ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ESMO REAL WORLD DATA AND DIGITAL ONCOLOGY

PERSPECTIVES

Advancing population-based cancer registries through the OHDSI OMOP Common Data Model: governance, validation, and interoperability for cancer research

M. M. H. van Swieten[1], B. Lima[2,3]*, T. Niemi[4], C. Backes[2,3], J. L. Bulliard[4], H. De Schutter[5], S. Larønningen[6] & E. Enerly[7]

[1]Department of Clinical Data Science, Comprehensive Cancer Organization, Eindhoven, The Netherlands; [2]Registre National du Cancer du Luxembourg (RNC); [3]Cancer Epidemiology and Prevention Group (EPICAN), Department of Precision Health (DoPH), Luxembourg Institute of Health (LIH), Strassen, Luxembourg; [4]Centre for Primary Care and Public Health (Unisanté), University of Lausanne, Lausanne, Switzerland; [5]Belgian Cancer Registry, Brussels, Belgium; Departments of [6]Registration; [7]Research, Cancer Registry of Norway, Norwegian Institute of Public Health, Oslo, Norway

Available online xxx

Population-based cancer registries (PBCRs) are essential for cancer surveillance and research but are limited in their analytical interoperability with other data sources because of heterogeneity in cancer-related definitions and database structure. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) can be used to improve interoperability between PBCRs and other health data. In this perspective, we provide a conceptual framework of how governance contributes to consistent mapping, validation, and sustainable use of OMOP CDM for PBCRs. We also demonstrate a newly developed R/Shiny validation tool that supports transparency and trust in the converted registry data. When OMOP CDMs are implemented with structured governance, it can support scalable, reproducible, multisite research and contribute to the readiness for the upcoming European Health Data Space regulation. Likewise, PBCRs advance cancer research and surveillance and contribute to the creation of a trusted, interoperable health data ecosystem that supports research and policy-making.

Key words: data reuse, governance, interoperability, OMOP Common Data Model, population-based cancer registries

INTRODUCTION

Population-based cancer registries (PBCRs) collect, code, consolidate, and link cancer data aligned to international standards for all new cancer cases diagnosed within a defined geographical region using data from multiple sources, providing exhaustive, continuous, systematic, and relevant cancer knowledge of the underlying population.[1] PBCR data and expertise are used for cancer surveillance and research, playing a fundamental role in cancer control, evaluating cancer care, supporting the design and implementation of evidence-based health interventions, and guiding cancer-related policy decisions.[2-4] These highly curated, quality-assured, and population-wide data enable analyses that hospital-based clinical databases cannot facilitate alone. Linking PBCR data with additional data sources expands the scope and use for research and enables further comprehensive population-based analyses.[5,6] However, such linkages remain challenging because of heterogeneity in data structures, variable definitions, coding systems, data quality, and national or regional regulatory constraints.[7] Aligned to their scope, PBCRs rely on international standards, including the International Classification of Diseases for Oncology (ICD-O)[8] and TNM (tumour—node—metastasis) classification,[9] as well as recommendations from the European Network of Cancer Registries (ENCR)[10] and the International Agency for Research on Cancer[11] to enable comparability between PBCRs at the international level. Nevertheless, these standards are not consistently applied in other clinical or administrative data sources. Consequently, limited harmonisation and heterogeneous data access procedures constrain analytical interoperability across institutions, regions, and countries. As a result, secondary use of PBCR data through linkage with other data is often limited by the need for labour-intensive data restructuring and harmonisation for each individual use case.[12-14] A Common Data Model (CDM), such as the Observational Medical Outcomes Partnership (OMOP) CDM developed by the Observational Health Data Sciences and Informatics (OHDSI)[15] collaboration, offers new opportunities to PBCRs and supports interoperability with

*Correspondence to: Bruno Lima, Department of Precision Health, Luxembourg Institute of Health, 1A-B, rue Thomas Edison, L-1445 Strassen, Luxembourg. Tel: +352 26970-876

E-mail: bruno.lima@lih.lu (B. Lima).

2949-8201/© 2026 The Authors. Published by Elsevier Ltd on behalf of European Society for Medical Oncology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

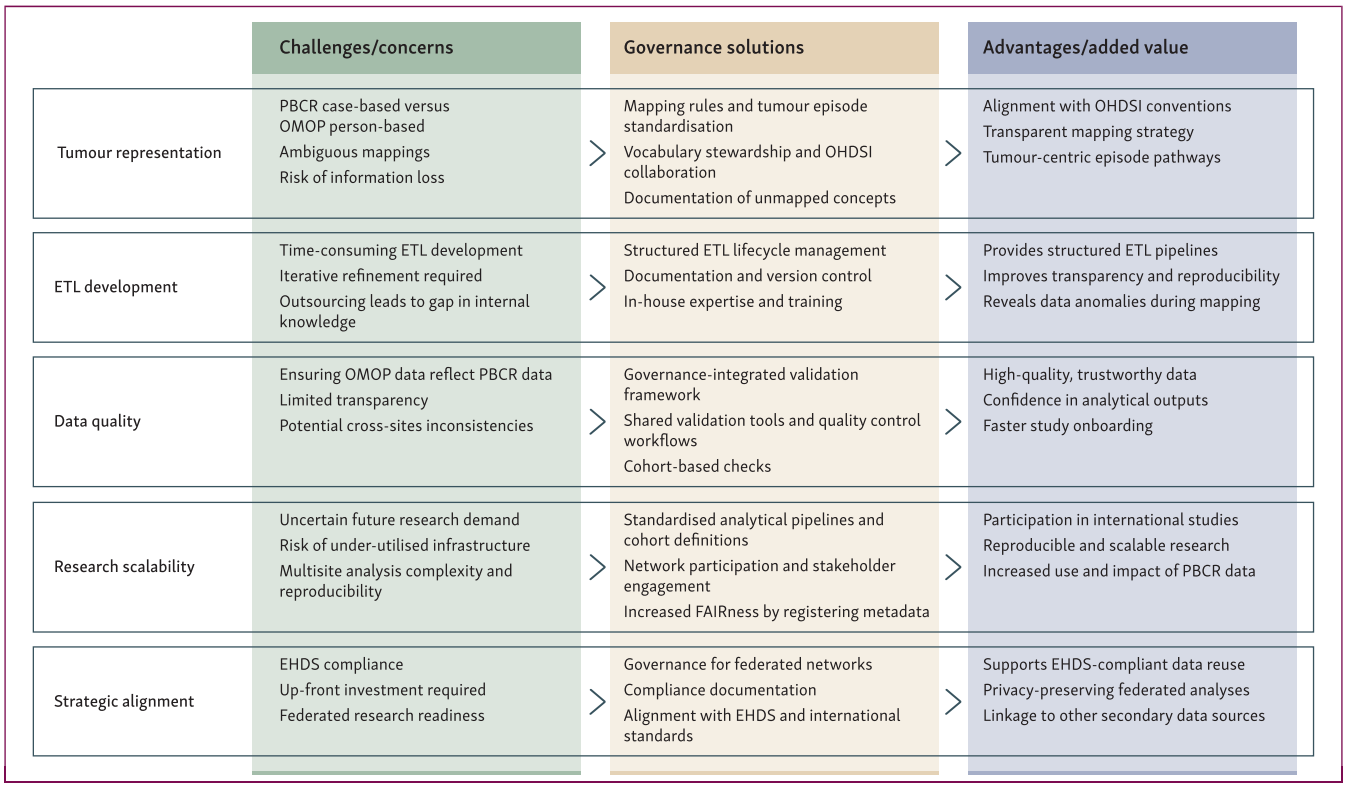

Figure 1. Schematic overview of the challenges, governance solutions, and achievable benefits of implementing the OMOP CDM for PBCRs. Structured governance addresses registry-specific challenges, enabling the realization of OMOP CDM's advantages, including standardisation, interoperability, and support for federated research.

CDM, Common Data Model; EHDS, European Health Data Space; ETL, Extract, Transform, and Load; FAIR, Findable, Accessible, Interoperable, and Reusable; OHDSI, Observational Health Data Sciences and Informatics; OMOP, Observational Medical Outcomes Partnership; PBCR, population-based cancer registry.

additional health data sources. It allows for the harmonisation of disparate data sources into a standardised, analytics-ready structure,[16] making registry data Findable, Accessible, Interoperable, and Reusable.[17]

Several publications have evaluated the OMOP CDM for oncology, identifying benefits for semantic interoperability, computational reproducibility, and large-scale research networks.[18,19] However, most implementations have been institution-specific and have relied on a limited set of variables and few have studied the conceptual implications of the OMOP CDM for PBCRs specifically. For PBCRs, the added value of the OMOP CDM extends beyond technical transformation. When implemented within an appropriated governance framework, it enables federated research,[20] facilitates collaboration, promotes the sustainable reuse of PBCR data, and strengthens preparedness for emerging initiatives, such as the European Health Data Space (EHDS),[21] which supports the OMOP CDM as one of the standards for secondary data use.[22] To realise this potential, consistent implementation of the OMOP CDM must be supported by a set of rules, policies, funding, and processes that guide how the underlying health data are managed, used, and overseen to ensure quality, accountability, and compliance, i.e. a robust governance.

In this article, we aim to provide (i) a conceptual framework of how governance serves as an enabling layer that addresses semantic, technical, and operational challenges to benefit from the advantages that the OMOP CDM can offer to PBCRs and (ii) a demonstration of how a newly developed validation environment, built using (R Core Team, Vienna, Austria) and Shiny (Posit Software, PBC, Boston, MA), can contribute to the governance in practice.

CONCEPTUAL FRAMEWORK

The benefits of implementing the OMOP CDM in PBCR contexts are conditional on the presence of structured governance mechanisms that address both external and internal challenges specific to cancer registration, as well as availability of resources. Figure 1 presents a governance-oriented framework that shows the OMOP CDM implementation in PBCRs as a threefold process: (i) challenges arising, partially related to applying a patient-centred, standardised data model to registry data where events are linked to cancer cases; (ii) governance solutions that mitigate these challenges; and (iii) the benefits that become achievable once governance is in place. With this framework, we show that proper governance supports the translation of the theoretical strengths of the OMOP CDM into sustainable, reproducible, and policy-relevant outputs. We discuss the core governance domains required for implementing the OMOP CDM in PBCRs in the following sections.

GOVERNANCE OF TUMOUR REPRESENTATION

Transforming PBCR data into the OMOP CDM involves several challenges. The main variables of PBCRs (e.g.

ENCR's recommendation for a minimum dataset[10]) can usually be mapped to standard vocabularies (e.g. Systematized Nomenclature of Medicine - Clinical Terms [SNOMED CT], Logical Observation Identifiers Names and Codes [LOINC], or OHDSI-maintained cancer modifiers) following OMOP CDM conventions but oncology-specific conventions are still inconsistently implemented across databases. More complex registry variables and study-specific mappings are often not covered by existing guidelines and in some cases, multiple valid mapping options exist. PBCR variables may also plausibly map to multiple domains (OBSERVATION, MEASUREMENT, or PROCEDURE), depending on the interpretation. Cancer-specific terminologies, including ICD-O classifications, are not always fully represented within OMOP CDM standardised vocabularies, creating risks of simplification or information loss.[7] Lastly, the mismatch between tumour-centric registry structures and the patient-centric OMOP CDM can also result in inconsistent representations, loss of detail, or complicated cohort definitions during the analytical stage.

Without centralized guidance, PBCRs may make different implementation choices while mapping to the OMOP CDM, thereby undermining comparability. Governance mechanisms, including mapping guides at (inter)-national and research-network levels (e.g. OHDSI[23] and DARWIN EU, the European Medicines Agency's Data Analysis and Real World Interrogation Network),[24] explicit domain allocation rules, and systemic documentation of mapping decisions, help ensure consistency and transparency while also providing a basis for audits, review, and future refinement.

The governance solutions proposed here allow PBCRs to maintain internal consistency and to avoid heterogeneous implementations that compromise downstream analyses. Ongoing vocabulary stewardship by tracking unmapped concepts, documenting compromises (e.g. custom concepts), and engaging with the OHDSI Oncology and Vocabulary Working Groups can address mapping ambiguities. Coordination with these initiatives allows registries to contribute to the development of, and sensitisation to, cancer-specific vocabularies and to ensure that local implementations align with community standards.

The OMOP CDM provides disease episodes (EPISODE domain) as a solution to preserve the tumour-centric logic.[25-27] This EPISODE domain has previously been underused or inconsistently implemented across sources owing to semantic ambiguities, heterogeneous source data, and limited implementation guidance. The recent release of the OHDSI Oncology Working Group's Oncology Extract, Transform, and Load (ETL) OnRamp[23] provides comprehensive recommendations for constructing the relevant episodes, such as disease and drug therapy episodes, as well as guidance on populating the EPISODE and EPISODE_EVENT tables. Although the completeness of episode representations will continue to depend on source data availability, these recommendations are expected to improve consistency and interoperability across OMOP CDM implementations.[28]

ETL LIFECYCLE GOVERNANCE AND TECHNICAL SUSTAINABILITY

The ETL process required to convert PBCR data into the OMOP CDM format is typically resource-intensive and iterative. Although ETL pipeline development may initially appear to be a one-time investment, in practice, it requires continuous updating as source data, vocabularies, and mappings change over time. Therefore, governance and funding should cover the changing nature of the ETL process. Version control, structured documentation, and peer review of the transformation logic are necessary. Registries that have outsourced the initial ETL development, owing to a lack of resources, require internal staff training to ensure knowledge transfer. By incorporating ETL governance within institutional processes, PBCRs can ensure that OMOP CDM implementations remain sustainable and adaptable to future developments while remaining reproducible over time.

DATA QUALITY GOVERNANCE AND VALIDATION

The OHDSI software stack provides tools for systematic data quality assessment, such as the Data Quality Dashboard. However, standard data quality tools primarily assess conformance, completeness, and plausibility within the OMOP CDM structure itself. For most data sources, including PBCRs, an additional layer of governance is required to ensure that the OMOP CDM representation properly reflects the source data. For PBCRs, this requires validation approaches that directly compare OMOP CDM—derived outputs with established registry statistics, such as incidence counts, survival estimates, or tumour distributions (e.g. the database of the Association of the Nordic Cancer Registries [NORDCAN],[14] European Cancer Information System [ECIS],[13] Global Cancer Observatory [GLOBOCAN],[29] and National Cancer Registries data). Governance mechanisms should define which indicators are used for validation, how discrepancies are investigated, and how corrections are implemented.

To illustrate how governance principles can be embedded directly into analytical workflows and quality assessments, we developed OMOPCancerIndicators, a new R package with Shiny interface[30] (Figure 2). This tool enables a direct comparison between OMOP CDM—derived outputs and established registry statistics. Simple cohorts based on combinations of cancer types, incidence periods, and selected measurements can be extracted from both the source registry and the OMOP CDM database, allowing discrepancies to be identified, visualised, and investigated directly. Although survival analysis is not currently available through OMOPCancerIndicators, we consider incorporating survival estimates in future versions as a key component for internal data validation. The tool has been built using standard OMOP CDM libraries such as DatabaseConnector, CohortGenerator, CohortDiagnostics, and SqlRender for OMOP-supported databases. The cohorts are defined as JSON files and created automatically on the basis of these

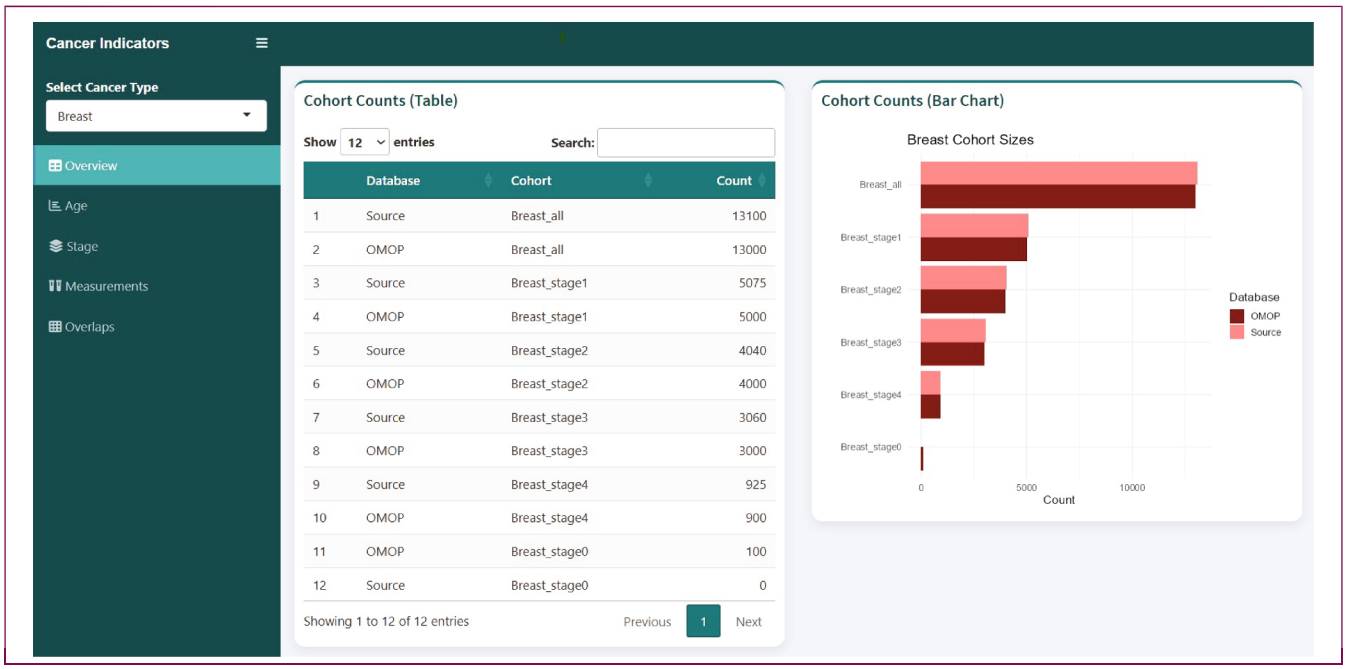

Figure 2. Shiny interface of the newly developed OMOPCancerIndicators R package for quality control and comparison with source data.[30] The tool visualises cohort count, age distributions per cohort, stage distributions in absolute counts and percentages, measurements (if defined) in absolute counts and percentages, and overlaps between cohorts in the OMOP CDM database. These overlaps can be used to identify implausible combinations of cohorts.
CDM, Common Data Model; OMOP, Observational Medical Outcomes Partnership.

definitions. The source data must be extracted separately for each registry database.

This approach serves multiple governance purposes. Firstly, it supports internal quality control by revealing mapping errors or unintended transformations that may not be apparent through generic checks. Secondly, it enhances transparency by making validation procedures reproducible and understandable to nontechnical stakeholders (through a Shiny user interface). Thirdly, it builds institutional trust in the OMOP CDM conversion, which is critical for enabling downstream research reuse and speeding up internal processes for data delivery (i.e. institutional review board approval) by making transparent which data will be used. Importantly, such tools also facilitate harmonised validation across multiple registries, supporting cross-site comparability within a transparent analytic framework. Although the current version of our package focuses on internal validation between source data and the OMOP CDM, future versions could incorporate external validation features by comparing transformed data against established public cancer statistics such as NORDCAN or ECIS.

RESEARCH SCALABILITY AND CROSS-INSTITUTIONAL COMPARABILITY

A key motivation for adopting the OMOP CDM is to enable participation in large multisite and international research initiatives. However, without harmonised governance, federated analyses may produce biased or incomparable results owing to differences in implementation choices, data completeness, or mapping strategies. PBCRs are uniquely positioned to play a coordinating role in this context. Through shared governance frameworks, registries can align OMOP CDM implementations across regions or countries, establish minimum standards for participation in federated studies,[31] and provide expertise in cancer-specific data interpretation. This role extends beyond technical alignment to include methodological guidance and quality assurance. In research studies that include both medical records and claims data, PBCRs can act as a benchmark for core variables related to incidence, prevalence, and survival statistics of cancer cohorts. Given that PBCRs typically also contain additional, more detailed, clinical information for certain cancer types, mapping these variables to the OMOP CDM allows PBCRs to participate in more health analysis studies and increases the impact of PBCR data.

STRATEGIC AND POLICY IMPLICATIONS

The governance of OMOP CDM implementation in PBCRs must be considered within a broader strategic and policy context. In Europe, the EHDS established a framework for secure, regulated secondary use of health data for research, public health, innovation, and policy-making, emphasising interoperability, privacy preservation, and accountability.[21,22] Based on their population coverage, high-quality data, and extensive knowledge and experience, PBCRs are well positioned to participate in EHDS-compliant initiatives and linkage of data across sources.

The OMOP CDM also aligns with these objectives by supporting standardised data representation and federated analytics, where data remain locally controlled while

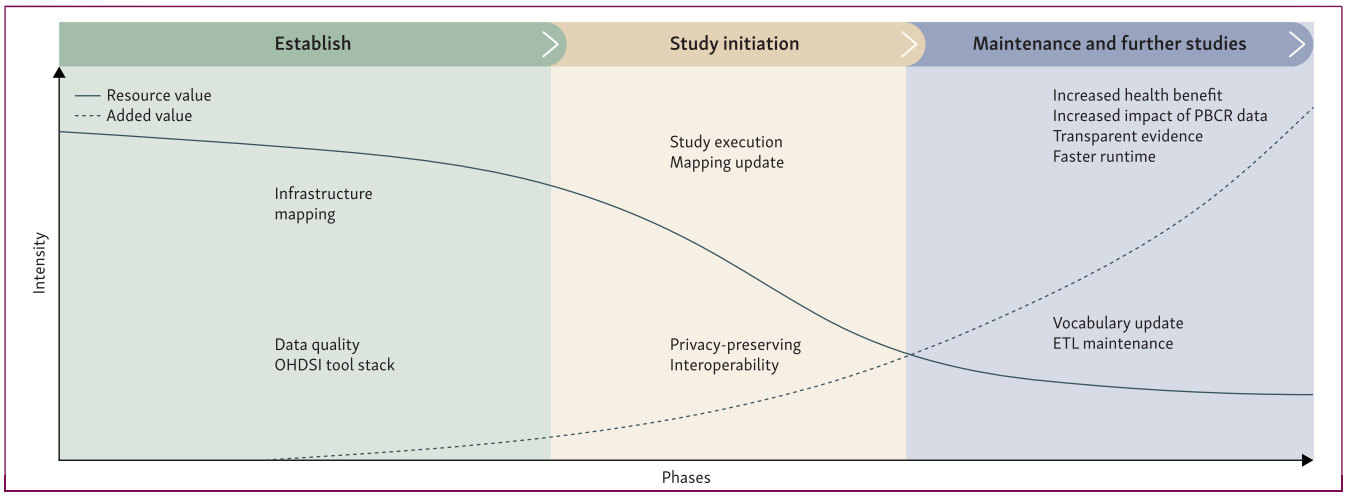

Figure 3. Efforts and value over time of OMOP CDM implementation by PBCRs.
CDM, Common Data Model; ETL, Extract, Transform, and Load; OHDSI, Observational Health Data Sciences and Informatics; OMOP, Observational Medical Outcomes Partnership; PBCR, population-based cancer registry.

analyses are conducted in a distributed manner. These federated analyses can be carried out in two ways: (i) by distributing ready-to-run analytic scripts, typically R-packages, to data partners, who run them locally on their database (e.g. in OHDSI network[15] and DARWIN EU[24] studies) or (ii) run them in a fully automated federated learning infrastructure (e.g. using vantage6 [IKNL, Utrecht, The Netherlands; Maastricht University Medical Centre+, Maastricht, The Netherlands][32] in European initiatives such as BlueBerry,[33] IDEA4RC [Intelligent Ecosystem to improve the governance, the sharing and the re-use of health Data for Rare Cancers],[34] etc). Robust governance is essential within such research networks to prevent biased or inconsistent results. By coordinating governance frameworks, validation practices, and shared methodological standards, registries can help ensure that federated analyses produce meaningful and comparable outcomes. PBCRs can also act as data stewards to support hospitals and other clinical data holders in aligning their data with registry-grade quality requirements. Thus, PBCRs not only advance cancer research and surveillance but also contribute to creating a trusted, interoperable health data ecosystem that supports research and policy-making.

DISCUSSION AND CONCLUSION

Investing in the implementation of the OMOP CDM offers substantial value for PBCRs in cancer research, enabling standardised, interoperable data, reproducible analyses; scalable, international multisite research; and alignment with federated initiatives, such as the upcoming EHDS.

Achieving these benefits can be challenging because of a lack of human, technical, and financial resources. The initial time, expertise, and infrastructure investments are nontrivial and may appear disproportionate when considering in relation to a single study. However, the benefits of OMOP CDM implementation accumulate over time (Figure 3). Once the foundational infrastructure, sustainable funding, and governance in the long term are in place, the additional cost per study is reduced, data preparation becomes more efficient, refinements are made per use case, and participation in additional studies becomes easier. Importantly, the investment value increases when the OMOP CDM is reused across multiple projects and when PBCR data are linked to other data sources within the EHDS.

Structured governance covering tumour representation, ETL life-cycle management, data quality, research scalability, and strategic alignment is needed to support these objectives. PBCRs can collaborate with partners such as OHDSI working groups, clinical networks, researchers, and other international research networks to develop and maintain governance frameworks.

PBCRs can enhance the secondary use of their data by adopting the OMOP CDM within a structured governance framework. By enabling interoperable linkage with other health data sources, this approach improves data quality transparency, facilitates reproducible and federated observational research, and supports the sustainable reuse of cancer registry data within evolving health data ecosystems.

ETHICAL APPROVAL

The authors declare that the study does not involve humans or animal subjects.

FUNDING

None declared.

DISCLOSURE

The authors have declared no conflicts of interest.

REFERENCES

1. Izquierdo JN, Schoenbach VJ. The potential and limitations of data from population-based state cancer registries. *Am J Public Health*. 2000;90(5):695-698.

2. Tucker TC, Durbin EB, McDowell JK, Huang B. Unlocking the potential of population-based cancer registries. *Cancer*. 2019;125(21):3729-3737.
3. Ahmadi N, Peng Y, Wolfien M, Zoch M, Sedlmayr M. OMOP CDM can facilitate data-driven studies for cancer prediction: a systematic review. *Int J Mol Sci*. 2022;23(19):11834.
4. Kempf E, Vaterkowski M, Leprovost D, et al. How to improve cancer patients enrollment in clinical trials from rEal-Life databases using the observational medical outcomes partnership oncology extension: results of the PENELOPE initiative in urologic cancers. *JCO Clin Cancer Inform*. 2023;7:e2200179.
5. Heins MJ, de Ligt KM, Verloop J, Siesling S, Korevaar JC. Opportunities and obstacles in linking large health care registries: the primary secondary cancer care registry - breast cancer. *BMC Med Res Methodol*. 2022;22(1):124.
6. Henke E, Zoch M, Peng Y, Reinecke I, Sedlmayr M, Bathelt F. Conceptual design of a generic data harmonization process for OMOP common data model. *BMC Med Inform Decis Mak*. 2024;24(1):58.
7. Lima B, Hasan F, Kannan P, et al. Optimizing data linkage for maximizing the potential of Luxembourg's national cancer registry: a comprehensive scoping review. *Front Oncol*. 2025;15:1679408.
8. Fritz A, Percy C, Jack A, et al. *ICD-O-3 International Classification of Diseases for Oncology*. Geneva: World Health Organization; 2013.
9. Brierley JD, Gospodarowicz MK, Wittekind C. *TNM Classification of Malignant Tumours*. 8th ed. Oxford: Wiley Blackwell; 2016.
10. European Network of Cancer Registries. ENCR recommendations. Available at https://www.encr.eu/ENCR-Recommendations. Accessed September 23, 2025.
11. Boyle P, Parkin DM. *Statistical methods for registries*. Cancer Registration: Principles and Methods. IARC Scientific Publications. Lyon, France: World Healh Organization; 1991. p. 126-158.
12. Bray FCM, Aitken JF, Bardot A, et al. *Cancer Incidence in Five Continents. XII. IARC Scientific Publications*. Lyon, France: Intrernational Agency for Research on Cancer; 2024.
13. European Commision. Joint Research Centre. ECIS — European Cancer Information System. Available at https://ecis.jrc.ec.europa.eu. Accessed May 2, 2026.
14. International Agency for Research on Cancer. Nordic cancer statistics web tool. Cancer incidence and mortality data: NORDCAN database. 2025. Availabe at https://nordcan.iarc.fr. Accessed January 14, 2026.
15. Hripcsak G, Duke JD, Shah NH, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform*. 2015;216:574-578.
16. Voss EA, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J Am Med Inform Assoc*. 2015;22(3):553-564.
17. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.
18. Wang L, Wen A, Fu S, et al. A scoping review of OMOP CDM adoption for cancer research using real world data. *NPJ Digit Med*. 2025;8(1):189.
19. Sibert NT, Soff J, La Ferla S, Quaranta M, Kremer A, Kowalski C. Transforming a large-scale prostate cancer outcomes dataset to the OMOP Common Data Model—experiences from a scientific data holder's perspective. *Cancers (Basel)*. 2024;16(11):2069.
20. Sinaci AA, Gencturk M, Alvarez-Romero C, et al. Privacy-preserving federated machine learning on FAIR health data: a real-world application. *Comput Struct Biotechnol J*. 2024;24:136-145.
21. European Commission. European Health Data Space Regulation (EHDS). Available at https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en. Accessed May 2, 2025.
22. Hussein R, Balaur I, Burmann A, et al. Getting ready for the European Health Data Space (EHDS): IDERHA's plan to align with the latest EHDS requirements for the secondary use of health data. *Open Res Eur*. 2024;4:160.
23. OHDSI Oncology Working Group. OnRamp OHDSI/OncologyWG Wiki. Available at https://github.com/OHDSI/OncologyWG/wiki/OnRamp. Accessed July 11, 2026.
24. European Medicines Agency. Methods — DARWIN EU: Data Analysis and Real World Interrogation Network. DARWIN EU® Coordination Centre. Available at https://darwin-eu.org/index.php/methods. Accessed May 2, 2025.
25. Carus J, Nürnberg S, Ückert F, Schlüter C, Bartels S. Mapping cancer registry data to the episode domain of the Observational Medical Outcomes Partnership Model (OMOP). *Appl Sci*. 2022;12(8):4010.
26. Belenkaya R, Gurley M, Dymshyts D, et al. Standardized observational cancer research using the OMOP CDM oncology module. *Stud Health Technol Inform*. 2019;264:1831-1832.
27. Belenkaya R, Gurley MJ, Golozar A, et al. Extending the OMOP Common Data Model and standardized vocabularies to support observational cancer research. *JCO Clin Cancer Inform*. 2021;5:12-20.
28. Observational Health Data Sciences and Informatics. OMOP Common Data Model v5.4. OHDSI. Available at https://ohdsi.github.io/CommonDataModel/index.html. Accessed January 14, 2026.
29. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2024;74(3):229-263.
30. van Swieten M. *IKNL/OMOPCancerIndicators v0.0.1*. Zenodo. Available at https://doi.org/10.5281/zenodo.18370294. Accessed January 25, 2026.
31. Blacketer C, Schuemie MJ, Moinat M, et al. Advancing real-world evidence through a federated health data network (EHDEN): descriptive study. *J Med Internet Res*. 2025;27:e74119.
32. Moncada-Torres A, Martin F, Sieswerda M, Van Soest J, Geleijnse G. VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange. *AMIA Annu Symp Proc*. 2021;2020:870-877.
33. BlueBerry Consortium. BlueBerry: co-creating a blueprint for building a sustainable, effective and scalable EURACAN rare cancer registry (2022—2024). Dutch Cancer Society (KWF)-funded infrastructure project. Available at https://www.fhi.no/en/projects/cancer/blueberry/. Accessed January 14, 2026.
34. IDEA4RC Consortium. IDEA4RC: Intelligent ecosystem to improve the governance, the sharing and the re-use of health data for rare cancers (2022—2026). Horizon Europe project 101057048. Available at https://www.idea4rc.eu. Accessed January 14, 2026.