Information Technology as Tools for Cancer Registry and Regional Cancer Network Integration

Fortunato Bianconi, *Member, IEEE*, Valerio Brunori, Paolo Valigi, *Member, IEEE*, Francesco La Rosa, and Fabrizio Stracci

***Abstract*—*Background*. With the publication of large studies from different health systems comparing survival probabilities, cancer registries are increasingly involved in clinical evaluation research. The changing role of registries strictly depends on the integration between the oncology system and proper information technology (IT) tools. IT is fundamental to improving validity and timeliness of data diffusion when both the number of sources linked and the number of variables registered are on the rise. *Aims*. In this paper, we present a modern web-based management system that allows to integrate different sources, validate and elaborate data thus providing a new evaluation system for the oncology network based on cancer registries. *Materials and methods*. We developed a Web 2.0 management system for the Umbria Cancer Registry (S.G.RTUP) based on AMPAX technology (Apache, Mysql, PHP, Ajax and XML) and object-oriented programming. ISO/IEC 27001:2005 standard is followed to ensure security access to the information. The S.G.RTUP architecture is modular and extensible and information consistency is guaranteed by entity-relationship principles. Cancer sites, topology, morphology, and behavior are coded according to the International Classification of Diseases. Classical epidemiological indices for a cancer registry are implemented: incidence, mortality, years of potential life lost, and cumulative risk. S.G.RTUP has tools to prepare data for trend analysis and relative survival analysis. Geographical analysis is also implemented. *Results*. S.G.RTUP is integrated with the Oncology Network and gives timely epidemiological indices for evaluation of oncological activities. The registration system that we developed can effectively manage different data sources. Automatic importing of routinely available data from pathology archives, screening services, and hospital discharge records will reduce the time needed to produce data and will also allow the expansion of registered information. Several services for data visualization and statistical analysis are implemented. A geographic information system based on Google maps API is used for geolocalization of cases and map plotting of incidence and mortality rates. We implemented Besag York and Mollie's algorithm for real-time smoothed maps. All services can be dynamically performed over a subset of data that the user can select through an innovative filtering system. *Discussion and conclusion*. IT contributed to shortening all phases of cancer registration, including linkage with external sources, coding, quality control, data management and analysis and publication of results. Integration in the oncology network and secure Web access allowed us to design with clinicians innovative population-based collaborative studies. Our geographic analysis system enables us to develop sophisticated dynamic geostatistic tools.**

***Index Terms*—Cancer registry, geographic information systems (GIS), health-care management, information technologies, oncology network evolution.**

I. Introduction

MEDICAL registries are defined as a systematic collection of a clearly defined set of health and demographic data for patients with specific health characteristics, held in a central database for a predefined purpose [1].

Drolet and Johnson expanded on the definition and indicate five features of medical registries: mergeable data (M) from multiple sources are combined into a standardized data set (D), according to a defined set of rules (R); moreover, patients in the registry have a unique identifier that allows follow-up over time (O) to assess the occurrence of specific outcomes of interest (K) [2].

A modern cancer registry is a medical registry based on a shared disease; cancer is the inclusion principle that meets all the above standards. Indeed, population-based registries collect information on all cancer cases diagnosed in a defined target population, generally residents in a specified geographic area. Cancer registries produce data according to accepted rules and methods [3].

The set of variables collected by registries differs but there is a core set common to all registries. The minimum data set defining a cancer registry includes variables essential to produce routine indicators (i.e., cancer incidence rates, prevalence, and survival probabilities) and basic indicators of data quality [4], [5].

Population-based cancer registries offer a unique potential to generate and test new hypotheses regarding cancer incidence and survival [6].

Cancer registration dates back to the first half of the 20th century [7]. The distinctive features of cancer made it a disease suitable for systematic registration. Cancer was a severe disease leading to hospitalization in most cases and surgery was the main treatment. The disease was often if not invariably fatal. Cancer registration was feasible where hospital records, pathology archives and causes of death certifications were available.

The first registries established were mainly concerned with cancer incidence; that is, monitoring the speed of appearance of new cases over time and comparing rates among geographic areas.

Manuscript received February 28, 2011; revised June 21, 2011 and September 13, 2011; accepted December 10, 2011. Date of current version October 12, 2012. This work was supported by the Department of Health, Regional Government of Umbria. This paper was recommended by Editor M. P. Fanti.

F. Bianconi, V. Brunori, and P. Valigi are with the Department of Electronic and Information Engineering, University of Perugia, 06125 Perugia, Italy (e-mail: fortunato.bianconi@diei.unipg.it; brunori.valerio@gmail.com; valigi@diei.unipg.it).

F. La Rosa and F. Stracci are with the Department of Surgical and Medical Specialties, and Public Health, University of Perugia, 06126 Perugia, Italy (e-mail: larosaf@unipg.it; fabs@unipg.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCA.2012.2210209

The population-based cancer registry thus has had a public health focus, dominated by a primary interest in cancer cause and prevention [8].

The EUROCARE study, published in 1995, is a comparison of population-based survival probabilities across European registries. After this date, cancer registries expanded their role to the evaluation of care processes, without losing their primary focus on surveillance and cancer causation [9].

Population-based survival was first measured without clinical detail by age, sex, and cancer site. Thus, it was proposed as an overall indicator of care provided by the oncology system. The future of cancer registries' survival is influenced both by access to specialized services and the quality of care provided [10].

Comparisons among registries shows wide and sometimes unexpected variability in survival rates. The health outcome for many types of cancers depends on the quality of care provided. The finding of differences in survival reported by cancer registries has had a profound influence on health policy in some countries, for example in the U.K. where the worst survival rates were reported for many types of cancers [11].

The availability of detailed clinical information on prognostic factors and treatment is increasingly used to produce quality of care studies coupling process indicators with health outcomes [12], [13]. The cancer registry is thus evolving toward a system for fine tuning evaluation of cancer control [8], [14].

The diffusion of screening interventions for many relevant cancers (e.g., breast, large bowel, the uterine cervix, prostate, and melanoma of the skin) gave cancer registries another important role in monitoring and evaluation of these complex interventions [15]–[17].

Because of the new evaluation goals and changes in cancer care, additional data sources are required for cancer registries to maintain completeness and validity of information. For instance, linkage with screening archives is useful to identify screen-detected cancers and improve the ability to evaluate screening activities. Similarly, ambulatory care, diagnostic tests, and drug prescription files are increasingly necessary due to the wide use of out-patient care and also to calculate quality of care indicators.

The evaluation of health outcomes produced by specific interventions and of the quality of care provided implies a significant increase in the number of recorded variables [8]. Thus, both the number of medical data sources acquired and information extracted from available archives are on the rise.

This important effort that is the evolution of the cancer registry into an intelligence unit for surveillance and evaluation of oncological care must be accomplished without losing timeliness of data diffusion. Indeed, since the usefulness of evaluation data decays much more quickly than cancer descriptive statistics, the reaching of evaluation goals requires an improvement in timeliness of cancer registries.

The registries access to an increasing number of sources can be achieved through information technology (IT) improvement [18]. Cancer registries presently have a major limitation, however, since they are bound to investigate routinely collected historical data registered in accessible health archives. This difficulty may heavily influence the data quality and/or time of data production [4]. Even though the final achievements in terms of data validity and timeliness depend on the level of automation of the various sources needed for cancer registration, the IT used by a cancer registry is also essential to make the best of the available data and to allow the rapid production of results.

Information systems have been widely adopted in the realization and management of cancer registries. The International Association of Cancer Registries (IACR) has an open-source tool to input, store, check, and analyze cancer registry data (CanReg ver 5) [19]. IACR software is a stand-alone application and implements consistency checks and basic analysis of data, but does not implement the integration with external sources. The National Cancer Registrars Association [20] offers Registry Plus, a suite of free software programs for collecting and processing cancer registry data [21]. In [20, Ch. 23] deals with database management systems and with development principles because cancer registries need to build ad hoc systems to store data. The Association of the Nordic Cancer Registries collects data into the NORDCAN database. The NORDCAN project has a web-based descriptive epidemiology tool that provides tabulations and graphs [22].

In addition to the above journal publications, the National Cancer Institute has several statistical tools described in [23]. Among these tools, the most significant ones are their software for Surveillance, Epidemiology, and End Results (SEER) [24] and some applications implementing statistical methods. Most of the SEER registries use the SEER Data Management System, SEER$*$DMS, to manage registry processes. An additional tool dealing with geographical information and related resources is also included in [23].

The aim of this paper is to show how integration between the oncology network and IT allow to fulfill the demanding tasks of a modern cancer registry without loss in completeness, validity, and timeliness. Moreover, IT integration in the oncology network allows new prospective population-based investigations on prognostic factors and the timely evaluation of new interventions. We present a modern web-based management system that allows to integrate different sources, validate, and elaborate data introducing a new evaluation system for the oncology network based on cancer registries. Quality of care indicators, agreed upon with clinical professionals, can also be routinely produced. Geographical analysis is considered in more detail since it is by far the most complex statistical tool fully implemented in the registry program. A preliminary version of this study was presented in [25]: here, we propose a more detailed discussion of the system, and we also included the presentation of the new statistical and geographical tools.

The paper is organized as follows. Section II presents materials and methods and is divided in six subsections. Section II-A system technology, Section II-B system architecture, Section II-C strategies for data integration, Section II-D software development life cycle, Section II-E coding standards and check data integrity, Section II-F statistical methods for epidemiology indices, trend analysis, geographical analysis, and relative survival analysis. Section III shows the results as: Section III-A services of the management system, Section III-B monitoring user activity, Section III-C example of data analysis in particular tables of standard indices and trend analysis,

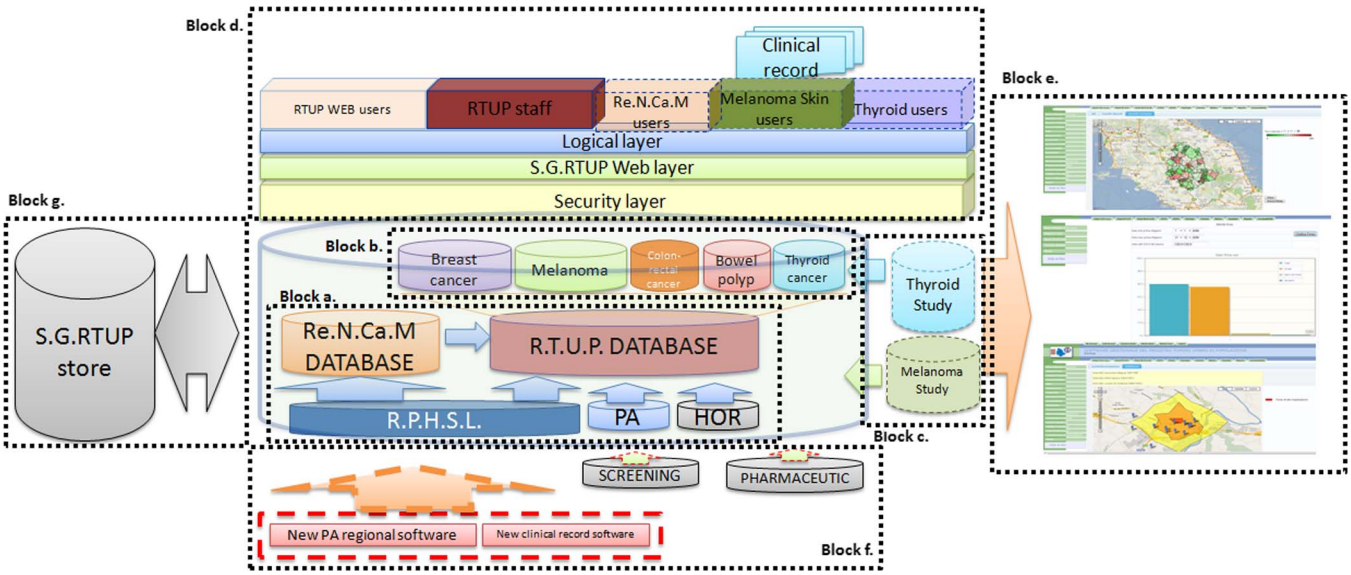


Fig. 1. Architecture of Management System for Umbria Cancer Registry (S.G.RTUP). Block a.: core registry. Block b.: extended specialist registry. Block c.: cooperative registry. Block e.: statistical elaborations. Block d.: layers for hierarchy of users. Block f.: integration with incoming new regional systems. Block g.: S.G.RTUP store for concomitant accesses.

geostatistics (local geolocalization, standard incidence ratio (SIR) and standardized mortality ratio (SMR) maps with Google API, Besag York and Mollie's (BYM) algorithm results), Section III-C3 Preparing data for relative survival analysis. Discussion and conclusions are presented in Sections IV and V, respectively.

II. Materials and Methods

A. IT

The management system for the Umbria Cancer Registry (S.G.RTUP) is designed and implemented by Web 2.0 technology [26] and open-source software [25]. Web 2.0 technology integrates different sources coming from different departmental structures and is also a good strategy to give different experts of the oncology network the opportunity to have a patient-management repository in real time. The server site and client site of our systems are both open source to allow a low maintenance cost and eliminate annual fees. Ubuntu Server Edition [27] is the server operating system and we call the Web application tools used AMPAX. AMPAX stands for a combination of different softwares: Apache, Mysql, PHP, Ajax and XML [28]–[32]. These work together managing the data easily and implementing different algorithms. Our programming strategy is based on object-oriented programming. This means that in any of the above languages, we use objects to represent functional parts of an application and real-life entities.

Our systems implement security protocol following ISO/IEC 27001:2005 [33]. The levels of security are:

- *physics level*: redundancy of server and access protocol to the server's rooms;
- *logic level*: hierarchy levels of user and access (Fig. 1);
- *software level*: encryption of data and certification of access to the systems;
- *channel communication level*: HyperText Transfer Protocol Secure protocol provides encryption securing the identification of the server;
- *client level*: using browser certificate to allow users access to the systems.

S.G.RTUP implementing ISO/IEC 27001:2005 completely conforms to the "Code for the Protection of Personal Data" [34] and its Annex B (Article 33 to 36) that follows standard ISO/IEC 17799:2005 (Information Technology Security Techniques) included in [33]. Hierarchical access is regulated in accordance with the principle of non-surplus and the data are exported only in anonymous format using an algorithm of MD5 Message-Digest Algorithm as a cryptographic hash function.

B. Architecture of S.G.RTUP

The S.G.RTUP repository is devised as a modular and extensible structure by means of the entity-relationship databases' principles. Tables are linked to each other by referential integrity constraints that ensure information consistency.

The general architecture of S.G.RTUP is essentially based on a core registry, an extended specialist registry and a cooperative registry (Fig. 1). The S.G.RTUP core registry database stores data concerning epidemiology surveillance of every type of cancer and it is directly linked to the Death Causes Register (Re.N.Ca.M) and to the Regional Public Health Subscribed List (R.P.H.S.L.). Re.N.Ca.M is the regional database where all information about deaths, including their causes, is stored. S.G.RTUP implements a particular tool to manage the insertion, update, and coding of causes of death. These features enable an automatic update of the follow-up of the patients inserted in the registry database without preventing the data import from an external death registry (Fig. 1, Block a.). This procedure is also less time consuming since it avoids double registration by cancer registry personal. Hospital outgoing record (HOR)

and pathological anatomy (PA) are searchable inside the application. This protocol of importing data is implemented in XML with an ability to integrate different sources. This online connection allows the users to obtain in real time a patient history inside the Department of Oncology of the Umbria Region. This is very useful for management of patients when they move from one node to another in the oncology network. Therefore, cancer registry users can make a sure classification taking into account all the information of an individual case. Previously, cancer registry operators were forced to look for information in a number of different archives with an increased probability of making mistakes or missing information.

The database structure extends to a specialist survey of the following cancer sites: breast, colon-rectal, thyroid, skin melanoma, and large bowel polyp. The specialist records contain some specific variables with a degree of detail that can be modified depending on the particular study (Fig. 1, Block b.). The specialist variables are archived and duplicable. This is a way to follow in time what the patients did, and it is a strategy to change the classical cancer registry approach. With this, the users of the oncology department area can update their patient information and at the same time encode the variables that can help evaluate new quality indices.

S.G.RTUP has initiated cooperative projects based on the joint data collected by the registry staff and health professionals through the controlled access to specialist files for skin melanoma and thyroid cancers (Fig. 1, Block c.). The key point of this module of the system is the hierarchy of users and the ability of different types of users (Fig. 1, Block d.). According to the group and the structures of the oncology network, the system shows a specific web interface and assigns a particular level of power to the user. S.G.RTUP public statistics and reports are available on the Web site (www.rtup.unipg.it) where there are epidemiological tools that provide tabulations with further user-specified options available (RTUP Web users in Fig. 1, Block d.). All reserved accesses are allowed by Virtual Private Network (VPN): Oncology network professionals can access the specialized interface for data management of cooperative projects (Melanoma Skin users and Thyroid users in Fig. 1, Block d.); Local Health Center operators access to the Re.N.Ca.M. interface (Re.N.Ca.M users in Fig. 1, Block d.); RTUP operators have a different level of operation with user-specified interface (RTUP staff in Fig. 1, Block d.).

S.G.RTUP implements a sophisticated strategy for monitoring users' activities: all database tables register the user and the date of the first insertion and the last updating. Also, implemented is a policy of duplication and storage for the management of conflicting user access to the same record (Fig. 1, Block g.).

C. Data Integration

S.G.RTUP allows the integration of different information flows of the health information system (HIS) as shown in Fig. 1, (Block a. and Block f.). The following strategies have been adopted for the automatic integration of data:

- regular exchange of data by extraction of information in a predefined format (CSV, Microsoft Excel or Microsoft Access). S.G.RUP implements specific algorithms to transform the data to be compatible with its structure data, and import them;
- databases can be viewed without replicating the information on any relationship, but making them accessible directly from the source that makes them available;
- communications-based data exchange (HL7 or XML), using the information to populate local relationships on its database instance and the files are stored in a special exchange storage.

All strategies are based on VPN to ensure the security of data exchange.

D. Development Strategy

The software development life cycle (SDLC) of S.G.RTUP integrates several model processes as described in ISO/IEC 12207 [35]. During the initial development phase begun in 2007 we adopted a *waterfall model* until the first operative release of the system on January 17, 2008 when the first records were inserted with the new system. In this phase, four operators (a cancer registry administrator, a registry coder specialist, a specialist in clinical anatomy/pathology, and a specialist in epidemiology) collaborated to develop the interface and the entity relationship digram with individual weekly meetings and monthly group meetings over a four-month period. During the individual meeting, we discussed *use cases* that describe user interactions with the system, the constraints of the systems and their opinion on graphical user interface (GUI). We produced a requirement document and four GUI prototypes that we presented and discussed in the monthly group meeting. We repeated those iterations four times, until we obtained the definitive requirements and the selected GUI.

We then started using a rapid prototyping exploratory incremental model, with an important involvement of users in the development of modules. Interaction with the users, and in particular with physicians, the registry staff, both at the top and operative level, and with the clinicians involved in the cooperative registration projects was conducted by means of a focus group to define requirements by incorporating a few representative users into the design team and by alpha and beta tests. In particular, the initial group of four operators has been extended with a clinical operator for each disease site of special interest (breast, colorectal, thyroid cancer, skin melanoma, and large bowel polyps) and one operator from each of the four local health units (LHU) collaborating with Re.N.Ca.M. During the alpha test, the selected specialist spent two weeks carrying out module testing and bugs identification. Subsequently, the beta test is opened to a subgroup of the specialists that used the module giving their feedback for about 20 days.

We refer to ISO/IEC 9126, together with ISO/IEC 14598, to ensure a proper level of software quality and product evaluation for the entire set of tools comprising S.G.RTUP [36], [37]. Among the quality characteristics, usability is the condition of the software application suited for human use (see [38, Ch.10]) and is strongly connected with the working environment. In this respect, S.G.RTUP adopted a layered and modular architecture (Fig. 1) that allows development of the user interfaces

independently from the core of the application (Fig. 1, Block c.). The Web interface templates can be further developed, maintained, and improved taking into account usability assessment without affecting the core system.

E. Coding Standards and Check Data Integrity

Classification of neoplasms involves their arrangement or distribution in classes according to a method or system. Neoplasms can be classified in many ways but, for cancer registry and clinician alike, the two most important items of information are the anatomical location of the tumor in the body and the morphology; i.e., the appearance of the tumor when examined under the microscope (histology and cytology), as this indicates its behavior (malignant, benign, *in situ*, and uncertain). Cancer registries endeavor, as a minimum, to classify each neoplasm according to its topography, morphology, and behavior, as well as to record details of the host. The principal manual for classifying diseases is the International Classification of Diseases (ICD) published by the World Health Organization [39].

The topography axis uses the ICD-10 classification of malignant neoplasms (except those categories which relate to secondary neoplasms and to specific morphological types of tumors) for all types of tumors, thereby providing greater site detail for non-malignant tumors than is provided by ICD-10.

The ICD for Oncology version 3 coding is used in cancer registries for coding the site (topography) and the histology (morphology) of neoplasms, usually obtained from a pathology report. It gives a multi-axial classification of the site, morphology, behavior, and grading of neoplasms. The morphology axis provides five-digit codes ranging from M-8000/0 to M-9989/3. The first four digits indicate the specific histological term. The fifth digit after the slash (/) is the behavior code, which indicates whether a tumor is malignant, benign, *in situ*, or uncertain (whether benign or malignant). A separate one-digit code is also provided for histologic grading (differentiation).

Integrity and consistency of data are checked using IARC checks in [40] and also implementing specific site checks for high resolution study [41], [42].

F. Statistical Methods

S.G.RTUP has internal algorithms to compute the standard statistics of a cancer registry. We developed a user-friendly interface to easily elaborate data in the DBMS and to obtain epidemiology indices, trend analysis, geographical analysis, and prepare data for relative survival analysis.

1) Epidemiology Indices: One of the fundamentals of cancer registry analysis is the comparison of basic cancer indicators such as incidence, mortality, years of potential life lost (YPLL) and cumulative risk [43].

An overall measure that does not take explicit account of the composition of the population is called "crude." Its value will be an average of the values for the individual subgroups, weighted by their relative sizes. The larger the subgroup, the more influence it will have on the crude measure. Suppose we consider a population of size P_S as consisting of K age groups, or strata. Each age stratum will have a specific number of people, say p_i $(i = 1, \ldots, K)$. During the following year, each stratum will experience a number e_i of occurrences for ith event (e.g., death or newly diagnosed case of some diseases). The total population size, P_S, is therefore $\sum_{i=1}^{K} p_i$, the total number of event, E, is $\sum_{i}^{K} e_i$, and the crude rate is

$$CR_P^E = \frac{E}{P_S}$$

which can also be written as a weighted average of the stratum-specific event rates, e_i/p_i, as follows:

$$CR = \frac{E}{P_S} = \frac{\sum_{i=1}^{K} e_i}{P_S} = \sum_{i}^{K} \frac{p_i}{P_S}\frac{e_i}{p_i} = \sum_{i}^{K} w_i \frac{e_i}{p_i}$$

where $w_i = p_i/P_S$ are the weights.

The crude rate is the simplest and most straightforward summary of the population experience. However, the event should be strongly related to age, so the stratum-specific event rates will differ greatly from one another. The summary provided by the crude rate glosses over this heterogeneity of stratum-specific mortality rates. However, the comparison of crude rates can sometimes be inadequate, particularly when the population structures are not comparable for factors such as age, sex, or socioeconomic level. These and other factors affect the magnitude of crude rates and may distort their interpretation in an effect called "confounding."

The process of rate standardization is a classic epidemiological method that removes the confounding effect of variables that we know differ in populations we wish to compare. The standardized rates are useful for information users, such as decision-makers, who prefer to use synthetic health indices in their activities. There are two main standardization methods, characterized by whether the standard used is a population distribution (direct method) or a set of specific rates (indirect method).

In *direct standardization* the stratum-specific rates of study populations are applied to the age distribution of a standard population. Computationally, direct standardization of rates DSR is straightforward

$$\text{DSR} = \frac{\sum_{i}^{K} r_i p_i}{P_S} = \sum_{i}^{K} r_i \frac{p_i}{P_S} = \sum_{i}^{K} r_i W_i$$

where r_i is the rate in ith stratum of the study population; p_i is the number of persons in ith stratum of the standard population; P_S is the total number of persons of the standard population $(\sum_{i}^{K} p_i)$; W_i is the weight for each stratum (equal to p_i/P_S). This formula shows that, when the same standard is used, if two study populations have the same age-specific rates (i.e., for each i their r_i's are equal) then their directly standardized rates will be identical, independent of the age distributions in the study populations [43].

Indirect standardization avoids the problem of imprecise estimates of stratum-specific rates in a study population by taking stratum-specific rates from a standard population of sufficient size and relevance. These rates are then averaged using as weights the stratum sizes of the study population. Thus,

the procedure is the mirror image of direct standardization. In direct standardization, the study population provides the rates, and the standard population provides the weights. In indirect standardization, the standard population provides the rates and the study population provides the weights. Indirect standardization can be thought of as taking the observed number of events in the study population and comparing that number to an "expected" number of events, i.e., the number of events that would be expected in the study population if its stratum-specific rates were the same as for the standard population. The ratio of observed to expected events is termed the SMR if death is the outcome or SIR if disease is the outcome. Hence, the indirect standardization of rates formulation is

$$\text{ISR} = \frac{\sum_i^K e_i}{\sum_i^K R_i p_i}$$

where e_i is the number of deaths in the ith stratum of the study population ("observed events"); p_i is the size of the ith stratum of the study population, and R_i is the event rate in the ith stratum of the standard population [43].

YPLL is a measure of the number of years not lived by each individual who died before reaching a predetermined age, usually 65 or 75. The YPLL for a population is computed as the sum of all the individual YPLL for individuals who died during a specific time period

$$\text{YPLL} = \sum_i^K (L - \bar{C}_i) e_i$$

where L is fixed age limit, $\bar{C}_i$ is the central value of the ith age class and e_i are the number of death for the ith age class [43].

Cumulative risk is the risk an individual would have of developing or dying of a particular cancer over a defined life span if that person were not to die beforehand from another cause. Cumulative risk is usually calculated using the following formula:

$$CumRisk = 1 - e^{\left(-\sum_i^K r_i\right) * yr}$$

where r_i is the rate in ith stratum of the study population and yr are number of years for the age class [43].

2) Trend Analysis: Our system performs trend analysis of incidence and mortality data. The joinpoint regression model, which is composed of a few continuous linear phases, is useful to describe changes in trend data. The joinpoint regression model for the observations $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, where x_i represent the time variables, e.g., calendar year and yi are the response variable, e.g., the annual age standardized rates or frequencies, can be written as

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i + \gamma_1 (x_i - \tau_1)^+ + \cdots + \gamma_n (x_i - \tau_n)^+$$

where $\beta_0, \beta_1, \gamma_1, \ldots, \gamma_n$ are regression coefficients and the τ_k, $k = 1, 2, \ldots, n$, $n < N$, are the unknown joinpoints in which $(x_i - \tau_k)^+ = (x_i - \tau_k)$ if $(x_i - \tau_k) > 0$, 0 otherwise. In this paper, we refer to the [44].

3) Geographical Analysis: The S.G.RTUP system implements geographical analysis by two ways: a crude plot and a smoothed plots of SIR and SMR. For the first one, the evaluation is easily obtained by computation of the indices over the area of interest. For the second method we implement a log-normal model with spatial effects. We consider the conditional autoregression (CAR) model for the log-relative risks [45]. Cressie and Chan in [46] use Gaussian modeling with random Markov field property. Let S_i denote the set of neighborhood of area $i (i = 1, \ldots, D)$. The CAR model is defined as

$$E(\beta_i, \beta_j) = \mu_i + \sum_{j \in S_i} C_{i,j} h_i(\beta_j), i = 1, \ldots, D. \quad (1)$$

Here, μ_i's are large-scale variation and C_{ij}'s, $j \in S_i$ are small-scale variation of spatial dependence model and $\beta_i = \log \theta_i (i = 1, \ldots, D)$ where θ_i is the relative risk of a death/disease in district area i (see [46]). Also, we take $\mu_i = \mu$ for $i = 1 \ldots, D$ and $C = (C_{ij}) = \rho((W_{ij}))$ where W is the adjacency matrix of the map, W_{ij}'s depend on the location of ith region $(i = 1, \ldots, D)$ and ρ is a spatial autocorrelation parameter that determines the size and nature (positive or negative) of the spatial neighborhood effect [46], [47]. Let $h_i(\beta_j) = \beta_j - \mu$ and $\text{Var}(\beta_i, \beta_j) = \nu^{-1}$. Let $1/\lambda_{\max}$ denote the maximum value of ρ in the CAR model.

Then, we assume the following BYM model, that is a Bayesian model commonly used in geographic analysis producing smoothed standardized mortality or incidence statistics (SIRs/SMRs) [48]:

- y_i the observed number of incidence events in district $i = 1, \ldots, D$;
- E_i the expected number of incidence events in district i;
- n_i the number of neighborhood areas of area i.

1) At the first level, a conditional likelihood function for observations y_i, is defined. Conditioning is performed on a random vector containing the relative risks of death/disease θ_i. This follows from the assumption that each of the observations y_i is subject to (conditionally on θ_i) Poisson distribution with an expected value of $\theta_i E_i$, where E_i is the expected (hypothetical) number of deaths/disease in a territorial unit S_i. Moreover, it is assumed that the random variables describing the occurrence of events are conditionally pairwise independent. This level can be expressed as follows:

$$y_i|\theta_i \approx Poisson(\theta_i E_i) = (\theta_i E_i)^{y_i} e^{-\theta_i E_i} = (E_i e^{\beta_i})^{y_i} e^{-(E_i e^{\beta_i})}.$$

2) At the second level, the relative risks of death/disease θ_i are defined. Their natural logarithms are explained through regression of two unobservable random effects, namely: the clustering effect (spatial dependence effect) ρ and the intercept μ, representing the average level of relative risk for the whole area

$$\beta_i = \log \theta_i = \mu + \rho_i.$$

3) The third level concerns the definition of the prior distribution for the free term in the regression equation and the distributions of the spatial dependence and spatial

heterogeneity effects. By the CAR model defined above, we have

$$\beta_i|\beta_j, j \in S_i$$

$$\rho, \nu \approx N\left[\mu + \rho \sum_{j \in S_i} W_{ij}(\beta_j - \mu), \nu^{-1} M_{i,j}\right]$$

ρ, μ, and ν are mutually independently distributed, respectively, with $\rho \approx U(0, 1/\lambda_{\max})$, $\mu \approx U(-\infty, +\infty)$ and $\nu \approx Gamma(a, b) = (b^a/\Gamma(a))e^{-b\nu}v^{a-1}$ $(a > 0$ and $b > 0)$.

Setting $\rho = 1$, $M_{ii} = 1/n_i$, $W_{ij} = 1/n_i$ if areas i and j are adjacent and $W_{ij} = 0$, otherwise [49], we get

$$\beta_i|\beta_j, j \in S_i$$

$$\rho, \nu \approx N\left[\frac{\sum_{j \in S_i} \beta_j}{n_i}, \frac{1}{\nu n_i}\right] = \frac{\sqrt{\nu n_i}}{\sqrt{2\pi}} e^{-\frac{\nu n_i}{2}\left(b_i - \frac{\Sigma_{j \in S_i}\beta_j}{n_i}\right)^2}.$$

The joint distribution is

$$f(y, \beta, \mu, \nu) = \prod_{i=1}^{D} \left[(E_i e^{\beta_i})^{y_i} e^{-(E_i e^{\beta_i})} \times \frac{\sqrt{\nu n_i}}{\sqrt{2\pi}} e^{-\frac{\nu n_i}{2}\left(b_i - \frac{\Sigma_{j \in S_i}\beta_j}{n_i}\right)^2} \right] \frac{b^a}{\Gamma(a)} e^{-b\nu} v^{a-1}.$$

The full-conditional distribution to implement the Gibbs sampling are

$$\lfloor \beta_i | y, \mu, \nu \rfloor \propto (e^{\beta_i})^{y_i} e^{-E_i E^{\beta_i}} e^{-\frac{\nu n_i}{2}\left(b_i - \frac{\sum_{j \in S_i}\beta_j}{n_i}\right)^2}$$

$$\nu | y, \beta, \mu \approx \prod_{i=1}^{D} \left[\sqrt{\nu} e^{-\frac{\nu n_i}{2}\left(b_i - \frac{\sum_{j \in S_i}\beta_j}{n_i}\right)^2} \right] e^{-b\nu} \nu^{a-1}$$

$$= Gamma\left(a + \frac{D}{2}, b + \frac{1}{1}\sum_{i=1}^{D} n_i \times \left(b_i - \frac{\sum_{j \in S_i}\beta_j}{n_i}\right)^2\right).$$

Since the CAR model defined above is improper (the overall mean of the S_i is not defined), it can only be used as a prior distribution for spatially distributed random effects, and not as a likelihood for data. It is often convenient to assume that such random effects have zero mean. Besag and Kooperberg in [50] show that constraining the random effects to sum to zero and specifying a separate intercept term with a location invariant $U(-\infty, +\infty)$ prior is equivalent to the unconstrained parameterization with no separate intercept. The generation of samples for the Gibbs sampling is easy for the distribution $\nu|y, \beta, \mu$. Instead, the conditional distribution $\lfloor \beta_i | y, \mu, \nu \rfloor$ is known only up to a multiplicative constant so we have to use the Metropolis–Hastings algorithm.

4) Relative Survival Analysis: S.G.RTUP prepare data to perform relative survival analysis. The main goal of survival analysis is to estimate the survival function $S(t) = P(T > t) = 1 - F(t)$ where $F(t)$ is the cumulative distribution function of T. Its graph is called the survival curve. Often we are interested in the conditional survival function $S(t|X)$ where X is a vector of covariates influencing the survival. This is usually done using one of the regression approaches, the Cox model [51] being the most common. The cumulative relative survival function is defined [52] as

$$r(t) = \frac{S_O(t)}{S_P(t)}$$

where $S_O(t)$ denotes observed survival and $S_P(t)$ stands for population or expected survival, which is estimated on the basis of population mortality tables. Correct calculation of the expected survival is not a straightforward task, and it is now generally accepted that the method of Hakulinen [53], gives the best results. In this paper, we refer to *strs* Stata tool which implements three relative survival methods: Ederer I, Ederer II, and Hakulinen [54], [55].

III. Results

Thanks to the management systems, S.G.RTUP is tightly integrated to the project of the Umbria Oncological Network in order to guarantee a well-timed epidemiological surveillance and also as an instrument for evaluating oncological activities. S.G.RTUP provides users with a number of services and with data management and analysis facilities. Within the more advanced collaborative projects, health professionals in the oncology network participate in the cancer registry cooperative projects on thyroid cancer and skin melanoma defining aims and variables, and also actively inserting data into the system through an agreed protocol. This is one of the key and innovative features of S.G.RTUP.

A. Services of the S.G.RTUP

S.G.RTUP implements a number of services that users can access according to their profile. Some of the very general services can be accessed by all types of authenticated users.

While inserting and updating data into the system, users take advantage of several tools to speed up and simplify data coding. Internal checks and cross variables controls ensure comparability among coding choices. Registry staff can quickly search and review the patient history in the HOR and the PA files. They can also automatically import, when available, case sheet data into the system. Data are marked to be reviewed when they are automatically imported or incompatible, or updated by a different center's user. When the source is not electronically recorded, S.G.RTUP users can upload a scanned copy of the paper for specific patients. This approach is extremely useful when there are hospital departments with a low level of IT particularly in oncology health patient records.

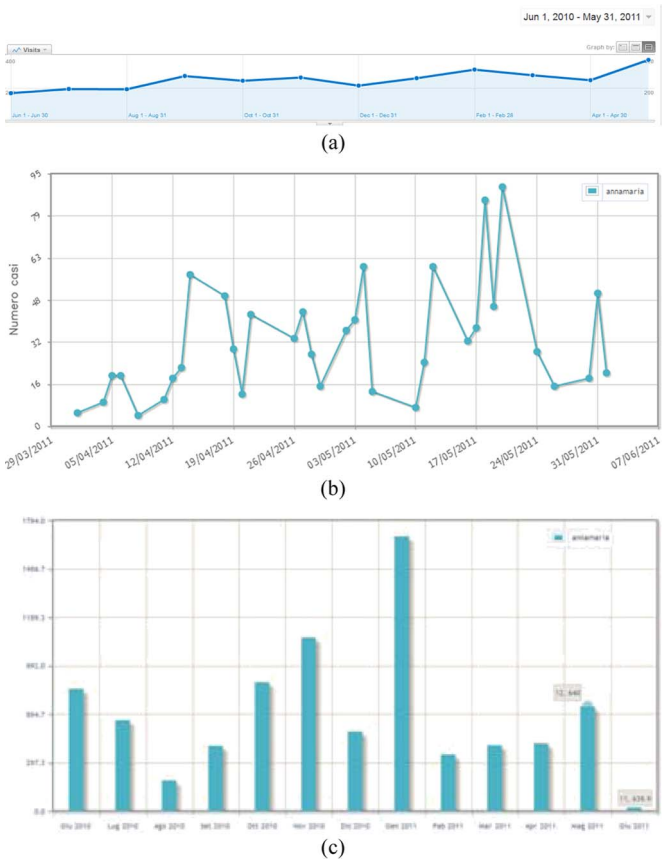


Fig. 2. Example of S.G.RTUP tool for monitoring user activity. (a) Monthly RTUP web user activity from April 2010 to May 2011. (b) Daily activity. (c) Monthly activity.

Users can use a filter to select cases according to both the standard variables such as sex, date/city of birth or cancer classification, and to the specialist criteria dynamically selectable based on the cancer type.

Users are allowed to carry out several statistical examinations of both incidence and mortality. In both cases, it is possible to analyze the number of cases, the annual rates and the standardized rates by age group, specific site, and the cumulative risks. As to the mortality, it is also possible to analyze every cause of death, classified by the ICD, visualized in a tree diagram, and the YPLL. S.G.RTUP also includes a tool to generate tables for joint point regression analysis as presented in Section II-F2 (Fig. 1, Block e.)

The system administrator can manage hierarchy of users with different grants, updating codification and incompatibility tables and also updating tables for coding and exporting to CSV/Excel/PDF format the results of filter research or statistical examination.

For users who do not have authentication, through the web site of Umbria Cancer Registry, it is possible to have access to statistics and reports of S.G.RTUP. This makes our system a useful tool for managers in the health system and for people who need data information [Fig. 2(a) shows the user activity report from April 2010 to May 2011]. The implemented services allow quick data export, analysis, and linkage with external sources. The controls introduced and user hierarchies warrant high data quality standards.

B. Cancer Registry Data Management and Users' Activity Monitoring

S.G.RTUP has been active since January 2008, and we imported from the old archive around 89 500 records produced from 1996 to 2007. From 2008 to 2010, RTUP operators jointly with cooperative project ones inserted around 32 000 new tumor records with an increased number of variables recorded for each cancer site including a number of high-resolution studies (i.e., studies based on the expansion of clinical information registered carried out for breast, colon-rectal, thyroid, skin melanoma, and large bowel polyp). A special database section was added for each high-resolution study. The number of RTUP operators involved in data entry for tumors not included in specialist files has not changed since 1996 and the introduction of the new system has increased the throughput of coded data and shortened the release delay (see Fig. 2(b) and (c) for an example of daily and monthly user data management). In 2007, cancer registry operators were processing 2004 data with a delay of 3 years. In spring 2011, they are processing 2010 data for standard tumors and 2011 data for cooperative projects. By the end of 2011, they will release the definitive 2009 cancer incidence data. As an example of the registry data throughput, the Re.N.Ca.M. are released with a delay of 6–8 months. This delay can be compared to the two years delay of the older system: this is one of the advantages of the cooperative approach adopted with S.G.RTUP.

C. Example Data Analysis

1) Tables of Standard Indices and Trend Analysis: Here are reported some examples of the statistical examinations that a user can carry out for both mortality and cancer incidence. The user can select the years of death/incidence for the interested population and for the indirect standardization reference population and the towns of residence or the districts. For the mortality indices, users can choose to visualize every death cause in a tree diagram or a single class of death causes. In incidence mode users can choose between malignant and benign cancer.

Moreover, users can analyze data about childhood incidence codified in the ICCC code. As regards, the Join Point regression analysis, in addition to the same criteria of the previous statistics, users can select age class interval. In Fig. 3, we show an example of trend analysis for lung cancer.

2) Geostatistics:

Local Geolocalization: By means of Google Maps API, we are able to provide a map of our cancer cases, geolocalizing patients' residence addresses. That is, by making use of such a Google tool we are able to automatically generate additional data for each resident person that can be used to automatically generate statistics over user-defined areas. Automatic geolocalization is not always available for every address or sometimes can be inaccurate. Patients' geolocalization results are checked by some *a priori* criteria, and when it is not immediately possible, addresses are conveniently elaborated by an algorithm that cleans them by increasing the number of successful geolocalizations. Users can select interval of incidence years and residence towns for cases to visualize in

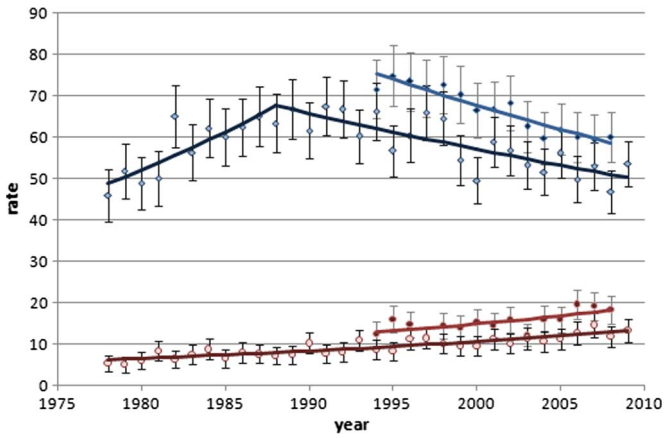


Fig. 3. Mortality trend analysis for lung cancer (C33.0–C34.9) in males (blue line) and females (red line) from 1978 to 2009. Incidence trend analysis for lung cancer (C33.0–C34.9) in males (blue line) and females (red line) from 1994 to 2008.

the map. Fig. 4(a) shows colon cancer (C18.0–C20.9) geolocalization from 2006 to 2007 in Umbria during the screening period.

The integration of Google Maps API in S.G.RTUP allows us to evaluate SIR/SMR over a user-defined area. The classical maps are residence town maps because cancer registries can directly take the benchmark populations from a national statistics institute like ISTAT in Italy. Our system performs the annual R.P.H.S.L. geolocalization and can evaluate the population and the incidence of cases over a user-defined area as shown in Fig. 4(b).

SIR and SMR Maps With Google API: SIR and SMR are calculated by selecting interval of incidence/death years and residence towns, and in addition to the table of results, it is possible to visualize SIR/SMR value of every township in a map of the region as is shown in Fig. 5(a) for colon cancer stage IV (C18.0–C20.9) in females from 2006 to 2007.

SIR and SMR can also be calculated for smaller areas (e.g., census section) to locate suspect disease clusters or to better define cancer risk for large areas. Fig. 6 shows an example of SIR census sections map in Perugia city for colon cancer in males from 2001 to 2008.

BYM Algorithm Results: The smoothed model described in Section II-F3 is evaluable on SIR and SMR. Selecting interval of incidence/death years and residence towns, it is possible to visualize smoothed SIR/SMR maps. In Fig. 6 we show stomach cancer incidence of (C16.0–C16.9) in males and females from 1994 to 2008.

In addition to SIR and SMR tables and mapping, we automatically generate suitably formatted data for geographical analysis in WinBUGS software [56]. By means of that software, it is possible to make some advanced analysis like Bayesian Smoothing of SIR/SMR.

3) Preparing Data for Relative Survival Analysis: To make easy relative survival analysis, our systems generate the table for the probabilities of surviving one year, stratified by age, sex, and calendar year. Also, by filtering the data, users can select the tumor sites population of interest and add to the classical variables the overall survival time and the life status of each case. Both tables for survival probability and selected population data can be exported in csv or txt format. By importing with standard routines of typical statistical toolboxes, relative survival analysis can be easily performed (Stata SE v.10.0 was used in our example [55]). The Fig. 7 shows an example of relative survival analysis by tumor stages for males in C18–C19 and in C20 sites.

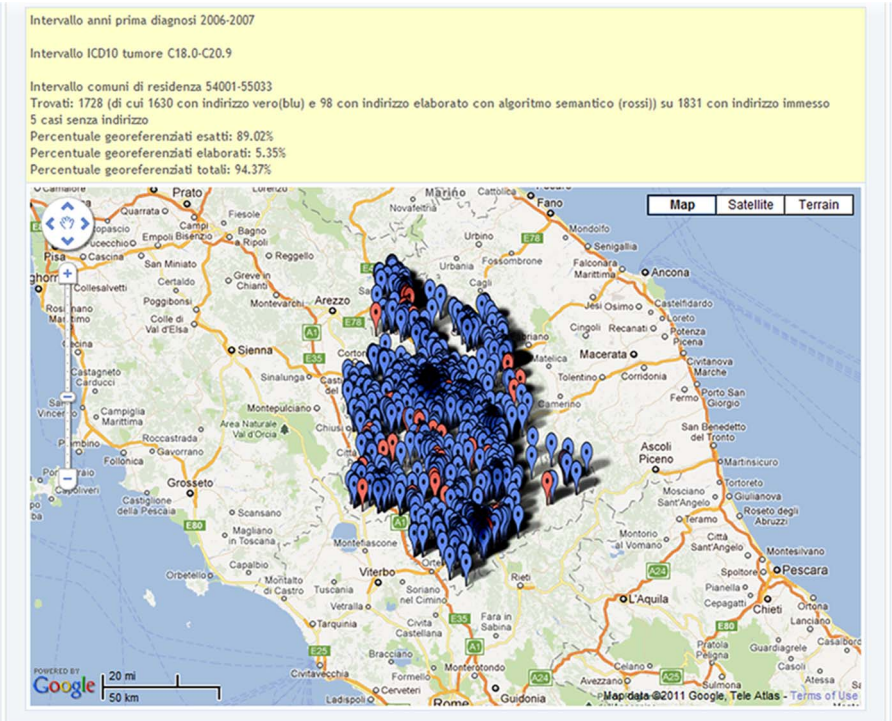
IV. Discussion

The population-based cancer registry of the Umbria region has a target population of approximately 900 000 inhabitants. The area covered is a whole Italian Region with an autonomous health system. The cancer registry is integrated in the recently established Regional Oncology Network, that is an oncology department articulated in hub and spoke centers. Within such a network, S.G.RTUP, a tool-based on extensive use of IT, allows the expansion of the role of the registry from epidemiologic surveillance only to evaluation and research [57].

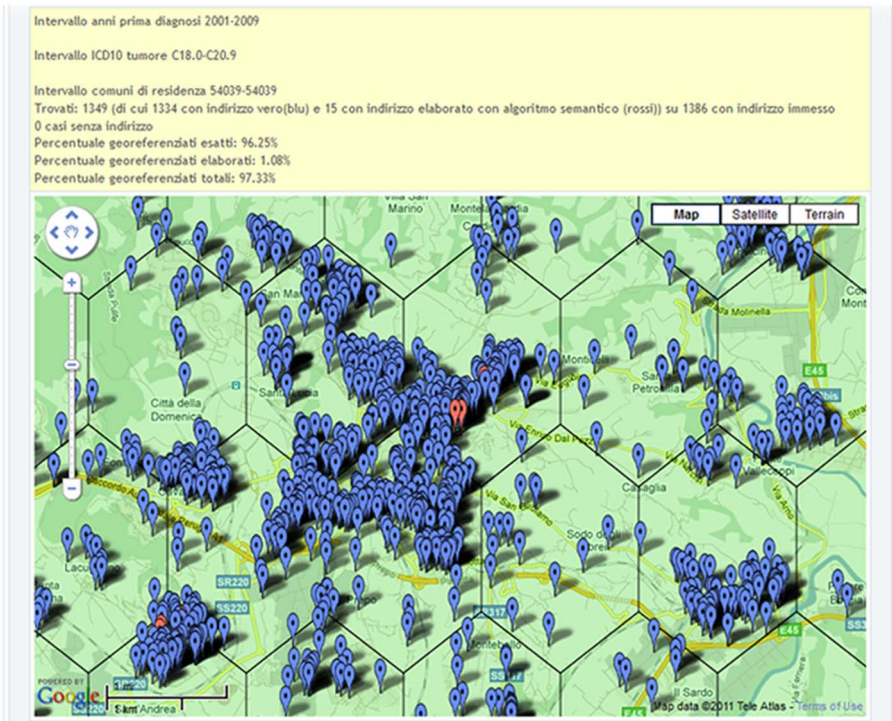
The newly introduced technology has capitalized on existing electronic databases despite their heterogeneity; e.g., HOR. Moreover, both the cancer registry and the oncology network exerted some influence on the regional policy strategies for IT diffusion in the health system. All screening services of the four regional LHU have adopted the same data management system. All seven regional pathology archives have adopted a unique data management system as well. Both the LHUs screening and the pathology archives systems are arranged for integration with the cancer registry's system. A unique electronic medical record system has already been introduced in the main oncology centers connecting hospital chemotherapy pharmacies, oncology, and radiotherapy services. All the above-listed electronic systems will be able to exchange information because of the adoption of international health standards for interoperability (XML and HL7). It is of note that the cancer registry system is the first fully operative web-based system introduced in the regional health system.

To maintain or even improve timeliness of data production and diffusion while expanding registered information many different services were introduced. The new services are important to shrink data production delays in almost every phase of cancer registration while keeping high data validity standards. All information abstracted from medical records or coded from health archives is linked to the individual registry record together with date of registration and identification of responsible registry personnel. Moreover, a signature system was introduced for high resolution studies (the feature of such study is the large number of variables collected on disease and treatment) to warrant a final expert supervision of each case. Easy access to all available information allows controls and re-abstracting. Accountability of registry staff is important to find individual errors and to ensure comparability among operators.

Automation of the statistical analyses ordinarily performed or included in periodic reports is important to further reduce data production times. Incidence, mortality and survival reports by cancer site, sex and period represent typical outputs of a cancer registry and useful "one-click" standard reports may be prepared. This facilitation leaves more time to data interpretation or to perform more refined analyses. The



(a)

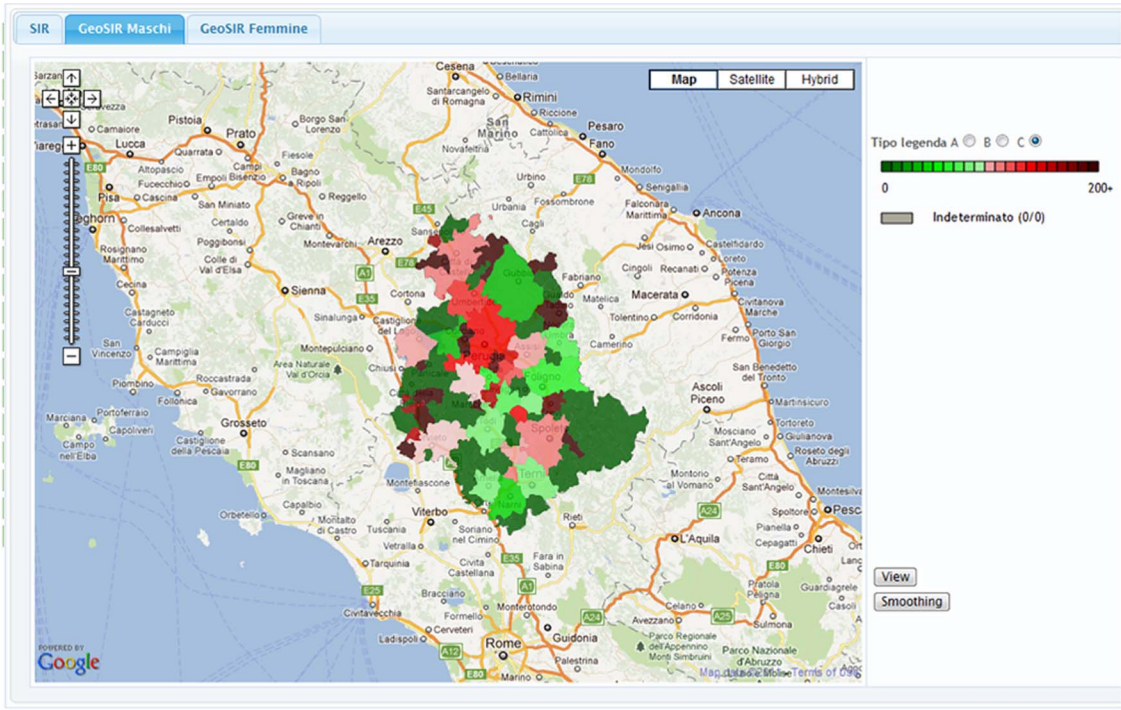


(b)

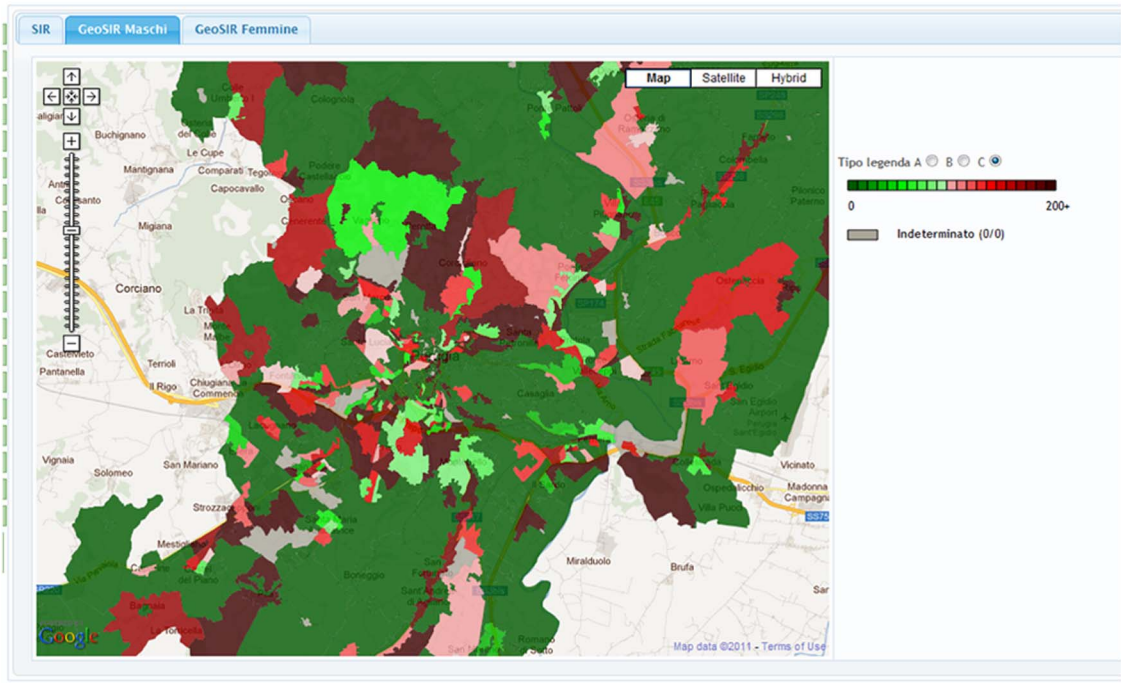
Fig. 4. Geolocalization with Google API for C18.0–C20.9 sites from 2006 to 2007. (a) Umbria overview. (b) Example of hexagonal area where reference population and incidence cases can be calculated.

availability of statistical analysis tools in the registry's system has another important feature, that is the possibility for external users to create personalized dynamic tabulation of interest on the cancer registry web site.

Web secure hierarchical access to the registry data set makes possible an entirely new kind of collaboration where clinicians directly feed the registry with their patients' data after shared decision on study aims and protocol. This collaboration is ongoing to evaluate the application and effectiveness of regional guidelines for diagnosis and treatment of thyroid cancer or to further investigate the increasing mortality trend reported for melanoma of the skin.



(a)



(b)

Fig. 5. SIR maps. (a) SIR residence town map for colon cancer stage IV in females from 2006 to 2007. (b) SIR Perugia census sections map for colon cancer in males from 2001 to 2008.

The increased number of variables registered forced the registry to acquire a number of new sources. For instance, registration of hormone treatment for breast cancer patients would not be feasible without the regional out-patient drug prescription file. Similarly, the identification of individual screening trajectories to evaluate the screening process and outcomes largely depends on linkage with screening archives. When cancer registry and screening services data are linked, it is possible to investigate effectiveness and quality of care of interventions [16], [58]. Thus, evaluation studies are carried out for programmed screenings only; i.e., large bowel, breast, and uterine cervix cancer. When the screening intervention is opportunistic (i.e., physicians advise on early diagnosis occasionally and separating symptomatic diagnoses from screening is not possible), as is the case for prostate cancer, only registry-based monitoring of incidence and mortality is possible.

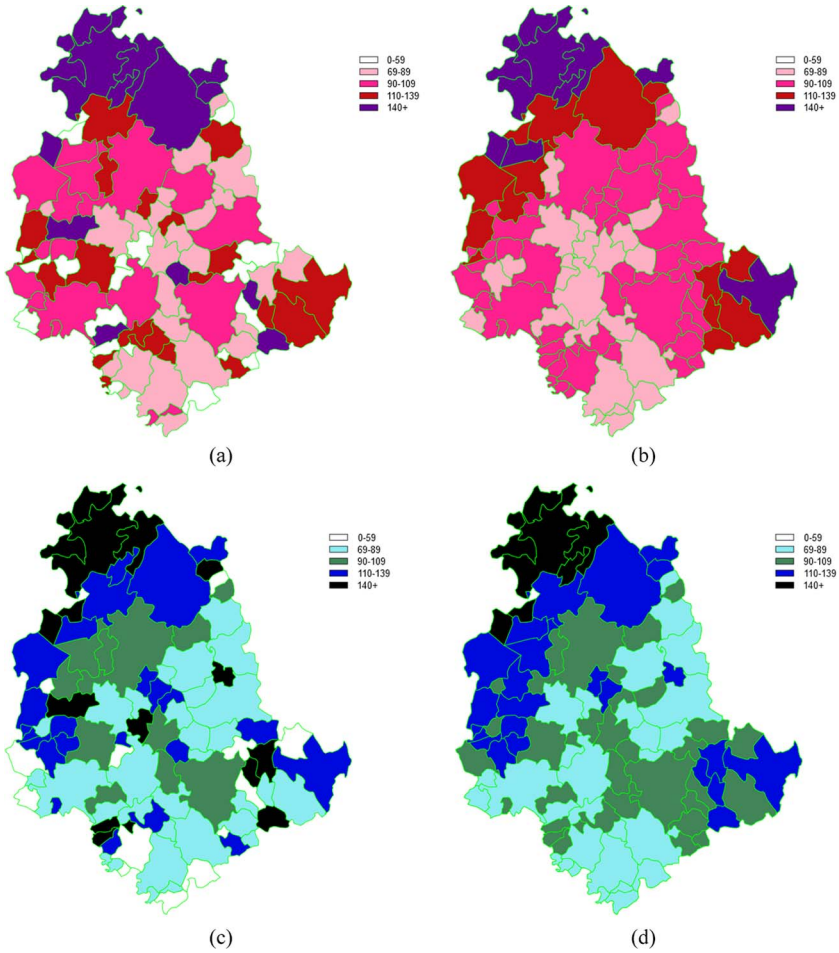


Fig. 6. C16 SIR maps during the period 1994–2008. (a) Crude SIR for female. (b) BYM SIR for female. (c) Crude SIR for male. (d) BYM SIR for male.

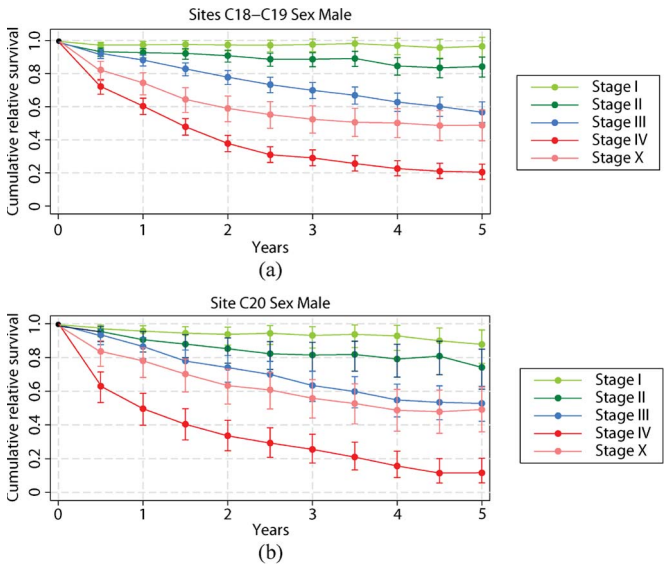


Fig. 7. Relative survival analysis for stages. (a) C18.0–C19.9 sites for males. (b) C20 site for males.

S.G.RTUP is a management system for the oncology network; it is highly integrable with external sources and dynamically performs geographical elaborations and statistics. All these features make our system distinctive among the IT applications of cancer registries presented in scientific journals. CanReg [19] is a desktop application that provides a form for data entry. The advantage of this application over other systems is that the IACR controls are implemented directly in the system, but the number of stored variables is fixed and is the same for all cancer sites, making CanReg not easily to use for specialized studies. High-resolution and evaluation studies require cancer specific information. Furthermore, the IACR application cannot integrate external sources of data and cannot produce statistics, while S.G.RTUP has both functionalities. Registry Plus [21] is released both in desktop and a Web-based version: similar to CanReg, it cannot be integrated with external sources, and it does not perform specific statistical processing except for very basic ones. The NORDCAN application is a tool to produce quality of data indicators and statistical analysis for cancer registries data. NORDCAN accepts one or more defined cancer registries data set as input, but it is not a management system for cancer registries [22]. Both SEER∗STAT [23] and NORDCAN are useful tools for the analysis of cancer registry data and derive their best utility from the ability to analyze data from a network of collaborating cancer registries, but they are independent of data production. Indeed Data for NORDCAN are delivered from the national cancer registries as data for SEER∗STAT are delivered from the SEER participating cancer

registries. Most of the SEER registries, however, share also a data management system, SEER∗DMS. Thus, the S.G.RTUP encompasses the function of SEER∗STAT and SEER∗DMS. SEER∗DMS seems a modern Web-based application supporting all core cancer registry functions. However, this application is available only to SEER registries and its architecture is neither published nor fully described in the manual on the web site, so that a detailed comparison with the present system is not possible [24].

The limited number of publications and tools for cancer registries in IT are probably the consequence of the historical role of the cancer registry and its activities within the health system. The cancer registries were limited to producing basic or surveillance statistics and their sources were nearly always not standardized if not electronic at all. Currently, evaluation purposes derived from cancer registries are growing, and IT is much more widespread in the health system than in the past. The change in cancer registries' role and context is guiding the change of their IT systems. Both scientific integration in the oncology department and IT integration in the HIS are needed to build a new evaluation system for the oncology network based on cancer registries.

Our system has been in use since 2008 and had a crucial role in increasing the ability and timeliness of RTUP operators to process data, as shown in Fig. 2. The cooperative approach adopted allows external users to have electronic patient records that are available for follow-up: at the same time, the information flow is used to populate the tumor registry database. The user interfaces developed by rapid prototyping exploratory incremental models were shared with the users in the design and development. Thanks to the modular and layered architecture of S.G.RTUP, personalization has become very flexible and can be improved by applying an assessment of usability [38]. Other approaches that can be used to improve healthcare systems, and in particular those where team communication has a central role, are based on cognitive work tools (see [59] and the references therein).

IT facilitates and expands the possibilities to execute traditional geographic analyses. Incidence comparisons among geographic areas have been performed since the establishment of the first cancer registries [7]. Geographic analysis has changed with the development of new statistical methods [60]–[62] and the improvement of personal data archives. Maps of smoothed cancer incidence, based on registry data, are frequently drawn and BYM is a commonly used model in geographic analysis [48]. Registry analyses sometimes are able to correlate cancer risk in small areas with point sources of pollution [63]. In Umbria, geographic analyses were able to identify high risk areas for gastric cancer and cancer of the upper aero digestive tract in the North of the Region [64]. The software WinBUGS was used to analyze data by municipality of residence.

Presently, S.G.RTUP implements an innovative and unique, to the best of our knowledge, geographic information system (GIS) based on Google Maps API. Our GIS is used both for population data and cancer cases maps as shown in Figs. 4 and 5. SIR and SMR can be evaluated over a user-defined area such as census sections thanks to S.G.RTUP R.P.H.S.L.s automatic geolocalization. This feature alone of our system greatly simplifies geographic analyses of cancer registry data such as the intra municipal study and local cluster scanning. Maps of Bayesian smoothed rates can also be drawn within the cancer registry software as shown in Fig. 6.

IT is fundamental to maintain the traditional role of registries in cancer surveillance and to allow for real integration in the oncology department. The integrated cancer registry is increasingly able to evaluate clinical intervention. Registries will thus be able to provide relevant information both on cancer burden and the quality and effectiveness of interventions to the stakeholders (health policy makers, clinicians, patients and the general population). The new Web-based architecture introduced in our registry makes it possible to design new projects in collaboration with clinical professionals in the oncology network. Shared decisions on research aims and information needs and direct registration of new cases by the clinicians are features of the collaborative project; such features should ensure timeliness of data production and diffusion and make population-based research much easier.

Sensitive data protection problems have been carefully considered during the realization of the registry through Web-oriented technology by setting permissions and having online professional people access policies. Current experience with thyroid cancer and skin melanoma has provided participation experience to the professionals involved at the regional health service level in defining aims and variables, and inserting data into the system through an agreed protocol. In both cases, the study is connected to the evaluation of guidelines/PDT shared at the regional level [65], [66]. In general, the bounding factor for achievement of the explained goals is the level of development of a certain source (e.g., regional program for organized screening and for a PA archive) and the development of a shared research project.

V. Conclusion

In conclusion, our registration system can manage various point sources acquired by a cancer registry. As to source type, both record linkage with periodic acquired sources (e.g., yearly linkage with hospital discharge records) and real-time data registration are allowed. Our approach is flexible since the cancer registry can receive information from an increasing number of sources and where features of the sources (e.g. automation, coding) and their content vary over time. Real-time registration is central to the realization of an effective integration within the oncology system. The registry tool has already allowed collaboration of oncologic multidisciplinary groups concerned with skin melanoma and thyroid cancers. This is the first time that that such a result has been achieved. Using real-time registration by health professionals, it will be possible to further expand the cancer registry's scope and to investigate the role of new prognostic factors or to contribute to the evaluation of health interventions with high-quality population-based data.

Our GIS system integrated in S.G.RTUP is an innovative example of the Web 2.0 application for cancer registries. This improvement opens interesting new possibilities for dynamical data visualization and statistical analysis. The search for local disease clusters and their correlation with socio-economic population data and environmental pollution models is a new and increasingly important task of cancer registries. These are evolving toward complex systems providing the oncology

network and public health policy makers with a number of indicators ranging from monitoring data to quality of care evaluation. IT has a pivotal role in the process since it greatly simplifies some of the cancer registry tasks and allows routine performance of other tasks such as active search for suspect local disease clusters that otherwise would be out of reach.

ACKNOWLEDGMENT

The authors would like to thank Dr. M. Tonato, coordinator of Umbria Regional Cancer Network, for fruitfully discussion and support.

REFERENCES

[1] D. Arts, N. D. Keizer, and G. Scheffer, "Defining and improving data quality in medical registries: A literature review, case study, generic framework," *J. Amer. Med. Inf. Assoc.*, vol. 9, no. 6, p. 600 611, Nov./Dec. 2002.

[2] B. Drolet and K. Johnson, "Categorizing the world of registries," *J. Biomed. Informat.*, vol. 41, no. 6, pp. 1009–1020, Dec. 2008.

[3] O. Jensen, D. Parkin, R. MacLennan, C. Muir, and R. Skeet, Eds., *Cancer Registration, Principles and Methods*. Lyon, France: IARC Sci. Publ., 1991, pp. 126–158.

[4] F. Bray and D. Parkin, "Evaluation of data quality in the cancer registry: Principles and methods, Part I: Comparability, validity and timeliness," *Eur. J. Cancer*, vol. 45, no. 5, pp. 747–755, Mar. 2009.

[5] D. M. Parkin and F. Bray, "Evaluation of data quality in the cancer registry: Principles and methods Part II. Completeness," *Eur. J. Cancer*, vol. 45, no. 5, pp. 756–764, Mar. 2009.

[6] M. Goodman, B. Hernandez, S. Hewitt, C. Lynch, T. Cot, H. J. Frierson, C. Moskaluk, J. Killeen, W. Cozen, C. Key, L. Clegg, M. Reichman, B. Hankey, and B. Edwards, "Tissues from population-based cancer registries: A novel approach to increasing research potential," *Hum. Pathol.*, vol. 36, no. 7, pp. 812–820, Jul. 2005.

[7] D. Parkin, "The evolution of the population-based cancer registry," *Nat. Rev. Cancer*, vol. 6, pp. 603–612, Aug. 2006.

[8] R. Sankila and J. Coebergh, "Cancer registries contribute to quality improvements in clinical care for all European cancer patients," *Eur. J. Cancer*, vol. 40, no. 5, pp. 635–637, Mar. 2004.

[9] F. Berrino, M. Sant, A. Verdecchia, R. Capocaccia, T. Hakulinen, and J. Estve, Eds., *Survival of Cancer Patients in Europe, The EUROCARE Study*. Lyon, France: IARC, 1995.

[10] A. Micheli, G. Gatta, and A. Verdecchia, "Studying survival of cancer patients in different populations: Its potential and role," *Tumori*, vol. 83, no. 1, pp. 3–8, Jan./Feb. 1997.

[11] G. Gatta, G. Zigon, T. Aareleid, E. Ardanaz, M. Bielska-Lasota, J. Galceran, S. Góźdź, T. Hakulinen, C. Martinez-Garcia, I. Plesko, M. Zakelj, J. Rachtan, G. Tagliabue, M. Vercelli, and J. Faivre, "Patterns of care for European colorectal cancer patients diagnosed 1996–1998: A Eurocare high resolution study," *Acta Oncol.*, vol. 49, no. 6, pp. 776–783, Aug. 2010.

[12] A. Talsma, A. Reedijk, R. Damhuis, P. Westenend, and W. Vles, "Rere-section rates after breast-conserving surgery as a performance indicator introduction of a case-mix model to allow comparison between dutch hospitals," *Eur. J. Surg. Oncol.*, vol. 37, no. 4, pp. 357–363, Apr. 2011.

[13] M. Elferink, S. Siesling, O. Visser, H. Rutten, J. van Krieken, R. Tollenaar, and V. Lemmens, "Large variation between hospitals and pathology laboratories in lymph node evaluation in colon cancer and its impact on survival, a nationwide population-based study in the Netherlands," *Ann. Oncol.*, vol. 22, no. 1, pp. 110–117, Jan. 2011.

[14] K. Y. Bilimoria, A. K. Stewart, D. P. Winchester, and C. Y. Ko, "The national cancer data base: A powerful initiative to improve cancer care in the United States," *Ann. Surg. Oncol.*, vol. 15, no. 3, pp. 683–690, Mar. 2008.

[15] R. Bennett, S. Sellars, and S. Moss, "Interval cancers in the NHS breast cancer screening programme in England, Wales and Northern Ireland," *Brit. J. Cancer*, vol. 104, no. 4, pp. 571–577, Feb. 2011.

[16] D. Puliti, G. Miccinesi, N. Collina, V. D. Lisi, M. Federico, S. Ferretti, A. Finarelli, F. Foca, L. Mangone, C. Naldoni, M. Petrella, A. Ponti, N. Segnan, A. Sigona, M. Zarcone, M. Zorzi, M. Zappa, E. Paci, and I. W. Group, "Effectiveness of service screening: A case-control study to assess breast cancer mortality reduction," *Brit. J. Cancer*, vol. 99, no. 3, pp. 423–427, Aug. 2008.

[17] F. Stracci, "Cancer screenings, diagnostic technology evolution, cancer control," *Methods Mol. Biol.*, vol. 471, pp. 107–136, 2009.

[18] S. Tognazzo, A. Andolfo, E. Bovo, A. Fiore, A. Greco, S. Guzzinati, D. Monetti, C. Stocco, and P. Zambon, "Quality control of automatically defined cancer cases by the automated registration system of the venetian tumour registry," *Eur. J. Public Health*, vol. 15, no. 6, pp. 657–664, Dec. 2005.

[19] O. M. Jensen, *Cancer Registration: Principles and Methods*. Lyon, France: IARC, Int. Assoc. Cancer Registries, 1991, ch. Appendix 4.

[20] National Cancer Registry Association, Cancer Registry Management: Principles & Practices for Hospitals and Central Registries, Dubuque, IA: Kendall Hunt, 2008.

[21] U.S. Dept. Health Human Services, Centers Disease Control Prevention, National Center Chronic Disease Prevention Health Promotion, Registry Plus, a Suite of Publicly Available Software Programs for Collecting and Processing Cancer Registry Data, Atlanta, GA, 2011. [Online]. Available: http://www.cdc.gov/cancer/npcr/

[22] G. Engholm, J. Ferlay, N. Christensen, F. Bray, M. L. Gjerstorff, A. Klint, J. E. Køtlum, E. Olafsd'ottir, E. Pukkala, and H. H. Storm, "NORDCAN—A Nordic tool for cancer information, planning, quality control and research," *Acta Oncol.*, vol. 49, no. 5, pp. 725–736, Jun. 2010.

[23] National Cancer Institute (NCI), Aug. 2011. [Online]. Available: http://www.cancer.gov/statistics/tools

[24] Surveillance, Epidemiology and End Results (SEER), Aug. 2011. [Online]. Available: http://seer.cancer.gov/

[25] F. Bianconi, V. Brunori, P. Valigi, F. Stracci, and F. L. Rosa, "Cancer registry and information technology: A new management system for integrating cancer registry and oncology departments," in *Proc. IEEE WHCM*, Feb. 2010, pp. 1–5.

[26] T. O'Reilly, What is Web 2.0?, Retrieved December 15, 2007. [Online]. Available: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[27] Ubuntu Server Edition, 2011. [Online]. Available: http://www.ubuntu.com/products/whatisubuntu/serveredition

[28] Apache http Server Version. [Online]. Available: http://httpd.apache.org/

[29] MySQL Documentation: MySQL Reference Manuals. [Online]. Available: http://dev.mysql.com/doc/

[30] Php Manual. [Online]. Available: http://www.php.net/docs.php

[31] J. J. Garrett, Ajax: A New Approach to Web Applications, 2005. [Online]. Available: http://www.adaptivepath.com/ideas/essays/archives/000385.php

[32] Extensible Markup Language (xml). [Online]. Available: http://www.w3.org/XML/

[33] *Information Technology—Security Techniques—Information Security Management Systems—Requirements*, ISO/IEC 27001:2005 Std.

[34] Dlgs 196/2003—Codice in materia di protezione dei dati personali, Gazzetta Ufficiale dello Stato Italiano n. 174 del 29 luglio 2003—Supplemento Ordinario n. 123, 2003.

[35] *Systems and Software Engineering Software Life Cycle Processes*, ISO/IEC Std. 12207:2008.

[36] *Software Engineering—Product Quality*, ISO/IEC Std. 9126.

[37] *Information Technology—Software Product Evaluation*, ISO/IEC Std. 14598.

[38] C. P. Nemeth, *Human Factors Methods for Design*, U. N. Stanton, Ed. New York: Taylor & Francis, 2004.

[39] *International Statistical Classification of Diseases and Related Health Problems*. [Online]. Available: http://www.who.int/classifications/icd/en/

[40] J. Ferlay, C. Burkhard, S. Whelan, and D. Parkin, "Check and Conversion Programs for Cancer Registries," IARC/IACR, Lyon, France, Tech. Rep., 2005.

[41] International Agency for Research on Cancer (IARC). [Online]. Available: http://www.iarc.fr/

[42] *European Cancer Registry-Based Study on Survival and Care of Cancer Patients*. [Online]. Available: http://www.eurocare.it/

[43] K. J. Rothman, S. Greenland, and T. L. Lash, *Modern Epidemiology,* 3rd ed. Philadelphia, PA: Lippincott, Williams, & Wilkins, 2008.

[44] H. Kim, M. Fay, E. Feuer, and D. Midthune, "Permutation Tests for Joinpoint Regression With Applications to Cancer Rates," *Stat. Med.*, vol. 19, no. 3, pp. 335–351, Feb. 2000.

[45] D. Clayton and J. Kaldor, "Empirical Bayes estimates of age-standardized relative risks for use in disease mapping," *Biometrics*, vol. 43, no. 3, pp. 671–681, Sep. 1987.

[46] N. Cressie and N. Chan, "Spatial modelling of regional variables," *J. Amer. Statist. Assoc.*, vol. 84, pp. 393–401, 1989.

[47] N. Cressie, *Statistics for Spatial Data*. New York: Wiley, 1993.
[48] R. Semenciw, N. D. Le, L. Marrett, D. Robson, D. Turner, and S. Walter, "Methodological issues in the development of the Canadian cancer incidence Atlas," *Stat. Med.*, vol. 19, no. 17/18, pp. 2437–2449, Sep. 2000.
[49] J. Besag, J. York, and A. Mollie, "Bayesian image restoration, with two applications in spatial statistics," *Ann. Inst. Stat. Math.*, vol. 43, pp. 1–59, 1991.
[50] J. Besag and C. Kooperberg, "On conditional and intrinsic autoregressions," *Biometika*, vol. 82, no. 4, pp. 733–746, Dec. 1995.
[51] D. Cox, "Regression models and life-tables," *J. R. Statist. Soc. Ser. B*, vol. 34, no. 2, pp. 187–220, 1972.
[52] F. Ederer, L. Axtell, and S. Cutler, "The relative survival rate: A statistical methodology," *Nat. Cancer Inst. Monogr.*, vol. 6, pp. 101–121, 1961.
[53] T. Hakulinen and L. Tenkanen, "Regression analysis of relative survival rates," *J. R. Stat. Soc. Ser.*, vol. C 36, pp. 309–317, 1987.
[54] P. Dickman, E. Coviello, and M. Hills. (2007) Estimating and modelling relative survival using stata. [Online]. Available: http://www.pauldickman.com/rsmodel/stata corecolon/
[55] *Stata: Data Analysis and Statistical Software*. [Online]. Available: http://www.stata.com/
[56] *The Bugs Project Welcome*. [Online]. Available: http://www.mrcbsu.cam.ac.uk/bugs/welcome.shtml
[57] F. Stracci, F. Bianconi, F. L. Rosa, and M. Tonato, "Cancer registry integration in the Umbria regional cancer network: Breast cancer as a model to build a supportive information system," *Ann. Oncol.*, vol. 21, p. suppl. 8, 2010.
[58] E. Paci, G. Miccinesi, D. Puliti, P. Baldazzi, V. D. Lisi, F. Falcini, C. Cirilli, S. Ferretti, L. Mangone, A. Finarelli, S. Rosso, N. Segnan, F. Stracci, A. Traina, R. Tumino, and M. Zorzi, "Estimate of overdiagnosis of breast cancer due to mammography after adjustment for lead time. A service screening study in Italy," *Breast Cancer Res.*, vol. 8, no. 6, p. R68, 2006.
[59] C. P. Nemeth, Ed., *Improving Healthcare Team Communication: Building on Lessons From Aviation and Aerospace*. Aldershot, U.K.: Ashgate, 280 pp.
[60] S. M. Cramb, K. L. Mengersen, and P. D. Baade, "Developing the atlas of cancer in Queensland: Methodological issues," *Int. J. Health Geograph.*, vol. 10, no. 1, p. 9, 2011.
[61] A. L. F. Cançado, A. R. Duarte, L. H. Duczmal, S. J. Ferreira, C. M. Fonseca, and E. C. D. M. Gontijo, "Penalized likelihood and multiobjective spatial scans for the detection and inference of irregular clusters," *Int. J. Health Geograph.*, vol. 9, no. 1, p. 55, Jan. 2010.
[62] C. Robertson and T. A. Nelson, "Review of software for space-time disease surveillance," *Int. J. Health Geograph.*, vol. 9, p. 16, Jan. 2010.
[63] S. Parodi, E. Stagnaro, C. Casella, A. Puppo, E. Daminelli, V. Fontana, F. Valerio, and M. Vercelli, "Lung cancer in an urban area in Northern Italy near a coke oven plant," *Lung Cancer*, vol. 47, no. 2, pp. 155–164, Feb. 2005.
[64] T. Cassetti, F. L. Rosa, L. Rossi, D. D'Al, and F. Stracci, "Cancer incidence in men: A cluster analysis of spatial patterns," *BMC Cancer*, vol. 8, p. 344, 2008.
[65] E. Puxeddu and G. O. M. R. per le Neoplasie Tiroidee, Linee guida per la gestione dei pazienti affetti da noduli tiroidei e carcinoma della tiroide (Italian), Perugia, Italy: Rete Oncologica Regionale dell'Umbria (ROR), 2008. [Online]. Available: http://www.reteoncologicaumbria.org
[66] S. Simonetti and G. O. M. R. per il melanoma cutaneo, Linee guida per la gestione dei pazienti affetti da melanoma cutaneo (Italian), Perugia, Italy: Rete Oncologica Regionale dell'Umbria (ROR), 2011. [Online]. Available: http://www.reteoncologicaumbria.org

Fortunato Bianconi (M'10) received the M.Sc. degree in information and communication technology engineering from University of Perugia, Perugia, Italy, where he has taken the Ph.D. degree, in 2010.

In the period 2008–2009, he was a Junior Specialist at the University of California, San Francisco at El Samad Systems Biology Lab. From February 2010, he is a Postdoctoral Researcher at the University of Perugia, Perugia. His research interests are in the field of systems biology, management systems in health applications, data analysis in medicine, and information technology applications in oncology.

Valerio Brunori was born in 1981. He graduated in information and telecommunication engineering from the University of Perugia, Perigua, Italy, in 2006. He received the Ph.D. degree in industrial engineering at the University of Perugia.

His main research interests are health management systems and feature detection in computer vision.

Paolo Valigi (M'96) received the Laurea degree from University of Rome La Sapienza, Rome, Italy, in 1986, and the Ph.D. degree from the University of Rome Tor Vergata, Rome, Italy, in 1991.

He was with Fondazione Ugo Bordoni, Rome, Italy, from 1990 to 1994. From 1994 to 1998, he was a Research Assistant at the University of Rome Tor Vergata, Rome. From 1998 to 2004, he has been an Associate Professor at the University of Perugia, Perugia, Italy, where since 2004, he is a Full Professor of System Theory, with the Department of Electronics and Information Engineering. His research interests are in the field of systems biology, robotics, and distributed control and optimization. He has authored or coauthored more than 100 hundred journal and conference papers and book chapters.

Francesco La Rosa received the degree in biological sciences at the University of Camerino, Macerata, Italy, in 1971. He graduated as a technician until 1980 at the Department of Hygiene of the University of Camerino, Macerata.

He was Teacher, on annual contract first, then permanent, from 1976 to 1980. He was an Associate Professor of microbiology from 1980 to 1987, an Associate Professor of hygiene from 1987 to 2002 at the University of Perugia, Perugia, Italy, and a Full Professor of hygiene since 2002 at the same university. He was charged with medical statistics and with several university courses and in many postgraduate specialization courses teaching since 1998 at the University of Perugia, Perugia. His scientific activity is documented by 110 publications appeared on scientific journals, 52 monographs, and several posters and congress relations. His main research area of is chronic diseases epidemiology. Presently, he is the Director of the Postgraduate School of Hygiene and Preventive Medicine, and the Director of the Umbrian Population Cancer Registry, of the Regional Mortality Registry and of the Umbrian Registry of Mesotheliomas.

Fabrizio Stracci graduated in medicine with first class honors at the University of Perugia, Perugia, Italy, in 1994, discussing a thesis on the cluster analysis of large bowel incident cancer cases. He received the Master's degree in public health at the same university, in 1998.

He worked at the establishment of the Cancer Registry of the Umbria Region, and presently, he is the Registry's Coordinator. He was the Scientific Coordinator of the epidemiology section of the Umbria Region Health Government. Presently, he works as an Assistant Professor at the Public Health Department of the Perugia University, Perugia. He is the author of over 50 scientific papers published on national and international peer-reviewed journals. His main research interests are in the field of epidemiology, particularly cancer epidemiology and public health.