# PRIN GENDATA 2020: Deliverable 3.2

# CLINICAL DATA MODELS AND APPLICATIONS

*Sergio Greco[1], Giuseppe Tradigo[1], Pierangelo Veltri[2]*

*(1) Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica*
*Università della Calabria*

*(2) Università degli Studi "Magna Grecia" di Catanzaro*

## Introduction

Recently, the research community has been showing increasing interest on the problem of integrating clinical and genomic data [1]. In literature, many data integration algorithms and approaches can be found. Nevertheless, applying them on clinical data is not a trivial task, due to its heterogeneity and inherent complexity. In fact, clinical data is usually persisted in a number of different data sources and is often unstructured and uncorrelated. One of the recent trends in clinical research, regarding clinical data integration, is the development of models on clinical quantities (e.g. pathologies, infections, pollution risk) related to their geographical extension.

The applications we present show how our clinical data integration system is able to discover spatial correlations among cardiovascular diseases and quality of waters in an Italian region [2, 3, 4]. The relevance of the topic, i.e. discovering new knowledge in clinical and genomic data, can have important implications in a better understanding of molecular insights about diseases and the identification of pathologies related to genes alterations. We studied and designed a framework able to integrate and analyze biological analytes able to relate biological data to diagnosis codes and to analyze integrated data towards areas of interest. The aim is to evidence correlation among patients characteristics (e.g. cluster of patients with similar profiles or outlier patient) and characteristics of the areas (e.g. presence of power grids).

In this section we report about research activities developed at University of Calabria - DIMES Department and at University of Catanzaro, Surgical and Medical Science Department, regarding the analysis of clinical and genomic data and their relation with respect to environmental distribution. Also, we report a generic model useful to represent data extracted from heterogeneous data sources including clinical and genomic data. Ideas underlying these activities, also reported in [2, 3, 4, 5, 6], have been performed to explore the information integration regarding health-related data thus allowing targeted public prevention programs and early disease detection.

## State-of-the-art

Personalized medicine is a novel trend in medicine which aims to create procedures and techniques able to accurately classify patients with respect to their peculiar needs in order to improve the diagnosis (i.e. getting a better one faster) and personalize the therapeutic path. To this end, genomic medicine is a data intensive field which attempts to explain the molecular basis of disease and to translate this knowledge into clinical practice by its integration with other diagnostic data. It uses information from genomes and their derivatives (e.g. RNA, proteins) to guide medical decision making [7].

Recently, the possibility to integrate geographical information about patients have gained a growing interest. Nevertheless, the integrated analysis of geographical and EMRs data is a challenging area causing 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.

Geographic clustering of health data has also been studying using data analytics methods and health data [8]; also recently geographic information has been related to genomic and proteomics data [9], trying to relate behavior, diseases and ethnicity. Data mining tools, such as Weka, present ad-hoc tools to analyze geographical data [10] and health data. Heart attack disease and mortality has been studied in a large context with respect to time to hospitalization, e.g. in [11] a study on a large area of Brazil is presented. Thyroidal pathologies and land distribution has also been studied in [2].

## Activity description

We present an application to improve cancer risk assessment models based on genetic tree predictive algorithms (see Figure 1) and an example of spatial correlation model (see Figures 5 and 6), both based on an architecture for clinical and genomic data integration and querying.

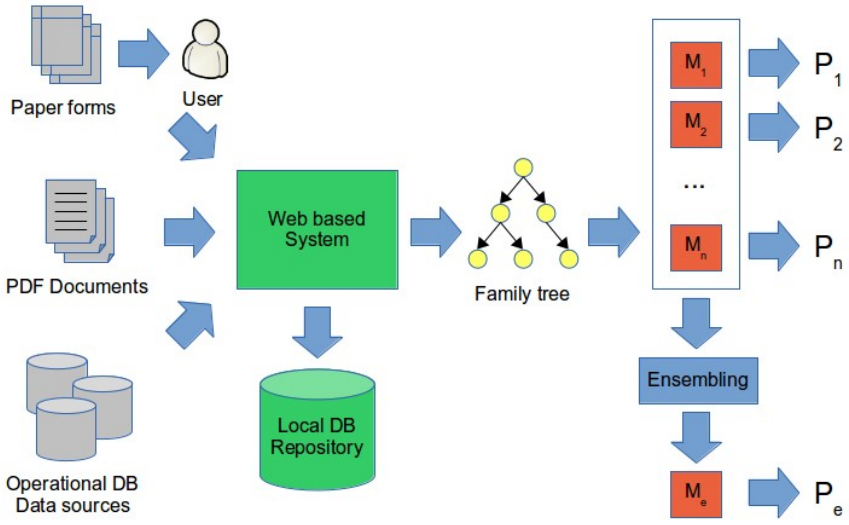



Figure 1 - System Architecture for clinical and genomic data integration.

Knowledge about patient data is extracted and organized by using an ontology, a fundamental tool for information extraction, to define cancer risk related to gene mutation. Such an information can be used for clinical strategies, such as monitoring and check frequencies, usually written in the patient record (see Figure 2).

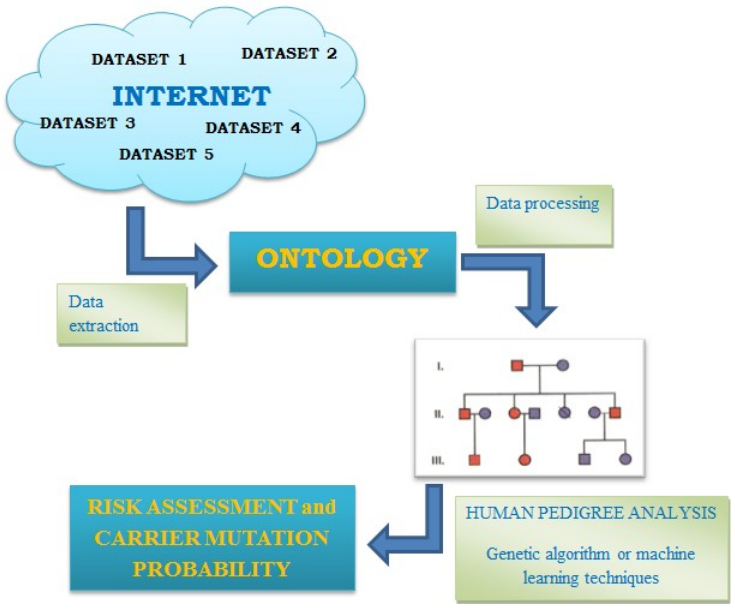



Figure 2 - Human pedigree analysis supporting the cancer risk assessment protocol.

In addition, more accurate predictors based on machine learning techniques (e.g., artificial neural networks, bayesian networks) can be built based on such information.

A possible application of the integrated system presented above, based on ClinicalGenomic Database, is represented by the genetic counseling clinical procedure. The purpose of genetic counseling procedure is the disease management for oncological patients (and their families) with susceptibility or showing high hereditary risks for cancer related diseases. It provides information about genetic transmission probabilities for a specific tumor, which is useful for early disease diagnosis and treatment.

The genetic counseling protocol consists of three main steps: pre-test, genetic test and post-test. The pre-test phase is composed of: (*i*) data collection of family and personal history of the subject, filling of family tree with age and health status of known relatives; (*ii*) verification of clinical data by retrieving medical records about all family cancer cases; (*iii*) computation of risk assessment; (*iv*) delivery of information material to patients and their families and, if indicated, proceed with blood sample analysis and genetic testing; (*v*) assessment of the psychological impact by a psychologist.

The genetic test phase regards molecular analysis of patient's DNA, to verify the presence of a mutation

associated with a high risk of developing a malignancy. The molecular study provides guidance on the possible presence of mutation in the gene of interest, the type of mutation with reference to its presence in the genomic database (e.g. GeneBank) and its presence in disease-specific databases (eg. Breast Cancer Information Core – BIC) [12].

The post-test phase consists of: (*i*) communication of the results by the medical geneticist; (*ii*) discussion of possible treatment options, monitoring and prevention; (*iii*) offer counseling and testing to family members in order to include them in the monitoring plans for high-risk individuals.

Thanks to genetic counseling it will be possible to enrich the ClinicalGenomic database, with data from probands and their families. Genealogical data previously acquired and, if available, genomic data obtained from molecular analysis that have been made, could serve a more accurate reconstruction of a patient's family history of cancer addressed to the counseling process. The risk assessment carried out by using already available models, will benefit from the larger information obtained by using ClinicalGenomic Database. In addition, an improvement of risk assessment models, today widely used, is conceivable. One of the main advantages of counseling controls resides in the improvement of prevention protocols and screening and, in an indirect manner, in the guided choice for a therapy.

Another important research topic is relating clinical and genomic information with geographical features, in order to enhance clinical knowledge and build sophisticated epidemiological models able to reason or predict the spread of particular diseases on the territory. One of our studies focuses on relating diagnosis and bio-analytes in geographic zones. The geographical area regards a southern Italian region with 2 million inhabitants. We consider the information integration of three different data sets joined with respect to common geographical areas. A first data set contains almost 20000 anonymized EMRs relating to biological analytes (such as glycemia, bilirubin, cholesterol); a second one is about diagnosis information of dismissed patients relating to a given geographical area in a one year observation time interval; and finally, anonymized residential data from patients joined with behavior information. The three datasets have been anonymized and joined by using equal values (i.e. patient anonymized IDs).

Finally, the methodology has been tested and validated for an Italian region but it can be extended to any area of interest by introducing additional layers such as water sources, water technological transport network or other facts about pathologies.

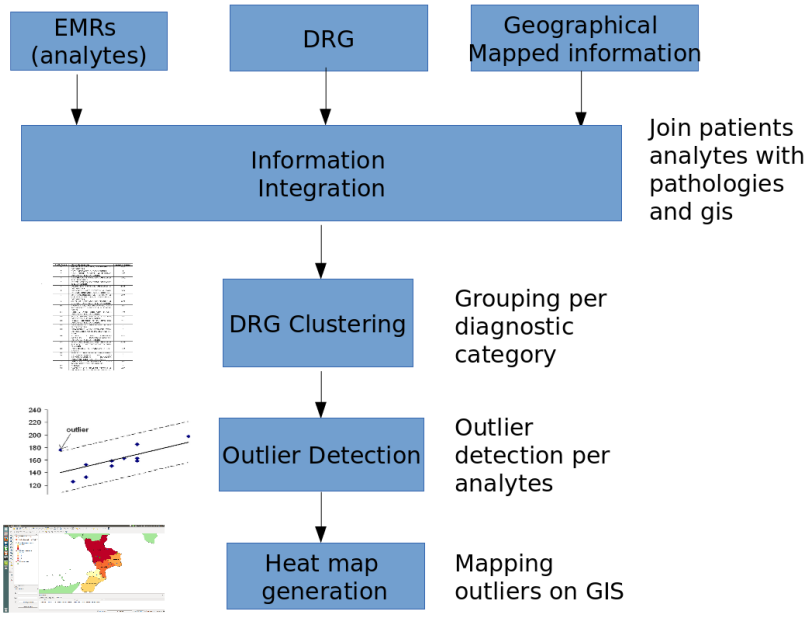



Figure 3 - The Wokflow of Data acquisition and processing for the mapping experiment.

In Figure 3 we report the experiment workflow showing the steps involved in data processing from source datasets towards the results projection on geographical maps. The upper part of the figure reports the data sets used for the experiments. The EPRs were extracted from biology department system by means of an ad hoc system (as reported in next section), and contain patients analytes.

DRGs have been obtained from administrative repository and regards the medical records and administrative information (such as costs) related to patients hosting periods. They contain medical information that are organized in medical classes diseases (MCD). Geographical data attains to geographical layers containing information about street and administrative information.

For this work we imported administrative and geopolitical layer which are used to map patient information. However, the module is able to import and merge additional information about environmental facts or geometries which could be related to pathologies (e.g. water sources, pollutants). Analytes have been extracted from a proprietary system, used to control the biological analysis processes, from blood samples of patients at the Magna Graecia University Medical Hospital. A parser has been implemented to extract analytes, which are then loaded into a MySQL database instance. Data access and browsing for domain experts (i.e. biologists and physicians) is made through a web-based interface, allowing also to perform statistics both on analytes and on single patients information.

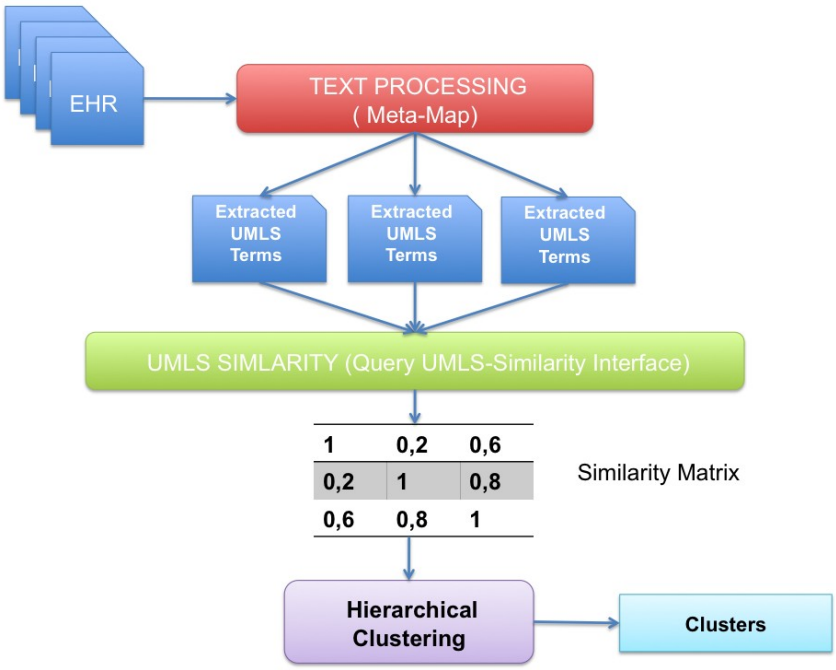



Figure 4 - The flow of Data into SHC-EHR.

We focused on the problem of extracting pathologies from DRGs collected by also considering geographical information regarding patients. Since DRGs refer to real cases, we filtered out administrative information (e.g. costs) treated them as Electronic Medical Records. The diagnoses have been obtained by mapping the DRG codes to pathology classes (also indicated as Medical Class Disease, MDC).

We considered 25 different MDC pathology classes and we clustered part of them by analyzing DRGs for one year observation time. We compared and grouped DRGs using the workflow depicted in Figure 4, where the semantic extraction and information clustering have been performed on DRGs treated as non-structured documents. In order to extract information about diagnoses, the UMLS Meta-thesaurus has been used (UMLS for Unified Medical Language System [13]). In particular, text contained in different DRGs is converted into a not-ambiguos signature using UMLS. For each patient we analyzed the related terms contained in UMLS and we mapped each term in the complete UMLS definition, obtaining a single representative file for each DRG. The extraction has been made using MetaMap, a tool for mapping a text into the UMLS Meta-thesaurus or to extract Meta-thesaurus from a text by using different text mining and Natural Language Processing methods [14]. We then compared two sets of UMLS terms on a semantic space by using semantic similarity functions able to evaluate the similarity or relatedness of two terms [15]. Similarly to other fields (e.g. computational biology [16]), in literature many examples of methods and tools for evaluating semantic similarity among UMLS terms can be found. Here we use UMLS-Similarity to calculate different semantic similarity measures among UMLS terms [17].

If we consider, for instance, *n* DRGs as input, we obtain an *n* x *n* matrix in which each element *(i, j)* represents the average semantic similarity among each pair of UMLS terms contained in the related DRG files. Finally we used hierarchical clustering to extract pathology clusters using the similarity matrix as a distance matrix. Data has been analyzed and queried using QGis, an open source platform for spatial data management and querying [18].

Analytes data has been extracted from the analytes databases and filtered with respect to diagnosis clustering classes (e.g., Lung Respiratory diseases). Once preprocessed, we searched for outliers and extreme values with unsupervised filtering analysis (performed in the Weka open source platform [19]) and we then mapped results in QGis.

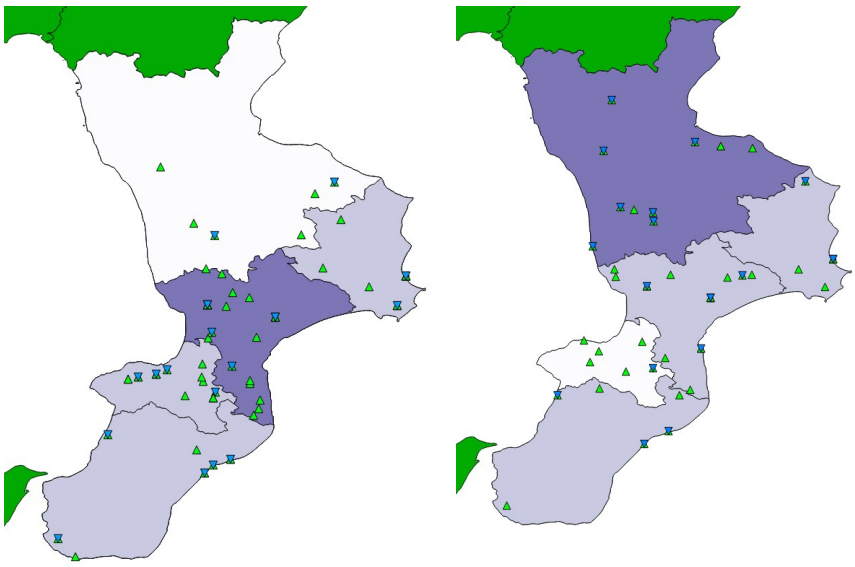
Table 1 reports the list of 25 MDC cluser classes categorizing the pathologies extracted from our DRG dataset. From this table we considered 17 out of 25 clusters, filtering out the ones having very low number of exams (e.g., 1 or 0).

| Code | MDC Description (Diseases And Disorders) | Num of Exams |
|---|---|---|
| 1 | Nervous system, diseases and disorders | 463 |
| 2 | Eye, Diseases And Disorders | 1 |
| 3 | Ear, Nose, Mouth Diseases And Disorders | 177 |
| 4 | Respiratory System, Dis. And Disorders | 1082 |
| 5 | Circulatory System, Dis. And Disorders | 0 |
| 6 | Digestive System, Dis. and Disorders | 2116 |
| 7 | Hepatobiliary System And Pancreas Disorders | 313 |
| 8 | Musculoskeletal System And Connective Tissue | 259 |
| 9 | Skin, Subcutaneous Tissue And Breast Disorders | 629 |
| 10 | Endocrine, Nutritional, And Metabolic Disorders | 61 |
| 11 | Kidney And Urinary Tract, Dis. And Disorders | 169 |
| 12 | Male Reproductive System, Dis. And Disorders | 12 |
| 13 | Female Reproductive System, Dis. And Disorders | 612 |
| 14 | Pregnancy, Childbirth, And The Puerperium | 0 |
| 15 | Newborns,Neonate Conditions In Perinatal Period | 0 |
| 16 | Blood Forming Organs,Immunological Disorders | 488 |
| 17 | Poorly Differentiated Neoplasms | 5796 |
| 18 | Infectious And Parasitic Diseases | 452 |
| 19 | Mental Diseases And Disorders | 0 |
| 20 | Alcohol Drug,Use And Induced Organic Mental | 0 |
| 21 | Injuries,Poisonings,Toxic Effects Drugs | 61 |
| 22 | Burns | 0 |
| 23 | Factors On Health Status and Health Services | 442 |
| 24 | Multiple Significant Trauma | 17 |
| 25 | Human Immunodeficiency Virus Infections | 0 |

Table 1 – MDC description

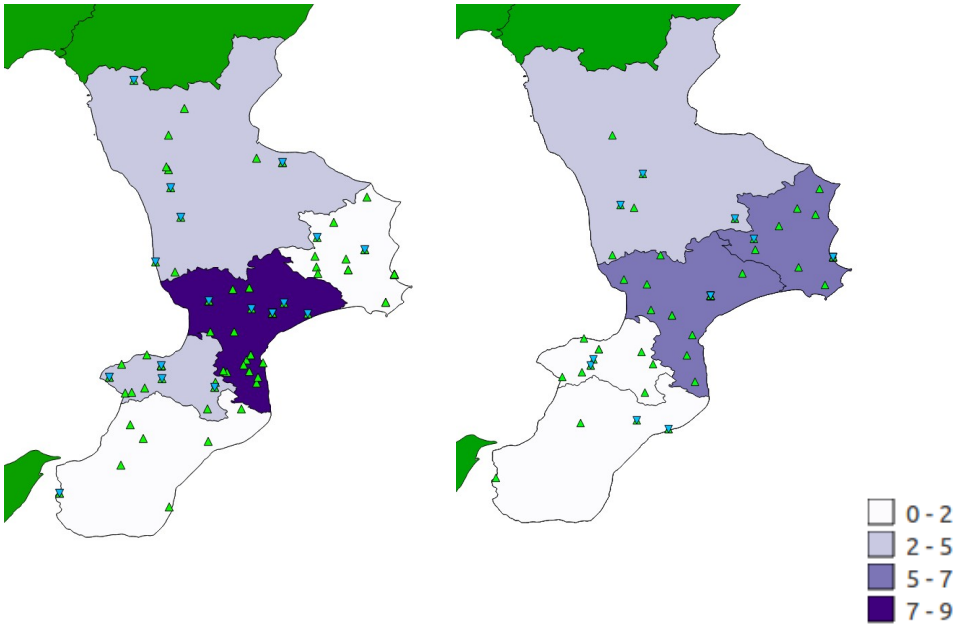
Such values have been then joined with analytes values obtained from EMRs patients, connecting pathologies with patients and with geographical areas. Biological analytes have been considered with respect to their pathology class, thus allowing to identify outlier patients (i.e. patients having analytes values remarkably different form other patients in the same class).

We performed experiments starting from the groups of MDC (Medical Disease Classes) reported in Table 1. We focused on 4 MDC classes, in particular we selected the following classes having the highest number of cases: (*i*) *Poorly Differentiated Neoplasms*, (*ii*) *Respiratory System, Dis. And Disorders*, (*iii*) *Skin, Subcutaneous Tissue and Breast Disorders* and (*iv*) *Digestive System, Dis. And Disorders*. Figure 5 reports four heat-maps calculated for the 4 MDC classes described above. We identified the outlier patients by considering the intra-cluster similarities among analytes. We then built a set of heat maps by spatially querying the number of outliers per areas of interest (i.e. provinces and municipalities). Visually we have large number of overlapping data points because our patients information has been furnished without exact residential details (e.g., building numbers for a given municipality).



(a) Digestive System, Dis. and Disorders (Digestive diseases) cluster

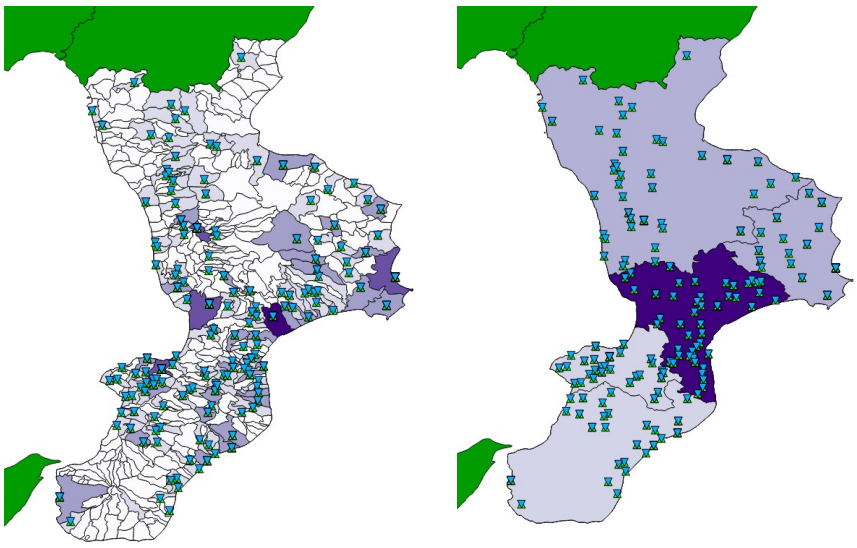
(b) Poorly Differentiated Neoplasms (Leukemia diseases) cluster




(c) Skin, Subcutaneous Tissue and Breast Disorders (Breast related diseases) cluster

(d) Respiratory System, Dis. and Disorders (Lung related diseases) cluster

Figure 5 - Heatmaps representing the number of outliers by area of interest (provinces of Calabria region). Legend reports the values ranges of outlier patients used to color the areas of interest.



(a) Pathologies by municipality

(b) Pathologies by province

Figure 6 - Heat-maps representing observed pathologies per municipalities and provinces

Nevertheless, the outlier accuracy does not depend on the exact position, because we consider the count of result points insisting on each area. Areas showing high counts are of high interest for health operators, especially considering relations with other land description layers (e.g. drinking water quality, land use, contaminated areas, pollution [20]). The four maps reported in Figure 5 depict both normal patients (green upward triangles) and outliers (blue downward triangles). Finally, Figure 6 shows all pathologies on two different area granularities, one by municipalities and the other by provinces.

## References


[1] S. Dey, R. Gupta, M. Steinbach, V. Kumar, Integration of Clinical and Genomic data: a Methodological Survey, *Technical Report, Department of Computer Science and Engineering University of Minnesota*, 2013

[2] G. Tradigo, P. Veltri, O. Marasco & G. Parlato. Studying human TSH distribution by using GIS, *ACM HealthGIS*, 2012

[3] Canino, G., Guzzi, P. H., Tradigo, G., Zhang, A., & Veltri, P., A system for Geoanalysis of Clinical and Geographical Data, In Proc. of *ACM HealthGIS*, 2014

[4] Tradigo, G., Pagliaro, C., Canino, G., Casalinuovo, F., Graziani, C., & Veltri, P., A model for the Geographical Analysis and monitoring of agricultural areas example and tests in south Italian regions, In Proc. of *ACM HealthGIS*, 2014

[5] Ginsburg, G. S., & Willard, H. F., Genomic and personalized medicine: foundations and applications. *Translational Research*, 154(6), 277-287, 2009

[6] Tradigo, G., Veneziano, C., Greco, S., & Veltri, P., An Architecture for integrating Genetic and Clinical data, In Proc. of *IEEE International Conference on Computational Science (ICCS)*, 2014

[7] Zeshan, F., & Mohamad, R., Medical Ontology in the Dynamic Healthcare Environment. *Procedia Computer Science*, 10, 340-348, 2012

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

[9] C.R. Williams-DeVane, D.M. Reif, E. Cohen 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

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

[11] 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, *ACM HealthGIS*, 2013

[12] Breast Cancer Information Core - BIC website, *http://research.nhgri.nih.gov/bic/*, 2014

[13] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, *Nucleic Acids Research*, vol. 32, no. 1, 2004

[14] A. R. Aronson, Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program,

In Proc. of *AMIA Annual Symposium*, 2001

[15] V. Garla, C. Brandt, Semantic similarity in the biomedical domain: an evaluation across knowledge sources, *BMC Bioinformatics*, vol. 13,n. 1, 2012

[16] P.H. Guzzi, M. Mina, C. Guerra, M. Cannataro, Semantic similarity analysis of protein data: assessment with biological features and issues , *Briefings in Bioinformatics*, vol. 13, n. 5, 2012

[17] T. Pedersen, S.V.S. Pakhomov, S. Patwardhan, C.G. Chute, Measures of semantic similarity and relatedness in the biomedical domain, *Journal of Biomed Information*, Vol 40, 2007.

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

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA Data Mining Software: An Update, *SIGKDD Explorations*, Volume 11, Issue 1, 2009

[20] M.P. Sauvant, D. Pepin, Geographic Variation of the Mortality from Cardiovascular Disease and Drinking Water in a French Small Area, *Environmental Research*, Section A, n. 84, 2000