Technical White Paper: Data Modeling for Colorectal and Lung Cancers

Luxembourg Institute of Health

April 2025

Project Coordinator: NTT Luxembourg PSF S.A.

1. Introduction

The Lux-X Dataspace4Health project, funded by the Ministry of the Economy under grant agreement number 20230505RDI170010392869, aims to develop a comprehensive interoperable data modeling framework for cancer. This technical white paper presents an approach for data modeling, highlighting the key entities, attributes, and relationships involved using colorectal and lung cancers as use cases.

Modern healthcare systems rely heavily on robust data models to manage and integrate patient information across various clinical domains. These models aim to enhance data management, data flow, integration, and analysis capabilities for healthcare providers and researchers. Data modeling processes for oncology cases pose challenges due to the dynamic nature of actors, participants, and the complexity of data.

Our data model is designed based on the oncology data model defined by the German Medical Informatics Initiative (MII) [1] to integrate patient demographics, medical history, treatment outcomes, and follow-up care, enabling advanced treatment and research applications. MII is a 500 million euro project funded by the German Federal Ministry of Education and Research (BMBF), enabling university hospitals, business, health insurers, and patient advocacy groups to collaboratively strengthen medical research and improve patient care. Numerous researchers in the German MII consortia endeavoured developing a comprehensive data model framework including the oncology data module. The data model framework has been validated and implemented across university hospitals in Germany demonstrating its robust design and utility with multiple case studies. Therefore, adopting the MII data model enables us to align the Dataspace4Health project with internationally recognized standards.

This white paper outlines a framework for applying the MII oncology model [2] to develop a standardized data model tailored to the Luxembourg healthcare system. By leveraging a comprehensive set of entities—including Histology, TNM Classification, Additional Classification, Residual Status, Metastases, General Performance, Operation, Radiotherapy, Chemotherapy, Side Effects, Tumor Board, Death, Genetic Variant, Study Participation, and Associated Data Entities—this model aims to enhance clinical decision-making, support outcome research, and facilitate interoperability across clinical and research settings.

2. Colorectal and Lung Cancer in Luxembourg

Colorectal and lung cancer rank among the most common and foremost causes of cancer deaths in Luxembourg [3]. According to recent data from the Global Cancer Observatory and OECD country profiles, colorectal and lung cancers are among the leading cancers in the nation [4], and in 2021

alone, lung cancer was the leading cause of death, accounting for 20% of cancer deaths, followed by colorectal cancer accounting for a further 11% of cancer deaths [5].
Colorectal cancer originates in the colon or rectum and often begins as benign polyps that can become malignant over time, influenced by a mix of genetic predispositions, lifestyle factors, and environmental exposures [6]. In Luxembourg, a national colorectal cancer screening programme targeting individuals aged 55 to 74 years has been implemented to reduce the burden of the disease. This organised screening, which uses stool blood tests followed by diagnostic colonoscopy when needed, is part of broader public health campaigns (often highlighted during the “Blue March” awareness month) that stress the importance of early detection and lifestyle modifications to reduce risk [7].

Lung cancer starts in the lung tissues and is one of the most lethal cancers, largely due to its late diagnosis and aggressive nature. Although smoking is a well-known risk factor, environmental pollutants and genetic factors also play key roles [8]. Lung cancer is the leading cause of cancer-related mortality in Luxembourg, particularly among men, but with an increasing prevalence in women [3]. In response, Luxembourg has adopted robust public health measures, including stringent anti-smoking campaigns and policies aimed at reducing air pollution, all of which are crucial to lowering the burden of lung cancer [9].

Due to the significance of these cancers, these two cancers have been targeted as use-cases for the Dataspace4Health project. Colorectal and lung cancers are of high public health and safety interest because they not only account for a significant burden of mortality and morbidity but also benefit from established screening programmes, standardized treatment guidelines, and comprehensive data collection systems that facilitate robust epidemiological research and continuous improvement in patient outcomes.

3. Data Modeling for Lung and Colorectal Cancer in Luxembourg

The primary objective in defining a data model is to illustrate how patient clinical data and genomic sequencing results flow between data holders and data users while ensuring security, pseudonymization, and compliance with data governance and regulations.

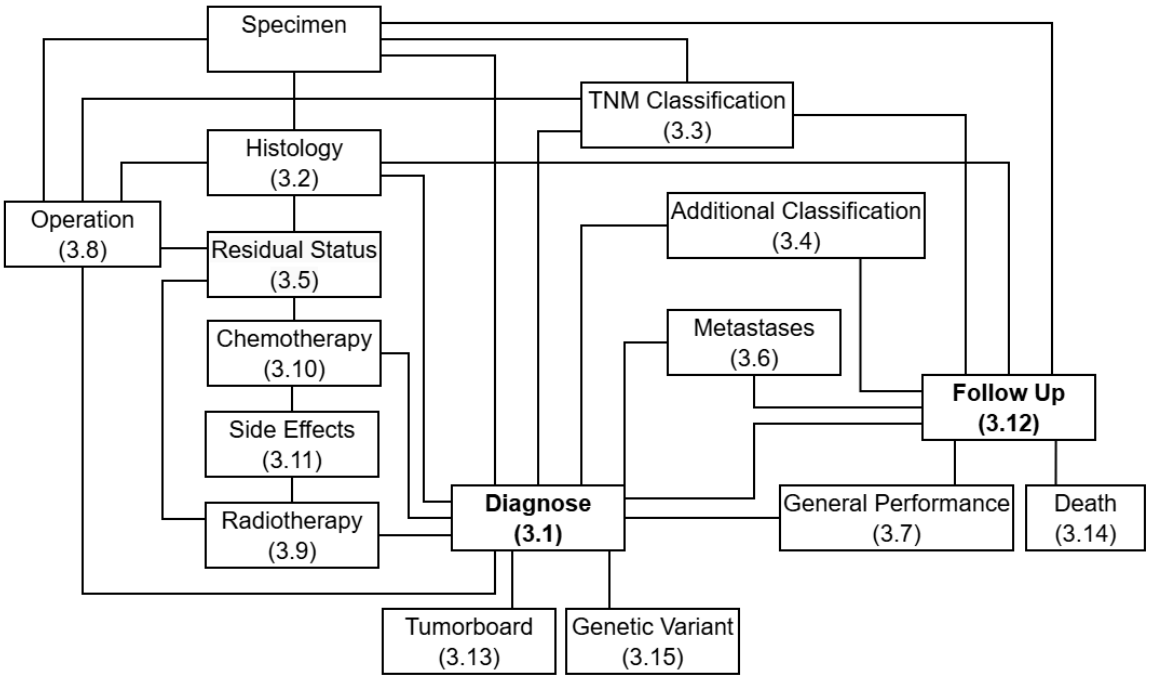

Figure 1. Overview of the relationship of the data entities for oncology.

Dataspace4Health adopts and adapts the oncology data model defined by the MII [2] to inform our data modeling procedures. The model adheres to global healthcare data standards such as the HL7 International healthcare standards organization [10], Fast Healthcare Interoperability Resources (FHIR) [11], Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) [12], and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [13], ensuring semantic consistency and data representation. The standardized healthcare data model is the key to enable interoperability (e.g., data exchange and service exchange). By adhering to the data standards, the MII data model framework eases research collaborations across healthcare entities and borders (e.g., participation in multinational clinical trials). Moreover, the MII data model aligns with European data protection regulations referred to as GDPR [14] and ethical guidelines for handling sensitive patient data [15]. This means that we ensure that our data modeling approach meets legal and ethical requirements for medical research from the outset. Given these advantages, the German MII oncology data model serves as the foundation for defining the data requirements for CRC and lung cancers in Luxembourg.

The proposed data model is designed to represent biological samples, diagnostics, follow-ups, histology, operations, and tumor-nodes-metastasis (TNM) classifications. These entities are linked through relationships that facilitate seamless data flow and analysis. The central nodes in this model are Diagnose and Follow Up, which connect to other entities such as Specimen, Histology, Operation, and TNM Classification. An overview of the data model can be seen in Figure 1.

The model includes several entities, covering a wide range of clinical dimensions, including:

- **3.1: Diagnosis:** Data on patient diagnosis, including cancer type and affected systems.

- **3.2 Histology**: Documentation of the tumor's cellular architecture and morphological features.
- **3.3 TNM Classification**: Defines the cancer's stage using a standardized system for staging cancer based on tumor size (T), lymph node involvement (N), and metastasis (M).
- **3.4 Additional Classification**: Incorporates supplementary staging systems such as UICC stages and molecular subtypes.
- **3.5 Residual Status**: Evaluation of the presence of residual tumor tissue post-treatment.
- **3.6 Metastases**: Details about the localization, size, and timing of distant metastatic sites.
- **3.7 General Performance**: Data on the patient's functional status using standardized scales such as ECOG.
- **3.8 Operation**: Any data on surgical interventions, including procedure codes and complication tracking.
- **3.9 Radiotherapy**: Records of radiation treatment details, including dosage, target areas, and application methods.
- **3.10 Chemotherapy (Systemic Therapy)**: Documentation of systemic treatment regimens including chemotherapeutic agents and their administration schedules.
- **3.11 Side Effects**: Systematic records of treatment-related toxicities and treatment affects.
- **3.12 Follow Up:** Records of patient outcomes over time, including disease progression, recurrence, and survival.
- **3.13 Tumor Board**: Records of meetings of multidisciplinary treatment planning decisions and deviations based on patient preferences.
- **3.14 Death**: Records of mortality data, including cause of death and ICD coding.
- **3.15 Genetic Variant**: Detail of genomic alterations critical for targeted therapies and precision medicine.
- **Specimen:** A HL7 FHIR resource type representing a patient's biological sample (e.g., tissue or blood) used for diagnostic or research purposes.

3.1. Key Features

The proposed data model has the following key features:

- Integrated Data Entities: The model encompasses a wide spectrum of clinical data—from diagnosis and histology to TNM classification, operations, and follow-up records. This comprehensive approach facilitates detailed insights into tumor biology, treatment responses, and patient outcomes.
- Centralized Nodes for Data Connectivity: The Diagnose and Follow Up entities serve as central nodes that link other entities (e.g., Specimen, Histology, Operation, TNM Classification). This interconnected structure promotes seamless data flow, enabling robust analysis and insights across various clinical dimensions.
- Standardization and Interoperability: Employing standardized terminologies and coding systems ensures consistency in data capture and reporting. The integration of HL7 FHIR for specimen data further supports interoperability between clinical systems and research platforms.

- Multidisciplinary Data Integration: The model not only captures clinical and pathological data but also incorporates genomic information (via the Genetic Variant entity) and treatment specifics (such as radiotherapy, chemotherapy, and tumor conference decisions). This multifaceted approach is pivotal for precision oncology and personalized treatment planning.
- Support for Diverse Use Cases: Designed to meet the needs of various stakeholders—bioinformaticians, molecular biologists, health data engineers, and data officers—the model provides a unified framework for both clinical decision support and advanced research. It ensures that patient data, even when pseudonymized for privacy, can be reliably linked to genomic and treatment data.
- Enhanced Data Governance and Privacy: With the involvement of a Data Officer/Coordinator and adherence to standards like the European Health Data Space (EHDS), the model emphasizes secure data processing, robust governance, and strict adherence to data privacy regulations.

4. Conclusion

The proposed data model successfully integrates a comprehensive range of clinical, pathological, genomic, and treatment-related data into a unified framework. By establishing central nodes such as Diagnose and Follow Up, the model facilitates seamless data connectivity across entities like Specimen, Histology, Operation, and TNM Classification. This integrated approach not only supports personalized treatment planning and outcome analysis but also enhances data interoperability and governance in line with current health data standards. By adapting this model for lung and colorectal cancer, healthcare providers and researchers can harness high-quality, interoperable data that enhances clinical decision-making, supports personalized treatment strategies, and drives research and quality improvement initiatives.

The Dataspace4Health initiative aims to leverage this data model to build harmonized data models and data flows to connect the various healthcare providers and research centers in Luxembourg. The Data Flow Diagrams will be used to provide a structured representation of the data flow between the project partners, and secure protocols will be leveraged toward a data space between participating partners. In particular, in future work, Lux-X Dataspace4Health will enable Gaia-X compliant individual data usage for individual use cases within not only the Luxembourgish healthcare industry but also the European Healthcare Industry as a whole.

For additional information on the project, please see the Dataspace4Health webpage.

References

1. Medizininformatik Initiative. (n.d.). Medizininformatik Initiative – Start. Retrieved March 21, 2025, from https://www.medizininformatik-initiative.de/en/start
2. Arbeitsgemeinschaft Deutscher Tumorzentren e. V. (ADT) & Gesellschaft der epidemiologischen Krebsregister in Deutschland e. V. (GEKID). (2021, May 10). Bekanntmachung: Aktualisierter einheitlicher onkologischer Basisdatensatz der Arbeitsgemeinschaft Deutscher Tumorzentren e. V. (ADT) und der Gesellschaft der epidemiologischen Krebsregister in Deutschland e. V. (GEKID) [PDF]. AG Daten (ADT/GEKID). https://www.basisdatensatz.de/download/Basisdatensatz12.7.pdf
3. Mafra, A., Weiss, J., Saleh, S., Weber, G., & Backes, C. (2024). Cancer mortality trends in Luxembourg: A 24-year descriptive study (1998–2021). *Cancer Epidemiology*, *93*, 102648.
4. OECD. (2025). *EU country cancer profile: Luxembourg 2025*. OECD Publishing. Retrieved March 21, 2025, from https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/02/eu-country-cancer-profile-luxembourg-2025_814b714e/4cc3ab0d-en.pdf
5. International Agency for Research on Cancer, World Health Organization. (2022). *Global Cancer Observatory: Luxembourg fact sheet* [PDF]. Retrieved March 21, 2025, from https://gco.iarc.who.int/media/globocan/factsheets/populations/442-luxembourg-fact-sheet.pdf
6. Mármol, I., Sánchez-de-Diego, C., Pradilla Dieste, A., Cerrada, E., & Rodriguez Yoldi, M. J. (2017). Colorectal carcinoma: a general overview and future perspectives in colorectal cancer. International journal of molecular sciences, 18(1), 197.
7. Government of Luxembourg. (2024, March 1). *1 mars – Bleu Cancer [Press release]*. Retrieved March 21, 2025, from https://gouvernement.lu/en/actualites/toutes_actualites/communiques/2024/03-mars/01-mars-bleu-cancer.html
8. de Groot, P. M., Wu, C. C., Carter, B. W., & Munden, R. F. (2018). The epidemiology of lung cancer. Translational lung cancer research, 7(3), 220.
9. Government of Luxembourg. (2024, February 2). *02 Journée Cancer [Press release]*. Retrieved March 21, 2025, from https://gouvernement.lu/en/gouvernement/martine-deprez/actualites.gouvernement2024%2Ben%2Bactualites%2Btoutes_actualites%2Bcommuniques%2B2024%2B02-fevrier%2B02-journee-cancer.html
10. HL7 International. (n.d.). *HL7 International [Homepage]*. Retrieved March 21, 2025, from https://www.hl7.org/
11. HL7 International. (n.d.). FHIR®: Fast Healthcare Interoperability Resources. Retrieved March 21, 2025, from https://www.hl7.org/fhir/index.html
12. OHDSI. (n.d.). Data Standardization. Retrieved March 21, 2025, from https://www.ohdsi.org/data-standardization/
13. Vuokko, R., Vakkuri, A., & Palojoki, S. (2023). Systematized Nomenclature of Medicine-Clinical Terminology (SNOMED CT) Clinical Use Cases in the Context of Electronic Health Record Systems: Systematic Literature Review. JMIR medical informatics, 11, e43750. https://doi.org/10.2196/43750
14. GDPR (General Data Protection Regulation): European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L119, 1–88. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679
15. Ethics and Data Protection Guidance (Horizon Europe): European Commission. (n.d.). *Ethics and data protection* [PDF]. Retrieved March 21, 2025, from https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-and-data-protection_he_en.pdf