# *Scientific report*

*on the implementation of the project in May - December 2018*

Project Title: **Next-Generation Non-Hodgkin Lymphoma Registry with an Ontology-Based Approach and Data Fusion**

**Overall objective of the project**. Development of a Non-Hodgkin Lymphoma (NHL) recording method based on combined ontology and data fusion techniques for integrating patient data with administrative and demographic data from multiple sources to create an associated dataset which could be queried in new ways.

**Specific Objective 1**: Develop a complete NHL recording solution capable of providing multisource data integration, semantic consistency, and data integrity in relation to ENCR recommendations and in line with the Cancer Registration Law in Romania.

**Activity 1.1 Obtain and collect information on key data sources by organizing a workshop on data quality assessment and quality verification with invited institutional representatives and important stakeholders from the NW part of Romania.**

The initial stage of the project was devoted to the development of the registration tool. The first step in software solution development was the identification and characterization of future data sources. This was done by organizing a data quality assessment and quality control workshop on November 14, 2018 with the participation of professionals interested in Non-Hodgkin's lymphoma field in the NW region of Romania. The workshop focused on obtaining and collecting information on key data sources: clinical records, hospital discharge data, hematological / oncological (electronic) ambulatory medical records, investigating laboratories (pathology, hematology, immunology, genetics and molecular), institutional or organizational databases, administrative databases and existing registers. Another emphasis was put on the assessment of diagnostic and therapy procedures and capacities of medical institutions for the management of NHL patients. We also conducted a questionnaire-based study to assess the reporting capacity of Non-Hodgkin lymphomas in the Northwest region of Romania. The questionnaire was distributed to workshop participants and was also published online on the EuSurvey platform of the European Commission, a platform that provides free online questionnaire hosting and allows the collection and centralization of results. The electronic questionnaire is available at:

*https://ec.europa.eu/eusurvey/runner/NextGenRel*

The agenda of the workshop included:

- The design and structure of a Non-Hodgkin lymphomas specialized registry based on ontology and data fusion;
- Presentation questionnaire on the usefulness of Non-Hodgkin lymphomas specialized registry, evaluating and checking the quality of data sources;
- Computer features of a registry based on ontology and data fusion;
- Free discussions.

The questionnaire contains 19 questions, was anonymous and was structured in 3 sections. In the first section we collected data on age, field of activity and level of knowledge of the medical registries issue (questions 1 to 6). The questions in the second section are dedicated to assessing the interaction of respondents with different types of medical registries (questions 7-10). The third section focused on assessing the perception of the usefulness of a specialized NHL registry (Questions 11-19):

**1. What is your age?**

..... ..

**2. Please select the county where you work:**

a. Bihor
b. Bistrita-Nasaud
c. Cluj
d. Maramures
e. Satu-Mare
f. Salaj
g. Other, please specify ..... .....

**3. What type of organization in which you work?**

a. City Hospital
b. Municipal Hospital
c. County Hospital
d. Ambulatory
e. Regional Institute
f. Higher Education Unit
g. Administrative Unit of Health.
h. Other ..... .....

**4. What is your position?**

a. Primary physician
b. Specialist
c. Resident physician
d. Teacher
e. Researcher
f. Administrative staff
g. Other ..... .....

**5. What is your domain / speciality?**

a. Hematology
b. Pathological anatomy
c. Oncology
d. Epidemiology
e. Laboratory medicine
f. Another medical specialty ..... ...
g. Another domain ..... .....

**6. Please describe the level of interactions with cancer registries:**

a. Active participation in the registration process in a cancer registry
b. I occasionally use the data provided by a cancer registry
c. There is an institutional cancer registry in place in my institution
d. I have very good theoretical knowledge about the topic of medical registries
e. I have medium / low theoretical knowledge about the subject of medical registries
f. I have no knowledge of the subject of medical registries.

**7. If applicable, please describe how you use the cancer registry data:**

a. I use cancer registry information to extract basic epidemiological data
b. I use cancer registry data to conduct studies in my field and related fields
c. I use cancer registry data to make informed decisions in current medical activity
d. Does not apply
e. Other, please specify ..... .....

**8. Please describe access to basic epidemiological information on cancer (incidence, prevalence, mortality) for your geographic region:**

a. I have easy access to this type of information because there is a functional cancer registry in the region that provides data on cancer pathology

b. I have easy access to this type of information because I access the data provided by the regional and national statistical institutions

c. I have difficult access to this type of information because I do not have enough information on how to access this data

d. I have a difficult access to this type of information because there is no functional cancer registry in the region that provides data on cancer pathology

e. I have a difficult access because I consider that insufficient data is collected at regional or national level for the calculation of these indicators

f. Other, please specify ..... ..... .....

**9. Please describe access to basic epidemiological information for the different categories of Non-Hodgkin's lymphoma (NHL) in your geographical area:**

a. I have access to information on the NHL category as a whole, but also to detailed information on the different types of NHL

b. I have access to information on the NHL category as a whole, but NOT to detailed information on the different types of NHL

c. I have no access to information on the NHL category as a whole nor to detailed information on the different types of NHL

d. Other, please specify ..... ..... ....

**10. Please describe the access to detailed epidemiological information (risk factors, response to treatment, evolution) for the different categories of Non-Hodgkin's lymphoma (NHL) in your geographical area:**

a. I have access to this kind of information for personal cases

b. I have access to this type of information for cases in my institution

c. I have access to this type of information for the NHL population in the region

d. I do not have access to this type of information for the NHL population in the region because this type of information is not collected at the population level in the region where I am active

e. I do not have access to this type of information for the population in the NHL region, because I do not know if this type of information is collected at the population level in my region

f. Other situations, please specify ..... ..... .....

**11. Please appreciate the usefulness of implementing a specialized NHL registry:**

a. Very useful

b. Useful

c. Less useful - registration of cases in the general cancer registry is sufficient

d. Useless

e. I have no opinion

**12. Please select the arguments that you consider valid in favour of implementing an NHL registry in your region**:

a. Data on basic epidemiological indicators for different types of NHL are needed

b. I suspect particular developments of some types of NHL in the region in which I operate

c. NHL is a very heterogeneous category of neoplasms requiring detailed information to be studied in a particular region

d. Globally variations in terms of incidence, prevalence, mortality, response to NHL treatment were observed

e. In my opinion, Romania is facing a lack of basic epidemiological information on NHL

f. Other, please specify .....

**13. Please describe the benefits that personal information provided by an NHL registry in your region would bring to you personally**:

a. It will allow me to make an informed decision about the diagnosis / treatment of patients with NHL

b. I will be able to evaluate my personal results compared to those in the region

c. I will be able to use the data to identify particular NHL developments in the region or in my cases

d. I will have access to detailed NHL information for use for research purposes

e. I will be able to identify areas that require intervention in my current personal medical practice

f. Other, please specify ..... ..... .....

**14. Please describe how the information provided by an NHL registry in your region could be used:**

a. Identifying particular developments of certain types of NHL in the population of the region

b. Identify areas where action is required that may have an impact on the incidence, prevalence and evolution of NHL-diagnosed cases

c. The formulation of public health measures in the field of NHL based on valid information

d. Identification of areas of interest for NHL research, particular to the region

e. Selection of NHL cases that may be included in clinical trials
e. Other, please specify ..... ..... .....

**15. Please appreciate the usefulness of standardizing anatomopathological diagnosis for NHL:**

a. Very useful
b. Moderately useful
c. Less useful - should be left to the pathologist
d. Unnecessary / unfeasible
e. I have no opinion

**16. Would you agree to record the data that you have access to for patients diagnosed with NHL in a Web platform available within a NHL specialist registry?**

a. Yes, without reservation
b. Yes, but with security and confidentiality guarantees
c. Basically yes, but I do not have time
d. I prefer to transfer the data on paper
e. No

**17. Please evaluate the percentage of cases diagnosed with NHL in your unit that are confirmed with histopathological diagnosis:**

a. Less than 25%
b. Between 25% and 50%
c. Between 51% and 75%
d. More than 75%
e. I do not know

**18. Please rate the percentage of cases diagnosed with NHL in your unit, whose diagnosis benefits from a second opinion:**

a. Less than 25%
b. Between 25% and 50%
c. Between 51% and 75%
d. More than 75%
e. I do not know

**19. Please rate the percentage of cases diagnosed with NHL in your unit that has a second opinion in your medical unit:**

a. Less than 25%
b. Between 25% and 50%
c. Between 51% and 75%
d. More than 75%
e. I do not know

Until this report was produced, we received 19 completed questionnaires. Their analysis is presented in the Report on Diagnosis, Therapy and Reporting Capacity of Non-Hodgkin Lymphomas in the NW region of Romania, which is available on the project site at:

*http://nwcanportal.iocn.ro/NextGenRel/Raport.pdf*

**Activity 1.2. Analysing the information obtained on key data sources, defining the list of mandatory and basic variables and their formats, defining the ontology for the NHL recording solution, defining the base data set and reporting routes.**

and

**Activity 1.4. Developing the beta version of the software solution (database and user interface)**

The analysis of the data obtained was used to prepare a preliminary list of mandatory and basic variables and to determine their formats. The next step was the development of an ontology-based storage format and recording format tailored to NHL cases. We developed an ontology model and a functional data model using Open Source Protégé 4.3 (http://protege.stanford.edu, Stanford

University) ontological editing software. At this time, a data harmonization algorithm has been developed to merge data from multiple sources to identify duplicate data and to select data of interest. Next steps in developing the NHL Recording Tool will focus on providing a structured but flexible interface to an electronic database that combines the benefits of fixed and predefined data items and adds data elements and catalogues new concepts.

The heterogeneity of NHL made it difficult to record these cases in the general cancer registry (CR). NHLs can be categorized based on several different properties, most of which do NOT resonate with the International Classification of Disease (ICD-10) coding methodology. For example, lymphoproliferative diseases, including chronic lymphocytic leukaemia and some plasma proliferation, although biologically related to NHL, are included in the leukaemia category by general registers. Within the EUROCARE 4 study has been noted that "the evolving classification and poor standardization of data collected on hematological malignancies vitiate the comparison of disease incidence and survival over time and across regions" (1), raising the concern on both clinical and research communities that traditional cancer registration systems are failing to ascertain a significant number of hematological malignancies. Improving the completeness and accuracy of the hematological malignancy registration process, it is also recognized as an important aspect of improving outcomes in hematological cancer.

We identified the main obstacles to collecting data on NHL patients:

- Data sources are very varied;

- Institutions use very different computer systems: sometimes even within the same institution there are systems and databases that store patient information separately, depending on the department and sometimes different classification systems (WHO 2016, ICD-O-3, ICD-10).

There is a need for a common semantic model well defined for customizable analyses and the connection of data with external resources. We have adopted an ontology-based method for integrating clinical and paraclinical diagnostic information from different sources and different electronic platforms into the NHL recording process. We started from identifying the main entities based on existing coding and classification systems (WHO 2016, ICD-O-3, ICD-10) and searching for existing ontologies containing classes representing these entities. In the next step, we propose the selection of the most appropriate ones (by our own criteria) and their extension with vocabulary entities consisting of simple keywords or key phrases (2, 3, 4).

**Taxonomies**:

- Are mostly unique, hierarchical classifications, contained in a subject
- Focus on an *is-a* relationship between classes
- Potential for limited deduction of terms due to the lack of expressiveness of the terms

**Ontologies:**

- Summon taxonomies
- Include cardinal attributes with restricted values
- Unlimited relationships among entities
- Increased potential for deduction due to relational expressivity

Ontology contains classes, properties, object properties, and logical axioms. Ontology covers the following classes:

*Patient*. Properties: gender, date of birth, demographics, date of diagnosis, last control

*The condition of the patient*. Properties: Reference date, age, performance index

*Diagnosis*. Properties: ICD-O-3 / ICD-10, classification, degree, stage, tumour type, structured histopathological data, including genetic and molecular data.

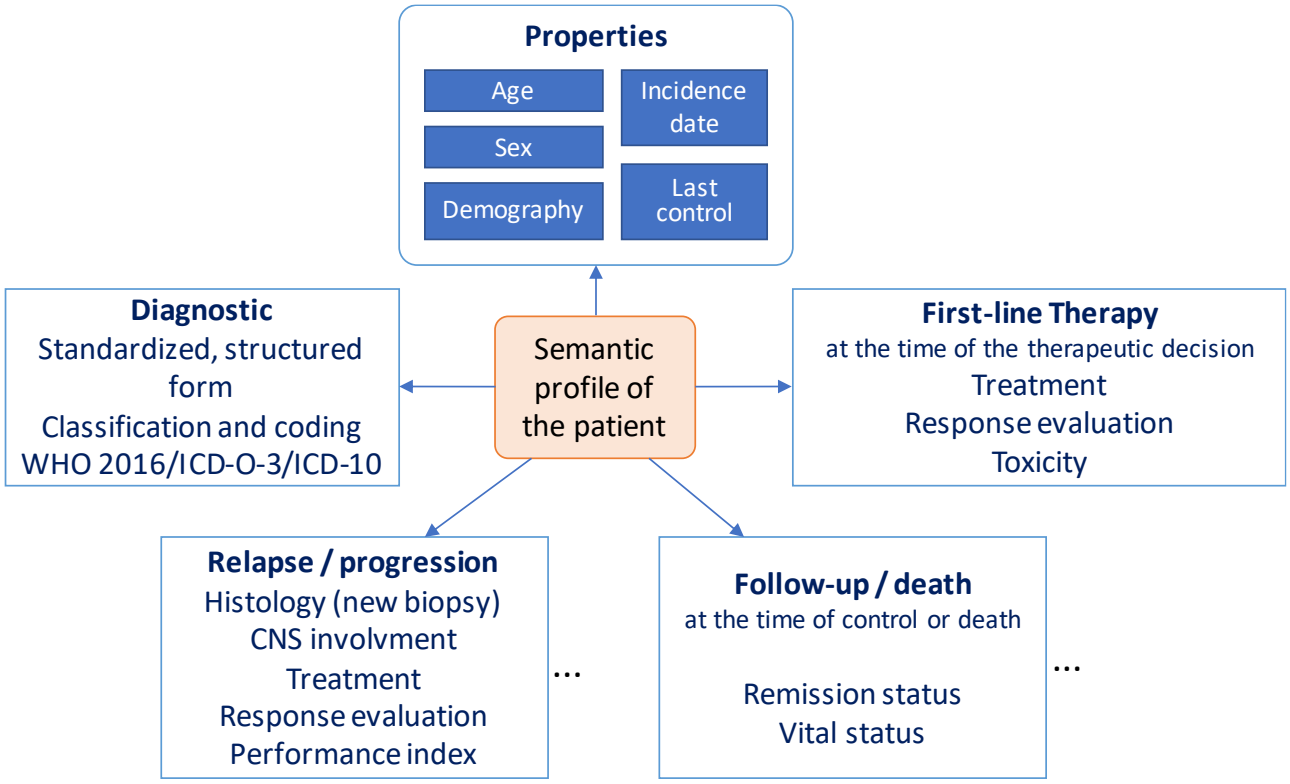
*Treatment*. The different types of therapy have been modelled in ontology as subclasses.

*Evolution of the disease*. Different types, such as complete remission, progression, relapse, were modelled in ontology as subclasses. Properties: Patient condition and date.

Ontology also includes some classes for *extranodal involvement*. These include anatomical entities in a hierarchical structure, e.g. primary tumour, regional lymph nodes and distant metastases.

We designed a semantic model for a regional registry (NW region) specialized for Non-Hodgkin's lymphomas. We aim to develop an automated computer classification and ontology-based NHL patient doubled by a semantic registration platform for the NHL dedicated registry, available to specialist groups, tailored to the needs of users; bringing together the information available on a case-by-case basis in a structured form appropriate to the collection of population data; developing an algorithm to identify NHL based on the diagnosis recorded in the anatomopathological reports.

Diagram of the semantic profile of a Non-Hodgkin's lymphoma:



The basic data set is structured into 4 modules:

**Diagnostic**
Date of diagnosis
Diagnosis according to WHO (2016) / ICD-O-3 / ICD-10
Diagnostic location
Nodal and extranodal involvement
Discordant lymphoma
Ann-Arbor stage
B symptoms
The largest diameter of the tumor
Performance status
Planned treatment
Enrollment in clinical trial
Laboratory Values *
* Hemoglobin, platelets, leucocytes, lymphocytes, albumin, calcium, bilirubin, alanine transaminase, alkaline phosphatase, lactate dehydrogenase, microglobulin beta and immunoglobulin A, G and M

**First-line treatment**
*at the time of the therapeutic decision*
Chemotherapy
Immunotherapy
Radio-immunotherapy
Radiotherapy
Major surgery
High dose therapy with autologous stem cell transplantation
Other treatments
Evaluation of the response
Degree of toxicity CTC III and IV

**Relapse / progression** (sequenced over time)
Relapse date
Histology (new biopsy)
Involvement of CNS at relapse
Treatment
Chemotherapy
Immunotherapy
Radio-immunotherapy
Radiotherapy
Major surgery
High dose therapy with autologous stem cell transplantation
Other treatments
Evaluation of the response
WHO performance status at the time of assessment

**Follow-up / Death** (sequenced over time)
*at the time of control or death*
Date of monitoring / death date
Status of remission
Vital status
Termination of outpatient monitoring

The solution adopted will allow:

- Platform independent and platform-accessible recording system;

- The complex chronology of disease at the patient's level can be clearly represented;

- The semantic structure facilitates aggregate analysis at both patient and population level;

- The system is designed to be adapted over time.

**Activity 1.3. Analysis of the obtained information on the diagnosis, therapy and reporting capacity of NHL cancer in the NW region of Romania.**

The deliverable planned for this activity is the Report on Diagnosis, Therapy and Reporting Capacity of Non-Hodgkin Lymphomas in the NW of Romania, which is available on the project site at *http://nwcanportal.iocn.ro/NextGenRel/Raport.pdf*

**Specific objective 2. Dissemination**

**Activity 1.5. Stage 1 - Participation in an international conference**

The dissemination of the results has begun even at this early stage, by presenting the results obtained under Activity 1.3. at the **19th International Conference of the European Association for Hematopathology**, which took place in Edinburgh on 29.09. - 04.10.2018. I presented a poster titled **Primary extranodal Non-Hodgkin's lymphoma in the North-Western Region of Romania: clinical and pathological features and survival**. The summary published in the Book of Abstracts of the Conference is available at:

*http://www.eahp-sh2018.com/wp-content/uploads/sites/81/Flip/epaper-EAHP2018/#132*

and the poster is available on the project site at:

*http://nwcanportal.iocn.ro/NextGenRel/PosterEAHP.pdf*

We also participated in the **XXVth National Conference of Clinical Hematology and Transfusion Medicine** (with international participation), held in Sinaia on October 10-14, 2018, where we had an oral presentation of the project: **A specialized Non-Hodgkin based on ontology lymphoma registry**, and where we organized an ad-hoc meeting with several specialists interested in the project.

**Specific objective 3. Management**

**Activity 1.6. Monitoring the project and preparing the necessary reports**

Within this activity, we aimed at the coordinated and coherent development of the entire project until reaching the stage objectives with the allocated budget. For this we have developed a monitoring plan for integrating all project implementation activities, risk and cost control, internal and external communication and ensuring results. We also developed the requirements for developing the project website and compiling the necessary documentation for purchasing the IT services provided in the budget. Thus, in September 2018 we launched the project website in a complex and complete structure, in English and Romanian, and at the same time we created the computer platform where the registry will be implemented. The website address of the project is:

*http://nwcanportal.iocn.ro/NextGenRel/Default.aspx*

## Conclusions

Traditional basic information collected by population general cancer registries (age, gender, topography, morphology, behaviour) no longer meets the needs of researchers and decision-makers in the field of hematological oncology. More and more, there is a demand for information on the clinical features of the disease (biomarkers, genetic mutations ...), patient characteristics (comorbidities, risk behaviours ...), detailed treatment and response information, and follow-up (recurrence, progression, death). At this stage we studied and adopted methods of using existing data integration tools to build a database for patients with Non-Hodgkin's lymphoma and create an ontology linking the WHO classification system of lymphoid neoplasms to the most recent revision of 2016, with ICD-O-3 and ICD-10 encoded data from the medical databases and the cancer registry.

The picture of Non-Hodgkin's lymphoma in the Northwest region of Romania shows a large and heterogeneous collection of cancers that reflects the complexity of hematopoietic systems and represents 2.8% of all new cancer cases from 2008 to 2013. Although these occur less frequently than some solid tumours, is a significant burden of the disease in the population. The analysis of the data recorded in the Northwest Regional Cancer Registry (which is a general cancer registry) revealed a significant lack of information on this disease. The analysis of the survival of these patients resulted in differences that cannot be explained by the available data. On the other hand, the study of reporting capacity among hematologists and pathologists highlighted their very encouraging compliance with the reporting of specific data to a Non-Hodgkin lymphoma specialty registry.

Director proiect, Dr. Bogdan FETICA

Mentor, Prof.Dr. Alexandru IRIMIE

## References

1. Sant M, Allemani C, Santaquilani M, et al. EUROCARE-4. Survival of cancer patients diagnosed in 1995-1999. Results and commentary. Eur J Cancer. 2009;45(6):931-91.

2. B Fetica, P Achimas-Cadariu, B Pop, D Dima, L Petrov, AM Perry, BN Nathwani, HK Müller-Hermelink, J Diebold, KA MacLennan, A Fulop, ML Blaga, D Coza, A Irimie, DD Weisenburger. Non-Hodgkin lymphoma in Romania: a single-centre experience. Hematological Oncology, 35: 198–205 (2017). doi: 10.1002/hon.2266

3. B Fetica, B Pop, ML Blaga, A Fulop, D Dima, MT Zdrenghea, CI Vlad, AS Bojan, P Achimas-Cadariu, CI Lisencu, A Irimie, DD Weisenburger. High prevalence of viral hepatitis in a series of splenic marginal zone lymphomas from Romania. Blood Cancer Journal , volume 6, page e498 (2016). DOI: 10.1038/bcj.2016.102

4. Bogdan Pop, Bogdan Fetica, Mihaiela Luminita Blaga, Dan Gheban, Patriciu Achimaş-Cadariu, Catalin Ioan Vlad, Andrei Achimaş-Cadariu. Ontology-Based Search Procedure to Identify Tissue Samples in an Autopsy Archive: A Pilot Study. Applied Medical Informatics , 40(3-4):45-53 (2018).