# Data from Population-based Cancer Registration for Secondary Data Analysis: Methodological Challenges and Perspectives

## Daten bevölkerungsbezogener Krebsregister für die Sekundärdatenanalyse: Methodische Herausforderungen und Perspektiven


Authors

**Volker Arndt[1, 2], Bernd Holleczek[3], Hiltraud Kajüter[4], Sabine Luttmann[5], Alice Nennecke[6], Sylke Ruth Zeissig[7], Klaus Kraywinkel[8], Alexander Katalinic[9, 10, 11]**

Affiliations

1 Epidemiologisches Krebsregister Baden-Württemberg, Deutsches Krebsforschungszentrum, Heidelberg, Germany
2 Deutsches Krebsforschungszentrum, AG Cancer Survivorship, Abt. Klinische Epidemiologie und Alternsforschung, Heidelberg, Germany
3 Krebsregister Saarland, Saarbrücken, Germany
4 Epidemiologisches Krebsregister NRW gGmbH, Bochum, Germany
5 Leibniz-Institut für Präventionsforschung und Epidemiologie - BIPS GmbH, Bremer Krebsregister, Bremen, Germany
6 Behörde für Gesundheit und Verbraucherschutz (BGV), Hamburgisches Krebsregister, Hamburg, Germany
7 Krebsregister Rheinland-Pfalz gGmbH, Mainz, Germany
8 Robert Koch-Institut, Zentrum für Krebsregisterdaten (ZfKD), Berlin, Germany
9 Universität zu Lübeck, Krebsregister Schleswig-Holstein, Institut für Krebsepidemiologie e.V., Lübeck, Germany
10 Universität zu Lübeck, Institut für Sozialmedizin und Epidemiologie, Lübeck, Germany
11 Gesellschaft epidemiologischer Krebsregister in Deutschland e.V. (GEKID e.V.), GEKID Working Group, Lübeck, Germany

Key words

cancer registration, epidemiology, secondary data, data quality, data sources

Schlüsselwörter

Krebsregistrierung, Epidemiologie, Sekundärdaten, Datenqualität, Datenquellen

Bibliography

**DOI** https://doi.org/10.1055/a-1009-6466
Online-Publikation: 29.10.2019
Gesundheitswesen 2020; 82 (Suppl. 1): S62–S71

ISSN 0949-7013

Correspondence

PD Dr. Volker Arndt
Epidemiologisches Krebsregister Baden-Württemberg
Deutsches Krebsforschungszentrum
Im Neuenheimer Feld 280
69120 Heidelberg
Germany
v.arndt@dkfz-heidelberg.de

**ABSTRACT**

Population-based cancer registries have a long-standing role in cancer monitoring. Scientific use of cancer registry data is one important purpose of cancer registration, but use of cancer registry data is not restricted to cancer registries. Cancer registration in Germany is currently heading towards population-based collection of detailed clinical data. This development together with additional options for record linkage and long-term follow-up will offer new opportunities for health services and outcome research. Both regional population-based registries and the German Centre for Cancer Registry Data (ZfKD) at the Robert Koch-Institute as well as international cancer registries and consortia or organizations may provide external researchers access to individual or aggregate level data for secondary data analysis. In this review, we elaborate on the access to cancer registry data for research purposes, availability of specific data items, and options for data linkage with external data sources. We also discuss as well as on limitations in data availability and quality, and describe typical biases in design and analysis.

**ZUSAMMENFASSUNG**

Bevölkerungsbezogene Krebsregister haben eine entscheidende Rolle in der Krebsbekämpfung. Die wissenschaftliche Verwendung von Krebsregisterdaten ist allerdings nicht allein den Krebsregistern vorbehalten. Die Krebsregistrierung in Deutschland entwickelt sich aktuell von einer primär epidemiologischen Registrierung hin zu einer klinisch-epidemiologischen Registrierung mit Erfassung detaillierter klinischer Daten. Diese Entwicklung zusammen mit weiteren Optionen für die Verknüpfung von Datensätzen und einer künftigen langfristigen Nachbeobachtung bieten neue Möglichkeiten für die Versorgungs- und Gesundheitssystemforschung. Sowohl regionale bevölkerungsbezogene Krebsregister als auch das Zentrum für Krebsregisterdaten (ZfKD) am Robert Koch-Institut sowie internationale kollaborative Projekte und Dachorganisationen der Krebsregister können Wissenschaftlern Individualdaten oder aggregierte Daten für die Sekundärdatenanalyse bereitstellen. In dieser Übersichtsarbeit erläutern wir Zugangsmöglichkeiten zu Krebsregisterdaten, die Verfügbarkeit spezifischer Daten und Möglichkeiten der Verknüpfung der Registerdaten mit externen Datenquellen sowie mögliche Einschränkungen in Bezug auf Datenverfügbarkeit und Datenqualität wie auch typische Fehlerquellen bei Studiendesign und Datenanalyse.

## Introduction

Population-based cancer registries (PBCRs) have a long-standing and pivotal role in the fight against cancer. For many years, the focus of cancer registration had been on epidemiological surveillance including trends and forecasting of cancer burden and on comparative studies on the variation of incidence and survival rates in time and place. PBCRs also play a crucial role in formulating cancer control plans, as well as in monitoring their success [1].

Scientific use of data collected by PBCRs is not limited to the research staff of the PBCRs. PBCRs on the federal state level, the German Centre for Cancer Registry Data (ZfKD) at the Robert Koch Institute as well as international resources (details see below) may provide external researchers with individual or aggregate level data for secondary data analysis. In this review, we will elaborate on access to PBCR data for research purposes, availability of specific data items, and options for data linkage with external data sources. The use of PBCR data for research purposes requires particular caution, however. Potential limitations in data availability and aspects of data quality will be addressed as well as specific biases which should be considered by any user of cancer registration data.

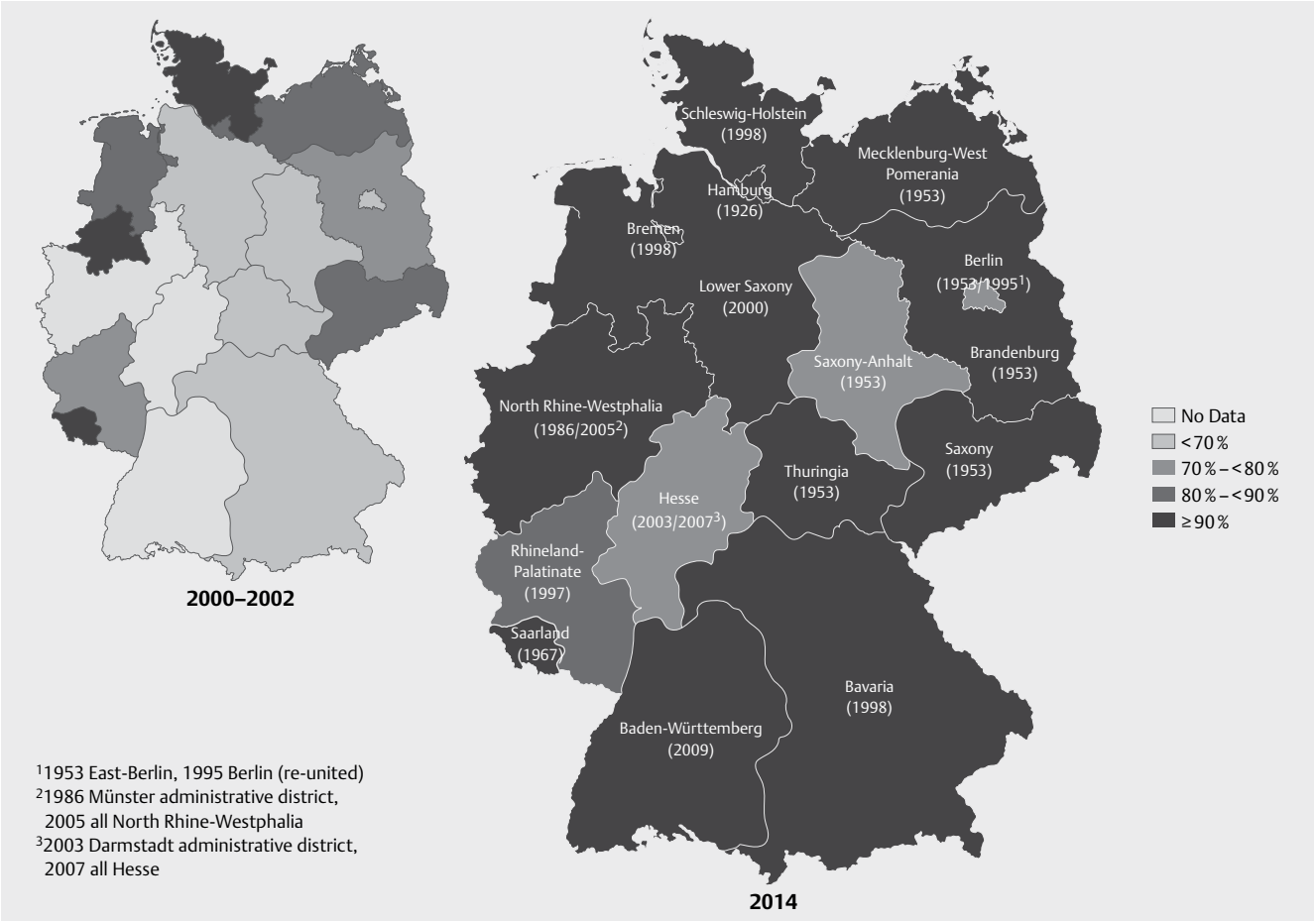
## Population-based Cancer Registration in Germany

One of the first PBCRs in the world was established in Hamburg in the late 1920s. It was for many years the only one in Germany. After registration in Hamburg was stopped during the turmoil of World War II, the National Cancer Registry of the German Democratic Republic (East Germany) started operation in 1953. More than a decade later, the Saarland Cancer Registry was established in 1967. In the late 1970s, the Hamburg Cancer Registry was critically reviewed due to the emerging discussion of data privacy issues and completeness of registration dropped. Therefore, the Saarland Cancer Registry became the only internationally acknowledged PBCR in West Germany for several years and its data were used to approximate cancer incidence in West Germany until the 1990s. The Hamburg Cancer Registry was reestablished in 1985 and the Münster registry in North Rhine-Westphalia was set up in 1986. The former National Cancer Registry (GKR) of the German Democratic Republic temporarily stopped operation due to changes in administration and legal basis during the process of the German reunification. The situation changed completely when a federal law on cancer registration (Krebsregistergesetz, KRG) [2] came into effect in 1995. All federal states were obligated by the KRG to set up PBCRs until 1999. As a result, coverage of PBCR constantly increased over the years and an increasing number of PBCRs has attained a high level of completeness (▶ **Fig. 1**). The Federal Cancer Register Data Act (Bundeskrebsregisterdatengesetz, BKRG), which came into force in 2009, obligated the federal states to ensure that the data from the PBCRs at the federal state level are collected comprehensively and transmitted in a uniform format to the Center for Cancer Registry Data (Zentrum für Krebsregisterdaten, ZfKD) at the Robert Koch-Institute [3]. The ZfKD compiles a record set including cancer data collected by all PBCRs in Germany, which is used for national cancer monitoring as well as for scientific purposes.

In April 2013, the cancer detection and registration law (Krebsfrüherkennungs- und -registergesetz, KFRG) [4] expanded the legal basis for cancer registration with respect to a nationwide uniform population-based clinical registration.

The consolidation of national data (e. g. German data pooled by the ZfKD) continues at the international level. At the European level, the European Network of Cancer Registries (ENCR) together with the Joint Research Center (JRC), a research institute of the European Commission, is responsible for joint evaluations. Data from all cancer registries worldwide that meet certain quality criteria are compiled by WHO's International Agency for Research on Cancer (IARC). IARC together with the International Association of Cancer Registries (IACR) publishes the report 'Cancer Incidence in Five Continents' (http://ci5.iarc.fr) every five years. In the first report of this series, which was published in 1966, Germany was represented by the Hamburg cancer registry only [5]. The national cancer registry of the former GDR was included in the second volume [6], the Saarland cancer registry in the third volume [7]. The current 11$^{th}$ volume covers the period 2008–2012 and lists nine registries from Germany (Bavaria, Bremen, Hamburg, Lower Saxony, Munich, North Rhine-Westphalia, Rhineland-Palatinate, Saarland, Schleswig-Holstein) [8].

In addition to the above-mentioned PBCRs, which collect cancer cases in all age-groups, the German Childhood Cancer Registry (GCCR) has to be mentioned. The GCCR was founded in 1980. It is hosted by the Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI) at the University Medical Center of the Johannes Gutenberg University Mainz. It registers cancer cases for all children




▶ **Fig. 1** Development of the estimated completeness of the population-based cancer registries in Germany 2000/2002 and 2014 by federal state or region (showing start of registration) [29].

under 15 years (since 2009: under 18 years) in all of Germany. Since 1991, children living in the area of the former GDR are included. The completeness of the GCCR for all over Germany is about 95 % and thus conforms to international requirements for a PBCR. About 1800 cases are reported every year from pediatric oncology units affiliated at the Society for Paediatric Oncology and Haematology.

The current structures, the beginning of nationwide population-based and clinical cancer registration, and the coverage rate of the PBCRs are shown in ▶ **Table 1**. It has to be mentioned that nowadays most PBCRs fulfill both population-based and clinical cancer registration (CCR) in one organization, so that the distinction between PBCR and CCR primarily reflects different kinds/perspectives of data usage rather than processes of cancer registration. However, this does not hold for hospital-based cancer registries, which only collect information on cancer patients treated in the institution concerned.

## What Data are Available in Population-based Cancer Registries in Germany?

Statewide operating PBCRs collect data on invasive malignant neoplasms, their preliminary stages as well as neoplasms of uncertain or unknown behavior and benign neoplasms of the central nervous system. Non-melanoma skin cancer is an exception because the registration of these tumors is not mandatory in all federal states and the amount of collected information varies across the states.

Traditionally, PBCRs collect data for patients living in the geographical catchment area of the PBCR, irrespective of place of diagnosis, treatment or death. The data contain personal identifiers, demographic items, information about the tumor including stage, site and extent of disease at the date of diagnosis as well as follow-up information of the patient including vital status and date and cause of death (▶ **Table 2**) [9, 10].

In contrast, new comprehensive clinical cancer registries according to the KFRG collect data of all patients treated in the assigned area of the CCR irrespective of place of residence [3]. In addition to the data items captured by PBCRs, these CCRs record information about the specific treatment provided, disease trajectories, and additional tumor specific items relevant for the classification of the tumor. The data collected by these comprehensive CCRs are defined by a common catalogue of items (so called 'ADT/GEKID Core data set' [ADT/GEKID-Basisdatensatz] and additional tumor specific modules) [11–14]. The collection of these items is mandatory for all CCRs in Germany, which operate on a statutory basis. As mentioned above, most PBCRs in Germany are now responsible for both population-based and clinical cancer registration.

▶ **Table 1** Structure of population-based and clinical cancer registration in Germany by state

| State | Structure of cancer registry | Initiation of registration * | | Population 2017 (Million) | Completeness * * (2013/14) |
|---|---|---|---|---|---|
| | | Population-based | Clinical | | |
| Baden-Württemberg | Integrated CCR/PBCR with distinct TC, RO, and PBCR | 2009 | 2009 | 11,0 | ≥ 90 % |
| Bavaria | Integrated CCR/PBCR based on 6 regional registration units | 1998 | 2017 | 13,0 | ≥ 90 % |
| Brandenburg | CCR (in conjunction with Berlin) based on several regional RO; PBCR via GKR | 1953 | 1953 | 2,5 | ≥ 90 % |
| Berlin | CCR (in conjunction with Brandenburg), PBCR via GKR | East: 1953<br>West:1995 | 1953<br>2016 | 3,6 | 70 % – < 80 % |
| Bremen | Integrated CCR/PBCR with consolidated TC and RO | 1998 | 2015 | 0,7 | ≥ 90 % |
| Hamburg | Integrated CCR/PBCR with consolidated TC and RO | 1926 | 2014 | 1,8 | ≥ 90 % |
| Hesse | Integrated CCR/ECT with TC and RO | 2007 | 2014 | 6,2 | 70 % – < 80 % |
| Mecklenburg-West Pomerania | CCR with TC and RO; PBCR via JCR/GKR | 1953 | 1953 | 1,6 | ≥ 90 % |
| Lower Saxony | Distinct CCR and PBCR with mutual TC | 2000 | 2017 | 8,0 | ≥ 90 % |
| North Rhine-Westphalia | Integrated CCR/PBCR with TC and RO | 2005 | 2016 | 17,9 | ≥ 90 % |
| Rhineland-Palatinate | Integrated CCR/PBCR with TC and RO | 1997 | 2016 | 4,1 | 80 % – < 90 % |
| Saarland | Integrated CCR/PBCR with TC and RO | 1967 | 2015 | 1,0 | ≥ 90 % |
| Saxony | State CCR based on 4 regional CCR; PBCR via GKR | 1953 | 1953 | 4,1 | ≥ 90 % |
| Saxony-Anhalt | State CCR based on 3 regional RO; PBCR via GKR | 1953 | 1953 | 2,2 | 70 % – < 80 % |
| Schleswig-Holstein | Integrated CCR/PBCR with TC and RO | 1998 | 2016 | 2,9 | ≥ 90 % |
| Thuringia | State CCR based on 5 regional RO; PBCR via GKR | 1953 | 1953 | 2,2 | ≥ 90 % |
| Joint Cancer Registry (JCR/GKR) of Berlin, Brandenburg, Mecklenburg-West Pomerania, Saxony, Saxony-Anhalt and Thuringia | PBCR for 6 federal states (former GDR) with TC and RO | 1953 | – | 16,2 | see state specific details |
| German Childhood Cancer Registry (GCCR) | Nationwide PBCR/CCR for children under 15 years (since 2009: under 18 years) in all of Germany | 1980 | 1980 | 13,5 | 95 % |

Abbreviations: CCR Clinical Cancer Registry; PBCR Population-based Cancer Registry; GDR German Democratic Republic; GKR Gemeinsames Krebsregister (Joint Cancer Registry of Berlin, Brandenburg, Mecklenburg-Vorpommern, Saxony, Saxony-Anhalt and Thuringia); RO Registration Office; TC Trust Center. * Start of statewide registration based on state specific legislation; * * Completeness based on estimates by German Centre for Cancer Registry Data (ZfKD) regarding all cancer sites combined (ICD-10: C00-C97 excl. C44)[29].

One of the major limitations of data currently collected by PBCRs is that information on comorbidities, important risk factors like smoking or occupational hazards, data on quality of life or socio-economic status are not routinely collected. This limitation is based on the fact that a valid and complete documentation is not feasible in daily clinical routine. Also, genetic information of the tumors or links to biobank data, which are of utmost interest in the era of personalized medicine and which might provide further insight in the tumorigenesis, are not collected. But linkage of PBCR data with data from other sources to address specific research questions is possible (details see below).

## Cancer Registration Data for Secondary Data Analysis

In principle, several different ways to access and use cancer registry data have to be distinguished:

1. Aggregated data
2. Anonymized individual data
3. Cohort linkage
4. Individual patient access

### Aggregated data

Individual data based on selected characteristics are aggregated into groups e. g. defined by year of incidence, age and sex. For the resulting groups, different indicators such as the number of people in the relevant group and the incidence rate may be reported. Aggregated data may be used for trend analysis, regional comparisons, e. g. based on regional socio-economic or geophysical indices. The only legal requirement is that no individual can be identified from the data set provided to the external researcher.

Aggregated national and international data on incidence, mortality, survival or prevalence are available on interactive websites (▶ **Table 3**). In addition, many state-level PBCRs listed in ▶ **Table 1** already provide aggregate level statistics on their websites. These resources usually include age-standardized and age-specific rates (in 5-year age groups) by sex, cancer, year, and geographical region (if applicable).

Provision of aggregated data usually follows a standard format (i. e. predefined age and tumor categories). For rare diagnoses, subtypes or tumor stage distribution, provision of aggregated data is possible on request. The individual PBCR or the ZfKD, respectively, approve specific data requests on the basis of feasibility, validity and completeness of the data, data privacy aspects, data thrift, and

▶ **Table 2** Individual level data items available in population-based and clinical cancer registries in Germany

| Data group | Items in PBCR and CCR | Items only in CCR |
|---|---|---|
| Identity data * | – Name<br>– Address<br>– Date of birth | – Health insurance number |
| Notification * | – Name of notifying institution/ professional<br>– Date of receipt<br>– Patients objection | |
| Demographic data | – Age<br>– Year of birth<br>– Area of residence | |
| Tumor data at time of diagnosis | – Date of diagnosis<br>– ICD-Code<br>– Topographie (ICD-O)<br>– Morphology (ICD-O)<br>– Grading<br>– Stage of disease (TNM, UICC, other classifications) | – Date of histological report<br>– Number of (sentinel-) lymph nodules explored and positive for cancer<br>– Distant metastasis (date and location)<br>– Performance status of the patient<br>– Additional markers for special cancer entities according to national guidelines (e. g. HER-2 status, hormon receptor status, tumor size in mm, Gleason-Score, KRAS-Gen and other) |
| Treatment data | – Operation (Yes/No)<br>– Chemotherapy (Yes/No)<br>– Radiotherapie (Yes/No)<br>– Hormone therapy (Yes/No)<br>– Immune therapy (Yes/No)<br>– Bone marrow transplantation (Yes/No) | – Kind of therapy<br>– Intention of therapy<br>– Relation to surgery<br>– Date of begin/end<br>– Complications/ side effects<br>– Surgery OPS-Code<br>– Radiotherapy target area, type of application, single and total radiation dose<br>– Chemotherapy protocol and substances,<br>– Additional items for special cancer entities according to national guidelines |
| Disease trajectory | – Vital status | – Residual status after primary therapy<br>– Disease status: progression/ recurrence of disease |
| Death | – Date of death<br>– Causes of death<br>– Death caused by tumor | |
| Other data | | – Tumor conference (date)<br>– Additional items for special cancer entities: social service contact, study participation |

PBCR = Population-based cancer registries. CCR = Clinical cancer registries. * Limited data access only under special preconditions (patients' consent). * * Active surveillance of patients only in some CCR.

required resources. Service fees may be charged depending on required resources. If more complex evaluations are required for scientific purposes, the option of a scientific collaboration project should be considered.

## Anonymized individual data

For specific purposes, anonymized individual data can be provided by PBCRs to external researchers. Individual features might be omitted or collapsed in broader categories in order to preclude re-identification of individual persons and to follow the concept of data thrift. Access to anonymized individual level data from state cancer registries and from the national data set is regulated in state-specific PBCR legislation and in the Federal Cancer Registry Data Act (BKRG), respectively. The list of available data is usually limited to general information about the individual (age, sex and place of residence), data on tumor diagnosis (ICD-10, tumor site, histology, tumor stage and classification, date of diagnosis, diagnostic basis), and survival time.

Requests for access to anonymized individual data should usually include the formulation of a research question, a description of the planned evaluation methods and a comprehensible listing of the required variables (under consideration of data thrift) and the definition of inclusion and exclusion criteria. In some states, the research projects have to fulfill certain requirements, e. g. the Hamburg Cancer Registry Act requires that the planned project contributes to the "improvement of cancer prevention or cancer control". Similar requirements, which explicitly demand a public interest in the anticipated research results, are also found in other state laws. Prior to the formal application, it is generally recommended to contact the registries or the ZfKD in advance in order to clarify specific details and the feasibility of the planned research project.

## Cohort linkage

The third possibility is to link individual data of an external cohort (e. g. diabetics enrolled in a statutory health insurance care plan) with the data stored in the cancer registry (e. g. [15]). Linkage for large scale cohort studies is usually applied on encrypted data using a pseudonymization key based on name, gender, date of birth, and place of residence. This method has been well established and was evaluated in the early 1990s [16, 17]. In general, the identity data of the study participants who have agreed to the cohort linkage are transmitted together with a study ID to the trust center of the

▶ **Table 3** International and national cancer registry resources for secondary data analysis (selection)

| | Access | Data base and data provider | Granularity of data | Population covered | Indicators | Covariates |
|---|---|---|---|---|---|---|
| **International data resources** | | | | | | |
| ECIS - European Cancer Information System (https://ecis.jrc.ec.europa.eu/) | interactive web-based platform | EU Joint Research Centre (JRC)<br>– ENCR-JRC<br>– Eurocare<br>– IARC CI5 | Aggregate level | Europe (40 countries) | – National Incidence and mortality estimates 2018<br>– Incidence and mortality historical data (by region/registry)<br>– Survival | Country, registry, cancer site, sex, age, year |
| Global Cancer Observatory<br>– Cancer Today<br>– Cancer Over Time<br>– Cancer Tomorrow<br>– Cancer Causes (http://gco.iarc.fr/) | interactive web-based platform | International Agency for Research on Cancer (IARC, WHO)<br>– GLOBOCAN<br>– Cancer Incidence in Five Continents (CI5)<br>– International Incidence of Childhood Cancer (IICC)<br>– Cancer Survival in Africa, Asia, the Caribbean and Central America (SurvCan) | Aggregate level | Global (185 countries) | – Incidence, mortality, and prevalence (national estimates)<br>– Trends in incidence and mortality (by registry)<br>– Predictions of future cancer incidence and mortality<br>– Estimated burden of cancer due to specific causes of cancer 2012 | Region, Country, cancer site, sex, age, year |
| NORDCAN Cancer Statistics for the Nordic Countries (http://www-dep.iarc.fr/NORDCAN/english/frame.asp) | interactive web-based platform | Association of Nordic Cancer Registries | Aggregate level | Denmark, Faroer Islands, Finland, Greenland, Iceland, Norway, Sweden | – Incidence<br>– Mortality<br>– Survival<br>– Prevalence | Region, Country, cancer site, sex, age, year |
| **National data resources** | | | | | | |
| German Centre for Cancer Registry Data (DE) (https://www.krebsdaten.de/) | interactive web-based platform | Zentrum für Krebsregisterdaten (ZfKD)<br>– All population-based cancer registries in Germany (see website for details) | Aggregate level | Germany (national estimates since 1998, nationwide coverage since 2009) | – Incidence<br>– Mortality<br>– Prevalence<br>– Survival | Cancer site, sex, age, year |
| | scientific use file (on request, earmarked) | | Individual level | | | See website for details |
| GEKID Atlas (http://atlas.gekid.de/CurrentVersion/atlas.html) | interactive web-based platform | Association of Population Based Cancer Registries in Germany (GEKID) | Aggregate level | Germany (federal state level, start of time series depending on deferral state) | – Incidence<br>– Mortality<br>– Survival | Cancer site, sex, age, year, state |
| Surveillance, Epidemiology, and End Results - SEER (USA) (https://seer.cancer.gov/) | interactive web-based platform | National Cancer Institute (USA)<br>– 9 registries (from 1973)<br>– 13 registries (from 1992)<br>– 18 registries (from 2000) | Aggregate level | USA (since 1973, partial population coverage 9%-35%) | – Incidence<br>– Mortality<br>– Survival<br>– Risk of Developing/ Dying | Cancer site, sex, age, race, ethnicity, stage[1] |
| | scientific use file (on request) | | Individual level | | | See website for details |
| SEER-Medicare Linked Database (https://healthcaredelivery.cancer.gov/seermedicare/) | scientific use file (on request, earmarked) | National Cancer Institute (USA)<br>– SEER Data<br>– Medicare Enrollment & Claims Data | Individual level | USA (since 1991) | – Incidence<br>– Survival<br>– Treatment data | See https://www.resdac.org/ |
| NICER – CH (www.nicer.org) | interactive web-based platform | National Institute for Cancer Epidemiology and Registration (NICER)<br>– All population-based cancer registries in Switzerland (see website for details) | Aggregate level | Switzerland (since 1980, population coverage 41–88%) | – Incidence<br>– Mortality<br>– Prevalence<br>– Survival | Canton, cancer site, sex, age, year, (histology and stage on request) |
| | specific statistics (on request) | | Aggregate level | | | |
| | scientific use file (on request, earmarked) | | Individual level | | | |

[1] Stage specific results are restricted to incidence and survival statistics.

PBCR. The trust center converts the identity data into unique tokens ("pseudonymization") similar to the procedure applied for routine case notifications. These tokens are then compared with the tokens already present in the registry using probabilistic linkage. In case of matching tokens, the corresponding internal registry case number and the study ID are transmitted to the registration office. After successful linkage with the desired information from the cancer registry, the data are transmitted to the research institution without disclosing the identity of the patients.

This type of probabilistic record linkage is not error free; and so called 'homonym errors' (linkage of data pertaining to two distinct case) and 'synonym errors' (disaggregation of data pertaining to one case) may occur [18, 19]. Nevertheless, an evaluation study showed that this technique is able to process large amounts of data with very high quality of record linkage [20].

Deterministic record linkage via social security number is a standard method in studies using data from Scandinavian registries but is not yet established in Germany. However, this will also be possible in Germany in future years for around 90 % of the population as cancer registries are now required (by KFRG) to record the health insurance number of patients enrolled in statutory health insurance plans. The quality of these numbers requires attention as mistyping due in case of manual entry may occur.

A special case of cohort linkage arises in the evaluation of organized screening programs [21]. Here, participants in screening programs will be linked on the basis of legal regulations with data from the cancer registry to identify interval cancers (i. e. discovered between two screening rounds).

Provision of individual data for specific individuals in the context of cohort studies ('cohort linkage') usually requires the corresponding consent of the individual and approval by ethics committee. Individual patient consent should ideally be obtained in advance by the study in question at recruitment or during follow-up. Some states (e. g. Rhineland-Palatinate) offer the possibility of cohort linkage without the necessity of individual consent under the condition of strict separation of personal identifiers and medical data. In some registries (e. g. Baden-Württemberg) no informed consent is required if only information on date and cause of death is transmitted. Cohort linkage is only possible at PBCRs but not at the ZfKD. As a consequence, a nationwide cohort linkage requires separate linkages with all PBCRs in Germany.

### Individual patient access

Direct (by name) access to individual data stored in the registers, e. g. for interviews or examinations of patients as part of a case-control study, is only possible with informed consent of the patient in accordance with the state-specific regulations of the registry and under the precondition, that the patient did not veto the storage of the encrypted name beforehand.

For studies requiring individual patient access, e. g. for recruiting cases within a case-control study, approval by the responsible ethics committee and/or advisory board, the cancer registry or its regulatory body (government department) is required. In most federal states, after approval of the study protocol, patients will be initially contacted by cancer registry or the reporting physicians to obtain patients' consent to be contacted by the research group for this specific research project. Once this has taken place, the names and addresses of the patients can be forwarded to the research group for further contact. Again, the state- specific requirements may vary, e. g. the cancer registration act of North Rhine-Westphalia requires that the financing of the planned project has to be secured and disclosed before access to individual patients can be granted.

## Data Quality Aspects

The value of the cancer registration data to contribute to cancer control and research relies heavily on the underlying quality of its data and the quality control procedures in place. Key aspects of data quality with respect to cancer registration data include **comparability**, **validity, timeliness**, and **completeness** [22, 23].

**Comparability** is the extent to which coding and classification procedures at a registry, together with the definitions of recording and reporting specific data items, adhere to agreed international guidelines [22]. A basic requirement is the standardization of practices concerning classification and coding of new cases, and consistency in basic definitions of incidence, such as rules for the recording and reporting of multiple primary cancers occurring in the same individual [9, 10].

**Validity** is defined as the proportion of cases in the registry with a given characteristic (e. g. cancer site, or age) which truly have this attribute [24]. Validity is examined via numerical indices which are either compared with other registries, or, within a registry, over time, or with respect to specified subsets of cases [22]. Common indicators of validity are the percentage of histologically verified (HV) cases as statement on the diagnostic accuracy, the proportion of DCO (case information based on 'death certificate only'), and the percentage of unknown or ill-defined primary site (PSU) due to missing information. Of high priority are continuous internal consistency checks that look for impossible codes and implausible combinations of different variables in the same record. Plausibility testing according to international and national rules [10, 25, 26] is routinely performed by means of electronical processes as well as manual editing.

**Timeliness** with regard to cancer registration defines the time span between the notifiable disease event and its publication within health reporting. There are no international guidelines for timeliness at present [22], but a recent survey among European cancer registries indicated wide variation regarding latency for completing 1 year of case ascertainment and releasing data to the public [27]. Physicians in Germany are required to submit their case notifications to the corresponding CCR within a range of four weeks to 6 months, depending state specific regulations. The CCRs are itself required to process and check all new incoming case notifications within a maximum of 6 weeks. By end of 2017, 6 out of 17 clinical cancer registries fulfilled this criteria [28].

According to the Federal Cancer Registry Data Act [3], German cancer registries have to submit their data to ZfKD within 2 calendar years after the year of diagnoses. Another aspect of actuality involves the completion of mortality follow-up. The processing of death certificates provided by local health authorities and the collection of vital status information from the residents' registration office may require up to two years.

**Completeness** is usually defined as the extent to which all of the incident cancers occurring in the population are included in the registry database [23]. It is also called **completeness of coverage** in order to distinguish it from **item completeness,** which corresponds to the availability of complete information regarding specific data item such as information regarding stage, treatment etc.

**Completeness of coverage** of German PBCRs is regularly assessed by the ZfKD. For this purpose, log-linear models are used assuming a largely constant mortality to incidence ratio (M: I) for certain cancers, age groups, sex, and calendar years across Germany. The estimated degree is equal to the ratio of observed and expected case figures accumulated across all age groups. The results are published within the joint publication of "Cancer in Germany" [29] by GEKID and the ZfKD, with 90 % being rated as sufficiently complete.

A high level of **item completeness (data accuracy)** means that all notifiable details concerning a case and its course of disease are recorded. Availability and completeness of the collected data depend on the 'age' of the particular PBCR. As described earlier, most German PBCRs were set up around the turn of the millennium. Yet for some regions, continuous data are available for longer time periods. Currently most of the German PBCRs present highly satisfactory information content on personal data, diagnoses, and vital status. On the other hand, the proportion of sufficient stage information varies widely between registries, cancer types, and years of diagnosis. As statewide clinical cancer registration was initiated in 2013, most CCRs are still in the build-up phase and have not yet attained sufficient item completeness regarding certain clinical features such as treatment information and course of disease [28]. Standard techniques and criteria regarding target values for certain variables are currently discussed. When planning a research project, the requestor should contact the CCR beforehand in order to clarify for which time period is sufficiently complete clinical data is available. In some regions, e. g. in Bavaria or Brandenburg, CCRs

▶ **Table 4** Examples of potential biases in the use of cancer registration data

| Bias | Definition (Description) | Example | Strategies to minimize bias |
|---|---|---|---|
| Incidence – prevalence bias (or Neyman Bias or „Survival Bias") | Occurs when the estimation of the risk of a disease is made by using data collected at a given point in time in a series of survivors (rather than being based on data collected during a time period). → Selection bias where the very sick or very well (or both) are erroneously excluded from a study. | Return to work rate in long-term cancer survivors [32] | Careful selection of study type: Prospective cohort studies rather than cross-sectional or case-control studies. |
| Confounding by indication | Occurs when the indication to treat is a confounder for the treatment-outcome relationship. | Assessing treatment effects in older breast cancer patients [33] | Controlling statistically for known confounders related to the indication. |
| Immortal time bias | Occurs in cohort studies when a period of 'immortal time' (time during which death or an outcome that determines end of follow-up cannot occur) is excluded from the analysis, e. g. the start of follow-up for the group receiving "new" treatment is defined by the start of the new treatment, which is later than that for the comparison group. | Beta blockers and cancer prognosis [34] | Using time-fixed analysis with exclusion of immortal time and adjustment for confounders at baseline and/or during follow-up periods. |
| Index event bias (syn. "collider stratification bias") | Risk of disease sequelae may be affected when multiple risk factors for sequelae (e. g. mortality) are also risk factors for having the disease in the first place ("collider"). | Obesity paradox in survival after cancer diagnosis [35] | Avoid stratification/ control of a collider. |
| Lead-time bias | Overestimation of survival time, due to the backward shift in the starting point for measuring survival when follow-up of groups does not begin at comparable stages in the natural history of a condition. | Apparent increase in survival of patients if screening merely advances diagnosis without increasing chances of cure [36] | Use mortality as outcome of interest. |
| Length-time bias | Apparent survival advantage of screen-detected cases due to an oversampling of slowly growing tumors by screening, particularly in case of long screening intervals. | Apparent superior overall survival in women with asymptomatic recurrence of early stage endometrial cancer compared to symptomatic recurrence [37] | Count all outcomes in each group regardless of method of detection. |
| Compliance bias | Compliant patients tend to have better prognosis regardless of screening and might be overrepresented in screening programs. | Assessment of overdiagnosis due to mammography screening by comparing attendees of a screening program with those not attending [38–39] | Compare outcome in RCT with control group and group offered screening. |
| Completeness and registration bias | Significant differences in patient, tumor and/or treatment related characteristics between registered and non-registered patients lead to incomparable results. | Assessment of completeness and registration bias in a cancer cohort with participation on a voluntary basis [40] | Registration on a mandatory basis, quality assessment of completeness of registered data. |
| Information bias due to record linkage errors | Errors in the record linkage:<br>– Homonym error (linkage of data pertaining to two distinct case)<br>– Synonym error (disaggregation of data pertaining to one case) | Cohort linkage, comparison of incidence and survival rates between registries [41] | Exclude registries with high DCO rates or extreme incidence rates. |

were already established in the 1990s, so that longer time series of clinical data might be available.

Evaluation of data quality prior to analysis is a critical issue for sound analysis, unbiased results, and meaningful interpretation. Guidelines for the use of data quality criteria and their reporting have been exemplarily described by the GEKID Cancer Survival Group [30, 31].

In addition to data quality issues, bias in design and analysis might also imperil proper analysis and interpretation of cancer registration data. ▶ **Table 4** lists some potential specific and general examples of biases which may occur in the use of cancer registration data. This list is neither exhaustive nor specific to the use of cancer registration data in secondary data analysis. It reflects real-world examples from the literature or from previous data requests.

## Summary and Outlook

Cancer registries are a vital source of information on cancer epidemiology and cancer care. As cancer registration in Germany is moving towards population-based registration of detailed clinical data, evaluation of quality of care is becoming a new task. This development, together with additional options for record linkage (e. g. via health insurance number) and long-term follow-up, will offer new opportunities for health services and outcome research. Information coming from genome-wide association studies are currently not included in PBCRs but genomic data will play a pivotal role in future understanding of carcinogenesis, drug effectiveness, drug resistance, and occurrence of side effects. Regulations are required on how to enable data collection of cancer patients' genetic profiles and/or linking with available biobanks. In summary, data from German PBCRs have a high potential for scientific use and the options for their use within secondary data analysis have substantially improved in recent years.

### Conflict of Interest

The authors declare that they have no conflict of interest.

### References


[1] Parkin DM. The role of cancer registries in cancer control. Int J Clin Oncol 2008; 13: 102–111. doi:10.1007/s10147-008-0762-6

[2] Krebsregistergesetz (KRG) Bundesgesetzblatt Teil I 1994; 79: 3351–3355

[3] Bundeskrebsregisterdatengesetz (BKRG) Bundesgesetzblatt Teil I 2009; 53: 2702–2707

[4] Krebsfrüherkennungs- und -registergesetz (KFRG) Bundesgesetzblatt Teil I 2013; 16: 617–623

[5] Doll R, Payne P, Waterhouse JAH. Cancer Incidence in Five Continents, Vol. I. Geneva: Union Internationale Contre le Cancer; 1966

[6] Doll R, Muir CS, Waterhouse JAH. Cancer Incidence in Five Continents, Vol. II. Geneva: Union Internationale Contre le Cancer; 1970

[7] Waterhouse J, Muir CS, Correa P et al. eds. (1976). Cancer Incidence in Five Continents, Vol. III. Lyon: IARC; 1976

[8] Bray F, Colombet M, Mery L et al. Cancer Incidence in Five Continents, Vol. XI (electronic version) URL http://ci5.iarc.fr Accessed: 16.06.2019

[9] Hentschel S, Katalinic A. Das Manual der epidemiologischen Krebsregistrierung. München, Wien, New York: W. Zuckschwerdt Verlag; 2008

[10] Stegmaier C, Hentschel S, Hofstädter F et al. Das Manual der Krebsregistrierung. Germering/München: W. Zuckschwerdt Verlag; 2019

[11] Arbeitsgemeinschaft Deutscher Tumorzentren e.V (ADT), Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. (GEKID). Einheitlicher onkologischer Basisdatensatz. BAnz 2014 AT 28.04.2014 B2 Anlage

[12] Arbeitsgemeinschaft Deutscher Tumorzentren e.V (ADT), Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. (GEKID). Organspezifisches Modul Mammakarzinom zum einheitlichen onkologischen Basisdatensatz von ADT/GEKID. BAnz 2015 AT 26.11.2015 B1 Anlage 1

[13] Arbeitsgemeinschaft Deutscher Tumorzentren e.V (ADT), Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. (GEKID). Organspezifisches Modul Kolorektales Karzinom zum einheitlichen onkologischen Basisdatensatz von ADT/GEKID. BAnz 2015 AT 26.11.2015 B1 Anlage 2

[14] Arbeitsgemeinschaft Deutscher Tumorzentren e.V (ADT), Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. (GEKID). Organspezifisches Modul Prostatakarzinom zum einheitlichen onkologischen Basisdatensatz von ADT/GEKID. BAnz 2015 AT 29.08.2017 B6 Anlage 1

[15] Kajüter H, Geier AS, Wellmann I et al. Kohortenstudie zur Krebsinzidenz bei Patienten mit Diabetes mellitus Typ 2. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2014; 57: 52–59. doi:10.1007/s00103-013-1880-5

[16] Pommerening K, Miller M, Schmidtmann I et al. Pseudonyms for cancer registries. Methods Inf Med 1996; 35: 112–121

[17] Michaelis J, Miller M, Pommerening K et al. A new concept to ensure data privacy and data security in cancer registries. Medinfo 1995; 8: Pt 1 661–665

[18] Brenner H, Schmidtmann I, Stegmaier C. Effects of record linkage errors on registry-based follow-up studies. Stat Med 1997; 16: 2633–2643

[19] Brenner H, Schmidtmann I. Effects of record linkage errors on disease registration. Methods Inf Med 1998; 37: 69–74

[20] Schmidtmann I, Sariyar M, Borg A et al. Quality of record linkage in a highly automated cancer registry that relies on encrypted identity data. GMS Med Inform Biom Epidemiol 2016; 12: Doc02. doi:10.3205/mibe000164

[21] Fuhs A, Bartholomaus S, Heidinger O et al. Evaluation der Auswirkungen des Mammographie-Screening-Programms auf die Brustkrebsmortalitat : Machbarkeitsstudie zur Verknüpfung verschiedener Datenquellen in Nordrhein-Westfalen. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2014; 57: 60–67. doi:10.1007/s00103-013-1870-7

[22] Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: Comparability, validity and timeliness. European Journal of Cancer 2009; 45: 747–755. doi:10.1016/j.ejca.2008.11.032

[23] Parkin DM, Bray F. Evaluation of data quality in the cancer registry: Principles and methods Part II. Completeness. European Journal of Cancer 2009; 45: 756–764. doi:10.1016/j.ejca.2008.11.033

[24] Parkin DM, Chen VW, Ferlay J et al. Comparability and quality in cancer registration. IARC Technical Report No. 19; Lyon: 1994

[25] Martos C, Crocetti E, Visser O et al. A proposal on cancer data quality checks: One common procedure for European cancer registries – version 1.1. Luxembourg: Publications Office of the European Union; 2018

[26] Ferlay J, Burkhard C, Whelan S et al. Check and conversion programs for cancer registries (IARC/IACR Tools). IARC Technical Report No. 42; Lyon: 2005

[27] Zanetti R, Schmidtmann I, Sacchetto L et al. Completeness and timeliness: Cancer registries could/should improve their performance. European Journal of Cancer 2015; 51: 1091–1098. doi:10.1016/j.ejca.2013.11.040

[28] Hölterhoff M, Löffler L, Ludwig L et al. Stand der klinischen Krebsregistrierung - Ergebnisse der Überprüfung der Förderkriterien zum 31.12.2017: Prognos AG 2018

[29] Robert Koch-Institut & Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V Krebs in Deutschland 2013/2014. Berlin 2017, doi:10.17886/rkipubl-2017-007

[30] Nennecke A, Brenner H, Eberle A et al. Überlebenschancen von Krebspatienten in Deutschland – auf dem Weg zu repräsentativen, vergleichbaren Aussagen. Gesundheitswesen 2010; 72: 692–699. doi:10.1055/s-0029-1242772

[31] Nennecke A, Barnes B, Brenner H et al. Datenqualität oder Unterschiede in der onkologischen Versorgung? - Berichtsstandards für Überlebenszeitanalysen mit Krebsregisterdaten. Gesundheitswesen 2013; 75: 94–98. doi:10.1055/s-0032-1311622

[32] Arndt V, Koch-Gallenkamp L, Bertram H et al. Return to work after cancer. A multi-regional population-based study from Germany. Acta Oncol 2019; 58: 811–818. doi:10.1080/0284186X.2018.1557341

[33] de Glas NA, Kiderlen M, de Craen AJ et al. Assessing treatment effects in older breast cancer patients: systematic review of observational research methods. Cancer Treatment Reviews 2015; 41: 254–261. doi:10.1016/j.ctrv.2014.12.014

[34] Weberpals J, Jansen L, Carr PR et al. Beta blockers and cancer prognosis - The role of immortal time bias: A systematic review and meta-analysis. Cancer Treatment Reviews 2016; 47: 1–11. doi:10.1016/j.ctrv.2016.04.004

[35] Mayeda ER, Glymour MM. The obesity paradox in survival after cancer diagnosis: tools for evaluation of potential bias. Cancer Epidemiol Biomarkers Prev 2017; 26: 17–20. doi:10.1158/1055-9965.EPI-16-0559

[36] Knight K, Oliphant R, Maxwell F et al. Colorectal cancer in the elderly and the influence of lead time bias: better survival does not equate with improved life expectancy. Int J Colorectal Dis 2016; 31: 553–559. doi:10.1007/s00384-015-2496-z

[37] Jeppesen MM, Mogensen O, Hansen DG et al. Detection of recurrence in early stage endometrial cancer - the role of symptoms and routine follow-up. Acta Oncol 2017; 56: 262–269. doi:10.1080/028418 6X.2016.1267396

[38] Kalager M, Loberg M, Fonnebo VM et al. Failure to account for selection-bias. Int J Cancer 2013; 133: 2751–2753. doi:10.1002/ijc.28244

[39] Falk RS, Hofvind S, Skaane P et al. Overdiagnosis among women attending a population-based mammography screening program. Int J Cancer 2013; 133: 705–712. doi:10.1002/ijc.28052

[40] Jegou D, Penninckx F, Vandendael T et al. Completeness and registration bias in PROCARE, a Belgian multidisciplinary project on cancer of the rectum with participation on a voluntary basis. European Journal of Cancer 2015; 51: 1099–1108. doi:10.1016/j.ejca.2014.02.025

[41] Andersen MR, Storm HH.Eurocourse Work Package G Cancer registration, public health and the reform of the European data protection framework: Abandoning or improving European public health research? European journal of cancer 2015; 51: 1028–1038. doi:10.1016/j.ejca.2013.09.005