frontiers | Frontiers in Oncology

TYPE Review
PUBLISHED 18 November 2025
DOI 10.3389/fonc.2025.1679408

OPEN ACCESS

EDITED BY
Dana Kristjansson,
Norwegian Institute of Public Health (NIPH), Norway

REVIEWED BY
Nina Van Goethem,
Sciensano, Belgium
Nicholas Nicholson,
Joint Research Centre, Italy

*CORRESPONDENCE
Claudine Backes
claudine.backes@lih.lu

†PRESENT ADDRESS
Pragathy Kannan,
Faculty of Medicine, University of Helsinki, Helsinki, Finland

RECEIVED 04 August 2025
ACCEPTED 27 October 2025
PUBLISHED 18 November 2025

CITATION
Lima B, Hasan F, Kannan P, Schnell M, Mafra A, Couffignal S and Backes C (2025) Optimizing data linkage for maximizing the potential of Luxembourg's national cancer registry: a comprehensive scoping review. *Front. Oncol.* 15:1679408. doi: 10.3389/fonc.2025.1679408

Optimizing data linkage for maximizing the potential of Luxembourg's national cancer registry: a comprehensive scoping review

Bruno Lima[1,2], Farah Hasan[2], Pragathy Kannan[2†], Michael Schnell[1,3], Allini Mafra[1,2], Sophie Couffignal[1,4] and Claudine Backes[1,2,4*]

[1]Registre National du Cancer du Luxembourg (RNC), Strassen, Luxembourg, [2]Cancer Epidemiology and Prevention Group (EPICAN), Department of Precision Health (DoPH), Luxembourg Institute of Health (LIH), Strassen Luxembourg, [3]Data Integration Center, Department of Medical Information (DMI), Luxembourg Institute of Health (LIH), Strassen, Luxembourg, [4]Public Health Expertise Unit (PHE), Luxembourg Institute of Health (LIH), Strassen, Luxembourg

Population-based cancer registries (PBCRs) provide international standardized indicators and evaluate public health actions and cancer care. Their research potential can be significantly enhanced through linkage with secondary data sources, such as biobanks, sociodemographic or genomic data. However, legal, ethical, and technical challenges often hinder such integration. This scoping review aims at identifying data linkage opportunities between cancer registries and secondary data sources, while describing the current state of the Luxembourg's National Cancer Registry (RNC). Ultimately, steps for linkages between cancer registries and biobanks and/or sociodemographic data are assessed to enhance cancer research and public health initiatives. A scoping review using PubMed and Embase databases was performed. English guidelines, reports, and qualitative and quantitative studies on hospital-based cancer registries, PBCRs, and site-specific registries were included. One thousand three hundred and twelve articles (n = 1312) were identified. After scanning titles and abstracts, 49 articles were examined for full-text reading, where fifteen articles met the inclusion criteria. Moreover, 13 articles were included following the snowball search approach (n = 28). Included articles report significant differences between countries in all avenues, including data availability and harmonization, confidentiality, access to data, exchange, and linkage methods. Results underline that PBCR's potential, efficiency, and cost-effectiveness are maximized thanks to linkage activities with secondary data sources such as biobanks or sociodemographic databases. In addition, the results of this scoping review enable the identification of key questions to address before establishing data linkage grouped into five domains being: (i) legal permission, (ii) data

availability assessment, (iii) data flow protocol, (iv) linkage key and (v) linkage method. In conclusion, addressing the five key domains identified in this review will support the development of robust, efficient, and ethically sound data linkage strategies, unlocking the full research potential of PBCRs and to aid decision making.

KEYWORDS

cancer, registries, population-based register, biological specimen banks, data linkage

Introduction

Cancer is a significant public health concern worldwide and has become a major barrier to increasing life expectancy (1, 2). In Europe, cancer ranks as the second leading cause of death across all age groups, resulting in substantial healthcare and social expenses (3, 4).

Luxembourg is one of the smallest European countries (672,050 inhabitants in 2024-01-01) and has the highest population growth rate (1.7%) in Europe, with foreigners representing almost half of the population (5). Cross-border workers commuting from neighboring countries (France, Belgium, and Germany) contribute to nearly half of the entire labor force. According to a 2023 Organization for Economic Co-operation and Development (OECD) report, more than one third of those covered by the Luxembourg's National Health Insurance Fund (35.8%) are cross-border employees (6). Moreover, in countries such as Luxembourg, where the age structure of the population has changed over the last decades, with notable increase in the proportions of individuals aged 40–64 and those aged 80 and above (5), the overall cancer burden has been steadily rising over time (7). Turning the tide against cancer and reducing its economic impact is a crucial objective in today's public health debate, as demonstrated by Europe's Beating Cancer Plan (8). The fight against cancer needs collaborative efforts involving numerous stakeholders, including clinicians, health workforces, researchers, epidemiologists, public health professionals, and policymakers, and requires accurate and comprehensive cancer data and registration (9).

Cancer registration is a continuous process of systematic, exhaustive, and non-redundant data collection, storage, analysis, interpretation, and reporting of cancer occurrences and characteristics (10). The World Health Organization (WHO) states that population-based cancer registries (PBCRs) are a core component of cancer control strategies and the gold standard for cancer surveillance in defined communities (11). Broadly speaking, population data must be collected in a uniform and systematic way, maintaining personalized information that will allow data from different sources to be consolidated into a record per patient to be followed over time. Furthermore, the usefulness of this data depends largely on its quality, i.e. its validity, comparability, timeliness and completeness (12, 13).

PBCRs are essential for describing cancer burden, examining cancer trends, and evaluating prevention measures and cancer care. For decades, PBCRs have been supporting countries in their actions against cancer by reflecting the most representative real-world cancer situation in a well-defined geolocation. They provide internationally standardized indicators, including incidence, prevalence, mortality, and survival rates and evaluate public health interventions such as prevention, screening, and quality of care (14). Systematic assessment of these cancer indicators provides reliable information for evidence-based science policies, support policies to minimize inequalities and improve healthcare (15). The International Agency for Research on Cancer (IARC) highlights the significance of PBCRs as crucial resources for developing and evaluating cancer control plans, as well as for enhancing epidemiological and clinical research and informing cancer-related health policies.

The Luxembourg's National Cancer Registry (in French: *Registre National du Cancer du Luxembourg*, RNC) is the official data collection of all new cancer cases diagnosed and/or treated in Luxembourg, for both residents and non-residents, in order to have an overview of all cancer patients being cared for in Luxembourg (Table 1). The RNC was established to monitor cancer incidence, mortality, and survival trends, to assess the effectiveness of treatments offered to patients, and to evaluate the efficiency of public health prevention efforts, screening programs, and to compare outcomes at international level (16). It also serves as an infrastructure to support epidemiological and clinical cancer research, aligning with national goals to enhance translational and precision cancer research (17). Created in 2013, the RNC is one of Europe's youngest PBCR. The RNC facilitates national and international comparisons and supports public authorities to plan and tailor health services aligned to the identified needs of the population.

Significant benefits and wider research scopes are obtained by linking PBCR data with appropriate secondary datasets such as biobanking data, socioeconomic data, or genomic datasets (18). In general, patients' data routinely collected for a variety of sources other than traditional clinical trials (Real-World Data RWD) can offer a significant potential for cancer research. These sources include: electronic records, administrative claims, disease registries, screening programs, vital statistics, or even wearable

TABLE 1 Example of data items collected by the RNC.

Data type	Examples of data items collected
Record Identification	- Registry Identification Number
Demographic data	- Age at diagnosis - Gender - Country of birth - Last known address
Death data	- Date of death - Cause of death - Autopsy
Tumor data	- Date of incidence - Topography (*ICD-O-3) - Morphology (ICD-O-3) - Basis of diagnosis - Clinical stages - Pathological stages - Metastases at time of diagnosis - Biopsy related data - Cytology related data - Histological prognostic factors - Tissue tumor markers and molecular alterations
Clinical data	- Circumstances of discovery - Comorbidity - Performance score (*ECOG)
Therapeutic management data	- Initial treatment - Surgery - Chemotherapy - Hormone therapy - Radiotherapy - Targeted therapy - Other treatments

*ICD-O-3, International classification of diseases for oncology (ICD-O) – 3rd edition, *ECOG, Eastern Cooperative Oncology Group.

digital health tools (19). As RWD can come from different sources and in different formats, it can also facilitate the characterization of health care provision, including health insurances, health providers and patients' characteristics (20).

These potential linkages enable a broader range of research questions and highly efficient use of PBCR data in cancer control and public health initiatives (21, 22). Some of the limitations of PBCR may include lack of information on longitudinal treatments, drug use, risk factors, comorbidities, quality of care and outcomes other than death. Linking cancer registries with other complementary data sources may provide a more comprehensive picture of disease development and management (23, 24).

The feasibility of data linkage relies on multiple factors including data availability, quality, and completeness of identifying information across data sources (25, 26). Data linkage methods are mainly categorized as deterministic or probabilistic, emphasizing the protection of patient privacy and ensuring compliance with confidentiality regulations (27).

In Luxembourg, potential secondary datasets may come from the Integrated Biobank of Luxembourg (IBBL) which offers biobanking services, including collecting, processing, analyzing, and storing biological samples and associated data (28), as well as the General Inspectorate of Social Security (IGSS), responsible for managing the national health insurance registry (29). Linking data from such sources with RNC may enhance cancer research, particularly in studies focusing on the socioeconomic characteristics of the population.

This study aims to identify opportunities for data linkage between cancer registries and secondary data sources, while outlining the current situation of the RNC. Ultimately, steps for linkages between cancer registries and biobanks and/or sociodemographic data are assessed to enhance cancer research and public health initiatives. To achieve this, a scoping review was conducted to explore published data linkage methodologies, ensuring data confidentiality, and to examine future opportunities for Luxembourg.

In our review, we focus on the questions:

a. What are the possibilities for data linkage between cancer registries and other data sources?
b. How RNC data linkage with other data sources would contribute to improving cancer research in Luxembourg?
c. What methodologies are available for carrying out data linkages ensuring data confidentiality?

Materials and methods

A scoping review: examining published methodologies

A scoping review using PubMed and Embase databases was performed utilizing the methodology developed by Arksey and O'Malley, and followed by the backward snowballing search approach in Google Scholar (30, 31). To structure our approach in addressing the research questions, we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) guidelines (32). Selection criteria were developed using the Population, Concept and Context approach as recommended by PRISMA (ScR). The population of this scoping review encompassed all cancer patients from all types of cancer registries (PBCR, hospital based cancer registries, and site specific registries). The main concept was data linkage performed between cancer registries and other data sources such as biobanks, sociodemographic surveys, census data or insurance claims. Our interest was primarily in data linkage methodologies and its feasibility on large scale population based studies. At last, for this review the context was to investigate which kind of data sources would be linkable to Luxembourg's RNC and how new linkages can be used to uncovering new insights for cancer research and enhancing public health interventions.

Therefore, the following algorithm was developed: ("Cancer registry" OR "Population-based cancer registry" OR "Cancer registries" OR "Population-based cancer registries") AND (Record link* OR Date link* [MeSH]) AND ((Biobanks OR biorepositories) OR (Sociodemographic OR "Census data")).

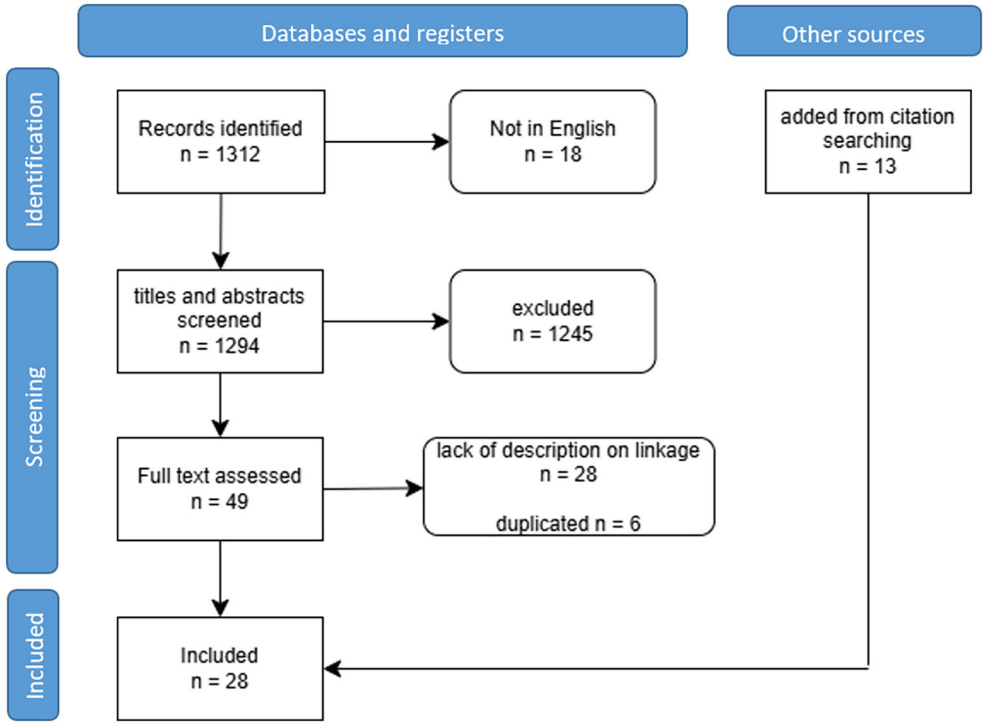

FIGURE 1
Flowchart of the scoping review process.

The same keywords were used for the backward snowballing approach to create a starter set for the search in Google Scholar. Afterwards, titles of articles from the reference lists were scanned, and pursued to assess the inclusion of relevant articles after full-text reading (30). We conduct a first search in June 2019 and an update in February 2025.

Inclusion criteria were restricted to English-language guidelines, reports, and qualitative and quantitative studies. Articles that did not describe data linkage were excluded. There were no restrictions imposed based on to geolocation and year.

For the selected studies, we extracted the information: authors, year, reference type, country of origin, data source, and title.

Results

In total, 1312 articles were initially identified, including 516 from PubMed and 796 from Embase. After scanning the titles and abstracts, 1245 articles were excluded, leaving 49 for full-text reading. Among these, six duplicate articles and 28 articles lacking data linkage descriptions were further excluded. Remaining fifteen articles (n=15) met the inclusion criteria (Figure 1).

Additionally, 13 articles were included by using the snowballing search approach, resulting in 28 articles identified (Supplementary Table). All 28 included articles involved cancer registries and examined data linkage at national or regional level. Studies were published between 1990 and 2022, and originated from various countries, including European countries (n=20) (33–47), the USA (n=6) (48–52), and Australia (n=2) (53) (Figure 2).

Among the included articles, 12 assessed data linkage between PBCRs and biobanks (33–35, 38, 40, 42–46, 49, 52), and fourteen explored data linkage between PBCRs and administrative, sociodemographic, and other health data (22, 36, 37, 41, 48, 50, 51, 53–59). Two articles described the linkage methodologies used, without specifying the source of the data (39, 47).

Findings were grouped into four themes: data availability and harmonization, addressing heterogeneity and interoperability; data confidentiality, covering legal and ethical requirements; data access and exchange, exploring centralized and federated models; and data linkage methods, outlining deterministic and probabilistic approaches. Within this framework, this scoping review aimed to understand how different studies have approached common obstacles for data linkage.

Data availability and harmonization

According to a large European study published in 2016, the ability to search for and get access to available data from samples stored at biobanks is essential to combine biomedical and clinical data (43). However, privacy, semantics (vocabularies), and technical heterogeneity challenges the search process for the sample's information. To formalize and evaluate a methodological

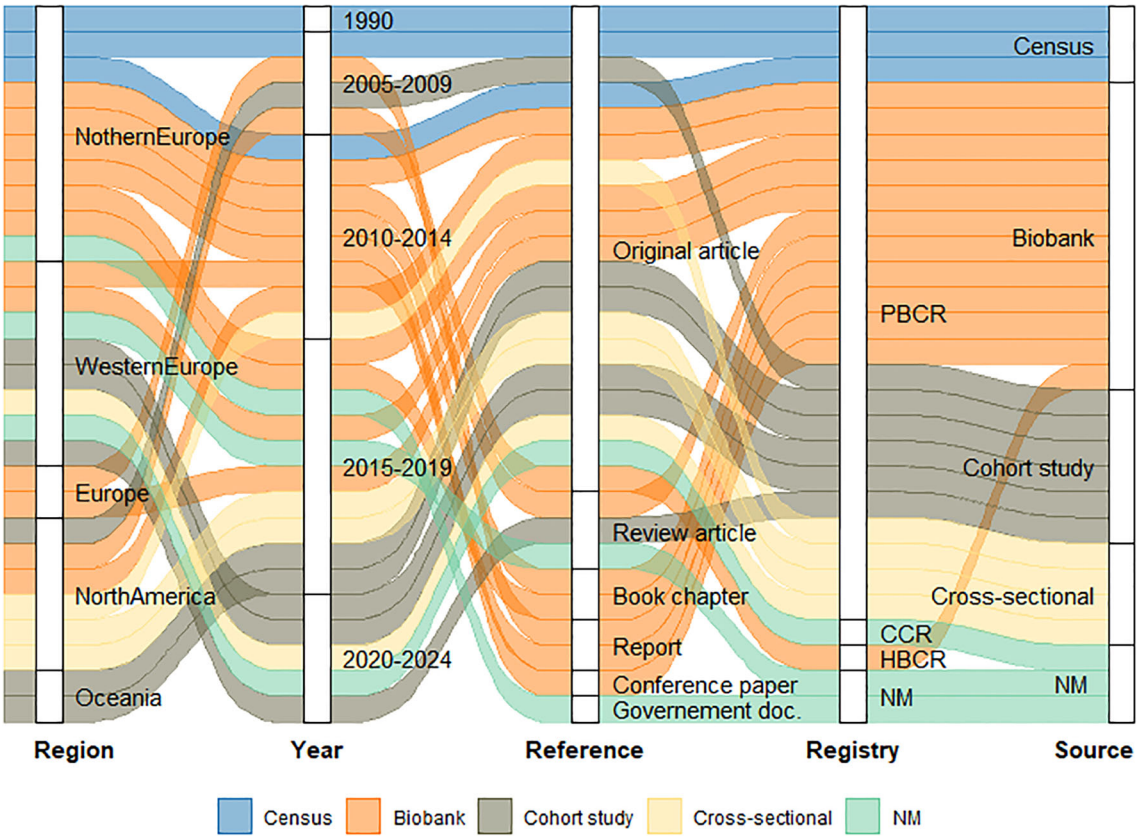

FIGURE 2
Scoping review results scheme. CCR – Clinical Cancer Registration; HBCR – Hospital Based Cancer Registry; PBCR – Population Based Cancer Registry; NM – Not Mentioned.

framework, the Sample avAILability (SAIL) method for data linkage between the Swedish biobank at Karolinska Institute and the Swedish national prostate cancer registry was used. SAIL operates on availability data (metadata; description of available data), which provides access to summary content at individual records level without disclosing its value, thus temporarily avoiding privacy issues. Retrospective data was harmonized by creating standardized vocabularies, data mapping, and integration. The SAIL web-based system provides an interface for the harmonization and submission of samples or phenotypic data. Furthermore, it is also a search tool to identify suitable cohort data for specific needs (46).

Similar challenges arising from datasets heterogeneity were described in a German study examining an established network between different medical institutions to facilitate translational cancer research architecture (47). To overcome these challenges, the study suggested collaborations among data sources to understand data elements, including vocabulary and semantics, and advocated for the creation of a centralized metadata repository (e.g., a data dictionary) accessible to all partners. This common framework was accompanied by tools supporting remote access and federated analysis, with limited data transfers and strict access control for local partners, promoting trust and data ownership.

Data confidentiality

In Europe, processing and sharing patient's data, with the exception of anonymized data, need to comply with national legal requirements and ethical guidelines, such as the European General Data Protection Regulation (GDPR) and country specific national laws (27, 47). In the United States, the privacy rules of the Health and Insurance Portability and Accountability Act (HIPAA) allow access to protected health information without patients permission if the provided information has been previously de-identified to prevent personal identifiers (52).

For biomedical research utilizing data from biobanks, related permissions from national data protection authorities, national or local ethical committees and from the biobank's organization are needed, often including the requirement of informed consent (35). However, basic consent standards and conditions vary widely depending on respective national laws. Legal requirements for PBCR linkage with sociodemographic data were only briefly discussed in included studies, but clearance by an ethics board was a prerequisite in all cases.

Patients' consent forms significantly differed between PBCR linkage protocols for biobanks or sociodemographic data.

The extent of patients informed consent could be defined as infinite, broad, or limited (60). Here are examples of consent practices observed in different countries:

1. According to the Belgian biobanks law, all in-patients are informed, thanks to the "hospital welcome brochure" about the potential use of their residual tissue for scientific research. Such a "presumed consent," turns into informed consent after explicit explanation and document signing (33).
2. According to the Estonian Human Genes Research Act, biobank participants sign a "broad consent," which permits to use collected samples for research, without previously identifying a specific project, and to retrieve additional participants' information from other databases (61). The "broad consent" was also been explored in Ireland. Following the recommendations of the Biobanks Irish Trust (BIT), the biobank consent was modified to include provisions for long-term storage and dissemination of both samples and de-identified data (38, 42).
3. According to the Finnish Personal Data Act, collecting and sharing health and social information are allowed only based on a patient's informed agreement unless the data is collected to be used in statistics, science, or historical research. Moreover, previously collected health data might be used for research without informed consent if the data is extensive or if seeking for informed consent is not possible. However, combining biological samples with registered data requires mandatory ethical board approval (35).
4. In the United States, linking sociodemographic data from the National Center of Health Statistics (NCHS) with cancer registry data requires approval from the NCHS ethics review board and the state's Department of Health Institutional Review Board (48). Similar requirements apply in Australia, where approval from the Population and Health Services Research Ethics Committee is mandatory to link cancer registry data with population sociodemographic data (53).

Data access and exchange

Data linkage is challenging in terms of data access and exchange due to confidentiality related restrictions. To facilitate data sharing both horizontal and vertical types of data integration can be employed (62). Horizontal integration includes data from different institutions that cover the same type of data:

- Interconnection between data sources was observed in Belgium for the virtual tumor bank, where all local biobanks were linked to a central database integrated within the Belgian National Cancer Registry. Biobanks export their local data in a standardized template to be uploaded in the central database. Upon request, the linkage with the PBCR is done by the cancer registry staff who has the authority to access that database and search across the biobanks datasets (33).
- In a study from California, PBCR data linkage with biobanks was conducted to establish a virtual biobank (49). Californian biobank data were extracted to Excel files and securely linked with the California Cancer Registry behind the cancer registry's firewalls.

In contrast, vertical integration combines different types of data:

- A decentralized system or federated system refers to the connection between heterogeneous databases from multiple sources into a unified framework, while the data is kept at its original source under the data owner's control, with the linkage occurring only when requested. This model was implemented in Estonia, where a national IT architecture was established, supported by a software called "X road". This open-source ecosystem solution enables unified and secured data exchange between different data services, through which registries and institutions can communicate safely (61). The German Cancer Consortium also explored a similar concept. Here an integration layer, "bridgehead", is provided for each data source to form a bridge between consortium members, thus creating a harmonized view of all data sources while data owners retain control over data access and are actively involved in all inquiries (47).
- A centralized model, named as data bank or data repository and analyzed in the United Kingdom, stores anonymized data from different sources in a single location. The Secure Anonymized Information Linkage (SAIL) Databank was established to store data from different health and non-health sources, with all data de-identified by a Trusted Third Party (TTP) - in this case, the National Health Service (NHS) (27). A linkage performed by a TTP enables the separation of the linkage and analysis processes, thereby ensuring patient privacy and enhancing data security (22).
- In other scenarios, exists a partnership agreement between cancer registries and biobanks (38, 42). For example, for data linkage between Northern Ireland Cancer Registry and the Northern Ireland Biobank, a full time staff member funded by the biobank is based at the cancer registry to facilitate data linkage procedures.
- For data linkage activities involving national statistics data, a Danish and a Lithuanian study report that the national sociodemographic data was preserved at the statistics governmental department and was not allowed to be exported out of its source (36, 37). The linkage process was mainly performed at the statistics bureau in a secure environment.

A main challenge in designing a data flow for linkage involving personal identifiers is to ensure that end users do not have access to these identifiers and that sensitive information is not transferred to institutions that also possess identifiable data, thereby mitigating

the risk of re-identification. However, the quality of the final data linkage remains a critical concern. It is therefore essential to determine the most appropriate format for sharing encrypted identifiers, recognizing that linkage based on non-numeric identifiers carries a higher risk of mismatches (57).

Data linkage methods

This scoping review identified the following data linkage methodologies, depending on the use of personal identifiable data to assess matching between pairs of records: deterministic record linkage (DRL); probabilistic record linkage (PRL); or a combination of both. PRL involves assigning weight record pairs and assessing the likelihood of these records representing a true match (45). This approach can be implemented sequentially, beginning with the specification of blocking variables, followed by the application of matching variables. Record pairs are retained only if their similarity score surpasses a predefined threshold (55).

The choice of the appropriate linkage method depends on the datasets, available variables and quality regarding accuracy, stability, and completeness (63). DRL uses a linkage key as a single reliable identifier [e.g., a unique personal identifier (34, 36, 41, 44) or Social Security Identification Number (SSIN) (33)] and here all linkage variables used are equally important. In case of absence, a linkage key can be created, by combining linking variables (e.g. first name, last name, date of birth, gender). The linkage key must be produced by an authorized staff from the data source or by a trusted third-party associate (52).

Several studies carried out a two-step data linkage approach combining DRL and PRL. For example, first is checked whether it is possible to make a direct match with an unique identifier (if it exists); otherwise a probabilistic linkage procedure is applied using variables such as name and date of birth (59). Other studies applied only PRL, using various software programs such as Link Plus, developed by the United States Center for Disease Control and Prevention (64), and Choicemaker (65), used by the Centre for Health Record Linkage in Wales, Australia (48, 51, 53, 54).

In Germany, cohort data are typically linked to cancer registry records using a pseudonymisation key based on name, sex, date of birth, and place of residence. A trusted center generates tokens from personal identifiers, which are then matched using probabilistic linkage. Only non-identifiable data are returned to the requesting institution. This probabilistic linkage process is subject to two main types of error: homonym errors, where records from two different individuals are erroneously linked, and synonym errors, where records belonging to the same individual are incorrectly treated as separate entities (56).

Another German study is also worth mentioning. This study employed a DRL using common indirect identifiers between insurance claims and cancer registries. This approach was found to be less expensive and faster than usual PRL. The authors concluded that, although it is possible to use a deterministic linkage with indirect identifiers, the sensitivity of this method is very low and recommended using standard probabilistic methods instead (58).

Key questions for PBCR data linkage with secondary data sources

Based on the results of this scoping review, key questions were identified as crucial for establishing data linkage between PBCR and secondary data sources. These questions are categorized into five domains: legal permission, data availability, dataflow protocol, linkage key, and linkage method (Figure 3).

Domain 1 - Legal permission: Understanding the legal framework and obtaining the necessary permissions and approvals for data linkage activities are crucial to ensure compliance with regulations and protect individuals' privacy.

Following questions are formulated:

- Do data providers have the permission to link their data?
- What approval processes are required for data linkage activities?

Domain 2 - Data availability assessment: Assessing the availability and accessibility of data in secondary sources is essential to ascertain the feasibility of data linkage and to identify potential data gaps that may impact the research outcomes.

Following questions are formulated:

- Is a general description of the data available to assess the feasibility of the linkage process? What is the quality of available data? Are datasets interoperable?
- Are data protection, data quality and data use assured and clear for all parties?

Domain 3 - Data flow protocol: Developing a robust data flow protocol is needed to establish efficient and secure processes for data exchange between PBCRs and secondary sources, which will ensure data integrity and confidentiality throughout the linkage process.

Following questions are formulated:

- Is it possible to export and transfer data at an individual level to external data sources? How will the dataflow be performed?
- Is virtual data transfer an option? Does the process need an intermediary acting as a trusted third party?

Domain 4 -Linkage key: Choosing an appropriate linkage key, whether it is a unique identifier or a combination of variables, is critical for accurately matching records between PBCRs and secondary sources, ensuring reliable linkage results.

Following questions are formulated:

- Is there a common unique identifier available for use, or should be created a linkage key?

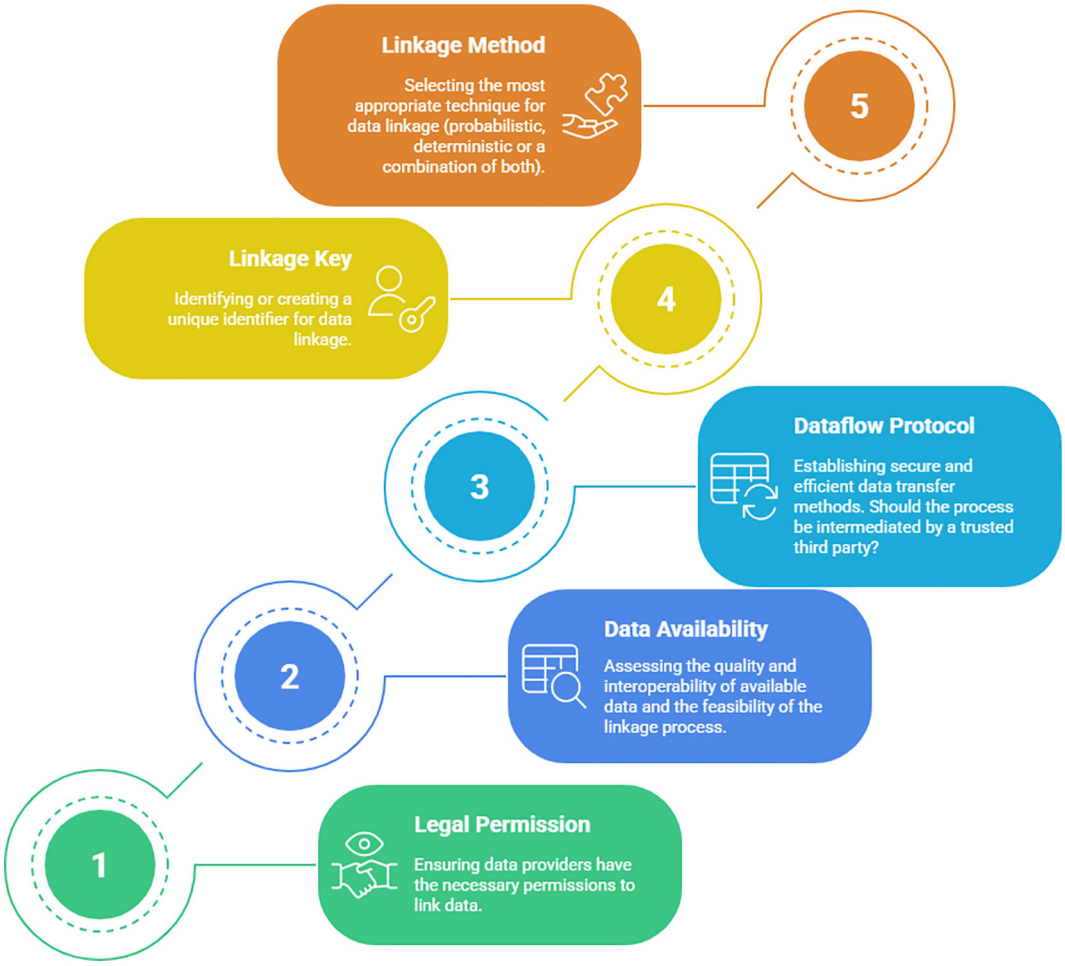

FIGURE 3
Five key domains and related questions to be considered, step by step (1 to 5), prior to establishing data linkage.

- Does the process require the involvement of a trusted third party?

Domain 5- Linkage method: Picking the most suitable linkage method, such as DRL, PRL, or a combination of both, depends on the quality and characteristics of the available datasets.

Following questions are formulated:

- Which linking technique (probabilistic, deterministic, or a combination of both) is most suited to apply, given the data quality, researcher's goals, and available resources?

Discussion

To the best of our knowledge, this is the first scoping review examining the procedures and methodologies used to link cancer registry data with secondary databases (respective biobank and administrative data sources). Evidence gathered from this scoping review suggests that main challenges in data linkage include high cost processes, privacy issues, dataset heterogeneity, and data quality. Five key domains and their related questions were identified as critical areas requiring clarification before establishing data linkage: legal permission, data availability assessment, dataflow protocol, linkage key, and linkage method.

The earliest study found in the literature review dates back to 1990, indicating that data linkage in the health sector is not a recently discovered activity (37). In this scoping review, we observed that performed linkage steps were rarely described in detail. Studies exploring cancer registry data linkage with biobanks were more extensively analyzed compared to those linking cancer registry data with sociodemographic, administrative, and health data. It is evident from this scoping review that linkage activities and procedures need to be adapted to the context of each country. Methodologies employed differ depending on available resources, the country's regulations, data quality, and the linkage's purpose.

According to a Nordic study, the high costs associated with the data linkage process are related to technical setups, data protection procedures (such as consent and ethics committees), or developing

key linkage in the lack of a unique national Personal Identity Code (PIC) (60). It was reported that the cost of linking records without PICs can be up to 50 times more expensive than linking records with PICs (35). Data confidentiality should be aligned with local regulations, and participant consent must be obtained when relevant and practical. In addition, a trusted third party can be involved in removing personal identifiers from respective datasets. To overcome dataset heterogeneity, a federated architecture was suggested as a successful approach to facilitate data linkage between heterogeneous data sources in Estonian and German studies (44, 47). However, according to a Swedish study, federated systems proved expensive and frequently impractical due to differences in underlying medical protocols and standard operating procedures (43).

To facilitate data linkage between national and international institutions and between heterogeneous data sources, international efforts and collaborations are essential. PBCRs are multiple source systems that collect all cancer cases in a well-defined geographical area. For decades, they have worked within strong national and international networks, conducting research and using international standards such as the International Classification of Diseases for Oncology (66) and the TNM classification (67), to guarantee quality indicators of completeness, comparability, validity and timeliness (12, 13). However, health data collected from secondary sources may use different standards, definitions, and levels of expertise, posing challenges for its integration. To address this, one of the initiatives to promote large-scale harmonization of health data is being led by the Observational Health Data Sciences and Informatics (OHDSI) consortium (68). Their goal is to facilitate the access to and analysis of health data. For that, they have created the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), a model that aims to standardize the representation of data (format) and its content (terminologies, vocabularies, coding schemes). This model has the capacity to accommodate data from diverse sources such as administrative claims, registries or electronic health records (69). The OHDSI's OMOP CDM aims to serve as a foundation for federated analytics and to support collaborative research. Supported by the European Health Data & Evidence Network (EHDEN) (70), the RNC and multiple registries and organizations across Europe, as well as others in the United States (US), have explored the OHDSI's OMOP CDM for a minimal set of variables and assessed its potential to enhance interoperability and support data sharing. Implementing this model may be challenging for some cancer registries with limited IT infrastructure, human resources, or established data-sharing frameworks. Moreover, European and US PBCRs have tested OMOP-CDM for cancer data, reporting improved interoperability but also a loss of data granularity that may limit clinical research. Despite these challenges, the OHDSI's OMOP CDM ecosystem has proven to be a successful support for cancer research, particularly in large scale collaborative studies (71).

Building on such models can significantly advance precision oncology by enabling the use of high-resolution cancer data to improve diagnosis, treatment, and outcomes. The upcoming European Health Data Space (EHDS) regulation further supports this vision by establishing a unified framework for electronic health data sharing and reuse across the European Union (72), positioning OHDSI's OMOP CDM initiative as a potential solution for adoption by PBCRs regarding the secondary use of data. However, it may also encourage PBCRs to embrace a different data modelling standard, such as openEHR (73) to support clinical care setting (i.e. primary use of data). Given that the EHDS regulation seeks to enable the reuse of certain data for purposes of public interest and scientific research, while also fostering a dedicated environment for health data within a unified market for digital health products and services, it will promote streamlined and harmonized data linkage within a robust regulatory environment.

Some countries have already introduced legislation to facilitate data linkage by enabling the usage of a PIC. The Nordic Occupational Cancer Study (NOCCA) is a prime example of a high-impact cohort study that collaborates across Nordic countries and links census data with PBCR data, transcending national boundaries to advance larger-scale research. This cohort study provides comprehensive insights into cancer incidence spanning up to 45 years, with a focus on occupational categories within Nordic populations. The study achieved this by linking individual records extracted from census data using the PIC utilized in all Nordic countries (41, 74).

In addition, the EUROCOURSE (Europe against Cancer: Optimization of Use of Registries for Scientific Excellence in Research) project has put forth guidelines, specifically within work package 7, to facilitate the linkage between cancer registries and biobanks. These guidelines are particularly valuable for translational cancer epidemiology and clinical research (75). To enable international operations, the European bio-banking platform (BBMRI) and EUROCOURSE have collaborated in developing a standardized minimum dataset for linking biobanks and cancer registries. This strategy advocates for a cost-effective and relatively uncomplicated approach that can be carried out annually while meeting the scientific expectations of researches (76). This work was pursued by the iPAAC (Innovative Partnership for Action Against Cancer), including in work package 7 the aim to advance PBCRs information to better support evidence-based cancer surveillance and care (77).

In Luxembourg, the second Plan National du Cancer (PNC2) (78) prioritizes, among other objectives, the digitalization and interoperability of health data, the expansion and integration of oncology data systems, the organization of oncology services into specialized competence networks, and the advancement of translational cancer research. In alignment with these objectives, and supported by the PNC2, the RELIANCE study, or "REaL-life cANCEr epidemiology to identify risk factors for cancer with a particular focus on prevention and care" (79), aims to evaluate cancer epidemiology for the first time in Luxembourg using longitudinal population data from the RNC. In addition, the RELIANCE study will investigate a range of research questions using RNC data, as well as exploring potential secondary data linkages with the RNC data. While the study is likely to expand to include different cancer sites in future, the first pilot study (RELIANCE – Breast Cancer) aims to evaluate breast cancer epidemiology in the Grand Duchy of Luxembourg.

Evidence selection bias may have been introduced because of the decision to restrict the study's search strategy to the PubMed and Embase databases and use a backward snowballing approach on Google Scholar (80). However, the scoping review only examined the linkage between PBCRs and biobanks and/or administrative or sociodemographic databases. This provides valuable insights into the linkage activities with respective sources. Although, a wider systematic review is required, including other potential secondary data sources (e.g., electronic health records, institutional or organizational databases, and several others).

Nevertheless, this study has laid the foundation for formulating a more targeted research question that emphasizes the linkage between PBCRs and prospective secondary data sources. This focus is particularly valuable for population-based studies, as it enables researchers to prioritize and explore the linkage activities and potential benefits associated with combining PBCR data with other datasets.

Perspectives and conclusions

The current state of cancer prevention, treatment, and care calls for a renewed commitment and adaptation to the rapid progress in oncology (51). The European Action Plan Against Cancer emphasizes the importance of addressing the entire cancer pathway. In this context, the RNC explores and evaluate existing secondary data sources and provides innovative solutions that meet the current cancer-related needs of the population. Moreover, the RNC actively support novel cancer research through digitalization, innovation, and interprofessional collaboration, to contribute to the development and implementation of comprehensive cancer control initiatives (81).

To enhance cancer research and innovation, RNC explores data linkage with suitable secondary data sources. The reviewed literature demonstrated the viability of achieving such a linkage when key questions are solved. These linkage efforts enhance the potential of cancer registries and biobank or socioeconomic databases by widening their research opportunities. Furthermore, to overcome the heterogeneity in datasets, RNC has applied OHDSI's OMOP CDM and successfully started the RELIANCE study. While data linkage can be performed independently of any data model, CDM improves data interoperability and facilitates the linkage process with other data sources in the same format. Potentially, this may lead to the use of clinical and epidemiological cancer data to improve patient outcomes both nationally and internationally.

PBCRs often face constraints in the range of data items collected for hypothesis-driven research, and collecting new variables can be time-consuming, costly, and susceptible to bias (82). Therefore, a promising avenue for future research involves linking PBCR data with population-based health surveys such as EHES-LUX, ORISCAV-LUX (waves 1 and 2) or similar surveys (51, 83, 84). Another illustrative example come from a Norwegian study that successfully demonstrated the linkage of a PBCR with a health interview survey to investigate cardiovascular risk factors in different cancer types (85). Linking cancer registry data with health interview surveys enables clinical cancer registry variables to be evaluated alongside individual risk factors, socioeconomic position, screening behaviors, and healthcare utilization variables, which are typically unavailable in cancer registries. Further research is needed to assess the feasibility and true potential of such linkages, but the prospects are promising, providing ample scope for future investigations. To move beyond *ad hoc* and project based initiatives, RNC data linkages should be formalized as part of a long-term national strategy. However, achieving this requires dedicated funding, clear governance structures and standardized procedures to ensure continuity, interoperability, and scalability over time.

The quality of PBCRs' data is generally assessed across four dimensions (12, 13). Comparability refers to the extent to which coding, classification, data recording and reporting definitions comply with international guidelines. Validity is defined as the proportion of cases in the registry that actually have a given characteristic. Timeliness measures how long it takes for registry information to be made available to professionals and researchers. Completeness is defined as the extent to which all incident cancers occurring in the population are included in the registry database. Linked datasets must also be evaluated according to these criteria to ensure that linkage itself does not compromise data integrity neither the quality of the PBCR data in any of its dimensions.

Strengthening linkage activities of PBCRs with secondary sources offers substantial benefits, enabling a more comprehensive exploration of cancer prevention, diagnosis, treatment, and prognostic factors. The findings of this study offer valuable insights into key questions that need to be addressed before establishing data linkage between PBCRs and secondary sources.

By addressing the key questions identified herein and considering the five aforementioned domains, researchers and policymakers can establish robust and effective data linkage strategies, thereby unlocking the potential of PBCRs and facilitating valuable research on cancer-related topics.

Author contributions

BL: Writing – original draft, Writing – review & editing, Formal analysis, Validation, Visualization. FH: Conceptualization, Formal analysis, Validation, Visualization, Writing – original draft, Writing – review & editing. PK: Validation, Writing – review & editing. MS: Validation, Writing – review & editing. AM: Formal analysis, Validation, Visualization, Writing – review & editing. SC: Validation, Writing – review & editing. CB: Conceptualization, Formal analysis, Funding acquisition, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research and/or publication of this article. This works was partially

funded by the Department of Precision Health of the Luxembourg Institute of Health as well as by Luxembourg's Health Directorate (Direction de la santé).

Acknowledgments

The authors warmly thank Coralie Dessens, Documentation Officer at the Luxembourg Institute of Health, for her excellence and support in producing this scoping review. Additionally, we would like to thank the Department of Precision Health of the LIH, as well as Luxembourg's Health Directorate (Direction de la santé) for funding parts of this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Generative AI was used in the creation of this manuscript. The authors used the generative AI tool NapkinAI (https://www.napkin.ai/) to assist in the creation of Figure 2. The tool was used to generate a conceptual visual representation based on the authors' input. The authors reviewed and verified the accuracy and appropriateness of the figure in relation to the manuscript content.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2025.1679408/full#supplementary-material

References

1. Jemal A, Torre LA, Soerjomataram I, Bray F. The burden of cancer. In: *The Cancer Atlas Third ed*. American Cancer Society, Inc, 250 Williams Street, Atlanta, Georgia 30303 USA (2019). p. 36–62.

2. Jemal A, Torre LA. *The Global Burden of Cancer*. Hoboken, New Jersey, USA: The American Cancer Society's Principles of Oncology (2018) p. 33–44.

3. Causes of death statistics: Major causes of death in the EU in 2022: EUROSTAT; Statistics Explained(2022). Available online at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Causes_of_death_statistics (Accessed March 5, 2025).

4. Dyba T, Randi G, Bray F, Martos C, Giusti F, Nicholson N, et al. The European cancer burden in 2020: Incidence and mortality estimates for 40 countries and 25 major cancers. *Eur J Cancer*. (2021) 157:308–47. doi: 10.1016/j.ejca.2021.07.039

5. Peltier FK,C. *La Demographie Luxembourgeoise en Chiffres*. Belvaux, Luxembourg: STATEC (2024).

6. OECD. *Luxembourg: Country Health Profile*. Paris, France: OECD Publishing. (2023).

7. Bray F, Soerjomataram I. The changing global burden of cancer: transitions in human development and implications for cancer prevention and control. *Cancer: Dis Control Priorities*. (2015) 3:23–44. doi: 10.1596/978-1-4648-0349-9_ch2

8. European Commission. *Europe's Beating Cancer Plan: A New EU Approach to Prevention, Treatment and Care*. Brussels, Belgium: European Commission (2021).

9. Beaupré M. *Le fichier des tumeurs du Quebec: un outil pour soutenir la surveillance du cancer et la recherche*. Québec City, Québec, Canada: Bulletin Epidémiologique Hebdomadaire (2006) p. 40–41,:302-5.

10. ECIS-European Cancer Information System. *Population-based cancer registries* (2025). European Union. Available online at: https://ecis.jrc.ec.europa.eu (Accessed March 5, 2025).

11. Bray F, Znaor A, Cueva P, Korir A, Swaminathan R, Ullrich A, et al. *Planning and Developing Population-Based Cancer Registration in Low- or Middle-Income Settings*. Lyon, France: IARC Technical Publications: International Agency for Research on Cancer, Lyon (FR (2014).

12. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. *Eur J Cancer*. (2009) 45:747–55. doi: 10.1016/j.ejca.2008.11.032

13. Parkin DM, Bray F. Evaluation of data quality in the cancer registry: principles and methods Part II. *Completeness Eur J Cancer*. (2009) 45:756–64. doi: 10.1016/j.ejca.2008.11.033

14. Silva I. *The role of cancer registries. Cancer Epidemiology: Principles and Methods*. Lyon: International Agency for Research on Cancer(IARC (1999) p. 385–404.

15. Smith TG, Dunn ME, Levin KY, Tsakraklides SP, Mitchell SA, van de Poll-Franse LV, et al. Cancer survivor perspectives on sharing patient-generated health data with central cancer registries. *Qual Life Res*. (2019) 28:2957–67. doi: 10.1007/s11136-019-02263-0

16. Registre National du Cancer Luxembourg (RNC). *Brochure patient (Patient leaflet)* (2020). Luxembourg Institute of Health (LIH. Available online at: https://www.rnc.lu/ (Accessed March 5, 2025).

17. Norgaard M, Skriver MV, Gregersen H, Pedersen G, Schonheyder HC, Sorensen HT. The data quality of haematological Malignancy ICD-10 diagnoses in a population-based hospital discharge registry. *Eur J Cancer Prev*. (2005) 14:201–6. doi: 10.1097/00008469-200506000-00002

18. Lix L, Smith M, Pitz M, Ahmed R, Quon H, Griffith J, et al. *Cancer Data Linkage in Manitoba: Expanding the Infrastructure for Research*. Canada: Faculty of Health Sciences, University of Manitoba (2016).

19. Wilson BE, Booth CM. Real-world data: bridging the gap between clinical trials and practice. *EClinicalMedicine*. (2024) 78:102915. doi: 10.1016/j.eclinm.2024.102915

20. Penberthy LT, Rivera DR, Lund JL, Bruno MA, Meyer AM. An overview of real-world data sources for oncology and considerations for research. *CA Cancer J Clin*. (2022) 72:287–300. doi: 10.3322/caac.21714

21. Parkin DM. The role of cancer registries in cancer control. *Int J Clin Oncol*. (2008) 13:102–11. doi: 10.1007/s10147-008-0762-6

22. 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) 22:124. doi: 10.1186/s12874-022-01601-0

23. Pearson C, Fraser J, Peake M, Valori R, Poirier V, Coupland VH, et al. Establishing population-based surveillance of diagnostic timeliness using linked cancer registry and administrative data for patients with colorectal and lung cancer. *Cancer Epidemiol*. (2019) 61:111–8. doi: 10.1016/j.canep.2019.05.010

24. Tham N, Skandarajah A, Hayes IP. Colorectal cancer databases and registries in Australia: what data is available? *ANZ J Surg*. (2022) 92:27–33. doi: 10.1111/ans.17221

25. Dusetzina SB TS, Meyer AM, Meyer A, Green L, Carpenter WR. *Linking Data for Health Services Research: A Framework and Instructional Guide [Internet]*. Rockville (MD: Agency for Healthcare Research and Quality (US (2014).

26. Newcombe HB, Smith ME, Howe GR, Mingay J, Strugnell A, Abbatt JD. Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers. *Comput Biol Med*. (1983) 13:157–69. doi: 10.1016/S0010-4825(83)80011-0

27. Jones k, Ford D. *Privacy, confidentiality and practicalities in data linkage*. Statistical N, editor. UK: Government Statistical Service (2018). p. 36.

28. Integrated Biobank of Luxembourg (IBBL) Luxembourg. Luxembourg institute of health(2022). Available online at: https://www.lih.lu/en/translational-medicine/translational-medicine-operations-hub/integrated-biobank-of-Luxembourg-ibbl/ (Accessed March 5, 2025).

29. Inspection générale de la sécurité sociale. *Luxembourg Microdata Platform on Labour and Social Protection: Le gouvernement Luxembourgeois* (2019). Available online at: https://igss.gouvernement.lu/fr/microdata-platform.html (Accessed March 5, 2025).

30. Wohlin C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. Association for Computing Machinery, London, England, United Kingdom (2014). p. Article 38.

31. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol*. (2005) 8:19–32. doi: 10.1080/1364557032000119616

32. Tricco A, Lillie EZW, O'Brien K, Colquhoun H, Levac D, Moher D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Internal Med*. (2018) 169:467–73. doi: 10.7326/M18-0850

33. Vande Loock K, van der Stock E, Debucquoy A, Emmerechts K, Van Damme N, Marbaix E. The Belgian virtual tumorbank: A tool for translational cancer research. *Front Med (Lausanne)*. (2019) 6:120. doi: 10.3389/fmed.2019.00120

34. Langseth H, Luostarinen T, Bray F, Dillner J. Ensuring quality in studies linking cancer registries and biobanks. *Acta Oncol*. (2010) 49:368–77. doi: 10.3109/02841860903447069

35. Pukkala E. *Biobanks and registers in epidemiologic research on cancer*. New York, USA: Methods in Biobanking (2011) p. 127–64.

36. Smailyte G, Jasilionis D, Kaceniene A, Krilaviciute A, Ambrozaitiene D, Stankuniene V. Suicides among cancer patients in Lithuania: A population-based census-linked study. *Cancer Epidemiol*. (2013) 37:714–8. doi: 10.1016/j.canep.2013.05.009

37. Guenel P, Engholm G, Lynge E. Laryngeal cancer in Denmark: a nationwide longitudinal study based on register linkage data. *Occup Environ Med*. (1990) 47:473–9. doi: 10.1136/oem.47.7.473

38. Lewis C, McQuaid S, Clark P, Murray P, McGuigan T, Greene C, et al. The Northern Ireland Biobank: a cancer focused repository of science. *Open J Bioresour*. (2018) 5:1–6. doi: 10.5334/ojb.47

39. Jones K, Ford D. *Privacy, confidentiality and practicalities in data linkage*. London, United Kingdom: National Statistician's Quality Review into Privacy and Data Confidentiality Methods, Government Statistical Service (2018).

40. Zika E, Schulte In den Bäumen T, Kaye J, Brand A, Ibarreta D. Sample, data use and protection in biobanking in Europe: legal issues. *Pharmacogenomics*. (2008) 9 (6):773–81. doi: 10.2217/14622416.9.6.773

41. Pukkala E, Martinsen JI, Lynge E, Gunnarsdottir HK, Sparén P, Tryggvadottir L, et al. Occupation and cancer–follow-up of 15 million people in five Nordic countries. *Acta Oncol*. (2009) 48:646–790. doi: 10.1080/02841860902913546

42. Mee B, Gaffney E, Glynn SA, Donatello S, Carroll P, Connolly E, et al. Development and progress of Ireland's biobank network: Ethical, legal, and social implications (ELSI), standardized documentation, sample and data release, and international perspective. *Biopreservation Biobanking*. (2013) 11:3–11. doi: 10.1089/bio.2012.0028

43. Spjuth O, Krestyaninova M, Hastings J, Shen H-Y, Heikkinen J, Waldenberger M, et al. Harmonising and linking biomedical and clinical data across disparate data archives to enable integrative cross-biobank research. *Eur J Hum Genet*. (2016) 24:521–8. doi: 10.1038/ejhg.2015.165

44. Leitsalu L , Metspalu A. From biobanking to precision medicine: the Estonian experience. In: *Genomic Precis Med*. Amsterdam, Netherlands: Elsevier (2017) 119–29. doi: 10.1016/B978-0-12-800681-8.00008-6

45. Ariel A, de Groot M, van Grootheest G, van der Laan J, Smit J, Verkerk B, et al. *Record linkage in health data: a simulation study*. The Hague/Heerlen: Statistics Netherlands (2014).

46. Spjuth O, Heikkinen J, Litton J-E, Palmgren J, Krestyaninova M eds. (2014). Data integration between Swedish national clinical health registries and biobanks using an availability system, in: *Data Integration in the Life Sciences: 10th International Conference, DILS 2014*, Lisbon, Portugal, July 17-18, 2014 Proceedings 10. Berlin, Germany: Springer, Lecture Notes in Computer Science.

47. Lablans M, Schmidt EE, Ückert F. An architecture for translational cancer research as exemplified by the German Cancer Consortium. *JCO Clin Cancer Informatics*. (2018) 1:1–8. doi: 10.1200/CCI.17.00062

48. Miller EA, Miller DM, Judson DH, He Y, Day HR, Zevallos K, et al. Linkage of 1986–2009 national health interview survey with 1981–2010 Florida cancer data system. *Vital Health Stat 2*. (2014) 2014:1–16.

49. McCusker ME, Cress RD, Allen M, Fernandez-Ami A, Gandour-Edwards R. Feasibility of linking population-based cancer registries and cancer center biorepositories. *Biopreserv Biobank*. (2012) 10:416–20. doi: 10.1089/bio.2012.0014

50. Clegg LX, Reichman ME, Miller BA, Hankey BF, Singh GK, Lin YD, et al. Impact of socioeconomic status on cancer incidence and stage at diagnosis: selected findings from the surveillance, epidemiology, and end results: National Longitudinal Mortality Study. *Cancer Causes Control*. (2009) 20:417–35. doi: 10.1007/s10552-008-9256-0

51. McClure LA, Miller EA, Tannenbaum SL, Hernandez MN, MacKinnon JA, He Y, et al. Linking the national health interview survey with the Florida cancer data system: A pilot study. *J Registry Manag*. (2016) 43:16.

52. Dhir R, Patel AA, Winters S, Bisceglia M, Swanson D, Aamodt R, et al. A multidisciplinary approach to honest broker services for tissue banks and clinical data: a pragmatic and practical model. *Cancer: Interdiscip Int J Am Cancer Society*. (2008) 113:1705–15. doi: 10.1002/cncr.23768

53. Creighton N, Purdie S, Soeberg M, Walton R, Baker D, Young J. Self-selection in a population-based cohort study: impact on health service use and survival for bowel and lung cancer assessed using data linkage. *BMC Med Res Methodol*. (2018) 18:84. doi: 10.1186/s12874-018-0537-3

54. Subramaniam K, Ang PW, Neeman T, Fadia M, Taupin D. Post-colonoscopy colorectal cancers identified by probabilistic and deterministic linkage: results in an Australian prospective cohort. *BMJ Open*. (2019) 9:e026138. doi: 10.1136/bmjopen-2018-026138

55. Gallaway MS, Huang B, Chen Q, Tucker T, McDowell J, Durbin E, et al. Identifying smoking status and smoking cessation using a data linkage between the Kentucky cancer registry and health claims data. *JCO Clin Cancer Inform*. (2019) 3:1–8. doi: 10.1200/CCI.19.00011

56. Arndt V, Holleczek B, Kajuter H, Luttmann S, Nennecke A, Zeissig SR, et al. Data from population-based cancer registration for secondary data analysis: methodological challenges and perspectives. *Gesundheitswesen*. (2020) 82:S62–71. doi: 10.1055/a-1009-6466

57. Langner I, Riedel O, Czwikla J, Heinze F, Rothgang H, Zeeb H, et al. Linkage of routine data to other data sources in Germany: A practical example illustrating challenges and solutions. *Gesundheitswesen*. (2020) 82:S117–S21. doi: 10.1055/a-0999-5509

58. Kollhorst B, Reinders T, Grill S, Eberle A, Intemann T, Kieschke J, et al. Record linkage of claims and cancer registries data-Evaluation of a deterministic linkage approach based on indirect personal identifiers. *Pharmacoepidemiol Drug Saf*. (2022) 31:1287–93. doi: 10.1002/pds.5545

59. Scheel K, Franke T, Weikert A, Dick M, Walter A, Zeidler J, et al. Record linkage in clinical cancer registration: experiences and findings from lower saxony. *Stud Health Technol Inform*. (2021) 278:101–9. doi: 10.3233/SHTI210057

60. Zika E, Schulte In den Bäumen T, Kaye J, Brand A, Ibarreta D. Sample, data use and protection in biobanking in Europe: legal issues. *Pharmacogenomics*. (2008) 9:773–81. doi: 10.2217/14622416.9.6.773

61. Leitsalu L, Metspalu A. From biobanking to precision medicine. In: Ginsburg GS, Willard HF, editors. *Genomic and Precision Medicine*. Academic Press, Boston (2017). p. 119–29.

62. Mihaylov I, Kandula M, Krachunov M, Vassilev D. A novel framework for horizontal and vertical data integration in cancer studies with application to survival time prediction models. *Biol Direct*. (2019) 14:22. doi: 10.1186/s13062-019-0249-6

63. Collaboration in Research and Methodology for Offical Statisitcs (CROS). *Statistical Data Warehouse Design Manual*. Luxembourg City, Luxembourg: European Commission, Eurostat (2017).

64. Centers for Disease Control and Prevention. Link plus(2024). Available online at: https://www.cdc.gov/national-program-cancer-registries/registry-plus/link-plus.html?CDC_AAref_Val=https://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm (Accessed March 5, 2025).

65. Borthwick A, Buechi M, Goldberg A eds. *Key Concepts in the ChoiceMaker 2 Record Matching System. Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*. Whashington, DC: Association for Computing Machinery (ACM) (2003).

66. World Health Organization. *International Classification of Diseases for Oncology*. Geneva, Switzerland: WHO (2008).

67. Brierley J, Gospodarowicz M, Wittekind C. *TNM Classification of Malignant Tumours. 8th ed*. Hoboken, New Jersey, USA: Wilwey Blackwell (2016).

68. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform*. (2015) 216:574–8. doi: 10.3233/978-1-61499-564-7-574

69. Observational Health Data Sciences and Informatics. *OHDSI's OMOP Common Data Model* (2025). OHDSI. Available online at: https://www.ohdsi.org/data-standardization/ (Accessed March 5, 2025).

70. European Health Data & Evidence Network. EHDEN(2024). Available online at: https://www.ehden.eu/ (Accessed March 5, 2025).

71. Wang L, Wen A, Fu S, Ruan X, Huang M, Li R, et al. A scoping review of OMOP CDM adoption for cancer research using real world data. *NPJ Digit Med*. (2025) 8:189. doi: 10.1038/s41746-025-01581-7

72. European Health Data Space Regulation (EHDS). European Commission(2025). Available online at: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en (Accessed March 5, 2025).

73. Kalra D, Beale T, Heard S. The openEHR Foundation. *Stud Health Technol Inform*. (2005) 115:153–73.

74. Pukkala E, Martinsen JI, Weiderpass E, Kjaerheim K, Lynge E, Tryggvadottir L, et al. Cancer incidence among firefighters: 45 years of follow-up in five Nordic countries. *Occup Environ Med*. (2014) 71:398–404. doi: 10.1136/oemed-2013-101803

75. Schouten LJ, Straatman H, Kiemeney LA, Gimbrere CH, Verbeek AL. The capture-recapture method for estimation of cancer registry completeness: a useful tool? *Int J Epidemiol*. (1994) 23:1111–6. doi: 10.1093/ije/23.6.1111

76. Dillner J. A basis for translational cancer research on aetiology, pathogenesis and prognosis: Guideline for standardised and population-based linkages of biobanks to cancer registries. *Eur J Cancer*. (2015) 51:1018–27. doi: 10.1016/j.ejca.2013.10.007

77. iPAAC Innovative Partnership for action against cancer. Workpackage 7 – cancer information and registries(2021). Available online at: https://www.ipaac.eu/en/work-packages/wp7/ (Accessed March 5, 2025).

78. Ministère de la Santé et de la Sécurité Sociale. Plan National Cancer 2020-2024 (prolongé jusqu'en 2026)(2025). Available online at: https://santesecu.public.lu/fr/espace-professionnel/plans-nationaux/plan-national-cancer.html (Accessed March 5, 2025).

79. Luxembourg Institue of Health. RELIANCE - Breast Cancer pilot study: LIH (2025). Available online at: https://www.lih.lu/en/reliance-breast-cancer-study/ (Accessed March 5, 2025).

80. Drucker AM, Fleming P, Chan A-W. Research techniques made simple: assessing risk of bias in systematic reviews. *J Invest Dermatol*. (2016) 136:e109–e14. doi: 10.1016/j.jid.2016.08.021

81. European Commission. Europe's Beating Cancer Plan: A new EU approach to prevention, treatment and care. *Factsheet. [press release]*. Brussels, Belgium (2021) Available online at: https://ec.europa.eu/commission/presscorner/detail/en/ip_21_342 (Accessed May 3, 2025).

82. Fahey PP, Page A, Stone G, Astell-Burt T. Augmenting cancer registry data with health survey data with no cases in common: the relationship between pre-diagnosis health behaviour and post-diagnosis survival in oesophageal cancer. *BMC Cancer*. (2020) 20:496. doi: 10.1186/s12885-020-06990-3

83. Bocquet V, Barré J, Couffignal S, d'Incau M, Delagardelle C, Michel G, et al. Study design and characteristics of the Luxembourg European Health Examination Survey (EHES-LUX). *BMC Public Health*. (2018) 18:1169. doi: 10.1186/s12889-018-6087-0

84. Alkerwi A, Sauvageot N, Donneau AF, Lair ML, Couffignal S, Beissel J, et al. First nationwide survey on cardiovascular risk factors in Grand-Duchy of Luxembourg (ORISCAV-LUX). *BMC Public Health*. (2010) 10:468. doi: 10.1186/1471-2458-10-468

85. Robsahm TE, Falk RS, Heir T, Sandvik L, Vos L, Erikssen J, et al. Cardiorespiratory fitness and risk of site-specific cancers: a long-term prospective cohort study. *Cancer Med*. (2017) 6:865–73. doi: 10.1002/cam4.1043