Z. Evid. Fortbild. Qual. Gesundh. wesen (ZEFQ) xxx (xxxx) xxx

Contents lists available at ScienceDirect

Z. Evid. Fortbild. Qual. Gesundh. wesen (ZEFQ)

journal homepage: http://www.elsevier.com/locate/zefq

Versorgungsforschung / Health Services Research

Routine practice data of three cancer entities: Comparison among cancer registry and health insurance data

Versorgungsnahe Daten dreier Krebsentitäten: Vergleich zwischen Krebsregisterdaten und Krankenversicherungsdaten

Lisa M. Lang [a], Christian Behr [b], Marion Ludwig [a,*], Jochen Walker [a], Hans Christian Lange [b], Frederike Basedow [a], Christina Justenhoven [b]

[a] *InGef – Institute for Applied Health Research Berlin GmbH, Berlin, Germany*
[b] *Cancer Registry of Rhineland-Palatinate gGmbH, Mainz, Germany*

ARTICLE INFO

Article History:
Received: 18 October 2022
Received in revised form: 12 December 2022
Accepted: 7 January 2023
Available online: xxxx

Keywords:
Breast cancer
Cancer registries
Administrative claims
Cancer staging
Cancer treatment

ABSTRACT

Introduction: Claims data and cancer registry data are valuable secondary data sources for addressing health service research questions. This study provides a thorough insight into the comparability of data from health insurance companies and cancer registries in Germany regarding breast, prostate, and lung cancer patients and their treatment.
Methods: For this study claims data of the InGef database and data of the Cancer Registry of Rhineland-Palatinate were used to identify patients living in Rhineland-Palatinate with an incident breast, prostate, or lung cancer diagnosis between Jan. 1, 2018 and Dec. 31, 2019. Both datasets were compared for patient and tumour characteristics as well as treatment strategy. For the descriptive analysis of tumour localisation and treatment all patients were followed up for a maximum of two years.
Results: A total of 1,470 incident cancer cases were identified in the InGef database and 1,694 in the Cancer Registry. Data on sex, age, and tumour localisation matched well for all cancer entities in the cohorts. Data for early UICC stages I+II varied between the cohorts for prostate cancer (84% InGef, 66% Cancer Registry) and lung cancer (29% InGef, 20% Cancer Registry). Larger deviations were found for antihormonal treatment (breast 54% vs. 44%, prostate 32% vs. 18%). Significant differences were found for surgery (breast and lung) and radiation (breast and prostate), respectively.
Discussion: Age at diagnosis, tumour localisation, and treatment for breast cancer was well documented in both databases. Tumour-specific deviations were observed for tumour localisations (lung cancer), UICC stage (prostate and lung cancer) and treatment options.
Conclusion: Both databases show very good completeness across cancer entities, but at the same time have minor limitations where they could readily complement each other. Individual linkage of claims and registry data could be an important step to improve oncological studies with routine practice data and to overcome the limitations identified.

ARTIKEL INFO

Artikel-Historie:
Eingegangen: 18. Oktober 2022
Revision eingegangen: 12. Dezember 2022
Akzeptiert: 7. Januar 2023
Online gestellt: xxxx

ZUSAMMENFASSUNG

Hintergrund: Abrechnungsdaten und Daten der Krebsregister sind wertvolle Sekundärdatenquellen zur Beantwortung versorgungsepidemiologischer Fragestellungen. Diese Studie gibt einen umfassenden Einblick in die Vergleichbarkeit der Daten von Krankenkassen und Krebsregistern in Deutschland in Bezug auf Brust-, Prostata- und Lungenkrebspatient*innen und deren Behandlung.
Methode: Für diese Studie wurden anonymisierte Daten der InGef-Forschungsdatenbank und Daten des Krebsregisters Rheinland-Pfalz verwendet. In Rheinland-Pfalz lebende Patientinnen und Patienten

Abbreviations: ICD-10-GM, International Classification of Diseases and Related Health Problems, 10th Revision, German Modification; InGef, Institute for Applied Health Research Berlin GmbH; M, metastasis status; N, nodal status; OPS, Operation and procedure codes; SHI, statutory health insurance; T, tumour status; UICC, Union of International Cancer Control

* Corresponding author. Dr. Marion Ludwig. InGef - Institute for Applied Health Research Berlin GmbH, Otto-Ostrowski-Straße 5, 10249 Berlin, Germany.
E-mail: Marion.Ludwig@ingef.de (M. Ludwig).

https://doi.org/10.1016/j.zefq.2023.01.001
1865-9217/

Please cite this article as: L.M. Lang, C. Behr, M. Ludwig et al., Routine practice data of three cancer entities: Comparison among cancer registry and health insurance data, Z. Evid. Fortbild. Qual. Gesundh. wesen (ZEFQ), https://doi.org/10.1016/j.zefq.2023.01.001

Schlüsselwörter:
Brustkrebs
Krebsregister
Abrechnungsdaten
Krebs-Staging
Krebsbehandlung

wurden identifiziert, bei denen zwischen dem 1.1.2018 und dem 31.12.2019 ein Brust-, Prostata- oder Lungenkarzinom diagnostiziert wurde. Beide Datensätze wurden hinsichtlich der Patienten- und Tumormerkmale sowie der Behandlungsstrategie verglichen. Für die deskriptive Analyse der Tumorlokalisation und der Behandlung wurden alle Patient*innen für maximal zwei Jahre nachverfolgt.
Ergebnisse: Insgesamt wurden 1.470 Fälle in der InGef-Datenbank und 1.694 Fälle im Krebsregister identifiziert. Die Daten zu Geschlecht, Alter und Tumorlokalisation stimmten für alle Krebsarten in den Kohorten gut überein. Die Daten für die frühen UICC-Stadien I+II variierten zwischen den Kohorten für Prostatakrebs (84% InGef, 66% Krebsregister) und Lungenkrebs (29% InGef, 20% Krebsregister). Größere Abweichungen wurden bei der antihormonellen Behandlung festgestellt (Brust 54% vs. 44%, Prostata 32% vs. 18%). Signifikante Unterschiede wurden bei Operationen (Brust und Lunge) bzw. Bestrahlungen (Brust und Prostata) ermittelt.
Diskussion: Das Alter bei Diagnose, die Tumorlokalisation und die Behandlung von Brustkrebs waren in beiden Datenbanken gut dokumentiert. Tumorspezifische Abweichungen wurden bei den Tumorlokalisationen (Lungenkrebs), dem UICC-Stadium (Prostata- und Lungenkrebs) und den Therapieoptionen beobachtet.
Schlussfolgerung: Beide Datenbanken weisen eine sehr gute Vollständigkeit über die verschiedenen Krebsentitäten hinweg auf, haben aber gleichzeitig geringfügige Einschränkungen, bei denen ein großes Ergänzungspotenzial besteht. Eine individuelle Verknüpfung von Abrechnungs- und Registerdaten könnte ein wichtiger Schritt sein, um insbesondere onkologische Studien mit Real-World Daten zu verbessern und die aufgezeigten Limitierungen zu überwinden.

Introduction

To increase the quality of health care while using resources efficiently, it is essential to assess and evaluate routine care. Oncological research of the past decades was mainly based on population-based studies and randomized clinical trials (RCT). However, certain patient groups, such as older as well as severely ill oncological patients, are less likely to be enrolled in clinical trials, which neglects their care situation. Here, routine practice data can complement RCTs by thoroughly analysing complex health care processes and structures under real-life conditions [1].

To provide reliable insights into the health care situation of cancer patients, routine practice data should represent a complete and comprehensive data collection that includes unselected information on patient characteristics, details of treatment as well as follow-up [2]. Beyond internal quality control, it is necessary to perform an independent evaluation with respect to completeness and distribution of variables in the respective dataset. For this purpose, comparison with an independently collected dataset is of highest value.

Suitable sources of routine practice data are clinical cancer registries or claims data from statutory health insurances (SHI). Population-based cancer registries were implemented in Germany to establish comprehensive documentation of all cancer cases including all levels of care. Aims of these registries were monitoring of cancer diagnosis, treatment and follow-up as well as conduction of clinical and epidemiological research [3]. Anonymized data of SHI provide cross-sector data content that is, unlike primary data, free of selection or recall bias. These data are available in a large sample size and offer the possibility of longitudinal analyses. The evaluation of claims data is used for health services research, including health economic or pharmaco-epidemiological studies and indication-specific analyses. Hence, using datasets of cancer registries or claims data for cross-sector analysis of treatment, treatment location, treatment options, and relevant outcomes has the potential to improve the current quality of care and contribute to innovative health care approaches that ultimately benefit tumour patients. Nevertheless, both datasets have data-related limitations for which awareness should be raised.

The reported study aims to compare information on the most frequent cancer types (breast, prostate, and lung cancer) for patients living in Rhineland-Palatinate, Germany, which was available from the Institute for Applied Health Research Berlin GmbH (InGef) research database with data collected by the Cancer Registry of Rhineland-Palatinate. To the best of our knowledge this is one of the first comprehensive studies which provides insight into the comparability of data from heath insurances and cancer registries in Germany.

Methods

Data sources

In the herein conducted investigation, accuracy, and completeness of two datasets were compared. These were the InGef research database and the dataset of the Cancer Registry of Rhineland-Palatinate. Both datasets are described in the following.

1) InGef research database: This database contains anonymized claims data on health care resource utilization and consumption of approx. 9 million who are insured with one of more than 60 German SHIs currently contributing data to the database and covers insurees from all federal states of Germany. In accordance with the obligations of the German health system, all medical diagnoses, as well as operation and procedures are documented for reimbursement purposes. Claims data are transferred directly from healthcare providers to a confidential data centre where all data are anonymized with respect to the individual insurant, healthcare providers and the respective SHI before entering the research database [4]. This complies with German data protection regulations as wells as German federal law and enabled the conduction of the present study without approval of an Ethics Committee. Data in the research database are longitudinally linked over a period of six years. In addition to sociodemographic information, the database contains information on ambulatory care and diagnoses; hospital data including admission dates, primary and secondary diagnoses as well as treatment; details on drug prescription; information on prescribed aids and remedies; and accrued costs with respect to these sectors. Clinical diagnoses are documented in the underlying SHI claims data according to the International Classification of Diseases and Related Health Problems (10th Revision, German Modification, ICD-10-GM) Information on the localisation of tumours was obtained from the fifth digit of the ICD-10-GM code [5].

The classification of surgeries and treatment was based on the Operation and Procedure Code (OPS) [6] and Diagnosis Related Groups (DRG) [7]. Prescriptions of medication were identified based on ATC codes [8] and OPS codes. In German claims data information on histopathological tumour characteristics such as tumour status (T), nodal status (N) or metastasis status (M) is not included, therefore, staging of tumour according to the recommendations of the Union of International Cancer Control (UICC) is hampered [9]. Moreover, claims data do not include information on the results of a diagnostic test or clinical procedure. Information on antihormonal treatment or immune therapy were based on prescription data.

2) Cancer Registry of Rhineland-Palatinate: Cancer registration in the German federal state Rhineland-Palatinate originally started on the level of epidemiological data, covering a population of approximately 4 million inhabitants. In 2013 a new federal legislation committed each federal state to perform registration of clinical data [10]. . All tasks of the former epidemiological cancer registration were transferred to the clinical Cancer Registry and the „Krebsregister Rheinland-Pfalz gGmbH" started registration on 01.01.2016. However, focus of the newly established facility was the collection of comprehensive data regarding diagnosis, histo-pathological characteristics of tumours, treatment and follow-up (course of disease, e.g., metastasis, relapse, progression, death) of cancer patients, resident and/or treated in Rhineland-Palatinate. Since datasets need to be comparable among cancer registries in Germany, data collection was standardized by introducing the nationwide uniform oncological dataset [11]. By now, personal and medical information as well as follow-up data of more than 200,000 patients are available in the dataset of the Cancer Registry of Rhineland-Palatinate.

Study population

Inclusion criteria were defined according to available parameters in both datasets.

The Cancer Registry of Rhineland-Palatinate selected patients based on the following criteria: (i) diagnosis of either breast cancer (ICD-10-GM C50), prostate cancer (ICD-10-GM C61) or lung cancer (ICD-10-GM C34) in the period of 01.01.2018 and 31.12.2019, (ii) insured in one of the SHIs (identity codes (IK-Number) of respective health insurances were provided by InGef), (iii) no prior cancer diagnosis documented in the five years before cohort entry, (iv) no further cancer diagnosis in the follow-up period (v) female sex for breast cancer and male sex for prostate cancer and vi) aged 18 years or older.

Similar criteria were used for the selection of patient from the InGef database: (i) at least two consecutive confirmed ambulatory diagnoses or one main or secondary hospital discharge diagnosis for breast, prostate or lung cancer (ICD-10-GM C50, C61 or C34) between 01.01.2018 and 31.12.2019, where the second diagnosis had to be documented by 30.06.2020 (ii) continuous insurance from index year until end of observation period or death, whatever comes first and in the five years prior to index diagnosis (iii) no prior cancer diagnosis (ICD-10-GM C*; ICD-10-GM D00- D09) in the five years before cohort entry (iv) no further cancer diagnosis (ICD-10-GM C*; ICD-10-GM D00- D09) in the follow-up period (v) female sex for breast cancer and male sex for prostate cancer (vi) living in Rhineland-Palatinate and vii) aged 18 years or older.

Data analysis

An observational study was conducted to assess the comparability and potential of mutual completion of two independent data sources. Information on age at diagnosis, tumour characteristics and treatment of breast, prostate, and lung cancer patients living in Rhineland-Palatinate was evaluated in both databases. Analysis of data was done separately by each partner and only aggregated data were compared; therefore, approval of an Ethics Committee was not required.

In detail, the total number of breast, prostate, and lung cancer cases (ICD-10-GM C50, C61 and C34, respectively) diagnosed between 01.01.2018 and 31.12.2019 was assessed in each dataset.

Age at diagnosis was calculated based on information on date of birth and date of diagnosis for patients from the Cancer Registry. In the InGef research database, age at diagnosis was calculated as the age in the quarter of the index diagnosis. Means, standard deviations, medians and the first and third quartiles (Q1-Q3) were determined for age. In addition, age was categorized into the following age groups: 18-34; 35-44; 45-54; 55-64; 65-74; 75-84; ≥85.

With respect to tumour characteristics information on the localisation of tumours was compared based on the fifths digit of the ICD-10-GM code (Appendix A, Table S1). For the InGef database information on localisation documented during the index quarter and the following quarter were included in the analysis. In the case of a specific localisation (ICD-10-GM C50.0-C50.8 (breast); C34.0-C34.8 (lung)) and a concurrent unspecified localisation (ICD-10-GM C50.9 (breast); C34.9 (lung)), only the specific localisation was counted. Otherwise, multiple entries were allowed.

Determination of tumour stages differed in both cohorts. In the Cancer Registry tumour stages were classified by available information on T (tumour status), N (nodal status) and M (metastasis status) according to the guidelines of the UICC [9]. In the InGef database no information on TNM is available, thus, classification according to the tumour stages defined by the UICC was estimated using the following criteria: UICC I and II: excluding patients with any secondary malignant neoplasms (ICD-10-GM C77-C79) in the index quarter and the following quarter; UICC III: excluding patients with secondary malignant neoplasms (ICD-10-GM C78-C79) and including patients with secondary malignant neoplasms only of the nodes (ICD-10-GM C77) in the index quarter and the following quarter; UICC IV: including patients with ICD-10-GM C78-C79 in the index quarter and the following quarter.

Of note, ICD-10-GM codes C77-C79 do not carry information about the primary tumour. In case of multiple cancer, the distant metastasis and affected lymph nodes might be caused by a different type of cancer. However, this was solved by excluding patients with any further C-diagnosis in the pre-observation period and in the follow-up period.

Moreover, information on non-assignable patients due to missing data (UICC N/A) is presented. Based on internal evaluations of the Cancer Registry, it could be assured that the number of missing data on the UICC stage is evenly distributed across all stages, which is why tumours with missing information (N/A) were excluded for the calculations of the relative proportion of this analysis.

The maximum follow-up period of two years (30.06.2021 or until death) was used to assess cancer treatment. The first occurrence of a documentation of surgery, radiation or antihormonal therapy was included in each therapy count. For all patients, each type of therapy was counted only once.

Data on surgery (based on OPS codes and DRG codes) and radiation (based on ICD, OPS, EBM and DRG codes) were available in both cohorts (reference codes to be found in Appendix A, Tables S2-S4).

In the Cancer Registry data on hormone receptor status are available. In addition, antihormonal and immune therapy, including applied drugs, are reported by medical facilities.

In contrast, information on the receptor status of the tumour is not reported to the InGef database, but treatment strategies could be concluded from prescription data and OPS codes for stationary application of medicines (compare Appendix A, Tables S2-S4).

For categorical variables absolute and relative frequencies (%) were reported. The proportion of patients with UICC stage, a specific therapy or a defined tumour localisation were calculated by dividing the number of breast, prostate, or lung cancer cases fulfilling the respective criteria by the number of persons in the respective cancer population. Cancer patients are usually treated with several therapies. Therefore, percentage for therapies was calculated based on the respective total number of cancer patients. Moreover, the proportion of breast cancer cases receiving breast-conserving therapy (BET) or mastectomy surgery were further based on the total number of breast cancer cases receiving surgery.

For absolute frequencies comprising less than five patients, the corresponding result is masked as "<5" (less than five) for data protection reasons. To detect significant differences in the distributions of the variables between the InGef and Cancer Registry data, a Chi-squared test or Fisher's test (for expected frequencies less than five) with a significance level of $p < 0.05$ was calculated for each variable except tumour localisation. Microsoft R Open software version 4.0.2 was used for all statistical analyses.

Results

Breast cancer

Patient and tumour characteristics of both breast cancer cohorts are displayed in Table 1. A total of 612 breast cancer cases diagnosed between 01.01.2018 and 31.12.2019 who were residing in Rhineland-Palatinate were identified in the InGef database. In the same period, 672 breast cancer patients, insured in one of the 60 SHIs of the InGef research database, were reported to the Cancer Registry of Rhineland-Palatinate.

Median age at breast cancer diagnosis was 60 years (Q1-Q3: 49-73 years) for InGef and 59 years (Q1-Q3: 50-69) for the Cancer Registry. Distribution of age at showed only slight deviations across age groups, with a percentage difference of 0.3-4% ($p = 0.04$).

Information on the localisation of tumours showed a higher variety among the two cohorts ranging from about 1 to 19%. Discrepancies were mainly based on the high number of tumours with unspecified localisation (C50.9) in the InGef cohort (31.4%). Data of the Cancer Registry documented only 12.9% tumours coded with C50.9. Despite this in both datasets tumours with specific coding were most frequently detected in the upper out quadrant (C50.4) (Table 1).

Missing information on UICC stages (N/A) accounted for 12.6% in the Cancer Registry. Excluding the missing values in the proportional calculation of UICC staging, the information from both cohorts matched almost exactly.

Surgery and radiation were reported for 73.4 and 58.8%, respectively, of breast cancer patients to the Cancer Registry and documented for 64.7 and 50.2%, respectively, of breast cancer patients in the InGef database (Table 1). In contrast information on antihormonal treatment was available for 44.2% of patients reported to the Cancer Registry and for 54.6% of patients in the InGef database ($p = 0.0003$). Immune therapies were reported for 8.3% of patients in both cohorts.

Surgical options for the treatment of breast cancer were differentiated between breast-conserving surgeries (BET) and ablatio/-mastectomy and showed significant differences between both cohorts ($p = 0.016$). For BET only minor differences were observed with 71.0% and 73.0% of patients in the databases of the Cancer Registry and InGef, respectively. With respect to ablatio/mastectomy a deviation of more the 5 % was observed with 26.6% in the InGef database and 21.2% of patients receiving surgery in the Cancer Registry database (Table 1).

Table 1
Characteristics and treatment of breast cancer patients diagnosed in the period 01.01.2018-31.12.2019 who were resident in Rhineland Palatinate, Germany, and insured in one of the 60 SHIs of the InGef research database.

	InGef	Cancer Registry	p-value
Incident breast cancer patients (*n*)	612	672	
Age at diagnosis in years, median (Q1-Q3)	60 (49-73)	59 (50-69)	
Age in years (n, %)			0.0414
18-34	35 (5.8)	14 (2.1)	
35-44	54 (8.8)	50 (7.4)	
45-54	150 (24.5)	171 (25.4)	
55-64	129 (21.1)	154 (22.9)	
65-74	102 (16.7)	122 (18.2)	
75-84	102 (16.7)	119 (17.7)	
>85	40 (6.5)	42 (6.2)	
Tumor localisation (n, %)			/
C50.0	23 (3.8)	8 (1.2)	
C50.1	49 (8.0)	34 (5.1)	
C50.2	55 (9.0)	89 (13.2)	
C50.3	38 (6.2)	49 (7.3)	
C50.4	179 (29.3)	262 (39)	
C50.5	34 (5.6)	47 (7)	
C50.6	13 (2.1)	<5	
C50.8	77 (12.6)	95 (14.1)	
C50.9	192 (31.4)	87 (12.9)	
UICC (n, %)			
I-II	473 (77.3)	454 (77.3)	0.9126
III	67 (11)	67 (11.4)	
IV	72 (11.8)	65 (11.1)	
N/A*	0	85 (12.6)	
Surgery (n, %)	396 (64.7)	493 (73.4)	0.1198
Breast-conserving surgery (BET) (n, %**)	289 (73.0)	350 (71.0)	0.0161
Ablatio/Mastectomy (n, %**)	84 (21.2)	131 (26.6)	
Other (n, %**)	23 (5.8)	12 (2.4)	
Radiation (n, %)	307 (50.2)	395 (58.8)	0.0707
Immune therapy (n, %)	51 (8.3)	56 (8.3)	0.9071
Antihormonal treatment (n, %)	334 (54.6)	297 (44.2)	0.0003

Characteristics were determined in the index quarter and tumor localisation and stage in the index quarter or follow-up quarter. Therapies were determined during the individual follow-up period of a maximum of two years after index diagnosis or until death. Calculation of p-value using chi-square independence tests or Fisher test for categorical data. Abbreviations: n: Number of cases, Q1: 1st quartile, Q3: 3rd quartile, %: Percentage relative to n.

* Excluded for calculation of percentage in consultation with the Cancer Registry.

** Percentage of patients who received surgery.

Prostate cancer

Patient and tumour characteristics of both prostate cancer cohorts are displayed in Table 2. In the period between 01.01.2018 and 31.12.2019 a total of 518 and 608 prostate cancer cases were identified in the InGef database and in the Cancer Registry, respectively.

Median age at prostate cancer diagnosis was 70 years (Q1-Q3: 63-77 years) for InGef and 69 years (Q1-Q3: 63-77) for the Cancer Registry. Distribution of age at diagnosis showed only slight deviations across age groups ($p = 0.567$).

Most patients received their diagnosis in UICC stage I-II in both cohorts (84.4% of InGef and 66% of Cancer Registry patients). Nevertheless, the proportions of patients per stage varied significantly between both cohorts ($p < 0.001$).

Prostate cancer patients most frequently received surgery as therapy, totaling 48.3% of patients documented in the InGef database and 33.9% of Cancer Registry patients. The proportion of patients receiving radiotherapy differed significantly by 3% ($p < 0.001$). In the InGef database, there were more patients who received antihormonal treatment (32% InGef, 18% Cancer Registry) ($p < 0.001$) (Table 2).

Table 2
Characteristics and treatment of prostate cancer patients diagnosed in the period 01.01.2018-31.12.2019 who were resident in Rhineland-Palatinate, Germany, and insured in one of the 60 SHIs of the InGef research database.

	InGef	Cancer Registry	p-value
Incident prostate cancer patients (*n*)	518	608	
Age at diagnosis in years, median (Q1-Q3)	70 (63-77)	69 (63-77)	
Age in years (n, %)			0.5674
18-34	0	0	
35-44	0	<5	
45-54	27 (5.2)	22 (3.6)	
55-64	129 (24.9)	142 (23.4)	
65-74	183 (35.3)	218 (35.9)	
75-84	149 (28.8)	194 (31.9)	
>85	30 (5.8)	31 (5.1)	
UICC (n, %)			<0.001
I-II	437 (84.4)	263 (66)	
III	14 (2.7)	68 (17.1)	
IV	67 (12.9)	67 (16.8)	
N/A*	0	210 (34.5)	
Surgery (n. %)	250 (48.3)	206 (33.9)	0.2687
Radiation (n, %)	125 (24.1)	164 (27)	<0.001
Antihormonal treatment (n, %)	164 (31.7)	107 (17.6)	0.0037

Characteristics were determined in the index quarter and tumor localisation and stage in the index quarter or follow-up quarter. Therapies were determined during the individual follow-up period of a maximum of two years after index diagnosis or until death. Calculation of p-value using chi-square independence tests or Fisher test for categorical data. Abbreviations: n: Number of cases, Q1: 1st quartile, Q3: 3rd quartile, %: Percentage relative to n.

* Excluded for calculation of percentage in consultation with the Cancer Registry.

Table 3
Characteristics and treatment of lung cancer patients diagnosed in the period 01.01.2018-31.12.2019 who were resident in Rhineland-Palatinate, Germany, and insured in one of the 60 SHIs of the InGef research database.

	InGef	Cancer Registry	p-value
Incident lung cancer patients	340	414	
Gender			0.728
female	132 (38.8)	167 (40.3)	
male	208 (61.2)	247 (59.7)	
Age at diagnosis in years (Median, Q1, Q3)	67 (61-76)	66 (60-72)	
Age in years (%)			0.648
18-34	0	0	
35-44	<5	<5	
45-54	25 (7.4)	28 (6.8)	
55-64	104 (30.6)	130 (31.4)	
65-74	112 (32.9)	153 (37)	
75-84	84 (24.7)	92 (22.2)	
>85	13 (3.8)	9 (2.2)	
Tumor localisation (n, %)			/
C34.0			
C34.1	48 (14.1)	26 (6.3)	
C34.2	152 (44.7)	191 (46.1)	
C34.3	13 (3.8)	10 (2.4)	
C34.8	74 (21.8)	100 (24.2)	
C34.9	33 (9.7)	19 (4.6)	
	36 (10.6)	68 (16.4)	
UICC (n, %)			0.016
I-II	97 (28.5)	63 (19.5)	
III	61 (17.9)	75 (23.2)	
IV	182 (53.5)	185 (57.3)	
N/A*	0	91 (22)	
Surgery (n, %)	87 (25.6)	94 (22.7)	<0,001
Radiation (n, %)	114 (33.6)	159 (38.4)	0.088
Immune therapy (n, %)	61 (18)	54 (13)	0.114

Characteristics were determined in the index quarter and tumor localisation and stage in the index quarter or follow-up quarter. Therapies were determined during the individual follow-up period of a maximum of two years after index diagnosis or until death. Calculation of p-value using chi-square independence tests or Fisher test for categorical data. Abbreviations: n: Number of cases, Q1: 1st quartile, Q3: 3rd quartile, %: Percentage relative to n.

* Excluded for calculation of percentage in consultation with the Cancer Registry.

Lung cancer

Patient and tumour characteristics of both lung cancer cohorts are displayed in Table 3. In the period between 01.01.2018 and 31.12.2019 a total of 340 and 414 lung cancer cases were identified in the InGef database and in the database of the Cancer Registry, respectively.

Median age at lung cancer diagnosis was 67 years (Q1-Q3: 61-76 years) for InGef and 66 years (Q1-Q3: 60-72) for the Cancer Registry. Distribution of age at diagnosis showed only slight deviations across age groups (p = 0.648).

The upper lobe of the lung (C34.1) was the most common tumour site in both cohorts (InGef: 44.7%, Cancer Registry: 46.1%). Contrary to breast and prostate cancer cases, coding of unspecific tumour localisation was more frequent in the Cancer Registry cohort (InGef: 10.6%, Cancer Registry: 16.4%).

Most patients received their diagnosis in UICC stage IV in both cohorts (53.5% of InGef and 57.3% of Cancer Registry patients). The proportions of patients per stage varied significantly between both cohorts (p = 0.016).

Radiation therapy was most frequently performed in both groups with 33.6% (InGef) and 38.4% (Cancer Registry), respectively. Regarding the frequency for surgery, both groups differed by 3% (p < 0.01) and for immune therapy by 5%.

Discussion

This study analysed the characteristics and treatments of three frequent cancer entities and compared two independent data sources. By selecting two gender-specific cancer entities and a gender non-specific type of cancer the strength and limitations of the two databases could be thoroughly explored.

The comparison of aggregated data of patient, tumour and therapy characteristics of the databases of InGef and the Cancer Registry Rhineland-Palatinate showed a high rate of concordance, especially for breast cancer. However, there were some cross-entity differences in both cohorts for the variables tumour locations, UICC stage, and treatment.

Demographics

The total number of cancer cases identified in both cohorts differs by 224 patients. This discrepancy might be based on slight differences in inclusion criteria and data availability. InGef included only patients with at least two consecutive, verified ambulatory or one stationary ICD-10-GM diagnoses in the observation period. Further, completeness of the database of the Cancer Registry depends on accurate reporting by oncological facilities.

The analysis on all three cancer types shows concordance with respect to age and sex between the two analyzed databases and in comparison, with reported figures [12]. The finding that the observed median ages in both breast cancer cohorts were 4 to 5 years less than the reported median age [13], might be biased by the characteristics of the insured persons of the 60 SHIs (mainly guild and company SHIs) included in the InGef database. Structural differences between health insurance funds have been reported before and might explain the found divergences in median age [14–16].

Tumour localisation and UICC

For breast cancer, the data on the most common tumour localisation (C50.4) and the classification into UICC stages matched very

well and closely approximated the data reported by the literature [17]. The observed discrepancy with respect to tumour localisation might be caused by more comprehensive information which is available in the Cancer Registry. Localisation of tumours is based on information from diagnosis, surgery and pathological reports. Claims data analysis on tumour localisation was based on the ICD-10-GM codes documented in the index and the first following quarter. The quarter following incident diagnosis was included to increase the number of specific localisations. However, documenting information on the localisation of the tumour in addition to the diagnosis is not mandatory for reimbursement and thus, might partly explain the higher percentage of unspecified localisations. In contrast to the Cancer Registry claims data do not include results of diagnostic test such as pathological examinations. Accordingly, UICC stages of tumours (based on TNM status) cannot be assessed directly. A strategy to overcome this limitation was to use information on the disease stage at initial diagnosis as proposed by Oppelt et al. [18]. In Germany, 80% of breast cancer patients and two third of prostate patients receive their diagnosis at UICC stage I+II [12]. This is in line with this study, however, there were discrepancies regarding discrimination of UICC stage III from I+II for prostate and lung cancer using the InGef database. For these two entities, the underlying algorithm, which relies on coding secondary malignant neoplasms of the nodes to classify a patient as UICC III, results in potentially misclassifying substantially more patients as UICC I+II.

For lung cancer, UICC stage IV is the most common at initial diagnosis [12]. In this comparative study, 53-57% of patients were initially diagnosed at this stage.

Hence, for breast cancer it is feasible to distinguish between advanced and early tumours for both databases. Regarding prostate and lung cancer, the applied strategy might be more suitable for claims data analyses focussing on advanced tumours (UICC IV). To achieve the precise classification necessary for many studies, either tumour-specific algorithms must be investigated for claims data or the databases available in Germany must be linked.

Therapy

The herein reported analysis showed good agreement for surgery, radiotherapy, and immune therapy between the two data sources and previously published studies. Main purpose of cancer surgeries is curative, which is no longer possible in the metastasized stage. In line with the low number of lung cancer patients in the early stages, only about a quarter of the patients in both cohorts are treated surgically. This is underscored by a systematic review in which Waser et al. (2022) reported that the proportion of nonmetastatic patients receiving surgery alone or with radiotherapy tended to decrease with an increase in tumour stage [19].

In further agreement with our study, Liu et al. (2019) described that radiation is performed in about 37% of all lung cancer patients depending on UICC stage [20] (radiation: 34% InGef, 38% Cancer Registry).

Breast cancer treatment statistics from the UK reported that usually about 80% of breast tumours are removed by surgery and about 60% of breast tumours are treated by radiation [21]. The corresponding shares of surgery (Cancer Registry: 73.4%; InGef: 64.7%) and radiation (Cancer Registry: 50.2%; InGef: 58.8%) were similar in both German databases, neatly reflecting these previously reported treatment proportions. Moreover, the most common types of breast cancer surgeries were compared. As of yet, breast-conserving therapy (BET) and ablatio/mastectomy remain the gold standards of breast cancer treatment with proportions of about 60% and 40%, respectively [22]. This is in line with our results and mirrors the recommendations of recent publications and the S3-Guidelines for breast cancer, suggesting that BET is preferable to mastectomy [23,24].

Literature also shows that prostate cancer is mainly treated with surgery or radiotherapy. A comprehensive report of German epidemiological cancer registries described that of 11,000 German men with locally advanced prostate cancer, radical prostatectomy was most commonly performed followed by radiotherapy [25]. Another study included 3,000 patients with clinically localized prostate cancer, of these approximately 40% received prostatectomy and 31% received radiotherapy [26]. These results are reflect-ed in our analysis with surgery and radiation therapy being most common for prostate cancer patients in both cohorts (surgery: 48.3% InGef, 33.9% Cancer Registry; radiation therapy: 24.1% InGef, 27% Cancer Registry).

In our study 54.6% (InGef) and 44.2% (Cancer Registry) of breast cancer patients received antihormonal therapy, respectively. Treatment with antihormonal therapy is prescribed when the tumour is hormone receptor positive indicating that its growth is hormone-dependent. In Germany, this is the case in about 80% of women with breast cancer [27]. With respect to prostate cancer, antihormonal therapy is indicated as a palliative form of therapy in patients with advanced tumours. 31.7% and 17.6% of cases received antihormonal treatment in the InGef database and Cancer Registry, respectively. Both shares are higher than previously reported by Schymura et al. (2010), where roughly 10% of prostate cancer patients received antihormonal treatment [26].

These results reveal that free support from a field crew and educational courses offered by the Cancer Registry have improved reporting of antihormonal therapy (internal communication). However, in an international comparison, antihormone therapies are also less accurately reported for cancer registries than in claims data [28]. This limitation arises from the fact that valid and complete documentation of all information is often not feasible in the daily clinical practice of physicians. The submission of data to the Cancer Registry must occur in addition to that to The National Association of Statutory Health Insurance Physicians. Although the Cancer Registry pays for complete reports, payment is not equivalent to the billings to the SHIs. On the other hand, claims data only include detailed information on drugs, if they were prescribed by practitioners, therefore, the remaining patients may have received drug therapy in a hospital or remained untreated. Nevertheless, antihormonal drug prescriptions are well represented in the InGef database.

Despite different approaches to identify immune therapy (approximation via ATC and OPS codes in the InGef database), both databases showed consistent results and reported equal percentages for immune therapy for breast cancer. For lung cancer immune therapy was slightly more often documented in the InGef database. However, the field of immune therapies is a rapidly developing and growing therapeutic area that profoundly impacts systems therapy for metastatic non-small cell lung cancer [29]. Moreover, if drugs are used off label (without approved indication or guideline), they are sometimes not reimbursed by the health insurer and patients must pay for them themselves, which in turn could explain why they are not reflected in claims data.

Limitations

In general, missing data is a common problem in health data, because they limit the size and completeness of patient cohorts and can lead to biased data sets. Cancer patients in particular are often treated by multiple departments within a hospital and by oncologists, radiation therapists, etc. in private practice, making comprehensive and complete coverage of all treatment pathways challenging. However, SHI routine data benefit from the standardized and uniform system for claims data. In general, the validity of

any variable stored in claims data depends on the quality of the information submitted by providers. Therefore, a comprehensive validation process is recommended for any analysis of claims data, following official guidelines to ensure the accuracy of the data source in the context of the research question at hand [30]. Matching SHI data with cancer registry data has rarely been used for this purpose in Germany. Although this method has already revealed relevant differences in the databases studied, the Cancer Registry data also generally showed good matches with the SHI data [18].

Another issue is that despite the application of the M2Q criterion for the InGef research database, we cannot exclude the possibility that coding for cancer could have occurred multiple times in the context of a diagnostic procedure that did not confirm the cancer diagnosis.

While comparing clinical literature data it should be noted that in this study it was neither investigated whether patients received multiple treatment regimens nor the sequence of treatment or patient subgroups.

Conclusions

With the herein performed study the value and quality of claims data and Cancer Registry data was demonstrated. Age at diagnosis, localisation of tumours, surgical treatment and radiation was well documented in both databases. However, both datasets still harbour minor limitations, where they could complement each other to increase the variable coverage and thus, the significance of a study. In detail, number of cases differs among the two cohorts which might be based on fidelity of exclusion criteria and completeness of reports. Information on UICC stage was more accurate in the Cancer Registry, while immune and antihormonal therapy was better documented in the InGef database. Further, sufficient discrimination of UICC stages using ICD-coding appeared to differ between cancer entities and should be particularly considered in study planning. Therefore, UICC staging, which is needed, among other things, to assess the appropriateness of a medical intervention, cannot always be adequately represented with claims data. Here, registry data could complement claims data well. Our results show that for a comprehensive description of utilization and treatment trajectories, both data sources are suitable. However, since cancer registries do not collect information on comorbidities or cancer-independent health care utilization, there is also potential for complementarity between the two data sources.

The often-problematic reporting behavior to registries could not be confirmed in our study and reporting is likely to increase qualitatively.

In summary, the herein performed analysis showed, that data of cancer registries are more suitable for studies including detailed information on histopathological factors, in contrast, claims data are more suitable for studies requiring additional information on clinical characteristics, such as comorbidities or hospitalizations. However, as widely discussed, a combination of both data sources or a future linkage of claims and registry data holds the potential to improve the value of routine practice data analyses and to overcome current limitations.

Data availability

InGef: The data used in this study cannot be made available in the manuscript, the supplemental files, or in a public repository due to German data protection laws (Bundesdatenschutzgesetz). To facilitate the replication of results, anonymized data used for this study are stored on a secure drive at the InGef - Institute for Applied Health Research Berlin GmbH. Access to the raw data used in this study can only be provided to external parties under the conditions of a cooperation contract and can be accessed upon request, after written approval (info@ingef.de), if required

Cancer registry

The datasets generated and analyzed during the current study are available from the senior author on reasonable request.

Ethics approval

Claims data are transferred directly from healthcare providers to a confidential data centre where all data are anonymized with respect to the individual insurant, healthcare providers and the respective SHI before entering the research database. This complies with German data protection regulations as wells as German federal law and enabled the conduction of the present study without approval of an Ethics Committee.

Conflict of interest

All authors declare that there is no conflict of interest.

CRediT author statement

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by FB, LL and HCL. The first draft of the manuscript was written by LL, ML and CJ and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.zefq.2023.01.001.

References

[1] Blonde L, Khunti K, Harris SB, Meizinger C, Skolnik NS. Interpretation and Impact of Real-World Clinical Data for the Practicing Clinician. Adv Ther 2018;35(11):1763–74. https://doi.org/10.1007/s12325-018-0805-y.

[2] Holleczek B, Katalinic A. Toward a comprehensive cancer registration in Germany. *Eur J Cancer Prev Off J Eur Cancer Prev Organ ECP* 2017:S132–8. https://doi.org/10.1097/CEJ.0000000000000388, Bd. 26 Joining forces for better cancer registration in Europe, S..

[3] Klinkhammer-Schalke M, Hofstädter F, Gerken M, Benz S. The contribution of clinical cancer registries to benefit assessments: Requirements and first results. Z Evidenz Fortbild Qual Im Gesundheitswesen 2016;112(Suppl 1): S3–S10. https://doi.org/10.1016/j.zefq.2016.04.008.

[4] Ludwig M, Enders D, Basedow F, Walker J, Jacob J. Sampling strategy, characteristics and representativeness of the InGef research database. *Public Health* 2022;Bd. 206:57–62. https://doi.org/10.1016/j.puhe.2022.02.013.

[5] Deutsches Institut für Medizinische Dokumentation und Informatik, „Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme 10. Revision German Modification Version 2020“, 2020. https://www.dimdi.de/static/de/klassifikationen/icd/icd-10-gm/kode-suche/htmlgm2020/ (zugegriffen 6. Juli 2020).

[6] Deutsches Institut für Medizinische Dokumentation und Informatik, „Operationen- und Prozedurenschlüssel Version 2020“, 2020. https://www.dimdi.de/static/de/klassifikationen/ops/kode-suche/opshtml2020/ (zugegriffen 6. Juli 2020).

[7] „Fallpauschalen-Katalog 2020, InEK GmbH“. https://www.g-drg.de/aG-DRG-System_2020/Fallpauschalen-Katalog/Fallpauschalen-Katalog_2020 (zugegriffen 5. August 2022).

[8] „ATC-Klassifikation“. https://www.dimdi.de/dynamic/de/arzneimittel/atc-klassifikation (zugegriffen 19. August 2021).

[9] C. Wittekind, *TNM Klassifikation maligner Tumoren*, 8. Aufl. Wiley-VCH, Weinheim, 2017. Zugegriffen: 18. August 2021. [Online]. Verfügbar unter: https://www.wiley-vch.de/de/fachgebiete/medizin-und-gesundheit/tnm-klassifikation-maligner-tumoren-978-3-527-34280-8

[10] „§ 65c SGB 5 - Einzelnorm“. https://www.gesetze-im-internet.de/sgb_5/__65c.html (zugegriffen 19. August 2021).

[11] „Amtliche Veröffentlichungen – Bundesanzeiger". https://www.bundesanzeiger.de/pub/de/amtliche-veroeffentlichung?1 (zugegriffen 18. August 2021).
[12] R. Koch-Institut, „Krebs in Deutschland für 2017/2018", S. 172.
[13] R. Robert-Koch-Institut, „Krebs in Deutschland | 2015/2016", S. 163.
[14] F. Hoffmann und A. Icks, „[Structural differences between health insurance funds and their impact on health services research: results from the Bertelsmann Health-Care Monitor]", *Gesundheitswesen Bundesverb. Arzte Offentlichen Gesundheitsdienstes Ger.*, Bd. 74, Nr. 5, S. 291–297, Mai 2012, doi: 10.1055/s-0031-1275711.
[15] Hoffmann F, Koller D. Verschiedene Regionen, verschiedene Versichertenpopulationen? Soziodemografische und gesundheitsbezogene Unterschiede zwischen Krankenkassen. Gesundheitswesen 2017;79(1):e1–9. https://doi.org/10.1055/s-0035-1564074.
[16] Jaunzeme J, Eberhard S, Geyer S. [How ‚representative' are SHI (statutory health insurance) data? Demographic and social differences and similarities between an SHI-insured population, the population of Lower Saxony, and that of the Federal Republic of Germany using the example of the AOK in Lower Saxony]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 2013;56(3):447–54. https://doi.org/10.1007/s00103-012-1626-9.
[17] Rummel S, Hueman MT, Costantino N, Shriver CD, Ellsworth RE. „Tumour location within the breast: Does tumour site have prognostic ability?", 13. Juli 2015. http://ecancer.org/en/journal/article/552-tumour-location-within-the-breast-does-tumour-site-have-prognostic-ability (zugegriffen 11. August 2022).
[18] Oppelt KA, Luttmann S, Kraywinkel K, Haug U. Incidence of advanced colorectal cancer in Germany: comparing claims data and cancer registry data. BMC Med Res Methodol 2019;19. https://doi.org/10.1186/s12874-019-0784-y.
[19] Waser N, Vo L, McKenna M, Penrod J, Goring S. Real-world treatment patterns in resectable (stages I-III) non-small-cell lung cancer: a systematic literature review. Future Oncol 2022;18(12):1519–30. https://doi.org/10.2217/fon-2021-1417.
[20] Liu W, Liu A, Chan J, Boldt RG, Munoz-Schuffenegger P, Louie AV. „What is the optimal radiotherapy utilization rate for lung cancer?—a systematic review", *Transl. Lung Cancer Res.*, Bd. 0, Nr. 0, Sep. 2019, Zugegriffen: 11. August 2022. [Online]. Verfügbar unter: https://tlcr.amegroups.com/article/view/31055
[21] „Breast cancer treatment statistics", *Cancer Research UK*, 14. Mai 2015. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/diagnosis-and-treatment (zugegriffen 14. September 2021)
[22] de Boniface J, Szulkin R, Johansson ALV. Survival After Breast Conservation vs Mastectomy Adjusted for Comorbidity and Socioeconomic Status: A Swedish National 6-Year Follow-up of 48 986 Women. JAMA Surg 2021;156(7):628–37. https://doi.org/10.1001/jamasurg.2021.1438.
[23] De Wilde RL, Devassy R, Torres-de la Roche LA, Krentel H, Tica V, Cezar C. Guidance and Standards for Breast Cancer Care in Europe. J Obstet Gynecol India 2020;70(5):330–6. https://doi.org/10.1007/s13224-020-01316-6.
[24] „S3-Leitlinie Mammakarzinom", S. 467, 2021.
[25] Hager B. Increasing use of radical prostatectomy for locally advanced prostate cancer in the USA and Germany: a comparative population-based study Art. Nr. 1. Prostate Cancer Prostatic Dis 2017;20(1). https://doi.org/10.1038/pcan.2016.43.
[26] Schymura MJ. Factors associated with initial treatment and survival for clinically localized prostate cancer: results from the CDC-NPCR Patterns of Care Study (PoC1). BMC Cancer 2010;10(1):152. https://doi.org/10.1186/1471-2407-10-152.
[27] Lumachi F, Santeufemia DA, Basso SM. Current medical treatment of estrogen receptor-positive breast cancer. World J Biol Chem 2015;6(3):231–9. https://doi.org/10.4331/wjbc.v6.i3.231.
[28] Anderson C. Validity of state cancer registry treatment information for adolescent and young adult women. Cancer Epidemiol 2020;64. https://doi.org/10.1016/j.canep.2019.101652.
[29] „S3-Leitlinie Lungenkarzinom", S. 417, 2018.
[30] Kreis K, Neubauer S, Klora M, Lange A, Zeidler J. Status and perspectives of claims data analyses in Germany-A systematic review. Health Policy Amst Neth 2016;120(2):213–26. https://doi.org/10.1016/j.healthpol.2016.01.007.