CREATING A ROMANIAN DATASET ON THYROID CANCER AND RADIOIODINE DOSAGE: CLINICAL IMPACT AND RESEARCH PERSPECTIVES

Irina-Oana LIXANDRU-PETRE[1,2],
Alexandru DIMA[2,3], Gratiela GRADISTEANU-PIRCALABIORU[4,5],
Florina ILIESCU[1], Dana Cristina TERZEA[6], Andrei GOLDSTEIN[6],
Mihai DASCALU[2,4,7,] Madalina MUSAT*[4,6,8], Ciprian ILIESCU[2,4]

Rezumat. *Întrucât bazele de date privind cancerul joacă un rol crucial în cercetare, crearea unui set de date privind cancerul tiroidian este esențială pentru dezvoltarea unor strategii moderne de gestionare a tratamentului și pentru îmbunătățirea înțelegerii bolii. Acest articol prezintă metodologia utilizată pentru a crea un set de date de N = 1,556 de probe extrase din dosarele medicale electronice ale pacienților oncologici diagnosticați la Institutul Național de Endocrinologie CI. Parhon între 2022 și 2024. Ulterior, este evidențiat modul în care aceste rezultate pot sprijini progresul medical, facilitând în același timp integrarea instrumentelor de învățare automată în endocrinologia clinică.*

Abstract. *As cancer databases play a crucial role in research, creating a thyroid cancer dataset is essential for developing modern treatment management strategies and improving disease understanding. This article presents the methodology used to create a dataset of N = 1,556 samples extracted from electronic medical records of oncology patients diagnosed at the CI. Parhon National Institute of Endocrinology between 2022 and 2024. Subsequently, it highlights how these results can support medical progress while facilitating the integration of Machine Learning tools into clinical endocrinology.*

Keywords: Thyroid Cancer; electronic medical records; dataset creation;

DOI 10.56082/annalsarsciinfo.2025.2.40

[1] Postdoc Researcher, eBio-Hub Research-Center, National University of Science and Technology "Politehnica" Bucharest (irina.petre@upb.ro; florina.iliescu@upb.ro).

[2] Academy of Romanian Scientists, Ilfov 3, 050044 Bucharest, Romania.

[3] PhD, Faculty of Automatic Control and Computer Science, National University of Science and Technology "Politehnica" Bucharest (alexandru.dima@upb.ro).

[4] Prof., Senior Researcher, eBio-Hub Research-Center, National University of Science and Technology "Politehnica" Bucharest (gratiela.gradisteanu@upb.ro; mihai.dascalu@upb.ro; ciprian.iliescu@upb.ro).

[5] Faculty of Biology, Department of Botany and Microbiology, University of Bucharest, Romania;

[6] Physician, C.I. Parhon National Institute of Endocrinology, Bucharest, Romania (madalina.musat@umfcd.ro; dana.terzea@gmail.com; andrei.goldstein@gmail.com).

[7] Prof., Faculty of Automatic Control and Computer Science, National University of Science and Technology "Politehnica" Bucharest.

[8] Assoc Prof., Carol Davila University of Medicine and Pharmacy, Bucharest, Romania.

* Corresponding Author madalina.musat@umfcd.ro.

1. Introduction

According to the United States National Cancer Institute [1], Thyroid Cancer (TC) is one of the cancer types diagnosed with the greatest frequency in the United States, with a total of 44,020 estimated cases and 2,290 deaths in 2025. However, after decades of increase, studies on incidence trends for thyroid cancer have shown a decline of 2% per year since 2014 [1], due to various changes in the disease management, including screening and incidental detection through imaging and integrated machine learning tools, which can detect the tumor in its early stages, without the need for thyroid removal [2]. The basic step in driving these advancements is establishing real-world patient cohorts and conducting extensive statistical and correlation analyses. Highly prevalent in adults over 40 years, especially in females, thyroid carcinoma also accounts for 12% of cancers in adolescents, and 2% of children [3]. The factors that influence the prognosis are mainly divided into: biological factors – the potential for tumor invasion and metastasis, the histological type of the tumor, demographic factors – age and gender of patients, the time factor – the period from the onset of the disease to the correctly established diagnosis, and the radicality of the treatment [4, 5].

The European Society for Medical Oncology (ESMO) Guidelines for thyroid cancer, establishes surgical treatment as the standard protocol - removal of the primary tumor mass (primary therapy), followed by staging and risk assessment in the Tumor-Nodule-Metastasis (TNM) system and histopathologic classification according to the World Health Organization (WHO) criteria [6]. Follow-up vary according to the tumour histotype, initial treatment, initial risk of persistent/recurrent disease and response to primary treatment and consist of assessment of thyroglobulin (Tg)/calcitonin, antithyroglobulin (ATG) antibodies, Carcinoembryonic antigen (CEA), Thyroid-Stimulating Hormone (TSH), neck ultrasound, whole body scintigraphy (WBS), Computed Tomography (CT)/ Magnetic Resonance Imaging (MRI) scans. TSH suppressive therapy and radioiodine (RAI) treatment (for intermediate/high risk follicular cell-derived neoplasms) aim to prevent/treat local/regional relapses and distant metastases in differentiated thyroid cancer, while lifelong case follow-up protocols establish the successive steps. Systemic therapies or a combination of chemo- and radiotherapies may be applied in advanced cancers [7].

Depending on the type of TC, the average prognosis and survival without recurrence is quite high. The 5-year relative survival rate for all cancers showed the highest contemporary survival for cancers of the thyroid (98%), prostate (97%), testis (95%), and melanoma (94%) [1].

The 5th edition of the World Health Organization classification of endocrine tumors, released in 2022, classified thyroid tumor types into benign, low-risk, and malignant neoplasms [6, 8]. The malignant category includes three subtypes: Well-differentiated thyroid carcinoma - DTC (Papillary Thyroid Carcinoma (PTC), Invasive encapsulated follicular variant PTC, Oncocytic Thyroid Carcinoma, and Follicular Thyroid Carcinoma (FTC)), Follicular-derived carcinoma high grade, and Anaplastic follicular cell-derived carcinoma, where the third subtype is the most aggressive [6].

In Romania, TC - the most common malignant tumor of the endocrine pathology, represents a small portion of all cancers diagnosed annually (particularly in women) [9]. According to the Global Cancer Observatory, in 2022, TC represented 1.7% of the new cancer cases in our country (after breast cancer, colorectal, prostate, or lung cancer) [10]. However, the lack of a National Registry of Cancer makes following these cases more difficult. Therefore, collecting a large dataset of TC cases from the renowned National Endocrinology Institute in Bucharest, Romania, would make it feasible and useful to investigate TC outcomes and support clinical case management, as doctors would be able to review, adjust, and compare their decisions based on real data. Statistical analysis of these data could help medical professionals choose more efficient treatments for each individual (in terms of quantities/doses), not to mention the potential to integrate various Machine Learning techniques into the data, for both medical and patient use.

2. Method

The purpose of this article is to present both the clinical impact and research perspectives derived from creating a dataset centered on histological types of differentiated thyroid carcinoma (DTC) and radioactive iodine therapy, using data from all patients hospitalized and monitored at the Romanian National Institute of Endocrinology CI. Parhon (CI. Parhon NIE) from January 2022 to May 2024.

CI. Parhon NIE has its own medical database containing electronic medical records for all patients, including those with TC who were consulted, treated, and followed at the institute. CI. Parhon NIE has also implemented a protocol for the treatment of differentiated thyroid cancer, which is applied across all departments. Since January 2011 (the year the institute began managing the medical documents in the Hipocrate database system) to date, there are over 17,000 open electronic medical records (for any visit to Parhon), totaling over 10,000 unique patients with the C73 code (malignant thyroid tumor) for Diagnostic Relation Groups (DRG), which can be accessed with the consent of the endocrinological institute, ensuring the confidentiality of all information.

Access to medical data was granted after obtaining the ethics approval of the Ethics Committee of the Scientific Council of the CI. Parhon, for a collaborative research project in thyroid cancer, with the broad consent of patients upon admission for 'participation in the education and research process' regarding the use of medical data for research purposes (in compliance with the legislation in force, GDPR), as well as the confidentiality commitment between the authors and the Institute regarding the confidentiality of all data. This thyroid cancer dataset from January 2022 to May 2024 involved manual querying of thousands of medical records, resulting in over 3,000 electronic files containing data on patients diagnosed with thyroid cancer, who were treated according to the CI. Parhon NIE protocol.

In the first phase, we anonymized the patients’ personal data and assigned each case a study code. Personal data of patients with matching codes remained under control at CI. Parhon NIE.

For this analysis, we extracted data related to the patient’s age at diagnosis/surgery, patient’s sex at birth, surgical treatment - removal of the primary tumor mass (primary therapy), location of the main tumor within the thyroid, histopathological diagnosis of the tumor, classification in the TNM system, cancer stage, and other classifications - such as degree of resection, vascular, lymphovascular, perineural invasion, postop levels of Tg, ATG, TSH, postop neck ultrasound and WBS assessment, and also the total and fractionated doses of radioactive iodine administered. Subsequently, from the histopathological reports, we extracted separate relevant information, such as cancer type, cancer subtype, tumor size, tumor location, or other comorbidities. Also, separate columns were created for each tumor characteristic including Primary tumor, Nodules, Metastases, Stage, Resection, Perineural invasion, Vascular invasion, and Lymphatic vessel extension; cases with missing information were noted.

A second phase included standardizing disease stage and TNM according to the 8th TNM Classification of Malignant Tumors [11], whereas the cancer subtypes were extracted from descriptive histopathological reports and classified according to the World Health Organization Classification [12]. This stage required the permanent, mandatory presence of the endocrinologist and the pathologist’s input for the detailed evaluation of each individual characteristic; all stages were completed under the endocrinologist’s supervision. Given the field’s specialized medical nature, without their applied knowledge, we would not have been able to correct and standardize the data to ensure the accuracy of the extracted information.

The challenges of this project were numerous, including inconsistent medical record standards over the years, changes in classification criteria, typographical errors, and the absence of records for certain parameters of interest. All these elements had to be addressed to standardize parameters for efficient data processing - a task performed in strong collaboration with the endocrinologist and pathologist whose medical knowledge was imperative for this step. The integrity and accuracy of the data are vital to the research's precision; verifying the correctness of the manually entered data was essential to meeting the proposed objectives.

Finally, our dataset is based on a total of 1,556 samples representing unique patients diagnosed with PTC, FTC or mixed variants (mixed PTC and FTC, or PTC and MTC (medullary thyroid carcinoma) of thyroid cancer, containing, not only demographic and histological parts, but also information about at least one round of radioactive iodine received, the results of ultrasound/scintigraphy performed after surgery, as well as postoperative hormonal values.

3. Results and Discussion

3.1. Clinical Impact

There are few databases in the world containing data on TC: „The Prostate, Lung, Colorectal, and Ovarian (PLCO) Thyroid dataset(s)" from the National Cancer Institute, with data on TC incidence and mortality analyses [13], the Medullary Thyroid Carcinoma (MTC) Registry Consortium [14], with a list of tracked and diagnosed MTC cases in the last decade, or „The Australian & New Zealand Thyroid Cancer Registry (ANZTCR)" – the official record with information about the diagnosis, treatment, and outcomes of individuals diagnosed with TC [15]. Romania also needs to build medical registries for analysis purposes and even personalized (specific) thyroid cancer treatment. An important step in this direction was made through the creation of the Biomat database [16], which contains data on thyroid cancer patients from the CI. Parhon NIE, Department of Nuclear Medicine. According to the authors, over 10,000 "classic" physical files dating back to 1965 have been digitized, preserving the consistency and integrity of the information. Unfortunately, this database is not publicly available. However, a collaboration with the database developers would be necessary for a more precise and effective analysis.

The clinical impact of performing various statistical analyses on existing parameters lies in the potential to increase efficiency and accuracy in determining precise treatment plans. Based on these data, we can track the total number of cases over a specified time interval, stratify patients by cancer type, subtype, age and sex, and monitor other clinical details relevant to the specialist. Table 1 and Table 2 present the main characteristics of patients diagnosed with TC whose data were used in this analysis.

Table 2. Descriptives on age and sex

Sex	*Count*
F	1233
M	323

Age	*Count*
<55	998
>=55	558

	Age	
Sex	<55	>=55
F	790	443
M	208	115

Table 3. Descriptives on TC type and subtype

Type	*Count*
PTC	1459
FTC	73
Mixed	24

Subtype	*Count*
Papillary microcarcinoma	418
Diffuse sclerosing variant of PTC	328
Papillary thyroid carcinoma (PTC)	316
Follicular variant of PTC	154
Encapsulated variant of PTC	65
Oncocytic variant of PTC	35
Minimally invasive FTC	32
Mixed variant	24
Solid-trabecular variant of PTC	20
Poorly differentiated thyroid carcinoma (PDTC)	18
Follicular thyroid carcinoma (FTC)	16
Encapsulated angioinvasive FTC	5
Folicular variant of PTC; Hurthle cell carcinoma (HCC)	4;4
Clear cell variant of PTC	2
Columnar cell variant of PTC; Cribriform-morular variant of PTC; Hurtle cell tumour	1;1;1

Regarding papillary and follicular cancer stages, patients were divided according to the age of 55 – the critical threshold in TC prognosis and staging -as follows: in patients up to 55 years, the stage of the disease is 1 (if patients do not have metastases) or stage 2 (if patients have metastases), regardless of the size of the tumor and lymph node invasion; in patients aged 55 years or older, the stages can range from 1 to 4, depending on the particularities and links between T, N, and M [11, 12]. The analysis showed that 63.82% of the TC patients were diagnosed with stage 1, 19.22% with stage 2, 0.71% with stage 3, and 1.8% with stage 4 (Figure 1). There were also 14.46% of cases in which the stage could not be determined due to missing values for the dimensions of the primary tumor, the presence of lymph node metastases, or distant metastases.

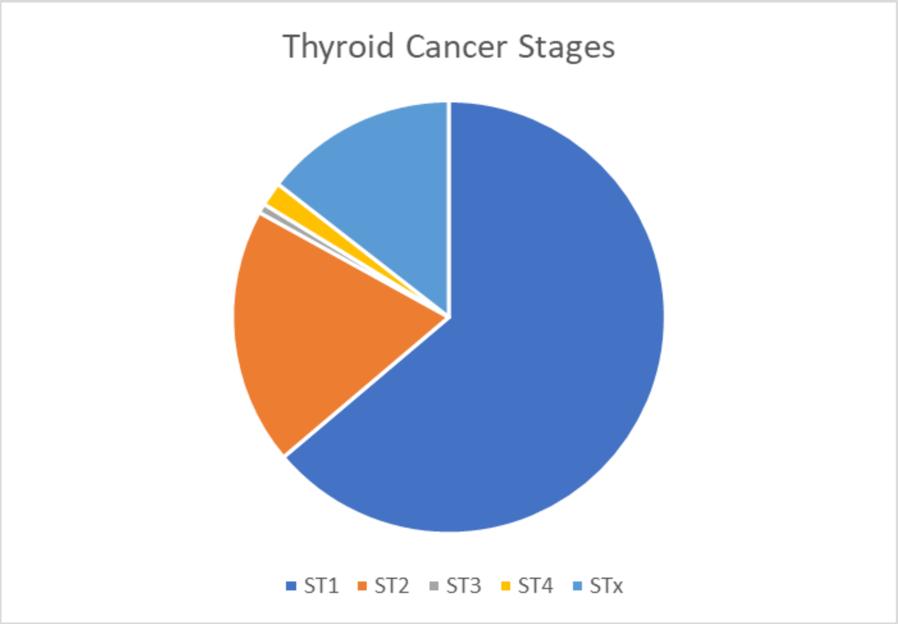

Fig. 6. Distribution of TC stages within the dataset

Relevant information relates to the dose(s) received by patients and their correlation with the type of cancer (Table 3).

Table 4. Descriptives on radioiod doses and radioiod-TC type

rI	**Count**
0-50	569
51-100	398
101-200	305
201-300	120
>300	164

	Type		
rI	**PTC**	**FTC**	**Mixed**
0-50	547	15	7
51-100	363	26	9
101-200	297	5	3
201-300	113	7	1
>300	139	20	4

These cleaned entries from our dataset provide essential information for doctors, enabling them to identify, for example, the optimal radioiodine dose range for a patient of a specific cancer subtype, age, sex, and disease stage. Furthermore, these findings offer valuable insights for long-term follow-up and can help endocrinologists improve precision and efficiency in treating future cases.

3.2. Research Perspectives

As advances in Artificial Intelligence-based methods for phenotypic personalized medicine continue to evolve, they will ultimately improve the efficacy and outcomes of clinical drug combination development and personalized medicine. Machine Learning (ML) holds significant promise in advancing personalized healthcare, which aims to tailor medical treatment and interventions to each patient's individual characteristics (including gender-specific factors), genetics, and lifestyle. In clinical practice, ML tools can help physicians make faster, more accurate diagnostic decisions, avoid unnecessary biopsies by better identifying patients who are truly at risk, and enable more personalized treatment plans tailored to each patient's unique characteristics. For example, predictive models integrated into electronic health records could alert physicians to high-risk patients, allowing for earlier intervention. In postoperative care, recurrence prediction models can guide follow-up intensity, optimize the use of medical resources, and provide patients with greater peace of mind. These applications not only improve the efficiency of clinical workflows but also promote more collaborative and informed decision-making between physicians and patients.

In our case, the analytical results obtained can serve as a foundation for these applications in thyroid cancer management. The research perspective involves exploring information to advance thyroid oncology management and medical data analysis, thereby improving understanding of TC prognosis in an efficient and personalized manner. For example, future research could focus on developing ML-based applications for TC to improve the quality of medical services in oncology by leveraging real historical medical data and correlations. Just as the Quadratic Phenotypic Optimization Platform (QPOP), developed by researchers at the National University of Singapore (NUS), delivers personalized drug combinations to improve patient prognosis [17], a similar software application could be developed to predict the optimal radioactive iodine dose for Romanian patients with thyroid cancer. Undoubtedly, a registry containing comprehensive national data would be optimal for this task, as would expanding the analysis to include new patients and ongoing follow-ups. Until then, definitive first steps in this direction have been taken through data collection process performed in this study, effectively launching the development of Romanian oncological clinical decision support systems.

4. Conclusions

Romania faces a need to build medical registries for in-depth analyses and the prediction of personalized (specific) cancer treatments. An important step in this direction has been the creation of this representative dataset of 1,556 samples from individuals hospitalized and monitored at CI. Parhon NIE between January

2022 and May 2024. Valuable research can be conducted using these data, particularly regarding statistical analyses, correlations, and ML-based applications. Therefore, having this national dataset as a starting point provides significant insights into this disease.

The originality of this project lies in transforming vast amounts of medical information into structured associations suitable for modern data analysis methods and mathematical modeling.

Acknowledgment

The authors acknowledge the European Union's Horizon Europe Research and Innovation framework programme 2021–2027, under the Coordination and Support Actions, HORIZON-WIDERA-2022-TALENTS-01 (grant agreement-101087007–eBio-hub), funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them. The authors also acknowledge support from the grant of the Academy of Romanian Scientists, AOSR-TEAMS-III Edition 2024-2025 "Digital Transformation in Sciences".

REFERENCES

[1] *National Cancer Institute*. Common cancer types. https://www.cancer.gov/types/common-cancers

[2] S. Moon, EK. Lee, H. Choi, SK. Park, YJ Park. *Survival Comparison of Incidentally Found versus Clinically Detected Thyroid Cancers: An Analysis of a Nationwide Cohort Study*. Endocrinol Metab. **38**(1):81-92, 2023, doi: 10.3803/EnM.2023.1668.

[3] BW. Corn BW and DB. Feldman. *Cancer statistics, 2025: A hinge moment for optimism to morph into hope?,* CA: A Cancer Journal for Clinicians, vol. **75**, 2025, https://doi.org/10.3322/caac.21871.

[4] M. Voichita M, M. Simona. *Clasic si modern in cancerul tiroidian diferentiat*. Jurnalul de Chirurgie **6**, 2010.

[5] Tibirna, G. Tibirna, I. Mereuta. *Cancerul glandei tiroide, conform stadializarii noi*. Buletinul Academiei de Stiinte a Moldovei, Stiinte Medicale **64**, 108–134, 2019.

[6] F. Basolo, E. Macerola, AM. Poma, L. Torregrossa. *The 5th edition of WHO classification of tumors of endocrine organs: changes in the diagnosis of follicular-derived thyroid carcinoma*. Endocrine. **80**(3):470-476, 2023, doi: 10.1007/s12020-023-03336-4.

[7] S. Filetti, C. Durante, D. Hartl et al. ESMO Guidelines Committee. *Thyroid cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†*. Ann Oncol **30**(12):1856-1883, 2019, doi: 10.1093/annonc/mdz400.

[8] IA. Cree. *Endocrine and neuroendocrine tumours*; Vol. **10**, WHO classification of tumours series, International Agency for Research on Cancer, 2025.

[9] CM. Buzdug, IC Pinzaru, C. Vulpoi et al. *Clinical and epidemiological profile of thyroid cancers in the north east region of Romania during 2005-2013*. Revista Romana de Anatomie functionala si clinica, macro- si microscopica si de Antropologie, vol. **XVI**, nr. 1, 2017.

[10] *Global Cancer Observatory. Romania*. Romanian facts. https://gco.iarc.who.int/media/globocan/factsheets/populations/642-romania-fact-sheet.pdf.

[11] J.D. Brierley, M.K. Gospodarowicz, C. Wittekind, Wiley Blackwell. *TNM Classification of Malignant Tumours -8th edition*, Union for International Cancer Control, 2017.

[12] Y. Bai, K. Kakudo, CK. Jung. *Updates in the Pathologic Classification of Thyroid Neoplasms: A Review of the World Health Organization Classification*. Endocrinol Metab (Seoul), **35**(4):696-715, 2020, doi: 10.3803/EnM.2020.807.

[13] *National Cancer Institute. Cancer Data Access System*. Thyroid Datasets - PLCO Thyroid datasets. https://cdas.cancer.gov/datasets/plco/8/.

[14] *Medullary Thyroid Carcinoma Registry Consortium (MTC). American Thyroid Association. Optimal Thyroid Health for All.* https://www.thyroid.org/professionals/partner-relations/medullary-thyroid-carcinoma-registry-consortium/.

[15] *Australian & New Zealand Thyroid Cancer Registry*. https://anztcr.org.au/.

[16] M. Purice, M. Radulescu, X. Pirvu et al. *The role of the BIOMAT database in the direction of precision (personalized) medicine: applied mathematical modeling in thyroid cancer*, Conference: Modern management of neuroendocrine disorders in the era of evidence based and precision medicine, Bucharest, 2022.

[17] MBMA. Rashid, TB. Toh, L. Hooi et al., *Optimizing drug combinations against multiple myeloma using a quadratic phenotypic optimization platform (QPOP),* Science Translational Medicine, Vol. **10**, No. 453, 2018, doi: https://doi.org/10.1126/scitranslmed.aan0941.