication No. 62, Helsinki (69 pages + CD)