A system for Geoanalysis of Clinical and Geographical Data

Giovanni Canino
Dept. of Clinical and Surgical Science, University Magna Græcia
Catanzaro, Italy
Dept. of Computer Science and Engineering, State University of New York at Buffalo
Buffalo, New York
canino@unicz.it

Pietro H.Guzzi
Dept. Clinical and Surgical Science, University Magna Græcia
Catanzaro, Italy
hguzzi@unicz.it

Giuseppe Tradigo
DIMES Dept., University of Calabria
Cosenza, Italy
tradigo@dimes.unical.it

Aidong Zhang
Dept. of Computer Science and Engineering, State University of New York at Buffalo
Buffalo, New York
azhang@buffalo.edu

Pierangelo Veltri
Dept. of Clinical and Surgical Science, University Magna Græcia
Catanzaro, Italy
veltri@unicz.it

ABSTRACT

Patients enrolled in clinical trials are regularly subject to biological analyses and related data is included in Electronic Medical Records (EMRs) to summarize patient health status and to support administrative information. Well defined protocols guide the bioanalytes studies on patients. Often EMRs also contain geographical data about patients, i.e. place of birth and place of living. The integration of geographical data and biological analytes may represent a meaningful way to extract hidden information from data. For instance, possible correlations among outlier patients and some feature of areas they live in.

In collaboration with the University Hospital of Catanzaro, we designed a framework able to integrate and analyze biological analytes. The system is able to relate biological data to diagnosis codes and to analyze integrated data against geographic areas of interest. The aim is to show correlations among patients features (e.g. cluster of patients with similar profiles or outlier patients) and areas features (e.g. presence of power grids or polluted sites). In addition we present a study on correlations between cardiovascular diseases and water quality in Calabria.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
HealtGIS'14, November 04-07 2014 Dallas/Fort Worth, TX, USA
Copyright 2014 ACM 978-1-4503-3136-4/14/11 ...$15.00
http://dx.doi.org/10.1145/2676629.2676635.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Data Analysys, Spatial Data

Keywords

EMR, Geographical Data Analysis, Clinical Diagnosis Analysis.

1. INTRODUCTION

Medical data, usually disseminated in different papers, may represent a meaningful resource of knowledge for both research and administrative purposes. The first step towards an efficient mining of such data has been represented by the introduction of methodologies and tools for the representation and management of medical data with computers [1]. The process then leaded to the definition and implementation of Electronic Medical Records (EMRs). As evidenced in literature, EMR may be defined as the *systematic collection of electronic health information about individual patients or populations*. Data contained in EMRs is usually highly heterogeneous, considering both structure and dimension. Example of typical data span from personal history of patients (e.g. results of laboratory tests, usually unstructured), to medical images (which may have huge dimension), to information about insurances. In fact, data is characterized by huge heterogeneity, both in terms of formats (e.g. text, images, time-series, biomedical signals) and representations. Moreover, a lot of information is usually stored in textual format (e.g. reports written by medical doctors) thus even simple operation, such as comparing two EMRs, is quite hard. Consequently, many research attempts

have been made in order to enable simpler extractions and integration of EMRs data [2].

More recently, the possibility to integrate geographical information about patients has attracted much interest. Nevertheless, the integrated analysis of geographical and EMRs data, is a challenging area which shows the need for the introduction of frameworks and tools able to gather information from different sources. Starting from these considerations, we designed and implemented a first software prototype able to relate diagnosis, bio-analytes of patients and characteristics of a given geographical area.
The prototype is based on different modules as depicted in Figure 1.

- *Integration module*, which collects data through a web based interface and integrates it into a single data model;
- *Analysis module*, able to analyze integrated data;
- *Bio-Analytes Data Analysis module*, performs intra cluster analysis of bio-analytes;
- *Geographical Analysis module*, it is responsible for analyzing the geographic information and represent it on a map.

The first prototype of this system has been implemented, and it has been used for the experiments presented in this article. We also discuss as a proof-of-concept the application of this prototype to the analysis of correlation between cardiovascular diseases and quality of public water, in a geographical area of the Calabria region, in the southern part of Italy. We correlate four different datasets, joined with respect to common alphanumeric or geographical features. We first consider a dataset containing almost 20 thousand anonymized EMRs relating to biological analytes (e.g. glycemia, bilirubin, cholesterol). All datasets are connected by unique code and,with dedicated queries, we extracted diagnosis and geographical informations . Then, for each given geographical area in one year observation time interval, we consider anonymized residential data from patients joined with some behavior information. We finally correlate with data about the quality of drinking water by considering public grid of transportation.

The goal is to show how heterogeneous information extracted from health data sources may be used to find possible correlations with similar diseases or water quality, as well as showing bio-analytes outliers and extreme values. In our experiments, health data sources have been: EMRs, bio-analytes data sets, administrative sources called DRGs (for Diagnosis Related Group), geographical information, water distribution grid. Finally,the proposed methodology has been tested and validated for a southern Italian region, but it can be applied to any area of interest, and can be extended by introducing additional layers, such as transport networks, to infer facts on pathologies. The paper is organized as follows. Section 2 reports some of the interesting references about ontologies and tools for health data integration. Section 3 reports the architecture of the system. Section 4 reports a case study on the correlation among the quality of drinking water and cardiovascular diseases. Finally Section 5 concludes the paper.

2. RELATED WORKS

Many studies have focused on semantic information extraction from EMRs and on definition of ontologies to include world wide approved terminology and health data description [3, 4]; similarly geo-epidemiology, i.e. the relation of healthy interesting information with environmental data, has been attracting many researchers are forming new communities [5, 6]. In non-health topics, such as business oriented and marketing information, cross analyses of heterogeneous data sources is often performed by using OLAP analysis. Nevertheless, few research has been currently presented regarding the integration and analysis of apparently unrelated health information to obtain prevention protocols (e.g. early disease detection).

Data extraction from EMRs is a known topic; for instance, ontologies have been defined to support such tasks. UMLS [7] was initially developed by the National Library of Medicine, which aimed at the standardization of terms in the biomedical domain. Current release of UMLS includes more than 100 controlled medical terminologies, like the *Systematized Nomenclature of Medicine - Clinical Terms* (SNOMED - CT) and the *Medical Subject Headings* (MeSH). The whole set of terminologies is used as a source for the UMLS Metathesaurus which integrates all of these data sources. Each term is labelled with Concept Unique Identifiers (CUIs) organized into is-a taxonomies.

A CUI may refer to different concepts belonging to different terminologies. Terms of terminologies are labeled with Atomic Unique Identifiers (AUIs). AUIs are structured into is-a taxonomies. Consequently a CUI refers to more than one AUIs. For example, the AUI Cold Temperature (*A15588749*) from MeSH and the AUI Low Temperature (*A3292554*) from SNOMED-CT are merged into a single CUI Cold Temperature (*C0009264*). Taxonomic relations within the Metathesaurus are stored using two tables: *MRREL* and *MRHIER*. The first one stores both hierarchical and non-hierarchical information among a pair of terms. The second one stores the full path-to-root from the sources and it is derived from the first one.

Many works on data integration are known and many medical terms have been defined. For instance, *GALEN* (Generalised Architecture for Languages, Encyclopedia and Nomenclature in Medicine) [8], *SNOMED* (Systematized Nomenclature of Medicine) [9], *GO* (Gene Ontology) [10], *Disease Ontology* [11] are examples of terms that can be used to integrate information. GALEN is designed to be a re-usable application-independent and language-independent model of medical concepts. SNOMED CT is a clinical healthcare terminology able to cross-map to other international standard. The Disease Ontology semantically integrates disease and medical vocabularies through extensive cross mapping of DO terms to MeSH, ICD, NCI's thesaurus, SNOMED and OMIM. Another diffusely adopted medical vocabulary is MeSH (Medical Subject Headings) [12]. MeSH is the controlled vocabulary thesaurus used for articles indexing on PubMed.

Geographic clustering of health data has also been studying using data analytics methods and health data [13]; also recently geographic information has been related to genomic and proteomics data [14], trying to relate behaviour, diseases and ethnicity. Data mining tools, such as Weka, present adhoc tools to analyze geographical and health [15]. Mortality of heart-related diseases has been studied for a long time;

e.g. in [16], where a study on a large area of Brazil is presented. Thyroidal pathologies and land distribution has also been studied in [17].

3. SYSTEM ARCHITECTURE

In Figure 1 we report the experiment workflow showing the macro-steps involved, from the acquisition of the original dataset towards the representation of the manipulated data on a geographical map.

The upper part of the figure shows the used data sets. EPRs were extracted from a biology department system by means of an ad-hoc system (as reported in the next section), and contain patients analytes. DRGs have been obtained from an administrative repository and regards the medical records and administrative information (such as costs) related to patients hosting periods. They contain medical information organized in medical classes diseases (MCD). Geographical data is related to geographical layers containing information about street and administrative information. For this work we import administrative and geopolitical layers which have been used to map patient information. The module is able to import and merge additional information about environmental facts or geometries that can be related to pathologies (e.g. water sources).

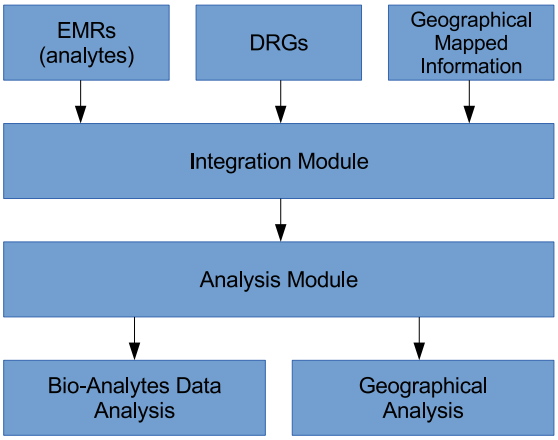


Figure 1: Architecture of the System

3.1 Integration Module

Analytes are extracted from a proprietary system used to control the biological analysis processes done on blood samples served at University Magna Græcia Medical Hospital. A simple parser Java has been implemented to extract 110 analytes that are then loaded into a MySQL database instance. A web-based system has been implemented to allow biologists and physicians to access data and perform statistics both on analytes, as well as on single patients information. This web system has been connected to our database of analytes.

The web-system has been implemented using the following technologies:

- PHP Server Side Language
- MySQL as the database management system (DBMS)
- Javascript,JQuery client-side language
- Twitter Bootstrap framework as a front-end

The PHP5 scripting language has been used for the implementation of the server-side scripts.

Finally an import/export interface has been implemented in order to import daily analytes from medical laboratories as well as to export analytes anonymized from personal data (see Figure 2). Thanks to a web interface, users can access analytes data, analyze them, visualize statistics and export relevant data (in a number of formats). By using the system biologists and physicians can have a more precise idea, for instance, of: (*i*) how analytes change over time in a single patient, or (*ii*) what are the statistics of blood values by sex and age. For example, one of our case study has been extracting patient data undergoing cardiac surgery for re-vascularization (coronary artery bypass). Understanding the evolution of data value, in patient hospitalization, has been relevant to find if come of cardiac damage related biomarkers (e.g. troponin, myoglobin, LDH, CK, proBNP) can be used as prognostic factors. Geographical locations are exported as standard WGS84 latitude and longitude to allow geographical information systems to store and manipulate data (see Figure 3).

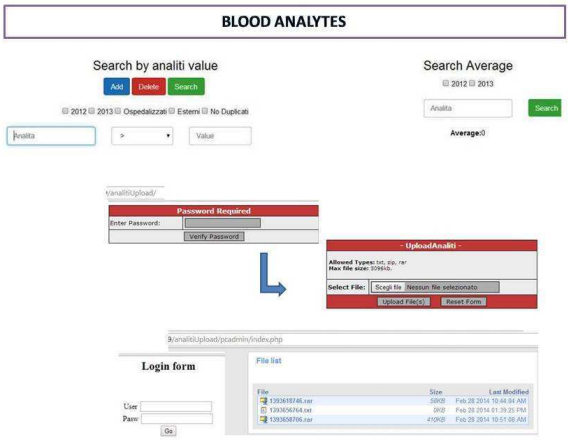


Figure 2: Website Blood Analites and Import/Export interface

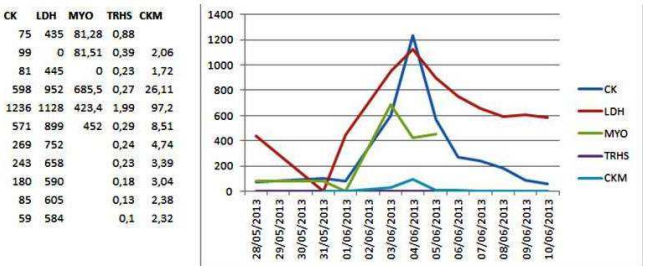


Figure 3: Evolution biomarkers of cardiac damage

3.2 Analysis Module

The analysis module is based on two main components: (*i*) an Analysis Module that analyzes textual information (e.g. DRG information) and bio-analytes, and (*ii*) Geographical Information Mapping and Analysis. The analysis module comprises two main modules: the textual analysis and the Weka [18] data-mining system. We implemented a software module which enables Weka to retrieve and store data directly from our database. Textual information analysis has been realized by wrapping Meta-Map [19] software to extract Diagnosis Codes from EMRs. Extracted EMRs are then used as input for the Weka platform. Geographical analyses have been implemented by integrating the QGis platform with our system [20].

3.3 DRG Analysis

A DRG includes information required by administration for costs identification and, where foreseen, the refunds. It contains different data about patients, such as: demographic data, personal history (e.g. medication, immunization, results of laboratory test), medical images, information about insurances. Diagnoses are mapped in a category code, each including classes of pathologies. Consequently, data are heterogeneous in terms of formats (textual, images, time-series, biomedical signals) and representations. There is an increasing focus on the research potential of both structured and textual data about patients (see for instance [2]). Some of these works deal strictly with structured and complex data, while some use text mining techniques to extract information from free text using natural language processing and ontologies.

We focus on the problem of extracting pathologies from DRGs, where diagnoses are obtained by mapping DRG codes into pathology classes, also known as MCD. We consider 25 different MCD pathology classes and we cluster part of them by analyzing DRGs on a one year observation time interval.

3.4 Geographical Information Mapping and Analysis

Data has been analyzed and queried using QGis [20], which is an open source tool for advanced spatial data management and querying and in this study the geographical area of interest is Calabria (a southern Italy region with more than 2 million inhabitants). Analytes data has been extracted from the analite databases, filtered with respect to diagnosis clustering (e.g., Lung Respiratory diseases), preprocessed and imported in Weka. An unsupervised filtering analysis has been performed to find outliers and extreme values.

4. CASE STUDY: RELATING CARDIOVASCULAR DISEASES TO DRINKING WATER QUALITY

Drinking water quality is an important factor for human health. Some previous studies demonstrated the presence of a correlation between cardiovascular diseases and geochemical factors, such as the level of calcium ions and magnesium present in drinking water [21]. Minerals found in the water are more bio-available to the body than those found in food. The assumption of appropriate concentrations of calcium and magnesium within drinking water allows to maintain appropriate physiological levels of these minerals in our body [22]. Adequate intake of magnesium allows you to maintain a lower cardio-metabolic risk in terms of reducing: (*i*) systolic and diastolic blood pressure, (*ii*) development of plaques of atherosclerosis, (*iii*) necrotic processes. Adequate intake of calcium modulates the levels of blood pressure and promotes natriuresis [23].

Starting from this scenario, the need arises for the introduction of frameworks and tools able to analyze water quality and diseases in an integrated way. Currently, data about diseases is available and stored in EMRs of hospitals. Data about drinking quality is available in regional government archives. Unfortunately these data is often managed (and analyzed) separately. The purpose of this study was to evaluate the correlation between levels of calcium and magnesium ions present in drinking water supplied by municipalities of Calabria with the large number of patients affected by heart-related diseases.

4.1 Input Dataset

Experiments performed in the present study have been conducted by using three main datasets:
(*i*) A dataset containing EMRs of patients from the University Hospital of Catanzaro. Data loaded in our system have been anonymized and then we extracted only analytes of patients with cardiovascular diseases (as reported in their DRG). For these patients we also considered their geographic information. This dataset has been used as reference model to identify average values and normal distribution of analytes in the region of interest;
(*ii*) A dataset (ASP dataset) containing data about patients with heart diseases in Cosenza province. For each patient we considered both geographical and analytes information. This dataset is used to mine surprisingly different distribution of analytes related to drinking water;
(*iii*) A dataset (Drinking Water dataset) about drinking water of the Cosenza province. It contains values of calcium, magnesium and hardness of waters, together with other useful information such as address, date and site of the levy as a source, fountain. Figure 4 shows an example of the database created for waters. All samples of drinking water have been acquired in the same year.

We were able to verify the geographic variation of cardiovascular diseases in the whole region and in the Cosenza province and the geographic variation of calcium, magnesium and hardness of the water in combination with the distribution of cardiovascular diseases.

Code_test	Label	Sample	City	Address	Latitude	Longitude	Hardness	Calcium	Magnesium
3.101_14	Acqua Potabile	Fontana Fellaro	Altomonte	Contrada Fellaro	39.7043463	16.1250493	21,1	54,84	17,3
7.52_14	Acqua Potabile	Fontana S. Francesco	Altomonte	Largo della Solidarietà	39.6984817	16.1289488	21	55,31	17,5
2.219_12	Acqua Potabile	Piazza Mancini	PIETRAPAOLA	Piazza Mancini	39.4861533	16.8146675	24,2	63	22,6
2.222_12	Analisi Acqua potabile	Uscita Fontana LOC. Pendino	Corigliano Calabro	via Manila	39.5941841	16.5201016	31	82,6	28,4
3.187_12	Acqua Potabile rete comunale	fontana loc. iacina	Corigliano Calabro	iacina	39.5941841	16.5201016	6	24	12,8
3.198_12	Acqua Potabile rete comunale	Fontana loc. Bonia	Corigliano Calabro	Bonia	39.5122358	16.5188196	6,6	18,4	5,35
4.187_12	Acqua Potabile rete comunale	Fontana Via Lucania	Corigliano Calabro	Via Lucania	39.6269050	16.5161700	27,3	82,1	18,3
4.198_12	Acqua Potabile rete comunale	Pozzo loc. Simonetti	Corigliano Calabro	Simonetti	39.5941841	16.5201016	9,72	24,8	10,12
4.222_12	Analisi Acqua potabile	Uscita fontana loc. Torricella inferiore	Corigliano Calabro	Torricella inferiore	39.5941841	16.5201016	22	48	24,3
5.187_12	Acqua Potabile rete comunale	Fontana Parco Periurbano	Corigliano Calabro	Parco Periurbano	39.5941841	16.5201016	5	25,89	11,92
5.198_12	Acqua Potabile rete comunale	loc. Bosco dell'acqua	Corigliano Calabro	Bosco dell'acqua	39.5941841	16.5201016	7	20	9,72
5.222_12	Analisi Acqua potabile	Uscita fontana loc Torricella superiore	Corigliano Calabro	Torricella superiore	39.5941841	16.5201016	21,6	47,6	23,6
6.187_12	Acqua Potabile rete comunale	fontana Via fontanella- corigliano scalo	Corigliano Calabro	Via fontanella	39.6265338	16.5097356	21,7	80,1	17,01
6.198_12	Acqua Potabile rete comunale	fontana Piana di Caruso	Corigliano Calabro	Piana di Caruso	39.5589363	16.5296824	6,7	16	6,56
6.222_12	Analisi Acqua potabile	uscita pozzo nuovo Cda S. Lucia	Corigliano Calabro	S. Lucia	39.6378839	16.5196808	18,1	43,2	17,74
6.223_12	Analisi Acqua potabile	Serbatoio villaggio Frassa	Corigliano Calabro	villaggio Frassa	39.5941841	16.5201016	26	80	14,58
7.198_12	Analisi Acque Reti Comunali	Serbatoio Cozzo Casale	Corigliano Calabro	Cozzo Casale	39.5941841	16.5201016	10	7,2	11
7.222_12	Analisi Acqua potabile	uscita pozzo Grande C.da S. Lucia	Corigliano Calabro	S. Lucia	39.6378839	16.5196808	19	45,6	19,44
7.223_12	Analisi Acqua potabile	Serbatoio Scalo	Corigliano Calabro	Scalo	39.6246707	16.5142169	26,4	80	15,52

Figure 4: Drinking Water dataset with value and geolocation sites

4.2 Relation between Calcium levels and patients Analytes

We studied the limit protection in cardiovascular disease from the literature for calcium, magnesium and water hardness. Figure 5 represents an example of calcium geoloca-

tion in the examined region. Images geolocation with QGIS show Calabria municipalities with calcium and magnesium ion concentrations higher than in other areas. These differences may be related to different geological characteristics and to waters rich in calcium and bicarbonates.

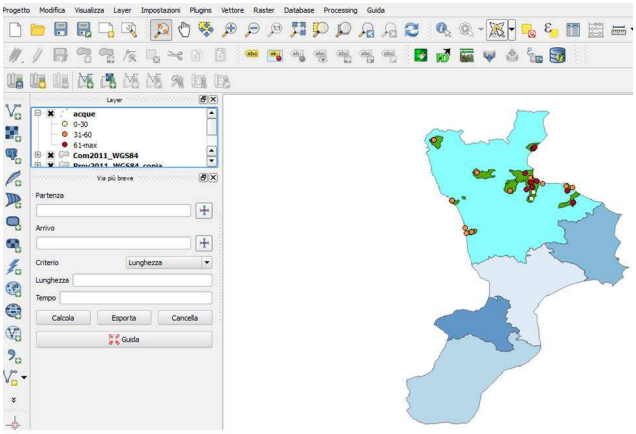
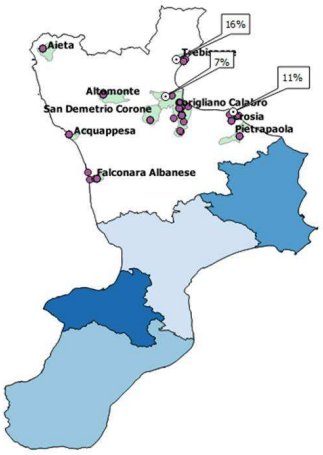


Figure 5: Calcium values

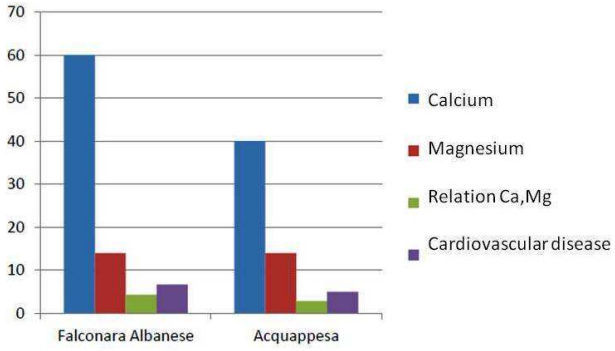
Concentrations of calcium and magnesium define the quality of drinking water [24], based on the composition of the mountain ranges or the type of rocks that the water passes through its way. The presence of these salts is a factor to be evaluated in terms of bioavailability of calcium and magnesium in drinking water. The presence of high chlorides or sulfates levels in water increases the urinary excretion of calcium and magnesium regardless of the hormonal changes associated with physiological values, such as PTH parathyroid hormone [25].

Using EMRs and comparing them with data in the Asp and the Drinking Water datasets, we derive the proportion of patients with heart disease. Where the proportion is higher, we have a higher percentage of heart disease and calcium concentration values, with magnesium and hardness exceeding thresholds minimums. Percentages are normalized with respect to population.

According to some studies, the increase of $1mg/l$ magnesium ion concentration in groundwater is associated with the decrease of 4.9% in the incidence of cardiovascular diseases, while a unit increment ratio Ca_2+/Mg_2+ is associated with an increase of 3.1% in the incidence of cardiovascular disease. An example of this is with the municipalities of *Falconara Albanese* and *Acquappesa* (shown in Figure 6(b)(b)). The ASP dataset indicates that the number of patients with cardiovascular disease is similar for the two municipalities. The values of calcium ion are different for the two municipalities. The concentration of magnesium ion are, however, similar and in the order of $15mg/l$. The ratio of $Ca_2 + /Mg_2+$ is the highest recorded for the town of *Falconara*, where we observed a relatively slight increase in patients with cardiovascular disease. Our study showed an inverse association of total hardness with the incidence of cardiovascular disease in the municipalities analyzed. Data shows that high concentrations of magnesium in drinking water may be associated with a smaller number of cases of cardiovascular disease. Based on this, the people living in geographical ar-



(a) Higher percentage of heart disease



(b) Cardiovascular disease detail values for two municipalities

Figure 6: Detail on result data about heart disease showing high risk

eas with low total hardness, but high Ca_2+/Mg_2+ ratio in groundwater, may have an increased risk of cardiovascular diseases.

5. CONCLUSION

The paper presents the design and first implementation of a framework able to integrate geographical data, EMRs, and laboratory tests. We are designing a framework to integrate clinical information related to patients (from clinical and family history to biological and omics) with environmental information.The final goal is to build a data warehouse hosting information about lifestyle of a population and monitor the environment status.We used the system to collect data from the University Hospital of Catanzaro. Moreover we also considered a dataset of the Cosenza province and the quality of drinking waters. The study is currently under evaluation by medical doctors. Future work regards the full implementation of the system and its deployment as a service.

Acknowledgment

The authors would like to thank Giovanni Cuda and Francesco S. Costanzo for their support on analytes data extraction and Assunta Gallo. This work has been partially funded by *MIUR PRIN 2010 - Gendata 2020* and by *MIUR PON-Staywell 2.0* projects.

6. REFERENCES

[1] F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Soeby, S.Bredkjr, A. Juul, T. Werge, L.J. Jensen, S. Brunak, Using electronic patient records to discover disease correlations and stratify patient cohorts, *PLoS Comput Biol*, 7(8), 2011

[2] S. Meystre, P. J. Haug, Natural language processing to extract medical problems from electronic clinical documents: performance evaluation,*Journal of biomedical informatics*, 39(6), 2006

[3] O. Bodenreider, Biomedical ontologies in action: role in knowledge management, data integration and decision support, *Yearb Med Inform*, 47(1), 2008

[4] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M. Storey, C. G. Chute, M. A. Musen, BioPortal: ontologies and integrated data resources at the click of a mouse, *Nucl. Acids Res.*, 37(2), 2002

[5] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, J. C. Denny, Application of information technology: MedEx: a medication information extraction system for clinical narratives, *J Am Med Inform Assoc*, 17(1), 19-24, 2010

[6] G. Hripcsak, C. Friedman, P.O. Alderson, Unlocking clinical data from narrative reports: a study of natural language processing, *Ann Intern Med*, 122, 681-8, 1995

[7] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, *Nucleic Acids Research*, 32(1), 2004

[8] J.E. Roger, Quality Assurance of Medical Ontologies, *Methods Inf Med* 45(3), 267-274, 2006

[9] International Health Terminology Standards Development Organisation, SNOMED website, *http://www.ihtsdo.org/snomed-ct/*, 2014

[10] Gene Ontology Consortium, The Gene Ontology (GO) database and informatics resource, *Nucl. Acids Res.*, 32(1), 2004

[11] L. M. Schriml, C. Arze, S. Nadendla, W. Chang, M. Mazaitis, V. Felix, G. Feng, W. Kibbe, Disease Ontology: a backbone for disease semantic integration, *Nucl. Acids Res.*, 40, 2012

[12] C. E. Lipscomb, Medical Subject Headings (MeSH), *Bull. Med. Libr. Assoc.*, 88(3), 2000 *http://www.nlm.nih.gov/mesh/*

[13] D. A. Moore, T. E. Carpenter, Spatial Analytical Methods and Geographic Information Systems: Use in Health Epidemiology, *Epidemiol Rev.*, 21(2), 1999

[14] C. R. Williams-DeVane, D. M. Reif, E. C. Hubal, P. R. Bushel, E. E. Hudgens, J. E. Gallagher, S. W. Edwards, Decision tree-based method for integrating gene expression, demographic, and clinical data to determine disease endotypes, *BMC Systems Biology*, 7(119), 2013

[15] S. K. David, A. T. M. Saeb, K. Al Rubeaan, Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics, *Computer Engineering and Intelligent Systems*, 4(13), 2013

[16] L. de Andrade, C. Lynch, E. M. Spiecker, M. D. de B. Carvalho, O. K. Nihei, Spatial Distribution of Ischemic Heart Disease Mortality in Rio Grande do Sul, Brazil, *2nd International ACM SIGSPATIAL Workshop on HealthGIS*, 2013.

[17] G. Tradigo, P. Veltri, O. Marasco, G. Scozzafava, G. Parlato, S. Greco, Studying neonatal TSH distribution by using GIS, *ACM HealthGIS*, 2012

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA Data Mining Software: An Update, *SIGKDD Explorations*,2009

[19] A. R. Aronson, Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program, *Proceedings of AMIA Annual Symposium*, 2001

[20] QGIS Development Team, QGIS Geographic Information System, *Open Source Geospatial Foundation*, 2009, http://qgis.osgeo.org

[21] P. R. Hunter, A. M. MacDonald, R. C. Carter, Water Supply and Health, *PLoS Med*, 7(11), 2010

[22] R. Maheswaran, S. Morris, S. Falconer, A. Grossinho, I. Perry, J. Wakefield, P. Elliott, Magnesium in drinking water supplies and mortality from acute myocardial infarction in north west England, *Heart* 82, 455-460, 1999

[23] A. Kousa, E. Moltchanova, M. Viik-Kajander, M. Rytkönen, J. Tuomilehto, T. Tarvainen, M. Karvonen, Geochemistry of ground water and the incidence of acute myocardial infarction in Finland, *Epidemiol Community Health*, 58, 136-139, 2004

[24] Rylander R, Arnaud M., Mineral water intake reduces blood pressure among subjects with low urinary magnesium and calcium levels, *BMC Public Health*, 4:56.2004.

[25] Cotruvo J, Bartram J, Calcium and Magnesium in Drinking-water : Public health significance. *World Health Organization*, 2009