Integration and Analysis of Clinical and Genomic Data of Neuroblastoma applying Conceptual Modeling

Sipan Arevshatyan1[0000-0001-8718-2211], José Fabián Reyes Román1[0000-0002-9598-1301], Verónica Burriel2, Adela Cañete3, Victoria Castel3, and Óscar Pastor1[0000-0002-1320-8471]

[1] PROS Research Center, Universitat Politècnica de València,
Camino Vera s/n. 46022, Valencia, Spain
[2] Department of Information and Computing Sciences, Utrecht University,
Domplein 29, 3512 JE, Utrecht, The Netherlands
[3] Pediatric Oncology Unit of Hospital Universitari i Politècnic La Fe, Valencia, Spain
siar5@doctor.upv.es, {jreyes|opastor}@pros.upv.es, v.burriel@uu.nl,
{canyete_ade|castel_vic}@gva.es

Abstract. Data management and analysis for risk assessment of rare and complex diseases such as *Neuroblastoma* require efficient management of multidisciplinary data. Recent advances in genomic testing are revealing new publicly available data whose storage and analysis with clinical and genomic data is becoming a big challenge. The use of *Conceptual Modeling* (CM) techniques helps to define and structure the *Neuroblastoma* domain, which serves as a basis to determine the information required for diagnosing the disease. It is important to highlight that a *Genomic Information System* (GeIS) based on a conceptual model allows improving the adaptation of new requirements of the domain, and greatly simplifies the integration and management of heterogeneous and homogeneous data. The main objectives of this work are: i) *to present a Conceptual Model of Neuroblastoma (CMN)*, which defines all elements involved in the clinical and genomic domain. ii) *to apply the SILE method*, in order to obtain all (*clinically*) relevant variations associated with *Neuroblastoma* from genomic data sources. The developed GeIS is intended to make the correct exploitation of the validated data set to provide an early and efficient risk assessment for patients with *Neuroblastoma*.

Keywords: Neuroblastoma, CM, GeIS, CMN, SILE Method, PM

1 Introduction

Since the first complete human genome was published in 2003, there has been an astonishing progress regarding speed and cost. This task took 13 years and an approximate expense of $3 billion [1]. With the barrier of $1000 per genome already broken, data acquirement is no longer a challenge for *Next-Generation Sequencing* (NGS) technologies. The storing, managing and making sense of the data has become the main issue since it requires a broad spectrum of specialists including *IT*, *computational biologists*, *genetic counsellors* or *pathologists* [2]. Once the raw sequence data is obtained, it is aligned to the reference sequence. According to a recent study performed on over 2,500 individuals, everyone differs at 4.1 to 5 million sites from that reference genome [3]. The combination of these genetic variations (or variants), known as "*genotype*", together with environmental factors, determines its host physical traits (known as "*phenotype*"). The main goal of genomic medicine is to understand the genotype-phenotype relationships. Considering that the reference sequence does not codify for any serious condition (*which is not clear is completely true* [4]), the genetic driver of a disease can be found among the group of differences between a patient's genome and the reference one. Therefore, *Whole Genome Sequencing* also known as WGS, may assist in differential diagnosis [5]. Intensive research has aimed to reveal associations between genetic variations and diseases through *Genome-Wide Association Studies* (GWAS) [6]. These associations are especially interesting in rare diseases in which hereditary studies are hindered by the high lethality in early childhood. Representing more than 7% of all pediatric cancers and causing around 15% oncology deaths, rarity and early lethality unfortunately fit with Neuroblastoma's definition.

Neuroblastoma is the most common extra-cranial solid tumor in childhood. It usually appears within the abdomen, neck, chest or pelvis, and its symptoms depend on the tumor's location [7]. The variability of clinical presentation and likelihood of cure are examples of one of the most remarkable hallmarks of Neuroblastoma: its clinical heterogeneity. Although some tumors undergo spontaneous regression, others progress despite aggressive therapy [8]. This disparate clinical behavior has been shown to be related to biological factors such as age at diagnosis, tumor histology or genetic aberrations. For many years, research groups have used different factors in order to stratify patients for risk-based clinical trials. These differences prevented results from being compared. To ease that analysis, several international efforts have resulted in standards for patient classification and staging, such as the *International Neuroblastoma Staging System* (INSS) [9], *International Neuroblastoma Risk Group Staging System* (INRGSS) [10] and *International Neuroblastoma Risk Group* (INRG) classification System [11].

The INSS consists of six stages (1, 2a, 2b, 3, 4 and 4S) according to the degree of surgical resection. Patients are assigned a stage post-surgery. Since that fact makes INSS not suitable for pre-treatment risk-based classification, a new INRGSS was defined. In contrast to INSS, INRGSS classifies patients in four stages (L1, L2, M and MS) [9] depending on clinical criteria and image-defined risk factors. The INRGSS together with other genetic features as tumor ploidy, chromosome 11q aberration, and "*v-myc avian myelocytomatosis viral oncogene neuroblastoma derived homolog*" (MYCN) amplification are considered when assigning patients to its corresponding INRG Classification System risk group.

The variations in the number and size of chromosomes as well as in the number of MYCN repetitions are not the only genetic aberrations associated with Neuroblastoma development and outcome. Also, smaller variations like substitutions in genes such as “*anaplastic lymphoma kinase*” (ALK) [12], “*pairedlike homeobox 2b*” (PHOX2B) [13] and “*kinesin family member 1B*” (KIF1B) [14] are shown to be related to the hereditary disease. Similarly, variations in regions between genes have been found to affect Neuroblastoma progression. Moreover, they have a cumulative effect allowing to assign patients to different risk-based groups, depending on the type of variations they carry. These variations are normally presented in scientific publications. Although many of them remain locked in unstructured sources, some researchers have focused on targeting genetic variations in order to transfer them into curated clinical databases [15]. Some of these text mining efforts have successfully located genetic variations within papers and loaded into databases (e.g., *ClinVar*[1]) with clinically relevant information. Scientists also have the option to directly load the variations they find in these databases or many others with different backgrounds.

ClinVar and *Human Gene Mutation Database* (HGMD, *www.hgmd.cf.ac.uk/*) are examples of clinical repositories gathering information about the effect that genetic variations have in humans, independently of the population or possible treatments for patients who carry them. Drugs targeting specific variations can be obtained by browsing the *Comparative Toxic genomics Database* (CTD, *ctdbase.org/*). On the other hand, the variation frequencies in different populations can be found in *dbSNP*[2], together with a small portion of the sequence containing the variation. The complete sequence of genes is stored in databases such as *Nucleotide*[3] or *Ensembl (https://www.ensembl.org/)*. These archives sometimes contain many sequences for a single gene. To facilitate study comparisons a *Reference Sequence* (RefSeq, *https://www.ncbi.nlm.nih.gov/refseq/*) database was created which stores a unique sequence per gene. This is only a small selection of databases showing the existing heterogeneity among the so-called “*genomic chaos*”. As seen above, valuable information in the decision-making process following a genetic test is dispersed over several sources. The same variation can be stored in different databases, which sometimes offers the same information. That fact leads in the best of the cases to *redundancy*. The data does not always match though; *incongruities* arise in these cases which can compromise the patient’s response to the prescribed treatment [16].

The main objective of the research is to create a conceptual model of Neuroblastoma which will integrate clinical and genomic data. Applying CM on the genomic domain is highly helpful since it allows to define it accurately, hence creating a robust structure from which data can be efficiently managed. According to Olivé [17] the activity which ultimately defines the general knowledge on which an *Information System* (IS) works is what we know as CM. Since an IS

[1] *https://www.ncbi.nlm.nih.gov/clinvar/*

[2] *https://www.ncbi.nlm.nih.gov/snp/*

[3] *https://www.ncbi.nlm.nih.gov/nucleotide/*

built upon a non-described knowledge is considered to be unpredictable, the obtaining of such a description is the main aim of CM. Once a genomic conceptual model is created, it is possible to design *Genomic Information Systems* (GeIS) [18], which play a key role in the biomedical domain. Taking into consideration the *heterogeneity*, *dispersion* and *redundancy* which characterize the genomic domain, the use of conceptual models structuring its basic features is of great use in the way towards *Precision Medicine* (PM) [19]. Personalizing the treatment, depending on the patient's genome and environment, will only be possible if there is a clear definition of the domain and a robust GeIS able to integrate and manage multidisciplinary data.

To achieve our goal following this research line, in Section 2 we firstly give a brief view of the current state of CM in the biological domain, as well as the contemporary condition of Neuroblastoma research. A Conceptual Model of Neuroblastoma (CMN) and an online tool built on it are introduced in Section 3. Section 4 shows the process of searching and identifying relevant genetic variations affecting Neuroblastoma development, and how to load them into our database, which will allow data exploitation. Finally, the conclusions and future works are presented in Section 5.

2 State of the Art

We study the state of the art in two different fields. Firstly, we describe how CM is applied to the genomic domain in order to provide deep knowledge to generate the CMN, and design and develop a system adapted to the needs of Pediatric Oncology department of the *Hospital Universitari i Politècnic La Fe (HUP/IIS La Fe)*. Secondly, we study the Neuroblastoma's domain and the available databases in use by the hospital. This was an important step of the research in order to figure it out how to create a conceptual model adapted to the needs of the department and additionally, implement advanced data management techniques in order to improve decision making processes.

2.1 Conceptual Modeling in Biology

Despite the large amount of publicly available genomic data repositories, it is not usual to find stable conceptual models underlying them. Since most of the accessible archives require storage improvements, CM in the biological domain has been used over the years. Initial proposals in the field were introduced by Paton [20], who presented several models defining cellular genetic components, products and their interactions. The work developed by Ram [21] was focused on the protein context and proved conceptual modeling usefulness in searching and comparing large volumes of 3D structural data. CM is not only an approach to describe and represent a specific domain but also aids in software production [34]. The models have been already used in bioinformatics. By applying this method Garwood et al. [22] created user interfaces for querying biological data repositories. Considering data archives designed so that variations associated to diseases are stored, a proper data management may assist in generating a diagnosis and recommending personalized treatments. Keeping in mind the principles of PM, this research group[4] developed GeIS in order to manage some diseases like *usher syndrome* [23], *breast cancer* [24], *alcohol sensitivity* [18], among others; based on its respective domain conceptualization. More generally, with the aim of creating a standardized vocabulary and unequivocally defining the genomic domain, the *Conceptual Model of the Human Genome* (CMHG) [25-26] was created representing all relevant information in the genomic domain as a whole. This model led to the development of a *Human Genome Database* (HGDB) [26], which can be exploited by using "*GenesLove.Me*" (*see more information in* [27]), an internally matured tool believed to be of great use in genetic diagnostics.

The use of CM is gaining momentum as a software development approach in the medical domain since it greatly improves *geneticists'*, *lab scientists'* and *physicians'* work. Through the creation of GeIS, diagnostic tools can be developed to allow the creation of accurate diagnosis once the right information is identified and collected.

2.2 Neuroblastoma

Given the relative rarity of Neuroblastoma, small volumes of data are readily available. Its correct management is then paramount in order to properly understand the molecular basis of the disease. To make Neuroblastoma data widely available and thus facilitate international, multi-disciplinary research, an INRG database [28] was created with the specific aims of enabling complex queries and linking their results to biological databanks and publicly available datasets. In 2013, after two years of work and when the most recent progress report was published, this database contained a total of thirty-four biological metrics on approximately 11,000 patients, including tumor stage, MYCN status, tissue availability, race, ethnicity or site of relapse [28]. Since several common SNP (*Single Nucleotide Polymorphism*) alleles have been demonstrated to be involved in Neuroblastoma tumorigenicity [29-30] genomic data stored in the *database of Genotypes and Phenotypes* (dbGaP, *https://www.ncbi.nlm.nih.gov/gap/*) the NGS and TARGET[5] data were planned to be added to INRG database.

The INRG international collaboration intends to compensate the relative lack of research focused on Neuroblastoma. This shortage was noticed after exhaustive

[4] Research Center on Software Production Methods (PROS), *http://www.pros.webs.upv.es*

[5] Therapeutically Applicable Research to Generate Effective Treatments, *https://ocg.cancer.gov/programs/target*

searches for scientific information prior to the project in collaboration with the *Pediatric Oncology Unit of HUP/IIS La Fe (http://www.hospital-lafe.com/)* in Valencia (*Spain*). Nowadays, the information related to Neuroblastoma available in databases is mainly clinical. The genetic tests performed on tumor cells in order to diagnose the disease analyses only a few variations for which there is a known treatment. This catalogue of druggable genetic variations is growing as new therapeutic targets are discovered and innovative genetic tests are developed. The information storage and analysis are becoming a challenge since it needs to meet the new requirements. This work is an effort to unify, in a single database, clinical and genetic information with the aim of assessing Neuroblastoma patient's risk. Building a GeIS on a database gathering interdisciplinary information will allow unveiling genetic patterns within Neuroblastoma cohorts. Together with recent technological advances in molecular testing, and also in the computer sciences (e.g., applying *Data Sciences*, *Artificial Intelligence*, and *Big Data* techniques), could help and improve our understanding about genetic basis of the disease, and would ultimately lead to an efficient diagnosis. A precise diagnosis will in turn assist in the process of developing more effective targeted therapies.

3 Domain Definition

The CM and ontological characterization of basic features making up a specific environment assure consistency, correction and an efficient exploitation of its specialized datasets. This is especially true for those repositories designed to store complex data coming from heterogeneous sources.

Neuroblastoma represents an excellent example of a disease whose treatment choice, application and traceability require advanced technologies able to gather and manage data coming from many different fields, ranging from histological tests to nuclear medicine. The users of these technologies fit quite different profiles encompassing pathologists, geneticists, lab assistants or bioinformaticians. A precise definition of the domain is highly advantageous in order to build applications able to store valuable information for each of them separately and, most important, to obtain worthwhile conclusions from its aggregated analysis. For that reason, the CMN (Fig. 1) was created in collaboration with experts of the *Clinical and Translational Cancer Research Group (GICT-Cancer) from HUP/IIS La Fe*. This CMN is the result of iterative meetings and discussions with the experts of the Research Group. Through the discussions and analyzes carried out, the CMN was generated, which facilitates the understanding of the domain of Pediatric Oncology.

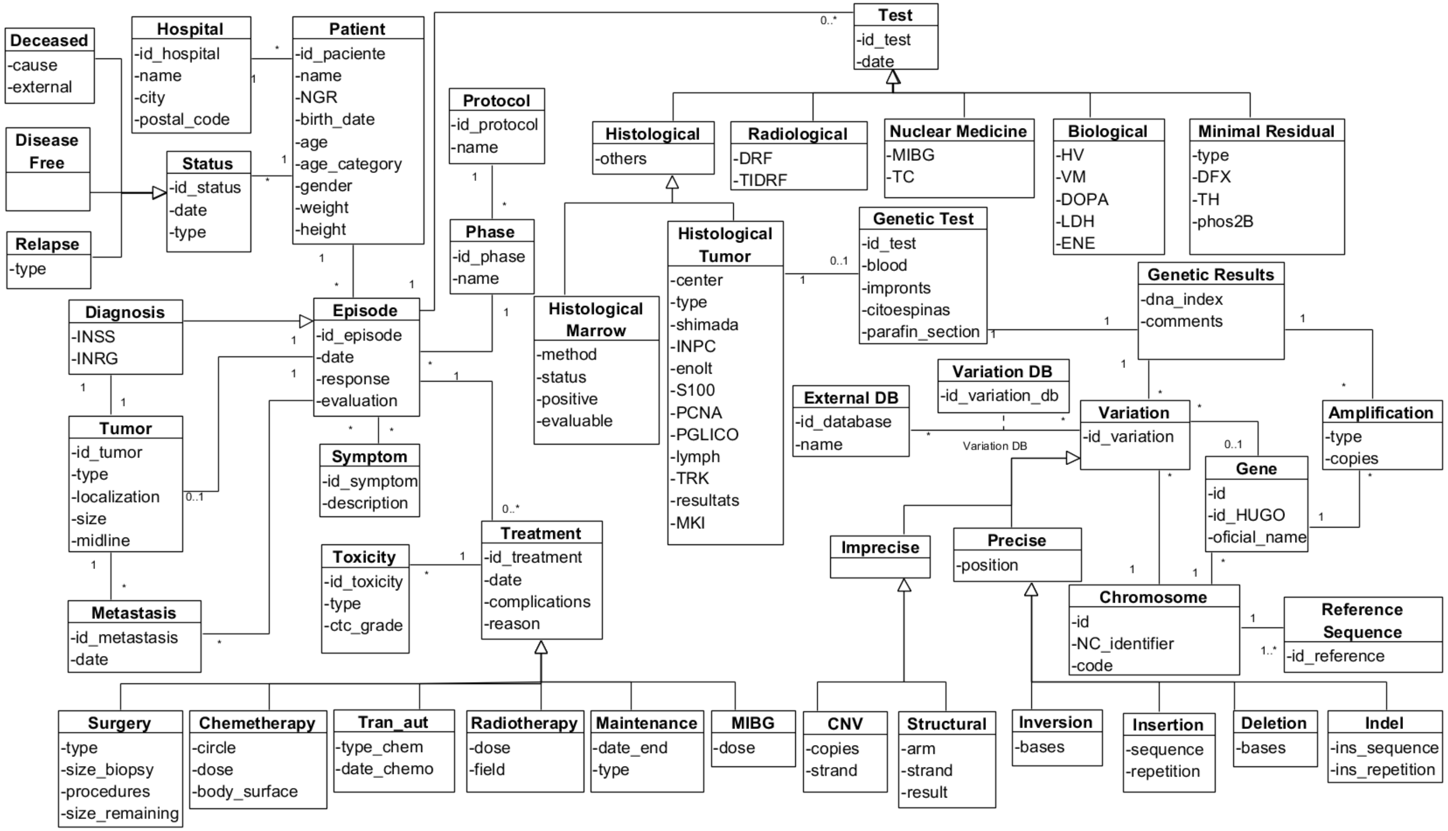


Fig. 1. The Conceptual Model of Neuroblastoma (CMN)

Next, the description of the CMN is presented, which contains all the elements of the previously defined domain. In the model the “*Patient*” class is characterized by attributes such as “*name*”, “*NRG*”, “*date of birth*” and “*date of diagnosis*”, “*age*” or “*weight*” and could be considered the center of the model. This class represents information about individuals attending a specific healthcare institution, represented in the model by the “*Hospital*” class. To start with, the patient is assigned a “*Status*” class, which is divided into “*Deceased*”, “*Relapse*” and “*Disease Free*” classes respectively. Given the Neuroblastoma heterogeneity, different diagnostic tests are performed depending on biological and genetic factors. Several standard procedures regulating tests were developed, which are represented by “*Protocol*” class. The disease development stages are also dependent on the protocol the patient is assigned to. They are represented by “*Phase*” class, whose intersection with patients in a specific phase results in the “*Episode*” class. The former class refers to tests performed, diagnosis arising from the tests and the treatment applied to the patient. One or several tests can be carried out during a single episode of the disease so its progression can be tracked. Since several tests coexist, with both common and unique attributes, “*Test*” class is divided into “*Histological*”, “*Radiological*”, “*Nuclear Medicine*”, “*Biological*” and “*Minimal Residual Disease*”. The “*Histological*” class has two classes which inherit from it. These classes are “*Histological Marrow*” class and “*Histological Tumor*” class, depending on the analyzed tissue.

The tumor tissue can be used for genetic analysis, represented in the model by “*Genetic Test*” class. The results from these tests are referred by “*Genetic Result*”. All the previous analysis is performed in order to characterize the “*Tumor*” class, whose size, location and possible dissemination (i.e., “*Metastasis*”) are recorded. The tumor features together with other significant test results assist in generating the “*Diagnosis*”, which ultimately determines the prescribed “*Treatment*”. Since there are several treatment options, this entity is specialized in the classes “*Chemotherapy*”, “*Surgery*”, “*Transplant Aut*”, “*Radiotherapy*”, “*MIBG*” and “*Maintenance*”. Some of these treatments involve toxic substances whose effects need to be tracked. For that reason, the “*Toxicity*” class was created which refers to the toxicity level. The “*Genetic Results*” class has a relationship with the “*Variation*” which represents every genomic variation in that patient. If those variations have been reported in scientific literature and are stored in a public genomic database, a relation will be defined with “*External DB*” class.

A “*Variation*” is related with the "*Chromosome*" in which it is located, and also it could be located inside a “*Gene*”. The relationship between the “*Gene*” and “*Chromosome*” classes shows the chromosome where the gene is located. The “*Variation*” class is divided into “*Precise*” and “*Imprecise*” classes to specify if the position of the mutation is known. In particular “*Deletion*”, “*Insertion*” and “*Inversion*” classes show the position of the deletion, inversion and insertion of nucleotide sequence in the DNA sequence of the chromosome while the “*Indel*” class represents consistent variations in insertions and deletions at the same time in the DNA sequence of the chromosome. In contrast to “*Precise*” class the “*Imprecise*” does not contain the position as it is unknown. This class could be specialized in 2 subclasses to define two kinds of imprecise variations, such as changes in the number of copies of certain DNA fragments (also called *Copy Number Variation*) - "*CNV*" class – and big changes affecting the structure of the chromosome – "*Structural*" class. Besides that, the "*Amplification*" class represents an increase in the expression of a specific "*Gene*", which is represented by a relationship between these two classes.

This model allows non-experts to understand the environment and easily use tools built on it. In particular, it was used as a basis to design an online tool meant for the management of clinical and genomic data collected from patients with Neuroblastoma. Furthermore, its structure does not limit its extension but leaves it open to evolution by adding possible diagnostic tests, drug targets or innovative treatments.

Several technologies were used in the development process of the web-based tool (*prototype*) aimed for Neuroblastoma clinical and genomic data management. Among these technologies we can highlight, *Java (https://www.java.com/en/)* using *Java Persistence API* [6] for the database loading process. *JavaScript (https://www.javascript.com/)* was used in the development of the tool itself. Specifically, the backend was designed with *Node.js (https://nodejs.org/en/)*, *Jade (https://jade.tilab.com/)* while *Bootstrap (https://getbootstrap.com/), jQuery (https://jquery.com/)* were used for frontend development.

An analysis of the available data storage technologies was carried out in order to determine the most appropriate one based on the type, quantity and subsequent use of the data to be stored. In this stage, the relational SQL technology was selected. However, in the future in case of the expansion of the data and according to the needs of the doctors the data could be migrated to NoSQL technologies (e.g. MongoDB or Neo4j among others).

This prototype is based on the conceptual model mentioned above and currently is being tested by the experts of the *GICT-Cancer from HUP/IIS La Fe*. In this stage, the prototype is being checked for errors, bugs with the aim of improving the prototype according to the needs of the doctors.

4 Applying the SILE Method to Neuroblastoma

In order to assess Neuroblastoma risk from NGS data, the SILE (*Search-Identification-Load-Exploitation*) method was followed [31]. Its main goal is to systematize the search and identification of genomic information to be loaded, analyzed and exploited by a GeIS based on the CMHG [26]. A summary of the activities taking place at each level of the method is defined in Table 1.

As previously stated, Neuroblastoma has been associated to/with a wide variety of genetic variations by means of different study methods. Since risk assessment in Neuroblastoma does not only rely on biological issues but also on clinical and

[6] *https://www.oracle.com/technetwork/java/javaee/tech/persistence-jsp-140049.html*

genetic features, the SILE method [35] represents a good choice to unify and efficiently use all the available information.

Table 1. Description of each level of the SILE method [31]

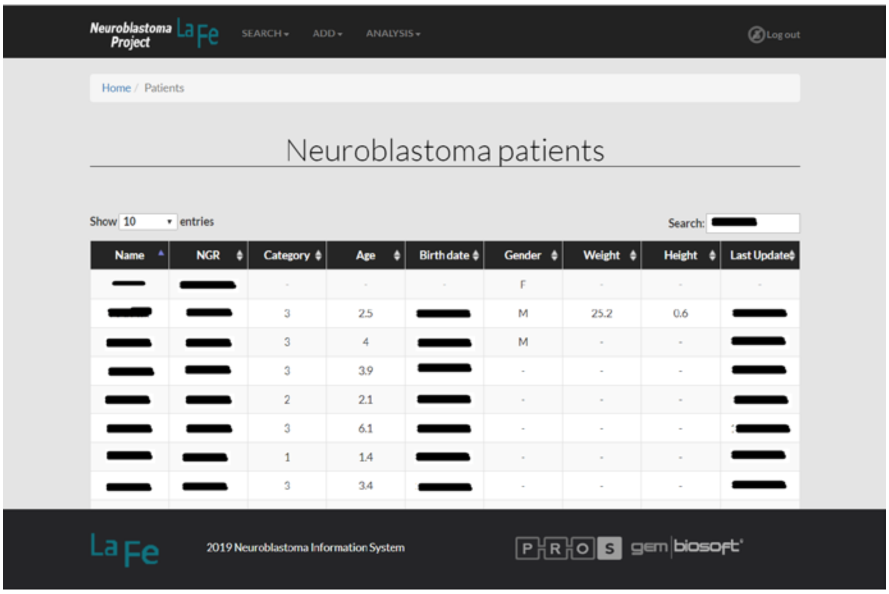
Level	Description
(S) Search	Determination of the information context, required to solve a concrete need, as well as the selection of data source from which to extract information
(I) Identification	Determination of a reliable and relevant dataset to be used to populate a database which structure is delimited by the CSHG
(L) Load	Population of the database with the data identified in the previous level
(E) Exploitation	Extraction of knowledge form the database by using tools to analyse and interpret genomic data

4.1 Results obtained from the SILE method

- ***Search.*** An exhaustive research on existing integrative databases was carried out in order to study other group's strategies (e.g., where to obtain variations from and how to properly annotate them). DisGeNET (*www.disgenet.org/*) was the major finding of this intensive search. Its creators define it as a discovery platform which integrates information on gene-disease associations (GDAs) from public data archives and the literature. Interestingly, DisGeNET classifies data as "*Curated*", "*Predicted*" and "*All*" depending on the original source it comes from. We also have used *ClinVar* and *dbGaP* genomic repositories to find that variations [32].
- ***Identification***. In this stage, the data which was collected from different data sources presented in the first stage of the SILE method, was analyzed in order to remove possible redundancies and other quality issues. The main aim of this step is to prepare the data for loading into the database. At the beginning there were 996 variations collected from different databases. After the Identification stage the complete validated dataset consisted of 375 clinically relevant variations annotated in order to allow a GeIS to efficiently locate and display the data [32-33]. Furthermore, *search* and *identification* processes allowed us to spot several data quality issues such as redundancy or inconsistency which must be solved before loading process.
- ***Load.*** A Database of Neuroblastoma (DBN) was developed in our research group in order to efficiently store the clinical and genomic data. It was based on the CMN mentioned above (Fig. 1). Both the DBN and CMN were developed with the aim of defining all features related to *research*, *diagnose* and *treatment* of Neuroblastoma and creating a strong structure on which a GeIS has been developed. Based on the GeIS previous studies, Neuroblastoma data was loaded into the DBN so that it

could be exploited using an online tool, the GeIS, which is presented in the "*Exploitation*" section.

- ***Exploitation.*** The exploitation will be carried out through the prototype (Fig. 2) developed for the *GICT-Cancer from HUP/IIS La Fe*. After the testing of the prototype by the experts it will be available for clinicians to use. The exploitation of the prototype will ultimately lead to a genetic risk assessment based on validated evidences available on public resources.



Neuroblastoma patients

Show 10 ▾ entries

Search: [redacted]

Name	NGR	Category	Age	Birth date	Gender	Weight	Height	Last Update
[redacted]	[redacted]	-	-	-	F	-	-	-
[redacted]	[redacted]	3	2.5	[redacted]	M	25.2	0.6	[redacted]
[redacted]	[redacted]	3	4	[redacted]	M	-	-	[redacted]
[redacted]	[redacted]	3	3.9	[redacted]	-	-	-	[redacted]
[redacted]	[redacted]	2	2.1	[redacted]	-	-	-	[redacted]
[redacted]	[redacted]	3	6.1	[redacted]	-	-	-	[redacted]
[redacted]	[redacted]	1	1.4	[redacted]	-	-	-	[redacted]
[redacted]	[redacted]	3	3.4	[redacted]	-	-	-	[redacted]

Fig. 2. The prototype developed to integrate and analyse clinical and genomic data

The design and the development of the prototype are based on emphasizing the *efficiency*, *usability*, and *security* aspects (e.g., the prototype will take into account all the current regulations of *Data Protection Law*[7]). A Web-based approach is the selected strategy to design the application as it has some well-reported advantages for the considered working domain. The web-based software gives the user flexible access to the all necessary tools or required data.

Although Model-Driven Engineering (MDE) tools can generate software code automatically by using the model as an artifact we used model-based archetypes [36] to specify user interaction strategies. Archetypes represent different views offered to the users. In order to design the interface of the application, some archetypes have been created based on the CM.

The prototype stores and manages patient's demographic information, episode description, complementary information, treatments, pathological and genomic

[7] *https://www.boe.es/buscar/doc.php?id=BOE-A-2018-16673*

information. Additionally, the prototype has an analysis section where the clinicians can define certain queries and get information from the GeIS.

5 Conclusions and Further Work

The main result of this work was the definition of the CMN, with the aim of making the disease widely understandable for any personal profile involved in its *diagnosis*, *treatment* and *traceability* process. The thorough analysis of the domain allowed us to spot which information was valuable for the disease diagnosis, its availability and current challenges regarding its management. In order to overcome the contemporary difficulties, a GeIS based on the CMN was built which was capable of efficiently managing Neuroblastoma-related clinical and genetic data. The implementation of the SILE method allowed to define a validated dataset of variations associated with Neuroblastoma, this group of variations was selectively loaded into our DBN, and finally the exploitation was carried out through the developed prototype, which is currently being tested.

The construction of the CMN has been proved to be highly convenient in assisting the development of the GeIS since it provides a robust structure supporting a correct data organization. Besides that, the model is also flexible in the way it could easily adapt to new diagnostic tests or treatment strategies, hence growing as Neuroblastoma related knowledge does. Although only curated databases have been browsed in order to obtain the validated set of variations, other databases might be useful for doctors in their decision-making process.

Following this research line, the CMN will be extended by characterizing new "*concepts*" or "*knowledge*" involved in the domain and adapting the developed prototype according to the new requirements. This study can serve as a basis or prototype for the risk assessment of many other genetic conditions. In the future it is foreseen to provide an IS that can be used in the *Hospitals of the Community of Valencia* (and even in the country) where all the studies and analyzes on Neuroblastoma can help researchers to continue advancing in the treatment and monitoring of this disease. It is important to highlight that the use of CM is very beneficial and efficient for the construction of GeISs, providing a strong structure for the data they are meant to manage, and capable of adapting heterogeneous, disperse, redundant and changing environment it ultimately defines.

Acknowledgment

The authors would like to thank the members of the PROS Research Center Genome group for the fruitful discussions regarding the application of CM in the medicine field. In addition, we would like to thank Dra. Desireé Ramal, Dra. Vanessa Segura, Dra. Blanca Martínez, and Dra. Yania Yáñez as experts in Neuroblastoma for their contribution to this research. This work was supported by the Spanish Ministry of Science and Innovation through Project *DataME* (ref: TIN2016-80811-P), the *Preparatory Action* - UPV / IIS La Fe (C07-ClinGenNBL,

2018) and the Generalitat Valenciana through project *GISPRO* (PROMETEO/2018/176).

References

1. Van Dijk, E.L. et al.: "Ten years of next-generation sequencing technology", Trends Genet., vol. 30, pp. 418-426 (2014).
2. Mardis, E.R.: "The $1,000 genome, the $100,000 analysis?", Genome Med., DOI: 10.1186/gm205 (2010).
3. Auton, A., Abecasis, G.: The 1000 Genomes Project Consortium, "A global reference for human genetic variation", Nature, vol. 526, pp. 68-74 (2015).
4. Chen, R., Butte, A.J.: "The reference human genome demonstrates high risk of type 1 diabetes and other disorders", Pac Symp Biocomput., pp. 231-242, Jan. 2011.
5. Gonzaga-Jauregui C., Lupski J.R., Gibbs R.A.: "Human Genome Sequencing in Health and Disease", Annu Rev Med, vol. 63, pp. 35-61 (2012).
6. Li, X. et al.: "Genome-wide association study identifies four SNPs associated with response to platinum-based neoadjuvant chemotherapy for cervical cancer", Sci Rep, vol. 7, DOI: 10.1038/srep41103 (2017).
7. Maris, J.M.: "Recent advances in neuroblastoma", N Engl J Med., vol. 362(23), pp. 2202-2211 (2010).
8. Cao, Y. et al.: "Research progress of neuroblastoma related gene variations", Oncotarget, vol. 5, DOI: 10.18632/oncotarget.14408 (2016).
9. Monclair, T. et al.: "The International Neuroblastoma Risk Group (INRG) staging system: an INRG Task Force report", J Clin Oncol., vol. 27, pp. 289-297 (2009)
10. Cohn, S.L. et al.: "The International Neuroblastoma Risk Group (INRG) classification system: An INRG Task Force Report", J Clin Oncol., vol. 27, pp. 289-297(2009).
11. Castel, V. et al.: "Prospective evaluation of the International Neuroblastoma Staging System (INSS) and the International Neuroblastoma Response Criteria (INRC) in a multicentre setting", Eur J Cancer., vol. 35, pp. 606-611 (1999).
12. Bourdeaut, F. et al.: "ALK germline mutations in patients with neuroblastoma: a rare and weakly penetrant syndrome", Eur J Hum Genet., vol. 20, pp. 291-297 (2012).
13. Van Limpt, V. et al.: "The Phox2B homeobox gene is mutated in sporadic neuroblastomas", Oncogene, vol. 23, pp. 9280-9288 (2004).
14. Yeh, I-T. et al.: "A germline mutation of the KIF1Bb gene on 1p36 in a family with neural and nonneural tumors", Hum Genet, vol. 124, pp. 279-285 (2008).
15. Jimeno Yepes, A., Verspoor, K.: "Literature mining of genetic variants for curation: quantifying the importance of supplementary material", Database (Oxford). DOI: 10.1093/database/bau003 (2014).
16. León, A. et al.: "Data Quality problems when integrating genomic information". Conceptual Modeling (ER2016): 35th International Conference, 3rd. Workshop Quality of Models and Models of Quality (QMMQ 2016), pp. 173-182 (2016)
17. Olivé, A.: "Conceptual Modeling of Information Systems", 1st ed, Springer-Verlag (2007).
18. Reyes Román, J.F., Pastor, Ó.: Use of GeIS for early diagnosis of alcohol sensitivity. In Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies (pp. 284-289). SCITEPRESS-Science and Technology Publications, Lda (2016).

19. Roden, D.M., Tyndale, R.F., "Genomic Medicine, Precision Medicine, Personalized Medicine: What's in a Name?" Clin Pharmacol Ther., vol. 94, pp. 169-172 (2013).
20. Paton, N. W. et al.: "Conceptual Modelling of Genomic Information", Bioinformatics., vol. 16, pp. 548-557 (2000).
21. Ram, S. Wei, W.: "Modeling the semantics of 3D protein structures", In Conceptual Modeling-ER 2004, Springer Berlin Heidelberg, pp. 696- 708 (2004).
22. Garwood, K. et al.: "Model-driven user interfaces for bioinformatics data resources: regenerating the wheel as an alternative to reinventing it," BMC bioinformatics., vol. 7, p. 532 (2006).
23. Burriel, V. et al.: "Design and Development of an Information System to Manage Clinical Data about Usher Syndrome Based on Conceptual Modeling", BIOTECHNO (2013).
24. Burriel, V., Pastor, Ó.: "Conceptual Schema of Breast Cancer: the background to design an efficient information system to manage data from diagnosis and treatment of breast cancer patients", IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) (2014).
25. Reyes Román, J.F., Pastor, Ó., Casamayor, J.C., Valverde, F.: Applying conceptual modeling to better understand the human genome. In International Conference on Conceptual Modeling (pp. 404-412). Springer, Cham (2016).
26. Reyes Román, J.F.: Diseño y Desarrollo de un Sistema de Información Genómica Basado en un Modelo Conceptual Holístico del Genoma Humano. PhD Thesis, Universitat Politècnica de València. *https://riunet.upv.es/handle/10251/99565* (2018).
27. Reyes Román, J.F., García, A., Rueda, U., Pastor, Ó.: GenesLove.Me 2.0: Improving the Prioritization of Genetic Variations. In: Damiani E., Spanoudakis G., Maciaszek L. (eds) Evaluation of Novel Approaches to Software Engineering. ENASE 2018. Communications in Computer and Information Science, vol 1023. Springer (2019).
28. Cohn, S.L.: "The interactive International Neuroblastoma Risk Group (INRG) database (db) 2nd year progress report" (2013).
29. Maris, J.M. et al.: "A genome-wide association study identifies a susceptibility locus to clinically aggressive neuroblastoma at 6p22", N Engl J Med. (358), pp. 2585-2593 (2008).
30. Capasso, M. et al.: "Replication of GWAS-identified neuroblastoma risk loci strengthens the role of BARD1 and affirms the cumulative effect of genetic variations on disease susceptibility", Carcinogenesis., vol. 34, pp. 605-611 (2013).
31. León, A., Pastor, Ó.: Smart Data for Genomic Information Systems: the SILE Method. Complex Systems Informatics and Modeling Quarterly, (17), pp.1-23 (2018).
32. Burriel, V. et al.: GeIS based on conceptual models for the risk assessment of neuroblastoma. In 2017 11th RCIS (pp. 451-452). IEEE (2017).
33. Soler, C.: Diseño de un sistema de información genómica para el diagnóstico del Neuroblastoma. https://riunet.upv.es/handle/10251/85716, Bachelor degree Project (2017).
34. Pastor, O., Molina, J.C.: Model-driven architecture in practice: a software production environment based on conceptual modeling. Springer Science & Business Media (2007).
35. León, A., Pastor, Ó.: From big data to smart data: a genomic information systems perspective. In 2018 12th RCIS, pp. 1-11. IEEE (2018).
36. Burriel, V.: Diseño y Desarrollo de un Sistema de Información para la Gestión de Información sobre Cáncer de Mama. PhD Thesis, Universitat Politècnica de València. https://riunet.upv.es/handle/10251/86158 (2017).