SPANDIDOS PUBLICATIONS

EXPERIMENTAL AND THERAPEUTIC MEDICINE 3: 540-546, 2012

CREST biorepository for translational studies on malignant mesothelioma, lung cancer and other respiratory tract diseases: Informatics infrastructure and standardized annotation

DONATELLA UGOLINI[1,2], MONICA NERI[3], LUCA BENNATI[1], PIER ALDO CANESSA[4], GEORGIA CASANOVA[1], CECILIA LANDO[5], GIACOMO LEONCINI[6], PAOLA MARRONI[7], BARBARA PARODI[8], CLAUDIO SIMONASSI[9] and STEFANO BONASSI[3]

[1]Department of Oncology, Biology and Genetics, University of Genoa; [2]Unit of Epidemiology, Biostatistics and Clinical Trials, National Cancer Research Institute, Genoa; [3]Unit of Clinical and Molecular Epidemiology, IRCCS San Raffaele Pisana, Rome; [4]Unit of Pneumology, Ospedale San Bartolomeo, Sarzana; [5]Unit of Molecular Epidemiology, National Cancer Research Institute; [6]Unit of Thoracic Surgery, Azienda Ospedaliera Universitaria San Martino; [7]Unit of Clinical Pathology, National Cancer Research Institute; [8]Biological Bank and Cell Factory, National Cancer Research Institute; [9]Unit of Pneumology, Azienda Ospedaliera Villa Scassi, Genoa, Italy

Received October 3, 2011; Accepted November 29, 2011

DOI: 10.3892/etm.2011.416

Abstract. Advances in molecular epidemiology and translational research have led to the need for biospecimen collection. The Cancer of the Respiratory Tract (CREST) biorepository is concerned with pleural malignant mesothelioma (MM) and lung cancer (LC). The biorepository staff has collected demographic and epidemiological data directly from consenting subjects using a structured questionnaire, in agreement with The Public Population Project in Genomics (P^3G). Clinical and follow-up data were collected. Sample data were also recorded. The architecture is based on a database designed with Microsoft Access. Data standardization was carried out to conform with established conventions or procedures. As from January 31, 2011, the overall number of recruited subjects was 1,857 (454 LC, 245 MM, 130 other cancers and 1,028 controls). Due to its infrastructure, CREST was able to join international projects, sharing samples and/or data with other research groups in the field. The data management system allows CREST to be involved, through a minimum data set, in the national project for the construction of the Italian network of Oncologic BioBanks (RIBBO), and in the infrastructure of a pan-European biobank network (BBMRI). The CREST biorepository is a valuable tool for translational studies on respiratory tract diseases, because of its simple and efficient infrastructure.

Correspondence to: Dr Donatella Ugolini, Department of Oncology, Biology and Genetics, University of Genoa, National Cancer Research Institute, Largo R. Benzi, 10-16132 Genoa, Italy
E-mail: donatella.ugolini@istge.it

Key words: molecular epidemiology, mesothelioma, lung neoplasms, specimen handling, translational research, Italy

Introduction

Lung cancer (LC) is the most common cause of cancer-related death in the world. In the US, over 90 million individuals are at risk for developing LC, and this disease is estimated to remain a major health problem for at least the next 50 years (1).

Malignant pleural mesothelioma (MPM) is an aggressive tumor generally attributable to asbestos exposure. Although this disease has a long latency period following exposure to the carcinogen, once detected it is rapidly fatal, with the median survival time being less than 1 year after diagnosis (2). The interest in MPM has recently increased, because of the expected peak of cancer incidence in the second decade of the new century (3).

The measurement of biomarkers in blood or tissue specimens has become an integral component of translational cancer research, with applications to studies of cancer aetiology, treatment and prognosis, including early cancer detection (4). These studies warrant the development of biorepositories capable of providing biological samples that meet the current demands of the cancer research community. Biorepositories, with an associated archive of epidemiological and clinical data, are among the most powerful and essential resources for molecular epidemiology and translational studies. They are also considered as fundamental research infrastructures in systems medicine approaches to cancer research (5). Biospecimen banking is an important complement to evaluate individual exposure to carcinogens and to assess genetic damage and individual susceptibility.

The project of establishing a biorepository dedicated to MPM and LC [Cancer of Respiratory Tract (CREST)] was

initiated several years ago at the National Cancer Research Institute in Genoa, Italy, a region with a high mesothelioma incidence due to asbestos exposure (6), which is still ongoing. Some of the main features of the CREST project have been previously published (7), but only a brief description of the informatics infrastructure and standardized annotation was provided. Here, we focus our attention on these aspects, with the aim to encourage dialogue with other biorepositories and to develop a wider network where the information can be easily exchanged for research purpose, in the frame of the European Biobanking and Biomolecular Resources Infrastructure (BBMRI) under construction (8).

Materials and methods

The CREST biorepository was established in 1996 as a part of the Biological Resource Centre (CRB-IST) of the National Cancer Research Institute, Genoa, Italy (IST).

Specimens archived in the CREST biorepository are classified according to the following categories of subjects: i) patients with MPM and LC; ii) patients with non-neoplastic respiratory conditions; and iii) referent subjects (hospital or population controls).

Whole blood, plasma, serum, lymphocytes, pleural fluid, saliva and tissue biopsies were collected before any medical treatments, including surgery. Sample collection, transportation, treatment and storage were performed according to international standards (9-14).

From January 31, 2011, the overall number of recruited subjects was 1,857 (454 LC, 245 MM, 130 other cancers and 1,028 controls). The biorepository included a total of 12,747 specimen aliquots (5,204 serum, 3,594 plasma, 2,352 whole blood for DNA extraction, 880 pleural fluid and 717 lymphocytes aliquots) in addition to 184 biopsies available for research purposes. The small number of available biopsies is mostly due to the previous use of these specimens for internal and collaborative research. Furthermore, it is not always possible to collect biopsies for research purposes during critical clinical procedures, such as thoracoscopy or bronchoscopy.

A questionnaire was administered to patients and controls. Before contributing any type of biological sample, all subjects were informed concerning the ethical code of CRB-IST, the aims of the project and concerning the use of the samples, and were requested to sign an informed consent.

Approval for recruitment was obtained from the IST Ethics Board. All ethically relevant procedures were planned in agreement with international guidelines for biorepositories. Individual data included in the CREST databank were treated in accordance with Italian privacy regulation (DL 196/2003).

Results

Questionnaire. A structured questionnaire was administered to collect information. The questionnaire was in agreement with the guidelines provided by the Public Population Project in Genomics (P^3G) and the Data Schema and Harmonization Platform for Epidemiological Research (DataSHaPER) (15). DataSHaPER is both a scientific approach and a suite of practical tools, aimed to facilitate the prospective harmonization of emerging biorepositories, to provide a template for retrospective synthesis and to support the development of questionnaires and information-collection devices, even when the pooling of data with other biorepositories is not foreseen (16). The questionnaire was structured as follows (Table I):

General data. This section of the questionnaire included the date of interview, hospital and department/unit (for cases and hospital controls) or recruitment location (for population controls) where the samples were collected, and demographic data of the subjects, including gender, age, place of birth, residence and education. Highest grade completed at school was used as a proxy for socio-economic level.

Occupational history. Participants were asked to describe all occupations held for at least 1 year. They provided the job title and described their major duties, equipment/materials/chemicals used while performing the job, and the type of work that the employer company did. They were also asked how long they had been employed in each job. Each specific job was coded using the Census Code of Occupational Classification (17). Occasional and/or spare time occupation was used to evaluate exposure unrelated to the job and was included only if done at least 6 months a year for 10 years. Information regarding heavy occupational asbestos exposure of cohabitant family members (typically father or husband) was also collected. The purpose of these questions was to obtain a rough estimate of exposure to asbestos or other toxic substances.

Since there is no universally approved method in the scientific literature to identify subjects exposed to asbestos and divide them in classes according to intensity of exposure, we used the Italian Mesothelioma Registry Job Exposure Matrix (JEM), and our experts classified asbestos exposure by type of job and nature of the manufacturing, taking into account possible non-occupational sources. Exposure was categorized as high, low and absent.

Smoking history. This section of the questionnaire included data on age at smoking initiation, number of cigarettes per day, duration of possible cessation intervals and computed pack-years. A cigarette smoker was defined as a person who had smoked at least 20 packs of cigarettes in a lifetime or at least 1 cigarette per day for at least 1 year. A former smoker was a person who had quit smoking at least 1 year before diagnosis (cancer cases) or the interview (respiratory disease patients and controls). A current smoker was someone who currently smoked or who had stopped less than 1 year before diagnosis or interview. Pack-years was calculated by multiplying the number of packs of cigarettes smoked per day by the number of years the person had smoked. Questions on passive smoking were also included.

Dietary pattern. This section included information on the weight and height of the subject, fruit and vegetable consumption (average per year), alcohol consumption and intake of dietary supplements (vitamins and minerals) in the last 6 months.

Anamnesis section. This section included questions of relevance to patient health. Family health history included data on the health of first-degree relatives for familial aggregation studies.

Clinical data collection and follow-up. Clinical features concerning all hospitalized patients were collected through

Table I. Structured questionnaire.

General
Subject ID number and Progressive number
Date of interview
Type (case or control)
First and last name
Hospital/location (where the collection took place)
Department/Unit (where the collection took place)
Gender
Date of birth, age
Place of birth
Highest school grade completed
Residence
From ... to ... (calendar year)
Address
Occupational history
From ... to ... (calendar year)
Duration (years)
Occupation
Occupation code (ISTAT code)
Name, address, activity of the company
Job description
Asbestos exposure code
Toxic substance code
Asbestos exposure of cohabitant family members
Occasional and/or spare time occupation
Have you ever worked in your spare time or do you have a hobby?
From ... to ... (calendar year)
Occupation
Asbestos exposure code
Toxic substance code
Smoking history
Have you ever smoked?
Age at smoking initiation/cessation - smoking duration
Cigarettes per day
Pack/years
Why was smoking stopped
Passive smoking at home
Passive smoking at workplace
Diet
Present weight
Normal weight
Height
Fruit and vegetable consumption
Dietary supplement consumption
Alcohol consumption
Anamnesis
Main respiratory diseases
Other main diseases
Family health history
Family members (list of all biological first-degree relatives)
Have any of your biological first-degree relatives ever had cancer?

Table I. Continued.

Clinical data and follow-up
Source
Medical record number
Diagnosis
Histology
Diagnostic modality
Staging/grading
Anamnesis
Laboratory test result
Performed treatments
Life status (alive/dead) and cause of death
Samples
Storage data
Aliquot number and location
Date of quality control
Serum
Plasma
Whole blood
Lymphocytes
Pleural fluid
Biopsies
Saliva

a linkage with institutional medical records. Histological parameters for patients with a cancer diagnosis were collected from the same source or directly from pathology records. To evaluate the impact of clinical parameters on disease progression and survival, subjects were actively followed up by searching clinical record archives. A collaboration was established with local cancer registries in the area where the CREST biorepository recruits subjects, namely the Genoa Cancer Registry and the Regional Mesothelioma Registry. The life status was requested by the registry office where the subject resided.

Biological sample storage. Samples were aliquoted and stored in ISO-certified laboratories. Specimen tubes were labeled by the CREST staff with a unique numeric code representing each subject identity code and ensuring anonymity and respect for privacy. The same code was also reported on the questionnaire forms and was the key for linking the specimen to the donor.

CREST data management: database structure. To store specimen information and to manage the link with clinical and epidemiologic data, a database was designed using Microsoft Access, the proprietary relational database management system for Microsoft Windows.

Defining the database structure was the primary task. In the relational model, the information was stored in tables, each dealing with a single class of information. Before designing the database, standard operating procedures (SOPs) for the management of work processes and data were prepared. The SOPs provided detailed written instructions for all aspects of

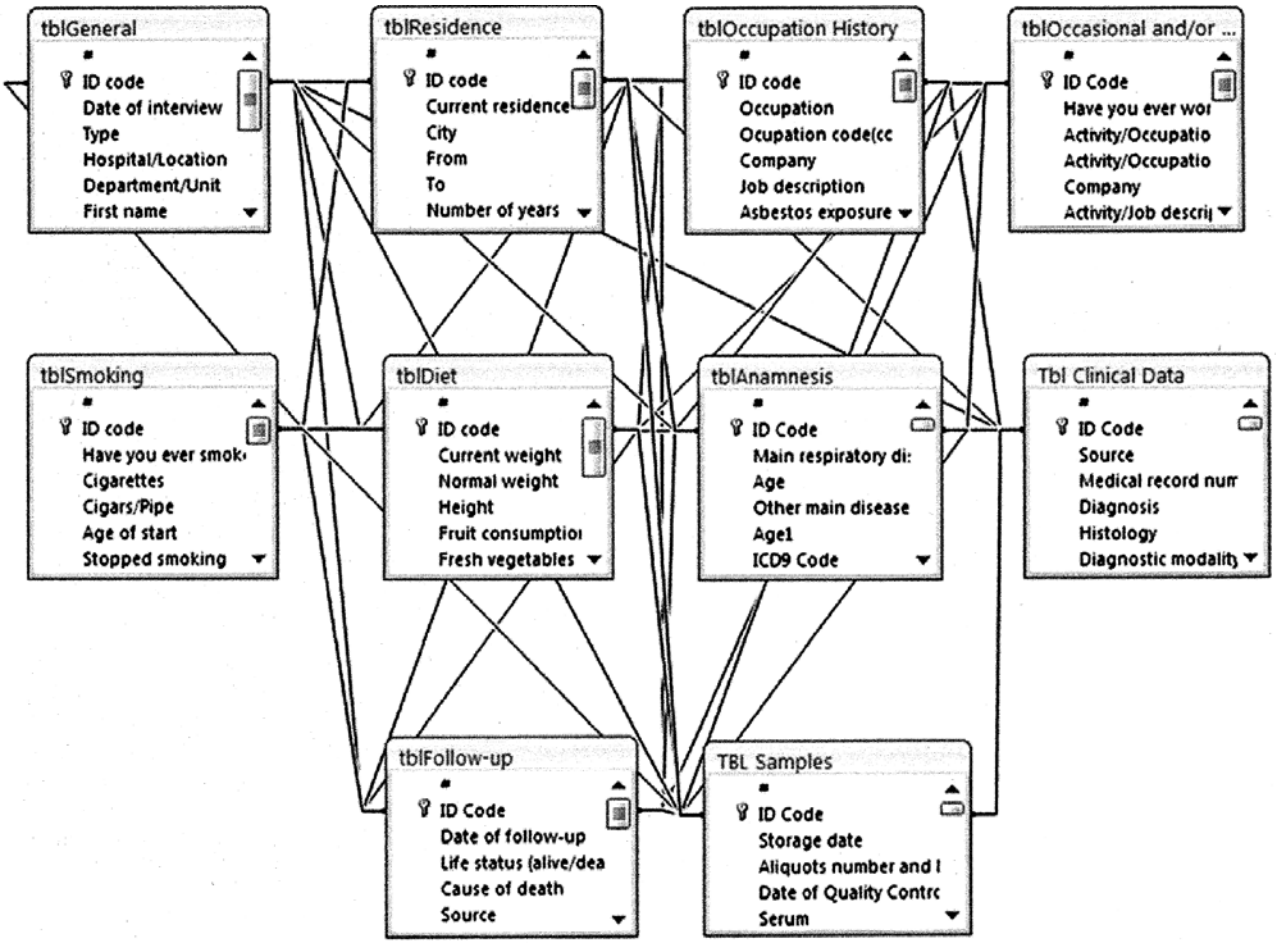

Figure 1. Relationship between the tables.

the process and a general framework for the implementation of all functions and activities and for data management.

After the analysis of data to manage, 10 tables were created containing four categories of information: general, lifestyle, clinical and management. Tables (e.g., General, Residence) were made up of one or more columns (or fields; e.g., subject ID, age, gender), and a given column (e.g., subject ID) may appear in more than one table in order to indicate a relationship between tables. The tables were inter-linked by the unique Subject ID number.

Microsoft Access provided a graphical method for creating relationships (Fig. 1).

After choosing the table structure, Microsoft Access was used to design a function to name the fields and define their properties. The program allows various field types (text, number, data/time). Each field may have its own set of properties. We derived the data in certain fields using mathematical functions, assigned default values or validation rules, defined input masks, or declared indices to permit rapid data input and search. Each table structure was saved with a unique name (e.g., tblGeneral, tblResidence).

It is possible to add or delete fields, modify field sizes, change data type without compromising the information already contained in the database.

Daily working with the data: masks, queries, reports, security features. We may use the datasheet view as the data entry screen, but a form-oriented screen (mask) was designed in order to aid the data manager to enter data correctly (Fig. 2). Microsoft Access masks provide a quick and easy way to modify and insert records into the databases. Command buttons, scrolling text boxes, image boxes, borders, were added to customize forms, with one of the many 'wizards' Microsoft Access provides to automate the process. Input masks provide a large amount of data validation features and prevent users from entering invalid data. Input masks also help ensure that users enter data in a consistent way. Data consistency is absolutely required not only for a safe database maintenance and for easy information retrieval, but particularly to allow unbiased and meaningful statistical analysis.

The user interface reflected the questionnaire and was intuitive and user-friendly. Data can be sorted in several ways, but the Query tool provides a powerful way to select, sort and calculate data from one or more tables. In addition, the Reports tool enables the data to be summarized from many fields and records even in different tables.

Security features include a password-based authentication of all database users, with the definition of specific rights for each user, and coding of data that allows identification of the

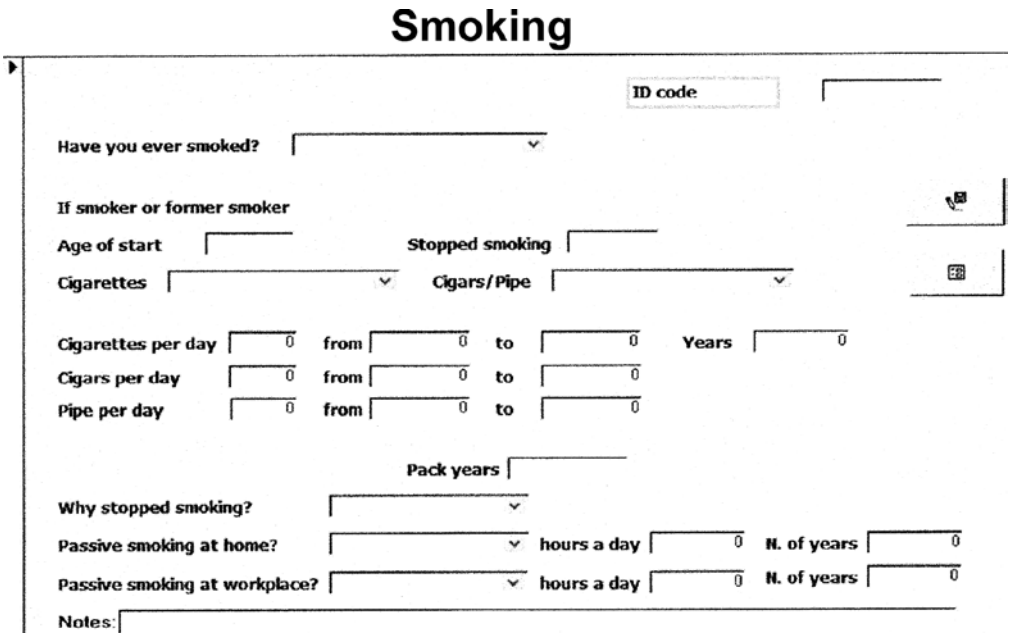

Figure 2. Example of a mask view.

subject. Surname and name of the subjects are stored in a different database, which can only be accessed by the responsible data security. Furthermore, the CREST staff must sign the CRB-IST Confidentiality Disclosure Agreement, based on a set of procedures recently approved by the local ethics committee.

Ensuring data consistency: procedure standardization. Data standardization, commonly called coding, is the process of ensuring that all data are of the same type or class and conform to an established convention or procedure to ensure consistency and to enable analysis. Coding data is the process of assigning numeric or alpha information to question responses that do not ordinarily return to the researcher in that format. Demographic data usually require coding, e.g., age and gender, but other information types can be codified as well.

In addition to assigning codes to answers returned from survey participants, we assigned codes to provide other questionnaire information, such as project code, interviewer code, date and time codes, and location.

Once we assigned the coding, we created the respective Microsoft Access table, so that in most cases, instead of typing a text description, it was possible to select the code from the list derived from the specific table that appeared in a pull down menu in order to avoid misspelling. We also developed a code book as an Appendix to the SOP. The code book provided other necessary information particularly useful for computer programs. This code map indicated where the columns are located on a data sheet, size of data fields and code type. A bit of thought and planning before actual coding saves significant time in data processing and analysis. Data normalization not only makes data more accurate, but also makes it easier to analyze.

To summarize, the strategy of normalization followed three different lines of action: i) we were strongly oriented towards normalization and the information or single elementary data. Multiple choice forms and the decomposition of information in simpler data forms are largely used. ii) We used, where possible, the codification of official classifications, both international (ICD9 classification relative to pathology, morphology) and national (activities and profession ISTAT 1991). iii) We gave particular attention to the codification of missing data. The code for the insertion of incomplete data is '-9'.

Discussion

The availability of a variety of epidemiological and clinical data and of biological samples, in different study groups, covering different time frames of the natural history of the disease, i.e., early stages, tumor development and survival, allows a wide range of investigations addressing the most important priorities in the field. Among these, include the role of genetic factors in the aetiology of disease and the impact of individual susceptibility, the importance of synergy between risk factors (e.g., asbestos and SV40), the development of new biomarkers for early diagnosis and the choice of new targets to improve therapeutic performance. Furthermore, the presence of stored biological material allows for the testing of future hypotheses concerning disease risk factors and early biomarkers of exposure. Indeed, recently, systems approaches have been advocated as pivotal in cancer research (18) and, in this setting, biorepositories such as CREST are recognized as an essential research tool (5).

Moreover, the added value of the CREST biorepository can be summarized in regards to three major issues. i) The collection, concurrently with the biological sample, of an extensive set of individual clinical and epidemiological data facilitates the realization of molecular epidemiology, translational and systems medicine studies. ii) The availability of epidemiological data and biological samples from different control groups, i.e., non-neoplastic diseases and controls, improves the potential to understand various steps in the cancer pathway and to implement primary and secondary prevention strategies.

iii) The selection of a homogeneous set of cancers, i.e., LC and MM, which share common risk factors and are treated in the same departments, allow the easier collection of specimens and administration of the questionnaire.

Much has been written concerning the collection and preservation of biological samples, while little exists regarding informatics infrastructure and standard annotation. Information systems and data management are an integral part of any biorepository, and efficient and effective information system support is mandatory for the viability of biorepositories.

Whereas most biorepositories are designed exclusively for managing biological samples, we planned to connect the sample information with environmental and clinical information.

Collecting biological specimens and samples and storing them in biorepositories plays an important role in the advancement of medical science, but equal or higher importance is information regarding the database design, planning and management. The model of the management of biorepositories presented herein is simple and efficient, and has enabled entry in network projects at the national and European levels.

Several molecular epidemiology and translational studies have been already performed based on the CREST biorepository. The endpoints measured in these studies included: levels of circulating proteins (HER-2/neu, mesothelin, osteopontin, mutated K-ras), growth factors (PDGF-AB, HGF, EGF, SCF) or antibodies (anti-p53), erythrocyte glycophorin-A variants, DNA adducts, micronuclei, DNA repair activities, SV40, single-nucleotide polymorphisms in metabolic, DNA repair or other genes, and microRNA. A list of published studies has been previously reported (7), while most recent publications have been added to the references (19-26).

In addition, due to our CREST infrastructure we are participating in some of the International Lung Cancer Consortium (ILCCO) pooled analysis. The ILCCO is an international group of lung cancer researchers, established in 2004, with the aim of sharing comparable data from ongoing lung cancer case-control and cohort studies (27). Questionnaire data from a total of 26,000 case-control pairs, and the biological samples from the majority of the subjects are available. The studies are from different geographical areas and ethnicities. Upon joining the ILCCO consortium and its research projects, the CREST biorepository has contributed to the achievement of greater research power, particularly for subgroup analyses, the reduction in the duplication of research efforts, replication of novel findings and substantial cost savings through large collaborative efforts.

Our biorepository is operating as part of an institutional network where individual banks can work together to develop consistent approaches to biobanking and, in some instances, the pooling of resources.

The developed data management system allows us, through a minimum data set, to enter the national project for the construction of the Italian Network of Oncologic BioBanks (RIBBO) initiated in 2007 and supported by the Italian Minister of Health in collaboration with Alliance Against Cancer (ACC), a federation of Italian Cancer Comprehensive Centers, and in the infrastructure of a pan-European Biobank Network (BBMRI), that will be integrated in an infrastructure for research on many diseases, including cancer. Major synergism of existing databases is reached by the realization of networks of biorepositories, which interlink the resources of the individual biorepositories in order to increase scientific excellence and efficacy of biomedical European research. This approach will allow the expansion and secure the competitiveness of European research and industry in a global context, particularly in the field of medicine and biology.

Acknowledgements

The authors thank Michela de Astis for the skilled technical help. The biorepository was supported, in part, by grants from the Fondazione Buzzi-Unicem per la Ricerca sul Mesotelioma, the Associazione Italiana per la Ricerca sul Cancro (AIRC), and the University of Genoa.

References

1. Jemal A, Siegel R, Xu J, *et al*: Cancer statistics, 2010. CA Cancer J Clin 60: 277-300, 2010.
2. Robinson BW and Lake RA: Advances in malignant mesothelioma. N Engl J Med 353: 1591-1603, 2005.
3. Robinson BWS, Musk AW and Lake RA: Malignant mesothelioma. Lancet 366: 397-408, 2005.
4. Bonassi S and Neri M: Genetic biomarkers in human population studies. In: Handbook of Genomic Medicine. Willard HF and Ginsburg GS (eds.). Elsevier, 2008.
5. Tramontano A and Valencia V: Education and research infrastructures. In: Cancer Systems Biology, Bioinformatics and Medicine. Cesario A and Marcus F (eds.). Springer, 2011.
6. Gennaro V, Ugolini D, Viarengo P, *et al*: Incidence of pleural mesothelioma in Liguria Region, Italy (1996-2002). Eur J Cancer 41: 2709-2714, 2005.
7. Ugolini D, Neri M, Canessa PA, *et al*: The CREST biorepository: a tool for molecular epidemiology and translational studies on malignant mesothelioma, lung cancer, and other respiratory tract diseases. Cancer Epidemiol Biomarkers Prev 17: 3013-3019, 2008.
8. Yuille M, van Ommen GJ, Bréchot C, *et al*: Biobanking for Europe. Brief Bioinform 9: 14-24, 2008.
9. National Cancer Institute, National Institute of Health, US Department of Health and Human Services: National Cancer Institute. Best practices for biospecimen resources. June 2007. http://biospecimens.cancer.gov/practices/.
10. ISBER (International Society for Biological and Environmental Repositories): Best practices for repositories: collection, storage, and retrieval of human biological materials for research. Cell Preserv Technol 3: 5-48, 2005.
11. Holland NT, Smith MT, Eskenazi B and Bastaki M: Biological sample collection and processing for molecular epidemiological studies. Mutat Res 543: 217-234, 2003.
12. Holland NT, Pfleger L, Berger E, Ho A and Bastaki M: Molecular epidemiology biomarkers – sample collection and processing considerations. Toxicol Appl Pharmacol 206: 261-268, 2005.
13. OECD Best Practice Guidelines for Biological Resource Centres: OECD, 2007. http://www.oecd.org/document/36/0,3343,en_2649_37407_38777060_1_1_1_37407,00.html.
14. Common Minimum Technical Standards and Protocols for Biological Resource Centres Dedicated to Cancer Research: Workgroup Report 2. Caboux E, Plymoth A and Hainaut P (eds.). IARC, 2007.
15. Data Schema and Harmonization Platform for Epidemiological Research. Montreal, Canada, 2011. http://www.datashaper.org/
16. Fortier I, Burton PR, Robson PJ, *et al*: Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. Int J Epidemiol 39: 1383-1393, 2010.
17. Istituto Centrale di Statistica: Classificazione delle Professioni, 2001.
18. Bousquet J, Anto J, Sterk P, *et al*: Systems Medicine and integrated care to combat non-communicable disease. Genome Med 3: 43, 2011.
19. Foss KM, Sima C, Ugolini D, *et al*: miR-1254 and miR-574-5p: serum-based microRNA biomarkers for early-stage non-small cell lung cancer. J Thorac Oncol 6: 482-488, 2011.
20. Gee GV, Koestler DC, Christensen BC, *et al*: Downregulated microRNAs in the differential diagnosis of malignant pleural mesothelioma. Int J Cancer 127: 2859-2869, 2010.

21. Gee GV, Stanifer ML, Christensen BC, *et al*: SV40 associated miRNAs are not detectable in mesotheliomas. Br J Cancer 103: 885-888, 2010.
22. Gemignani F, Neri M, Bottari F, *et al*: Risk of malignant pleural mesothelioma and polymorphisms in genes involved in the genome stability and xenobiotics metabolism. Mutat Res 671: 76-83, 2009.
23. Betti M, Neri M, Ferrante D, *et al*: Pooled analysis of NAT2 genotypes as risk factors for asbestos-related malignant mesothelioma. Int J Hyg Environ Health 212: 322-329, 2009.
24. Magistrelli P, Neri M, Granone P, *et al*: K-ras mutations in circulating DNA from pancreatic and lung cancers: bridging methodology for a common validation of the molecular diagnosis value. Pancreas 37: 101-102, 2008.
25. Landi S, Gemignani F, Neri M, *et al*: Polymorphisms of glutathione-S-transferase M1 and manganese superoxide dismutase are associated with the risk of malignant pleural mesothelioma. Int J Cancer 120: 2739-2743, 2007.
26. Cristaudo A, Foddis R, Vivaldi A, *et al*: Clinical significance of serum mesothelin in patients with mesothelioma and lung cancer. Clin Cancer Res 13: 5076-5081, 2007.
27. IARC: International Lung Cancer Consortium. Lyon, France, 2011. http://ilcco.iarc.fr