Advancing Cancer Treatment by Strategically Integrating Multimodal Data: Challenges and Opportunities

Haridimos Kondylakis[1,2,3,a], Angelina Kouroubali[2,3,b], Dimitrios G. Katehakis[2,3,c]

[1] Department of Computer Science, University of Crete (UOC), [2]Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH-ICS), [3] Hellenic Digital Health Cluster (HDHC

[a]kondylak@uoc.gr, [b]kouroub@ics.forth.gr, [c]katehaki@hdhc.gr

Introduction

This position paper explores the transformative potential of multimodal data integration in cancer treatment. It investigates the strategic utilization of diverse datasets, including genomic, proteomic, imaging, and clinical data, to advance personalized medicine and enable tailored therapies for individual patients. The integration of heterogeneous data sources poses significant challenges, particularly in semantic interoperability. The semantic gap, the disparity between raw data and meaningful interpretation, is identified as a critical barrier to effective data utilization. The obstacles in harmonizing semantics across diverse datasets and the development of robust ontologies to bridge this gap are discussed. The analysis highlights innovative approaches, including machine learning algorithms, to address these challenges and enhance semantic integration. Interdisciplinary collaboration, involving experts from oncology, data science, and bioinformatics, plays a crucial role in creating a unified framework for data interpretation. By overcoming the discussed challenges, the full potential of big data in oncology is unlocked, leading to more effective, efficient, and personalized cancer care [1]. This paper underscores the transformative impact of multimodal data integration in revolutionizing cancer treatment and paving the way for a new era of precision oncology.

The Landscape of Cancer Treatment: Multimodal Data Integration and the critical role of semantic interoperability

Cancer is one of the main priorities of the European Commission in the health domain. In 2020, 2.7 million people in the European Union were diagnosed with cancer, and another 1.3 million people lost their lives to it, including over 2,000 young people. The overall economic impact of cancer in Europe is €100 billion annually. Evidence shows that 40% of cancers are preventable[1]. Cancer is a multifaceted disease, with many diverse factors involved in cancer prevention, progression, treatment, and follow-up. For developing decision support tools and AI models for cancer management and prediction, access to large cohorts with multilevel and multimodal data is essential [2-5]. However, currently, the required data remain fragmented in various health silos. Besides their fragmentation, they remain heterogeneous and dirty which prevents their further usage [6].

EU Initiatives and the VELES project

MyHealth@EU is a functional, robust European health data infrastructure that connects 25 member states[2] and will expand to cover the entire EU + Norway and Iceland. Current cross-border services include Patient Summary and ePrescription, but will be expanded to lab results, hospital discharge reports, medical images and more. The HealthData@EU infrastructure will be piloted by a 2-year project facilitating the cross-

[1] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/promoting-our-european-way-life/european-health-union/cancer-plan-europe_en

[2] https://health.ec.europa.eu/ehealth-digital-health-and-care/electronic-cross-border-health-services_en#which-services-are-available-in-which-countries

border use of health data for secondary purposes like research, policy making, regulatory activities, and innovation (HealthData@EU Pilot).

Joint Action Towards the European Health Data Space – TEHDASCancer: was a two-year initiative co-funded by the European Commission with the support of 25 European Union member states. The goal of the project was to develop and promote concepts related to data sharing for citizens' health, public health, as well as health research & innovation in Europe, necessary for the development and implementation of the European Health Data Space. EUropean Federation for CAncer IMages (EUCAIM) is the cornerstone of the European Commission-initiated European Cancer Imaging Initiative, a flagship of Europe's Beating Cancer Plan (EBCP), which aims to foster innovation and deployment of digital technologies in cancer treatment and care to achieve more precise and faster clinical decision making, diagnostics, treatment and predictive medicine for cancer patients. The Genomic Data Infrastructure (GDI) project is a European initiative, which brings together national authorities, research organizations, and technology providers in 20 countries to create a cross-border interconnected network of national genome collections in conjunction with other relevant data.

The VELES Excellence Hub[3] is a significant initiative funded by the European Commission[4] with the goal of strengthening smart health innovation in Southeast Europe. It aims to raise the level of innovation excellence by creating a sustainable, place-based innovation ecosystem. This is enabled through the development of a Regional Smart Health Data Space, which includes a novel transformational framework, research and innovation (R&I) strategies, and an investment action plan for the development and adoption of innovative, secure digital solutions that underpin the delivery of sustainable healthcare. The VELES Excellence Hub represents a strategic effort to integrate and optimize health data across borders, aiming to improve outcomes in critical areas of public health, including amongst others, cancer treatment as one of the Regional Smart Health Data Space pilot projects it focuses upon.

Semantic interoperability in healthcare

The importance of semantic interoperability in healthcare is multifold. It contributes to:

- Improved patient care as clinicians have access to comprehensive and accurate patient information, leading to better diagnosis and treatment.
- Enhanced data exchange through smooth and meaningful data exchange across various healthcare systems and organizations, including hospitals laboratories, and clinics.
- Support of research and analytics through high-quality, standardized data which is crucial for clinical research, public health monitoring, and advanced analytics.
- Reduction of redundant testing and administrative costs by ensuring accurate data transfer and understanding, increasing efficiency, and reducing costs.
- Compliance of healthcare organizations with regulations that require accurate and consistent health data reporting.

Ontology-based frameworks utilize ontologies, which are structured sets of terms and concepts representing a domain of knowledge, to facilitate the integration of heterogeneous data sources [7]. Successful implementations of semantic integration demonstrate the practical benefits, such as improved data interoperability and enhanced decision-making in healthcare settings [8]. Semantic workflows and tools are used to automate the process of integrating data from various sources [9], ensuring that the combined dataset is semantically coherent and ready for analysis [10]. Ontology-guided frameworks guide

[3] https://veleshub.eu/
[4] https://cordis.europa.eu/project/id/101087483

the integration process by providing a semantic structure to which data from different sources can be mapped, enhancing the clarity and usability of the integrated data [11]. Semantic integration plays a crucial role in clinical decision support systems, particularly in complex care scenarios such as breast cancer treatment, where it can lead to more personalized and effective patient care [12-14]. Semantic interoperability ensures that data from diverse healthcare systems can be combined and used effectively, leading to better patient outcomes and more efficient care delivery. Ongoing advancements in semantic technologies and standards are expected to further enhance the ability to exchange and integrate data across different platforms and domains, paving the way for more sophisticated and scalable data ecosystems [15]. The future of data exchange looks promising with the continuous evolution of standards and technologies [16].

Terminologies

SNOMED CT (Systematized nomenclature of medicine – clinical terms): A comprehensive clinical terminology that covers diseases, clinical findings, procedures, microorganisms, and more.
LOINC (Logical Observation Identifiers Names and Codes): Standardized terms for laboratory and clinical observations.
ICD (International Classification of Diseases): A system used for diagnosing and classifying diseases and health conditions.

Data models and structures

HL7 FHIR (Fast Healthcare Interoperability Resources): A standard describing data formats and elements (known as "resources") and an API for exchanging electronic health records.

CDA (Clinical Document Architecture): An HL7 standard for the structure and semantics of clinical documents.

openEHR: An open standard specification for electronic health records and data interoperability.

OMOP-CDM (Observational Medical Outcomes Partnership Common Data Model): Harmonizes disparate coding systems —with minimal information loss—into a standardized vocabulary. It is an open community data standard, designed to standardize the structure and content of observational data and to enable efficient analyses that can produce reliable evidence. A central component of the OMOP CDM is the OHDSI standardized vocabularies.

Implementation and frameworks

Implementation guides provide detailed instructions on how to implement standards in specific contexts, ensuring consistent adoption. Frameworks like the IHE (Integrating the Healthcare Enterprise) provide practical tools for achieving interoperability through profiles and integration patterns.

Ontologies and integration strategies are needed to overcome the semantic gap. Advanced analytics and machine learning can only be realized based on data that is curated for semantic interoperability.

Challenges to Semantic Interoperability in Cancer Data

Data heterogeneity in terms of both syntax and semantics: Needs better data curation, better tools respecting established standards, and common data models. Effective ETL tools.

Inconsistent, poor quality, and/or missing data: needs better data curation, and more rules when data are recorded. eCRF forms are more user-friendly and automatically completed.

Clinicians use hand-writing notes: needs clinicians to use voice to text or write the text on a PC. Even better clinicians do not write any free text and only make selections.

Discussion and Conclusions

The FAIRification of multimodal cancer data focuses on data collection, storage, and accessibility of data from various sources, whereas potential problems include governance and regulation. To this end, a data innovation canvas will be created, as part of the foreseen cancer treatment pilot in VELES, that maps stakeholders, data sources, customers, users, culture, data skills and available tools, data channels, added value, impact, and business value. This will be a useful tool for the creation of a data space that will enable the collection, cleaning, and FAIRification of the collected data.

Easily accessible, homogenized data; regulated access to the available data; Browsable, findable and reusable cancer data that can be used for modeling and Decision Support tools; Generate data-driven insights; More efficient cancer care planning and decision making.

The current needs related to medical data are:

- Effective common data models.
- Methods and tools to enable the harmonization and the integration of data.
- AI developers with clinical understanding.

Organizations are willing and capable to work together towards a collaborative medical data ecosystem, but:

- appropriate infrastructures should ensure that they remain owners of their data
- can identify who is using their dataset and why
- have an appropriate and sustainable pricing model
- they are properly acknowledged in the developed AI models

Key Features of a data ecosystem

- Clinicians to contribute their datasets, hosted on their own premises if needed or pushed to a central repository if there is the capability to do so.
- All sources adopt the same common data model
- A meta-data catalog will allow search on the meta-data of the available datasets and issue requests for access
- Data owners can accept or reject data access requests
- A pricing model will be later implemented
- Upon data access is granted AI models can be sent to be trained in the datasets

The position paper concludes that shared, controlled vocabulary and clear semantic relationships among data from different sources can significantly enhance the integration process, leading to more effective and personalized cancer treatments, whereas governance and regulation problems are equally important and should not be neglected.

References

[1] Jiang, Peng, Sanju Sinha, Kenneth Aldape, Sridhar Hannenhalli, Cenk Sahinalp, and Eytan Ruppin. "Big data in basic and translational cancer research." Nature Reviews Cancer 22, no. 11 (2022): 625-639.

[2] Stahlberg, Eric A., Mohamed Abdel-Rahman, Boris Aguilar, Alireza Asadpoure, Robert A. Beckman, Lynn L. Borkon, Jeffrey N. Bryan et al. "Exploring approaches for predictive cancer patient digital twins: Opportunities for collaboration and innovation." *Frontiers in Digital Health* 4 (2022): 1007784.

[3] Salvi, Massimo, Hui Wen Loh, Silvia Seoni, Prabal Datta Barua, Salvador García, Filippo Molinari, and U. Rajendra Acharya. "Multi-modality approaches for medical support systems: A systematic review of the last decade." Information Fusion (2023): 102134.

[4] Darda, Pooja, and Nikita Matta. "The Nexus of Healthcare and Technology: A Thematic Analysis of Digital Transformation Through Artificial Intelligence." In Transformative Approaches to Patient Literacy and Healthcare Innovation, pp. 261-282. IGI Global, 2024.

[5] Triantafyllidis, Andreas, Haridimos Kondylakis, Dimitrios Katehakis, Angelina Kouroubali, Lefteris Koumakis, Kostas Marias, Anastasios Alexiadis, Konstantinos Votis, and Dimitrios Tzovaras. "Deep learning in mHealth for cardiovascular disease, diabetes, and cancer: systematic review." JMIR mHealth and uHealth 10, no. 4 (2022): e32344.

[6] Kondylakis, Haridimos, Varvara Kalokyri, Stelios Sfakianakis, Kostas Marias, Manolis Tsiknakis, Ana Jimenez-Pastor, Eduardo Camacho-Ramos et al. "Data infrastructures for AI in medical imaging: a report on the experiences of five EU projects." European radiology experimental 7, no. 1 (2023): 20.

[7] Tripathi, Aakash, Asim Waqas, Kavya Venkatesan, Yasin Yilmaz, and Ghulam Rasool. "Building Flexible, Scalable, and Machine Learning-ready Multimodal Oncology Datasets." Sensors 24, no. 5 (2024): 1634.

[8] Zhang, H., Guo, Y., Li, Q., George, T. J., Shenkman, E., Modave, F., & Bian, J. (2018). An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. BMC medical informatics and decision making, 18, 129-147.

[9] De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., & Rosati, R. (2018). Using ontologies for semantic data integration. A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years, 187-202.

[10] Hassan, Mubashir, Faryal Mehwish Awan, Anam Naz, Enrique J. deAndrés-Galiana, Oscar Alvarez, Ana Cernea, Lucas Fernández-Brillet, Juan Luis Fernández-Martínez, and Andrzej Kloczkowski. "Innovations in genomics and big data analytics for personalized medicine and health care: A review." *International journal of molecular Sciences* 23, no. 9 (2022): 4645.

[11] Zhang, H., Guo, Y., Li, Q., George, T. J., Shenkman, E., Modave, F., & Bian, J. (2018). An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. BMC medical informatics and decision making, 18, 129-147.

[12] Kondylakis, Haridimos, Cristian Axenie, Dhundy Bastola, Dimitrios G. Katehakis, Angelina Kouroubali, Daria Kurz, Nekane Larburu et al. "Status and recommendations of technological and data-driven innovations in cancer care: Focus group study." *Journal of medical Internet research* 22, no. 12 (2020): e22034.

[13] Kilintzis, Vassilis, Varvara Kalokyri, Haridimos Kondylakis, Smriti Joshi, Katerina Nikiforaki, Oliver Díaz, Karim Lekadir, Manolis Tsiknakis, and Kostas Marias. "Public data homogenization for AI model development in breast cancer." European Radiology Experimental 8, no. 1 (2024): 42.

[14] Lezcano, L., Sicilia, M. Á., & Rivero, E. (2013). Semantic Integration of Patient Data for Clinical Decision Support in Breast Cancer Care. In Interoperability in Healthcare Information Systems: Standards, Management, and Technology (pp. 250-267). IGI Global.

[15] Becha, H., Schröder, M., Voorspuij, J., Frazier, T., & Lind, M. (2020). Global data exchange standards: The basis for future smart container digital services. In Maritime informatics (pp. 293-307). Cham: Springer International Publishing.

[16] Katehakis, Dimitrios G., and Angelina Kouroubali. "The EHR as an Instrument for Effective Digital Transformation in the Post COVID-19 Era." In SWH@ ISWC, pp. 8-19. 2021.