Acta Oncologica, 2010; 49: 368–377

informa
healthcare

ORIGINAL ARTICLE

Ensuring quality in studies linking cancer registries and biobanks

HILDE LANGSETH[1], TAPIO LUOSTARINEN[2], FREDDIE BRAY[1,3] & JOAKIM DILLNER[4,5]

[1]*The Cancer Registry of Norway, Institute of Population-based Cancer Research, Oslo, Norway,* [2]*Finnish Cancer Registry, Institute of Statistical and Epidemiological Cancer Research, Helsinki, Finland,* [3]*Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway,* [4]*Department of Laboratory Medicine, Lund University, Malmo, Sweden and* [5]*Department of Laboratory Medicine, Karolinska Institute, Stockholm, Sweden*

Abstract

The Nordic countries have a long tradition of providing comparable and high quality cancer data through the national population-based cancer registries and the capability to link the diverse large-scale biobanks currently in operation. The joining of these two infrastructural resources can provide a study base for large-scale studies of etiology, treatment and early detection of cancer. Research projects based on combined data from cancer registries and biobanks provides great opportunities, but also presents major challenges. Biorepositories have become an important resource in molecular epidemiology, and the increased interest in performing etiological, clinical and gene-environment-interaction studies, involving information from biological samples linked to population-based cancer registries, warrants a joint evaluation of the quality aspects of the two resources, as well as an assessment of whether the resources can be successfully combined into a high quality study.

While the quality of biospecimen handling and analysis is commonly considered in different studies, the logistics of data handling including the linkage of the biobank with the cancer registry is an overlooked aspect of a biobank-based study. It is thus the aim of this paper to describe recommendations on data handling, in particular the linkage of biobank material to cancer registry data and the quality aspects thereof, based on the experience of Nordic collaborative projects combining data from cancer registries and biobanks.

We propose a standard documentation with respect to the following topics: the quality control aspects of cancer registration, the identification of cases and controls, the identification and use of data confounders, the stability of serum components, historical storage conditions, aliquoting history, the number of freeze/thaw cycles and available volumes.

The Nordic countries have a long tradition in cancer registration as well as the collection of biological samples from both healthy and diseased individuals. More than 50 years of experience in the registration, storage and analyses of cancer data provides an unparalleled platform for cancer research in the Nordic countries. Linking the cancer registry and biobank material together provides detailed information on cancer diagnoses among persons who have donated samples to a biobank.

The traditional function of population-based cancer registries – as producers predominantly of descriptive information on cancer incidence has changed markedly over the course of several decades, particularly in higher-resource countries. In the Nordic countries, the Registries have continued to develop their research programmes, and have been involved in many collaborative molecular epidemiologic studies that link biological samples to case records held at the Registries. Evidently, results derived from such a combination of registry and biological material will only be of optimal scientific value if the quality of both sources can jointly be assured, although at present there are few examples of such an evaluation. One recent study was able to report a high degree of completeness of cervix cancer registrations at the Cancer Registry of Norway as well as for the biological samples (paraffin blocks) stored in selected laboratories [1].

Certainly the principles and methods for the evaluation of data quality at a Cancer Registry are well-established, for which there are four quality indicators: comparability, completeness, validity and timeliness of the data. A two-part review examining the practical aspects and techniques for addressing these aspects was recently published [2,3]. The quality and completeness of cancer registration in the Nordic countries

Correspondence: Hilde Langseth, The Cancer Registry of Norway, Institute of population-based Cancer Research, Department of Etiological Research, Postbox 5313 Majorstuen, 0304 Oslo, Norway. Tel: +47 22451300. Fax: +47 22451370. E-mail: hilde.langseth@kreftregisteret.no

(Received 16 August 2009; 27 October 2009)

ISSN 0284-186X print/ISSN 1651-226X online © 2010 Informa UK Ltd. (Informa Healthcare, Taylor & Francis AS)
DOI: 10.3109/02841860903447069

has been evaluated in a number of studies over the years and the Nordic Cancer Registries are considered reasonably complete for most cancer sites [4–7].

The potential for Nordic biological specimen banks as the basis for research studies in the domains of cancer aetiology and cancer control has been reviewed previously [8]. More than 100 000 malignant neoplasms have occurred among two million sample donors contributing approximately 25 million person years.

The linkages of cancer registry data and biobank material create a study base with long-term follow-up and an increasing number of cancer cases among the donors. The personal identification number established in the 1960s, or earlier, in the Nordic countries allows one to link data from different sources and to follow a person from early life until a cancer diagnosis, survival, death or in rare cases loss to follow-up. Pre-diagnostic samples are valuable in identifying biomarkers for early detection of cancer and to investigate possible associations between life course exposures and risk of cancer, as illustrated in Figure 1. The time between the initiation of disease and cancer diagnosis is the time window where novel markers of early diagnosis can be used. Biobank-based epidemiological studies typically aim to estimate the strength of association of the hypothesised causes of a given cancer, to distinguish between the determinants of disease and confounders or mere correlates, and to report the specific interactions between different causes and/or confounders. The parameter of interest of these etiological studies is thus biologically meaningful risk estimates.

In biobank-based studies, data quality concerns are present in each phase of the study - from the donation of the biological specimen, to the long-term storage of study files. The following overview deals mainly with study logistics in interfaces between biobanks and cancer registries, rather than quality assurance of the biobank material itself. The description of sample handling, storage protocols and quality indicators for the collection, processing and archiving has been reviewed elsewhere [9–13].

The aim of this paper is thus to describe the linkage of biobank material and cancer registry data and quality aspects thereof, and give an overview of study design and logistical considerations, based on experiences from joint collaborative studies between the Nordic cancer registries and biobanks.

Design and methods in biobank- and registry-based studies

The choice of study design depends on the proposed outcome of interest. The most widely used and recommended study design in biobank-based epidemiological research is the nested case-control design. A case-control study is an epidemiological study in which, rather than measuring the experience of an entire population to obtain rates, controls are sampled from the source population of cases, and risk ratios or odds ratios can be estimated. In the nested case-control design controls are selected from the cohort members who are still disease-free at the time a case occurs (insidence density sampling). The control group provides an estimate of the exposure distribution in the source population for cases and is a substitute for the denominators of rates or risks [14]. An efficient case-control study design enables the calculation of risk estimates of not only common but also rare diseases, as well as diseases with a long induction phase, as is the case for most cancer forms.

When case-control studies are nested within the prospectively-followed cohorts, the study design preserves the validity of a cohort study. The nested case-control design is superior for the study of biomarkers where the measurements may be influenced by analytic batch, long-term storage and/or freeze-thaw cycles.

The case-cohort sampling design allows controls to be selected from a random sample of the whole cohort at the start of the follow-up. The sub-cohort

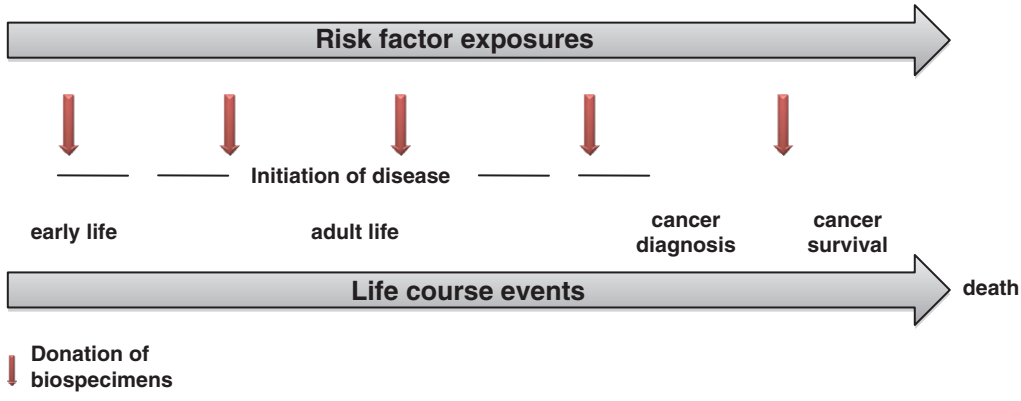

Figure 1. Illustration of potential risk factor exposures during life course and intermittent donation of biospecimens to biorepositories.

is representative of the full cohort and not only non-cases, in that a person ascertained as a case may also be selected as a control, and vice versa. Every person in the source population has the same chance of being included as a control, regardless of how much time that person has contributed to the person-time of the cohort. All subcohort members at risk at the time of the cases' diagnoses serve as controls. In cancer epidemiology studies, sampling may have to be stratified by important confounders such as age at serum sampling. This is a commonly-used design for the purpose of biobank studies in the Nordic countries [8,16,17]. Some laboratory analysis may be very expensive and necessitate keeping the number of controls to a minimum. In the case-cohort study design the same controls can be used for more than one case. The drawback is multiple use of non-renewable material. The case-cohort design is particularly suitable when there are several outcomes of interest, given that the measurements on stored materials remain sufficiently stable during the study.

Identification of cases

Biobank studies deal with non-renewable resources, and thus it is critical to select appropriate groups of cases as fitting for the specific aims of the study. All the donors of eligible cohorts, who were diagnosed with the disease of interest with relevant lagtime between specimen donation and cancer diagnosis are included as cases. In most Nordic biobank studies the cases are identified restricted to first cancer diagnosis. Case identification by second or third cancer is rare. The cases are identified by linking the personal identification number in the biobank to the cancer registry. This linkage has to be performed in every new study since the cancer registry is a dynamic database and the number of new cases among the donors will increase over time. Case ascertainment will be incomplete if not all specimens are recorded, or cannot be identified for other reasons. This may cause selection biases and reduces precision.

In some research projects it may be feasible to select samples from one specific point in time when the exposure of interest was highest and samples from all other time periods are not relevant due to lack of desired serum component. Samples from a specific point in time may be particularly relevant in studies on environmental, dietary or medication exposures.

However, the most typical approach is to include all cancer cases of a specific site during the follow-up period and select either the first available pre-diagnostic sample of the cases or the sample closest in time to the diagnosis. One criterion in choosing relevant lag time is that the disease of interest should not have any major effect on the exposures to be measured. If the purpose of the study is to investigate early detection of cancer one will benefit from using the sample closest in time to diagnosis. If the purpose is to investigate lifetime exposures it will be better to use samples with a long time between sample collection and cancer diagnosis. Another approach is to select serial samples from each case, where the aim of the study is to investigate changes in biomarkers at different time points before the cases are diagnosed.

In every research project it is important to be in close dialogue with the principle investigator (PI) of the study when defining the selection criteria. The first selection of cases is made using the specific criteria in the study protocol. For quality assurance purposes the PI can be given the opportunity to check that the first selection seems reasonable. They may, for example, wish to check that all cases have the cancer of interest registered as the first diagnosis. In this case, the files are thereafter returned to the responsible registry that completes the selection.

Selection of controls

The selection of controls should be by random sampling among all those eligible at the time of the diagnosis of the case, not just those that appear suitable for matching. The practice of selecting "best fitting controls" (the controls that among eligibles have values of matching variables closest to those of cases) and/or convenience sampling (selecting the controls that are stored adjacent to the cases and/or listed next to the cases on sample inventories) [15] is not recommended. When using such a sample of controls, it is not possible to make generalizations about the total cohort as the samples are not drawn at random from the eligable samples in the cohort. A person selected as a control that remains in the study population at risk after selection should remain eligible for further selection as a control. Moreover, a person selected as a control that later develops the disease, and then selected as a case, should be included in the study both as a control and as a case. Sampling with replacement, which allows a control specimen to occur more than once as a control, is in harmony with the incidence density sampling for matched case-control studies [14]. In biobank studies it is common to identify 1-2 substitutes for each control that fulfil the same matching criteria as all controls. Those will be used in case the control specimen cannot be located or the serum volume is too low.

The controls have to be at risk for the outcome, in other words, alive at the time of diagnosis. Therefore vital status is an important variable. Follow-up should include cancer occurrence, death and emigration keeping the loss to follow-up to a minimum.

Most biobanks are followed up through national population-based registries, but some are currently followed up within a region. Follow-up should not continue beyond the date of emigration even if the subjects immigrate again later; follow-up is not valid because events occurring outside of the country will not have been captured. If a project is set up as a multicentre study it is recommendable that equal sampling principles are followed at all study centres.

In some cases it might be adequate to reuse the controls from earlier studies. Reusing an existing control group as reference subjects for cases of another disease is reasonable only if the necessary risk factors, confounders, and modifiers have been measured both on the old controls and the new cases, and when similar quality assessments have been made [18]. This requires a high degree of stability of the analyte in the sample and high reproducibility of assays. If re-use of controls is considered, a new control group should also be selected, and data from the old and new control groups tested for homogeneity.

Matching

Matching refers to the selection of control series with respect to the distribution of potentially confounding factors. Matching is a procedure whereby controls are selected in such a way that the distribution of potential confounders among them will be identical to those of the cases. It is usually chosen as a means of increasing efficiency, although it does not necessarily do so. Increased efficiency will be the case only if the matching factors are strong risk factors and related to the exposure. Matching on a non-confounder associated with exposure status but not disease, reduces efficiency. Matching may introduce selection biases unless the matching variables are stratified on (the strength of the association is measured separately within each well defined and homogeneous category (stratum) of the confounding variable). Common matching variables are sex, age, date of diagnosis, date of blood collection and batch type. Other relevant matching variables may be storage time and thawing status. Studies with a high number of matching variables may reduce substantively the number of eligible controls. For example, in studies utilizing serial samples, it may be impossible to find a control with the same number of samples, so that the other variables are satisfactorily matched on. In such instances a categorized number of serum samples are preferable in the study protocol to limit the length of the time window. Matching for a limited number of variables has been recommended in a recent paper reviewing Nordic biobank studies [8]. Actual sampling dates at the level year-month-date (yymmdd) instead of approximate ones are favourable. For instance seasonal matching may be important in micronutrient studies where detailed information on date is essential in evaluating seasonal variations of vitamin levels [19,20].

When matching controls, special concern should be placed on defining eligibility. In a cohort in use, there is always a trade-off between the need for saving rare cases for future use and the need to match controls from the entire cohort to avoid selection bias. Historically, two main strategies have been used. The first is to use the whole cohort, pick out extra controls and then exclude controls not eligible. In the second approach, non-eligible controls are excluded before the matching of controls; for instance if healthy controls developed relatively rare diseases such as liver cancer or multiple myeloma at a later point in time, it seems reasonable to not select those samples, because they probably are valuable for future studies. The exclusion, from the reference group, of subjects who develop other diseases, which may be relevant to the exposure of interest, have been reported to involve quite small bias [21]. In Nordic biobank studies this is assumed to cause only a negligible theoretical error, since the number of eligible controls is large [8]. We assume that this issue has hitherto not been a scientific problem, as the selection of such valuable samples has been probalistically unlikely, on account of their rarity. However, with the number of studies nested in biobanks greatly increasing, the phenomenon has become sufficiently common to have epidemiological implications: controls saved for future use as cases may have different risk factor profiles than the controls used. The biobank will therefore need to prioritize between the validity of the controls in one study against feasibility of certain future studies. Irrespective of how the prioritization is done, it should be documented and the possible epidemiological implications of the decision evaluated and discussed. Computing an estimate of the proportion of saved cases, and how this may have influenced on the risk estimates in different studies would be valuable information to give the magnitude of this problem.

If the aim of the study is to elaborate upon the disease risk related to exposure in the general population, biobanks selected by health behaviour and/or risk factors and/or sickness, should be avoided. As population representativeness in practice commonly differs from that intended, the use of estimating standardized incidence ratios for a number of different cancer sites as a proxy for estimating population representativeness is highly recommended [8]. An example of donors not representative of the general population are the healthy Red Cross blood donors that constitute 10% of the Janus Serum Bank of Norway [13]. However, it should be noted that for many etiological studies, population representativeness is not necessary for

unbiased results. As prospectively-followed biobanks have a high within-cohort validity of studies, highly selected cohorts may also be highly informative. As long as the biobank holds a relatively large number of disease cases and eligible controls, they are in any case valuable in research projects. Biobanks often contains different sample collections with different selection and sampling criteria, and in such instances, matching on sample collection (sub-biobanks with the same selection and sampling criteria) is recommended. One should, for example, always match population-based cases to population-based controls, blood donor cases to blood donor controls and so on. Since blood donors are supposed to be selected from the healthiest fraction of the population, their use as controls for population-based cases is likely to result in an overestimation of risk.

Quality aspects of historical repositories

Pre-analytical sample handling is very important for the quality of the biological samples. A study that aimed to evaluate the frequency and types of mistakes in a laboratory found that pre-analytic mistakes constituted 68% of all errors [22].

Specimens in a biobank are commonly used several times for new studies and analyses. In more elderly biobanks, the samples are often not aliquoted and the same samples may need to be thawed several times for different studies. Some have tried to assess the impact of thawing status on the samples, and found this of minor importance for specific components [23,24]. The extent to which repeated thawing affects the quality of the specimens depends both on the analyte under study and on the thawing procedure. One commonly-used procedure for thawing samples (thawing at +4°C) gives adequate results for most analytes, but was found to give gross errors in the analysis of a specific analyte (von Willebrand factor) that cryoprecipitated under these conditions (Göran Hallmans, pers. comm.) and procedures had to be changed. We recommend that the number of thawings should not differ markedly and must not differ systematically between the cases and the controls. Matching for the number of freeze-thaw cycles is an option that should be considered if this information is recorded in the biobank of interest. However, the best way to prevent the effect of freeze-thaw cycle is to ensure a high degree of standardization due to biological samples from both cases and controls in a given study [25].

Stability studies are very important in assessing the quality of biobanks, especially in older samples. Long-term storage may introduce a considerable bias for valuable components due to degradation, and the ability to measure the effect of the storage time is thus of great importance. Careful matching on storage time might be essential for the comparison of biobank material, and large variations in the stability between selected proteins have been reported. Immunoglobulins seem to be relatively stable while some enzymes are particularly vulnerable [26]. Also, differences in sample handling before freezing may introduce bias on vulnerable components in serum. When an analysis of components with unknown stability is intended, we advise that this should only be done when there is evidence that the variability is sufficiently limited that the study can be informative. Measurements of analytes with substantial variability (either biological variability and/or instability on handling, storage and measurement) is scientifically meaningless, because studies will give null results (because of regression dilution bias), regardless of whether an association truly exists or not. Analyte stability studies can therefore be said to be the "biological equivalent to a statistical power calculation" as they will determine whether a study can answer the question or not. Pilot studies measuring the analyte of interest in serial samples from the same subjects are particularly informative since they measure both the biological variability and stability on handling, storage and measurement. Pilot studies to measure the level of components in samples with different storage time, and exploration of their distribution are also informative [27]. Such studies support the adoption of matching cases and controls with respect to storage time as routine practice in epidemiological studies where biobank material is used. An example of this is a study from the Finnish Maternity Cohort that showed that serum samples, stored long-term, can be used to study hormone-disease associations, and provided a close matching on storage time [28].

The PIs of studies often want to exclude possible false diagnoses and confirm the histology of cases which may have been diagnosed decades ago. Paraffin blocks are identified using data at the cancer registry, searched for in the pathological laboratory archives and diagnoses are verified or otherwise by pathologists on re-review. Such histological verification may sometimes lead to a change in the diagnosis registered in the cancer registry database, and is a good example of the principle that an active scientific research programme at a cancer registry itself promotes and enhances the quality of the stored data.

Linkages and logistics

In most biobank studies, several data sources are involved. It is therefore necessary that all the collaborators have a joint understanding of the logistics concerning them and that it is possible to monitor the logistics and the quality indicators over the entire project period of the study. All studies should have a

simple *"Logistics Scheme"* that outlines exactly who does what and specifies the required documentation and quality indicators at each step of the study. There has to be a very good co-operation between the PI and the investigators from the different cancer registries (potentially with different coding practices) and biobanks, in order that unwanted material from unsuitable cases will not be sent for analyses. The Logistic Scheme should name a Coordinator who should coordinate and monitor project progress. The Coordinator is sometimes the same person as the person charged with the scientific responsibility (the Principal Investigator), but experience indicates that the appointment of a Coordinator at the postdoctoral or postgraduate student level tends to provide more rapid progress in research projects. The Logistics Schemes must name the number of aliquots, the aliquot volumes and the analyzing laboratories who should receive the sample aliquots. The Logistics Schemes must also give a description of the unique study code to be used on each individual case and control. The definition of this code is essential and is the responsibility of the PI. The study code must not contain patient numbers or identities, running numbers or numbers revealing case-control sets. Design of study code numbers must be coordinated and must not be the responsibility of the biobanks participating in joint studies as there is a definite risk that codes assigned by different biobanks may overlap, resulting in non-unique IDs (e.g. if running numbers without a biobank ID are used). Study codes should not reveal any information with respect to identity or case-control status to the analyzing laboratories; – this is an essential quality criterion for ethical, high-quality and credible studies. The codes should, however, contain information enabling the identifying as to which study subjects belong to and to which biobank they originate from. Also linkage to the personal identity code is necessary for control purposes and for future use. Biobanks must follow the Logistics Schemes when assigning study codes and should send a file with the list of assigned study codes to the analyzing laboratories, hence minimising the risk of typing errors.

In all biobank research projects an essential aspect of quality is that the laboratory analysis is carried out blinded to avoid potential biases. In the Janus serum bank of Norway a code-keeping system has been in routine practice for many decades. The code of the case-control status is revealed only after completion of all the laboratory analyses. After the case-control status is revealed, no additional analysis of the samples is allowed. The laboratory data is deposited with the biobank before the case-control status is revealed. This is an important procedure to prevent any data manipulation, as it is possible to check whether the scientific papers only contain the data obtained by the analysis of coded samples. The code-keeping system also ensures a high degree of confidentiality and patients' names and personal identification numbers are never disclosed.

The first step in the research study is to link the biobank to the cancer registry to identify cases and controls. This gives information on the cancer diagnosis, sample collection date and the location of the sample at the storage facility. Before retrieving the samples a pathologist should verify all cancer diagnosis, since cases may have been diagnosed over a wide time period with different diagnostic routines and coding practices. After retrieval of all samples, a complete list of study codes and the coded biological samples are sent to the laboratories. When analyses are completed, the code list and the data from the analyses of the samples are returned to the biobank. The biobank is then tasked with adding all the information required for the study (e.g. the case-control status) and forwarding the file with the study data to the PI, or to a statistical analysis centre specified in the Logistics Scheme. It is important to ensure that the data file released to the PI does not contain accessory information on subjects at a detail that identification of individual subjects would be possible. While we recommend use of exact dates of diagnosis, birth date, sampling date etc. in study design and case-control selection, it is sometimes prudent to release only year and month data in the study file to the PI.

In some biobank projects, it is necessary to link data from different data sources to get information on confounding variables such as occupation, diet and smoking habits. In such circumstances, it is recommended that the list of personal identification numbers is sent to those institutions whom have the required data and that they themselves do the linkage to obtain the necessary additional information. For instance National Bureaus of Statistics or Public Health are used commonly as external institutions that store and give access to these type of data. If the unique study code is sent from the biobank together with the personal identifiers, the external institutions can then release the coded data (without identifiers) directly to the PI. As this means that not only the biobank but also an external institute has access to the code. An alternative approach is one that includes one extra administrative step, but avoid ever releasing the code to reveal personal identities from the biobank. The approach involves having the external institute return the study-specific data to the biobank that can then send this on to the PI in a single delivery. During the linkages it is important that each person in the study population should be assigned a unique code. A flowchart of the study logistics and data sources normally applied in Nordic biobank studies is illustrated in Figure 2.

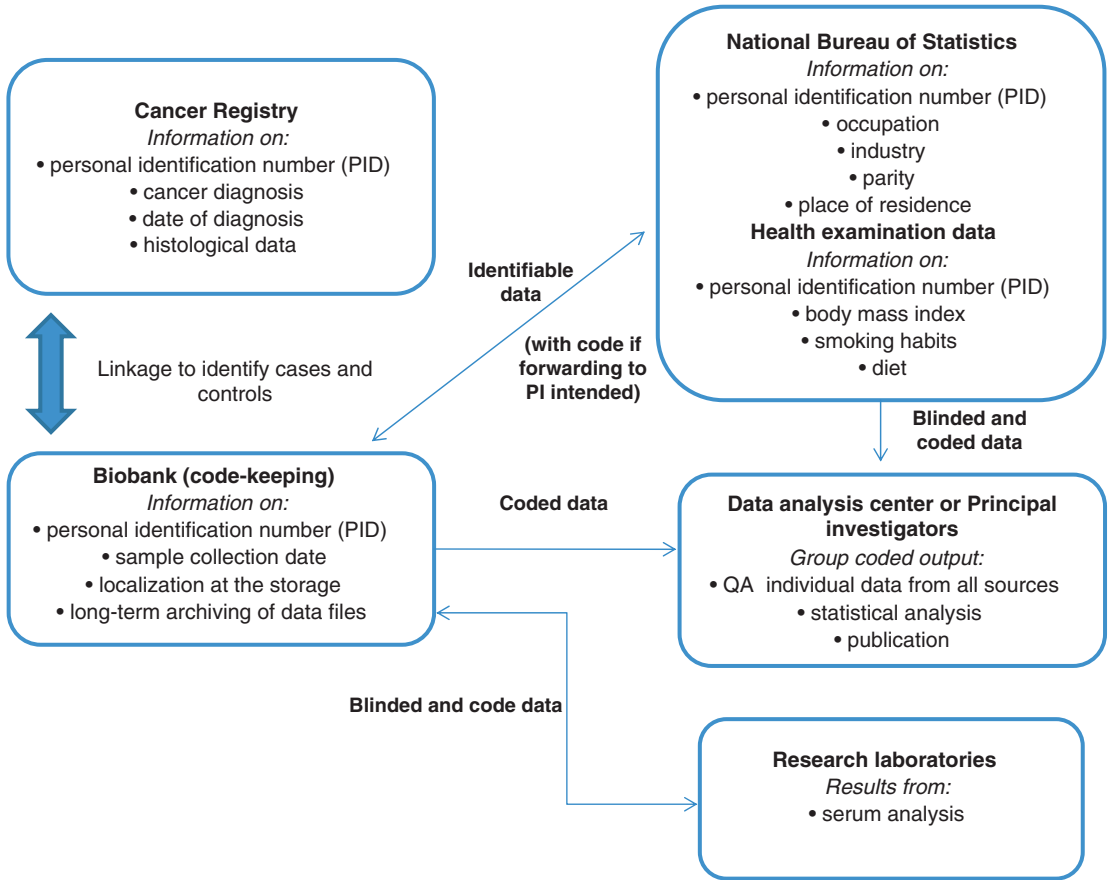

Figure 2. Flowchart providing an example of the study logistics and data sources in a joint Nordic Cancer Registry-Biobank study.

Laboratory methods

Appointing a coordinating laboratory, which takes care of well organised storage, aliquoting, labeling and distribution of the samples, is recommended. This is particularly important in multi-centre studies. The coordinating laboratory has to ensure that specimen boxes or single specimens will not be lost, that specimens' exposure to light, heat and other deleterious physical conditions are minimised. To avoid the possibility for differential errors in handling between potential cases and controls, the coordinating laboratory should be blinded to the identity of samples and they should not be sorted in any systematic order. The laboratories should be accredited and as few as possible. The biobanks should aliquote the samples, if there is not a coordinating laboratory. It is particularly important to organize the samples in analytical batches so that all samples of the matched sets are on the same panel in a randomized order.

To minimise potential biases due to any changes over time in handling or analytic procedures, the order of the sample batches should be randomized. Precision and power can be increased by ensuring that a case and its matched control are handled and analysed closely together. It is recommended to organize the samples in analytical batches where all samples of the matched case-control sets are in the same batch. The order of the samples in such batches must be randomized. Each batch of samples sent out should contain coded quality control samples as well as blinded repeat samples to estimate intra-assay variability. The blinded repeat samples often use other materials than the unique biobank material (e.g. routine blood donor samples), which is only valid if the repeat samples are comparable to the study samples (e.g. they have similar levels of the substance analysed). Files with laboratory analysis results must clearly distinguish between data below detection limit and missing data, so that results on empty tubes will not be interpreted as negative findings. It should be recommended that any sample residues from research projects are to be returned to the respective biobank or destructed.

Data management and statistical analyses

The data providers should send files of the requested, or at least convertible, file type. The data file should follow the file specification in the Logistic Scheme where possible, but as some deviation is often necessary, specifications should always be enclosed when

data is sent. Study codes other than those of the requested type cause delays and increase the likelihood of errors in the data management phase. Keys to data coding systems, e.g. histology codes, should be attached.

The writing of an explicit *"Data Management Report"* should become common practice and an important element of the study protocol, as this is essential for ensuring accuracy in describing the procedures undertaken, particularly when writing the scientific papers. The person responsible for data handling should write the report, especially if statistical analyses are conducted elsewhere. The report should contain information on i) identification of cases; ii) cases loss to follow-up; iii) matching criteria; iv) identification of controls; v) controls loss to follow-up; vi) final numbers of cases and controls; vii) dates of serum sample thawing (data on exposure to light, temperature); viii) missing data stratified by variable and by biobank and ix) any data errors found. The file record specification should be enclosed as an appendix.

A statistician must be involved in the study design phase in order to avoid errors, which may be impossible to correct for in later statistical analyses. Specific requests for data analyses should be made *a priori*, i.e. they should be formulated before the data is obtained and be a part of the study protocol/ Logistics Scheme describing which statistical methods and how they will be applied. Exploratory analyses are subjective.

Laboratory methods often have imperfect sensitivity and specificity. Statistical analyses should be run both with and without correction for misclassification. It must be emphasized that misclassification correction rests on assumptions and should never be interpreted as having removed bias due to misclassification. Such analyses are useful mostly to explore whether the conclusion would have been materially affected by measurement errors [29].

Historical data base system for biobank samples

There are still many biobanks where recorded specimen information is, at least partly, paper-based. High priority should be given to build data-base systems for recording identifiable specimen data. The computerization of available volumes is of benefit to the researchers and to the quality of samples themselves. In biobank studies it is common that the number of study subjects is smaller than originally planned because of inadequate serum volumes, which may jeopardise the power of the study. Tracking of the volumes not only ensures more realistic study sizes but also decreases wasted searches in the freezer.

The Alpha-Tocopherol-Beta-Carotene (ATBC) cancer prevention study and the Janus Serum Bank records information on serum volume, but have older specimens without volume information, and there is a clear need for more complete data-bases for many biobanks. In old repositories with a large number of samples it is of significant costs to register the specimens' volume all at a time. One alternative then is to index the volume at time of sample retrieval in different projects. Software for increasing biobank documentation and protection for biobank-based studies are available. The Laboratory Information Management System (LIMS) is a computer software that is used in the laboratory for the management of samples, laboratory users, instruments, standards and other laboratory functions such as invoicing and work flow automation. Systems that are even more tailor-made for biobanks (Biobank Information Management Systems (BIMS)) are in development.

Long-term archiving of the study files

Long-term archiving of the entire study file containing information from all participating biobanks should be organized. The institute responsible for the long-term storage should be stated on the logistics scheme. The long-term archive file should be based on the unique study code and should not contain any personal identifiers. The code(s) that may be used to link the unique study code to the personal identifiers should be archived at each biobank that originated the samples. It is common that there may be requests for actual crude data for clarifications, new research hypotheses that require further laboratory analyses adjusting for the previously-analyzed exposures, or requests from collaborators involved in pooling studies that require information from a study published many years earlier prevents fraud as personal identifiers allows rechecking. It should be possible to respond to such requests in an efficient manner. Because of the lengthy time spans involved and the requirement for unified data storage formats, it is recommended that the custody of the long-term archive is the responsibility of the host institution, and that they should continue storing the data after the study is completed. In contrast, requests for stored data that are linked to identifiable individuals are exceptionally rare, and are mostly restricted to follow-up (longitudinal) investigations of the same study and to specific fraud investigations. It is therefore sufficient from a practical and ethical point of view that codes enabling linking to personal identifiers remains in the biobank. Permissions to use data collected in previous studies for new purposes, should be given by the institution owning the data, and by the national Ethical Committees.

The person responsible for data handling should, at the end of the study, stratify the study file according to biobank and should then send these lists to

the biobanks for long-term storage. The official, valid and correct study file, is the one stored at the datacentre, and contains the relevant information on all biobanks collaborating in the specific study. The biobanks who provided the samples are formally responsible for blocking further use of records of donors who may retract informed consent (clear guidelines on this issue are needed via legislation in some countries) and can do linkage to the same subject in follow-up studies. It is therefore necessary in future studies that a file listing the personal identifiers, specimen codes, and study codes of the study subjects are retained at the biobank.

Summary and concluding remarks

The present paper has underlined the value of research projects based on combining data from cancer registries and biobanks. While such linkages offer unique study opportunities, they also create substantial logistical challenges. The paper gives recommendations on important quality aspects like matching procedures. Matching on confounders should be avoided, while matching on variables known to affect the measurement of the specimens are recommended. In general, the number of matching variables should be kept at a minimum. Biorepositories have become an important resource in molecular epidemiology and have led to an increased interest in etiological, clinical and gene-environment-interaction studies [30]. In linking information from biological samples to population-based cancer registries, we have emphasised the need to jointly evaluate the quality of the sources and to adhere to documented logistic processes that optimise their value in cancer research.

Use of biological material and the data attached to them are subjects to strict ethical and legal restrictions. This paper has focused on procedures required to obtain high quality scientific data (maximizing usefulness), but it must be emphasized that procedures must also consider the ethico-legal and societal aspects involved. An ethical framework for previously-collected biobank samples and data has been elaborated on in previous articles [31,32].

It is well established that generations of excellent biobank-based studies require the involvement of researchers qualified to respond to chemical, medical, biological, statistical, genetic as well as ethical questions. Expertise in data handling and quality assurance systems is increasingly emerging as key requirements for scientists involved in biobanking. Explicit documentation of good practice is required to ensure the high quality of scientific output, and to make it possible to compare results from different biobanks. The peer-reviewed publication of scientific studies, and the sharing of experiences and common practices in biobanking are the drivers that will enable the research community to achieve this goal.

Acknowledgements

This work was funded by European Union Sixth Framework Programme "Cancer Control using Population-based Registries and Biobanks" (Contract nr. LSHC-CT-2004-503465). We gratefully acknowledge Göran Hallmans, Matti Hakama, Timo Hakulinen and Matti Lehtinen for their valuable comments to the manuscript.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

[1] Bilet EF, Langseth H, Thoresen SO, Bray F. Completeness of invasive cervical cancer at the Cancer Registry of Norway. Acta Oncol 2009;48(7):1070–3.

[2] Bray F, Parkin DM. Evaluation of data quality in the cancer registry: Principles and methods. Part I. Comparability, validity and timeliness. Eur J Cancer 2009;45(5):747–55.

[3] Parkin DM, Bray F. Evaluation of data quality in the cancer registry: Principles and methods Part II. Completeness. Eur J Cancer 2009;45(5):756–64.

[4] Barlow L, Westergren K, Holmberg L, Talback M. The completeness of the Swedish Cancer Register: A sample survey for year 1998. Acta Oncol 2009;48:27–33.

[5] Larsen IK, Smastuen M, Johannesen TB, Langmark F, Parkin DM, Bray F, et al. Data quality at the Cancer Registry of Norway: An overview of comparability, completeness, validity and timeliness. Eur J Cancer 2009;45(7):1218–31.

[6] Teppo L, Pukkala E, Lehtonen M. Data quality and quality control of a population-based cancer registry. Experience in Finland. Acta Oncol 1994;33:365–9.

[7] Storm HH. Completeness of cancer registration in Denmark 1943-1966 and efficacy of record linkage procedures. Int J Epidemiol 1988;17:44–9.

[8] Pukkala E, Andersen A, Berglund G, Gislefoss R, Gudnason V, Hallmans G, et al. 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 2007;46:286–307.

[9] Elliott P, Peakman TC. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. Int J Epidemiol 2008; 37:234–44.

[10] Sjoholm MI, Dillner J, Carlson J. Assessing quality and functionality of DNA from fresh and archival dried blood spots and recommendations for quality control guidelines. Clin Chem 2007;53:1401–7.

[11] Ericsson C, Franzen B, Nister M. Frozen tissue biobanks. Tissue handling, cryopreservation, extraction, and use for proteomic analysis. Acta Oncol 2006;45:643–61.

[12] Holland NT, Smith MT, Eskenazi B, Bastaki M. Biological sample collection and processing for molecular epidemiological studies. Mutat Res 2003;543:217–34.

[13] Jellum E, Andersen A, Lund-Larsen P, Theodorsen L, Orjasaeter H. Experiences of the Janus Serum Bank in Norway. Environ Health Perspect 1995;103(Suppl 3):85–8.

[14] Rothman KJ. Epidemiology. An introduction. Oxford University Press, Oxford 2002.

[15] Stolt A, Kjellin M, Sasnauskas K, Luostarinen T, Koskela P, Lehtinen M, et al. Maternal human polyomavirus infection and risk of neuroblastoma in the child. Int J Cancer 2005;113:393–6.

[16] Rundle AG, Vineis P, Ahsan H. Design options for molecular epidemiology research within cohort studies. Cancer Epidemiol Biomarkers Prev 2005;14:1899–907.
[17] Kulathinal S, Karvanen J, Saarela O, Kuulasmaa K. Case-cohort design in practice - experiences from the MORGAM Project. Epidemiol Perspect Innov 2007; 4:15.
[18] Saarela O, Kulathinal S, Arjas E, Laara E. Nested case-control data utilized for multiple outcomes: A likelihood approach and alternatives. Stat Med 2008;27: 5991–6008.
[19] Tretli S, Hernes E, Berg JP, Hestvik UE, Robsahm TE. Association between serum 25(OH)D and death from prostate cancer. Br J Cancer 2009;100:450–4.
[20] Tuohimaa P, Tenkanen L, Ahonen M, Lumme S, Jellum E, Hallmans G, et al. 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 2004;108:104–8.
[21] Lubin JH, Gail MH. Biased selection of controls for case-control analyses of cohort studies. Biometrics 1984;40: 63–75.
[22] Plebani M, Carraro P. Mistakes in a stat laboratory: Types and frequency. Clin Chem 1997;43:1348–51.
[23] Peakman TC, Elliott P. The UK Biobank sample handling and storage validation studies. Int J Epidemiol 2008;37 (Suppl 1):i2–i6.
[24] Hakama M, Hakulinen T, Kenward MG, Aaran RK, Aromaa A, Knekt P, et al. Blood biochemistry and the risk of cancer. Acta Oncol 2004;43:667–74.
[25] Riboli E, Hunt KJ, 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 2002;5:1113–24.
[26] Gislefoss RE, Grimsrud TK, Morkrid L. Stability of selected serum proteins after long-term storage in the Janus Serum Bank. Clin Chem Lab Med 2009;47:596–603.
[27] Gislefoss RE, Grimsrud TK, Morkrid L. Long-term stability of serum components in the Janus Serum Bank. Scand J Clin Lab Invest 2008;68:402–9.
[28] Holl K, Lundin E, Kaasila M, Grankvist K, Afanasyeva Y, Hallmans G, et al. Effect of long-term storage on hormone measurements in samples from pregnant women: The experience of the Finnish Maternity Cohort. Acta Oncol 2008;47:406–12.
[29] Kuha J, Skinner C, Benichou J. Misclassification error. In Encyclopedia of Epdiemiological Methods. Gail MH, Benichou J, editors. Wiley & Sons, Chichester 2000. pp. 578–585.
[30] Wild CP. Environmental exposure measurement in cancer epidemiology. Mutagenesis 2009;24:117–25.
[31] Helgesson G, Dillner J, Carlson J, Bartram CR, Hansson MG. Ethical framework for previously collected biobank samples. Nat Biotechnol 2007;25:973–6.
[32] Hansson MG. Ethics and biobanks. Br J Cancer 2009; 100:8–12.