BMJ Open

Implementation of an algorithm for the identification of breast cancer deaths in German health insurance claims data: a validation study based on a record linkage with administrative mortality data

Ingo Langner,[1] Christoph Ohlmeier,[2] Ulrike Haug,[1,3] Hans Werner Hense,[4,5] Jonas Czwikla,[3,6] Hajo Zeeb[1,3]

To cite: Langner I, Ohlmeier C, Haug U, *et al*. Implementation of an algorithm for the identification of breast cancer deaths in German health insurance claims data: a validation study based on a record linkage with administrative mortality data. *BMJ Open* 2019;**9**:e026834. doi:10.1136/bmjopen-2018-026834

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-026834).

Received 21 September 2018
Revised 28 May 2019
Accepted 2 July 2019

For numbered affiliations see end of article.

Correspondence to
Dr Ingo Langner;
langner@leibniz-bips.de

ABSTRACT

Objective To adapt a Canadian algorithm for the identification of female cases of breast cancer (BC) deaths to German health insurance claims data and to test and validate the algorithm by comparing results with official cause of death (CoD) data on the individual and the population level.

Design Validation study, secondary data, medical claims.

Setting Claims data of two statutory health insurance providers (SHIs) for inpatient and outpatient care, CoD added via record linkage with epidemiological cancer registry (ECR). **Participants** All women insured with the two SHIs and who deceased in the period 2006–2013, were residents of North Rhine Westphalia (NRW) and were linked with ECR data: n=22 413.

Main outcome measures Based on inpatient and outpatient diagnoses in the year before death, six algorithms were derived and the accordance of the algorithm-based CoD with the official CoD was evaluated calculating specificity, sensitivity, negative and positive predictive values (NPV, PPV). Furthermore, algorithm-based age-specific BC mortality rates covering several calendar years were calculated for the entire insured female population and compared with official national rates.

Results Our final algorithm, derived from the NRW subsample, comprised codes indicating the presence of BC, metastases, a terminal illness phase and the absence of codes for other tumours. Overall, specificity, sensitivity, NPV and PPV of this algorithm were 97.4%, 91.3%, 98.9% and 81.7%, respectively. In the age range 40–80 years, sensitivity and PPV slightly decreased with increasing age. Algorithm-based age-specific BC mortality rates agreed well with official rates except for the age group 85 years and older.

Conclusions The algorithm-based identification of BC deaths in German claims data is feasible and valid, except for higher ages. The algorithm to ascertain BC mortality rates in an epidemiological study seems applicable when information on the official CoD is not available in the original database.

Strengths and limitations of this study

- This is the first study evaluating whether an algorithm for the identification of breast cancer deaths in Canadian claims data can be adapted to German health insurance claims data.
- Causes of death classifications of different algorithm versions were directly compared with the official cause of death on the individual level and indirectly compared with official vital statistics on the population level.
- The sample for testing and validating the algorithm on the individual level was based on a high-quality record linkage and included 22 413 women who died in the period 2006–2013.
- The test and validation sample was restricted to the age range 40–80 years.
- The study results are based on German claims data, however, the procedures are considered transferable to other settings with some adaptation.

INTRODUCTION

Cause-specific mortality is a major outcome in epidemiological cohort studies. Especially for studies with rare outcomes, where a great amount of cumulative follow-up years with a consistent exposure ascertainment is essential, claims data constitute a promising data source. However, in German claims data, the official cause of death (CoD) is not included and individual record linkage with data sources providing this information is limited due to data protection regulations. Such record linkage is most likely feasible only on a regional level due to the regional character of suitable registries comprising CoD information which, however, restricts the usable study population.

For the intended nationwide monitoring and evaluation of the impact of the German mammography screening programme (MSP)[1] on breast cancer (BC) mortality,[2] claims data as included in the German Pharmacoepidemiological Research Database (GePaRD) represent one of the most suitable data sources. The data comprise demographic information, inpatient and outpatient diagnoses, health services, and detailed information concerning the participation in the MSP. Further, deceased individuals can be identified in GePaRD because 'death' is coded either as the reason for the end of insurance coverage or for the discharge from hospital. The lack of information on the official CoD, however, represents a limitation of this data source.

As CoD information is essential for epidemiological cohort studies such as the intended nationwide monitoring and evaluation of BC mortality in the MSP, there is a need to add at least the information 'death due to BC (yes/no)'. In order to mimic official CoD information, we, therefore, aimed to implement, optimise and validate a claims data-based algorithm for the identification of BC deaths among deceased females in GePaRD, inspired by the approach developed by Gagnon *et al*[3] for Canadian administrative data.

METHODS

Data source

Analyses were based on the GePaRD database, which has been described elsewhere.[4–6] In brief, GePaRD includes pseudonymised claims data from four statutory health insurance providers (SHIs) and contains information on about 20 million individuals from all over Germany who have been insured at one of the participating SHIs since 2004 or later. The database contains information on demographic characteristics, the start and end of insurance periods, hospital stays, outpatient physician visits and outpatient prescriptions. The hospital data comprise information on the dates of admission and discharge, admission diagnoses, one main discharge diagnosis (which specifies the disease causing the hospital stay), further main and secondary hospital diagnoses, diagnostic and therapeutic procedures with their respective dates, as well as the reason for hospital discharge (eg, 'treatment terminated regularly', 'transfer to another hospital', 'deceased' and others). Outpatient data contain information on outpatient diagnoses, treatments and procedures. Outpatient diagnoses are recorded per quarter and are distinguishable into different types, for example, 'confirmed'. Diagnoses are coded according to the German modification of the International Classification of Diseases, 10th Revision (ICD-10-GM). Information related to the death of individuals can be obtained from the variable specifying the reason for the end of insurance (eg, 'change to another insurance company', 'deceased' and others) and for subjects dying in hospital also from the respective variable specifying the cause of the hospital discharge (about 50% of all deaths in GePaRD).

The algorithm-based individual assignment of a death by BC in GePaRD was compared with the individual official CoD provided by the State Cancer Registry of North Rhine Westphalia (CR-NRW) that holds information on the official causes of each deceased person in NRW since 2006. This information was linked via a probabilistic record linkage to the claims data on the person level (for details see reference [7]). In addition, to assess the performance of the algorithm-based classification on the population level, age-specific BC death rates estimated in the entire GePaRD data base were compared with data from the German Centre for Cancer Registry Data (ZfKD) that provided data on the national BC mortality rates.

Patient and public involvement

Patients or public were not involved.

Study populations

Individual-level analysis

All death cases occurring between 2006 and 2013 among female GePaRD residents of NRW (as identified in GePaRD) who had been continuously insured in the year before death and whose GePaRD data could be successfully linked with the CR-NRW records, formed the study population for the algorithm validation on the individual level (see also online supplementary data 1). Data year restrictions were due to limitations imposed by legal authorities. The age range of women included in this subsample of GePaRD was limited to 40–80 years encompassing the age range of eligibility for the MSP (50–69 years) and the adjacent decades.

Population-level analysis

For the calculation of claims data-based age-specific BC mortality rates at the population level, all women in GePaRD irrespective of place of residence and age insured in 2007, 2010 and 2012, respectively, were included in the study population. Deceased women were identified via a corresponding data entry on the reason for the end of insurance coverage or for the discharge from hospital. The date of death was defined accordingly as the date of the end of insurance or the date of hospital discharge.

Algorithm for the identification of BC deaths

Based on Canadian routine health data, Gagnon *et al*[3] developed an algorithm for the identification of women who had died of BC. The identification was based on ICD-9-coded inpatient diagnoses from all hospital stays of a single person ever recorded at provincial hospital discharge databases. Gagnon *et al* applied three criteria to identify BC as the CoD. First, women had to be diagnosed with BC (ICD-9: 174.0–174.9) with or without regional metastases (ICD-9: 196.0–196.9). Second, remote metastases had to be documented (ICD-9: 197.0–199.0). Third, the respective women had to have diagnoses indicating terminal illness (eg, septicaemia, pathological fractures or gastrointestinal haemorrhage; for details see reference[3]).

To adapt the algorithm for use in German claims data, the diagnostic criteria were implemented according to the ICD-10-GM. We considered a 1-year period before death to evaluate the criteria for the identification of BC deaths.

As in the Canadian approach, we considered information from hospitalisations. Furthermore, in extension of Gagnon *et al*,[3] we also considered outpatient diagnoses and differentiated between hospital main discharge diagnoses and hospital secondary diagnoses. The latter facilitated further options to adapt the algorithm. The criteria used by Gagnon *et al*[3] were expanded by information on comorbidity. Overall, we ended up with eight criteria (Cr1–8) (online supplementary data 2). Cr1–5 were set a priori in which Cr1–4 with the intention to mimic the algorithm criteria of Gagnon *et al*.[3] Cr6–8 were set after additional considerations following a descriptive analysis of the official causes of death of those cases selected as false positives by algorithm version C to evaluate which were the most prominent groups of cancer causes of death among these cases. Different combination of all criteria resulted in six different versions of the algorithm (A–F) as shown in figure 1.

Statistical analyses

To evaluate the performance of each algorithm, we randomly divided the subsample for individual-level

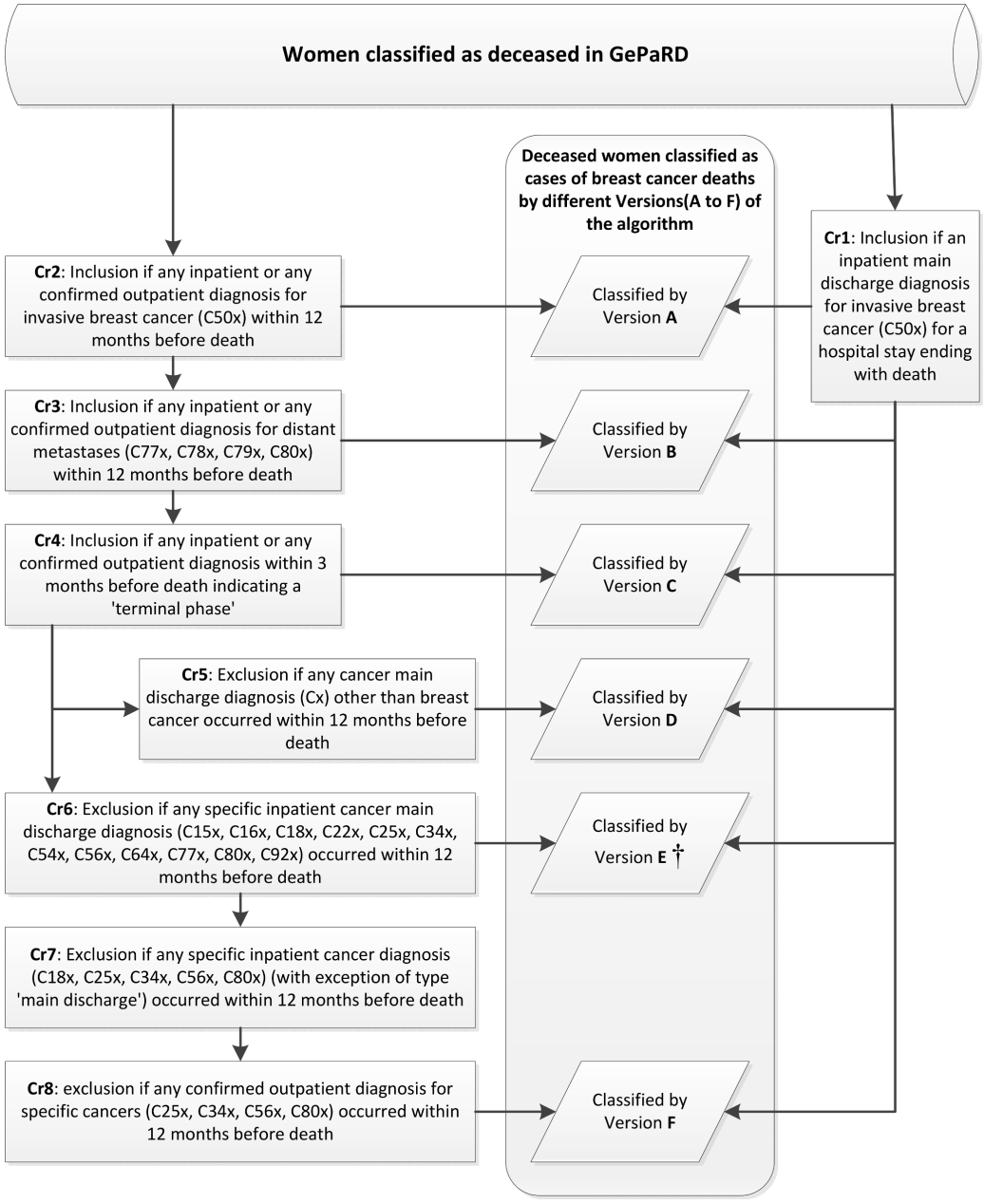

Figure 1 Data flow chart showing the criteria (Cr1–Cr8) and decision paths used in the different algorithm versions (A–F) to identify potential cases of breast cancer death among deceased women in the German Pharmacoepidemiological Research Database (GePaRD); cases fulfilling Cr1 were selected in each algorithm version; cases selected via the left decision path had to fulfil all criteria depicted on the path to the respective algorithm version (eg, version E, left path: Cr2 and Cr3 and Cr4 and Cr6); diagnosis codes listed for the criteria are given in ICD-10-GM format (only three digit ICD codes are shown). †Selected for further evaluation. ICD-10-GM, German modification of the International Classification of Diseases, 10th Revision.

analyses into 10 disjunct equally sized test samples (similar to a cross-validation but without using training samples for the optimisation of algorithm parameters). For each test sample, we calculated sensitivity, specificity, negative predictive value (NPV) and positive predictive value (PPV) for each version of the algorithm, using the official CoD information from the CR-NRW as gold standard. The results of these test samples are presented as median values and the corresponding variances in boxplot graphs, with the aim of identifying an algorithm that exhibited the best measures of accordance. In contrast to the study population used by Gagnon *et al*,[3] our study sample included a high proportion of women who had died from causes other than BC. Therefore, to avoid high numbers of false positives, we put special emphasis on a high PPV in the selection of the most promising algorithm version, maintaining comparatively high values for specificity and sensitivity. Measures of accordance stratified by age were also calculated for the algorithm version selected on the basis of these criteria.

For population-level analyses, annual age-specific BC mortality rates were calculated by using the number of BC deaths according to the selected algorithm version as the numerator and the female population of GePaRD ever insured in the respective calendar year as denominator. These algorithm-based BC mortality rates were then compared with the official rates for the total German population as provided by the ZfKD.

For the BC mortality rates and the measures of accordance, corresponding 95% CIs were calculated according to the methods recommended by Newcombe and Altman.[8]

All analyses were carried out with SAS V.9.3.

RESULTS

Validation of algorithm-based CoD on the individual level

The study subsample comprised 22 413 deceased females, whose records were successfully linked with CR-NRW records and the official CoD information was added. The majority of this sample was aged 60–79 years (65.7%). BC was documented as the official CoD in 10.9% (table 1). Females with BC as the official CoD were younger than those dying from other diseases.

Algorithm A, which required at least one BC diagnosis during the twelve months before death, reached a sensitivity of 97.5% and a specificity of 93.0%. The NPV was 99.7% and the PPV was 62.9% (table 2, figure 2). The additional requirement of documented distant metastases during the 12 months before death (algorithm B) led to a lower sensitivity (94.3%), a higher specificity (95.5%), a comparable NPV (99.3%) and a higher PPV (72.4%). The additional consideration of diagnoses indicating a terminal phase of the disease during the 3 months before death (algorithm C) resulted in only minor changes of the validity measures. The additional requirement of the absence of other cancer main discharge diagnoses during the 12 months before death (algorithm D) led to a further reduction of the sensitivity (89.9%) and the NPV (98.8%) as well as a further increase of the specificity (97.6%) and the PPV (82.2%). Relaxing this criterion by considering main discharge diagnoses only for those cancer types which were the most frequent cancer CoDs among the false positives classified by algorithm C only slightly changed the quality measures (algorithm E). As the CoD 'cancer other than BC' was still frequent among the false positives of algorithm E, we expanded

Table 1 Official cause of death (breast cancer; other causes) among women of the study population deceased between 2006 and 2013 in North Rhine Westphalia, by age group

	Official cause of death				
	Other causes		Breast cancer (German modification of the International Classification of Diseases, 10th Revision: C50)		Proportion of breast cancer deaths among all deaths
	N	%	N	%	N
Age at death (years)					
40 to <45	54	0.27	8	0.33	12.90
45 to <50	763	3.82	124	5.06	13.98
50 to <55	1755	8.79	290	11.84	14.18
55 to <60	2475	12.40	357	14.57	12.61
60 to <65	2984	14.95	411	16.78	12.11
65 to <70	4503	22.56	555	22.65	10.97
70 to <75	5684	28.47	583	23.80	9.30
75 to <80	1745	8.74	122	4.98	6.53
All	19 963	100.00	2450	100.00	10.93

Records of the epidemiological cancer registry of North Rhine Westphalia.

Table 2 Numbers and derived measures of accordance between the official cause of death (CoD) as the 'gold standard' and the CoD classification based on different versions of the CoD algorithm for breast cancer deaths

Algorithm version	Criteria used in the algorithm version*	Measures of accordance (%): Median resulting from 10 test-samples			
		Sensitivity	Specificity	Negative predictive value	Positive predictive value
(A)	Cr1, Cr2	97.5	93.0	99.7	62.9
(B)	Cr1–Cr3	94.3	95.5	99.3	72.4
(C)	Cr1–Cr4	94.1	95.5	99.3	72.5
(D)	Cr1–Cr5	89.9	97.6	98.8	82.2
(E)	Cr1–Cr4, Cr6	91.3	97.4	98.9	81.7
(F)	Cr1–Cr4, Cr6–Cr8	66.2	98.4	96.0	83.8

Female residents of North Rhine Westphalia included in the German Pharmacoepidemiological Research Database who died between 2006 and 2013.
*For more details see figure 1 and online supplementary data 2.

the exclusion criteria by also considering inpatient secondary and outpatient diagnoses (algorithm F). Although we restricted this approach to cancers with a high fatality rate and a high prevalence in the study population to reduce false negatives, the sensitivity decreased to 66.2% while the specificity (98.4%), NPV (96.0%) and PPV (83.8%) changed only slightly when compared with the results of algorithm E.

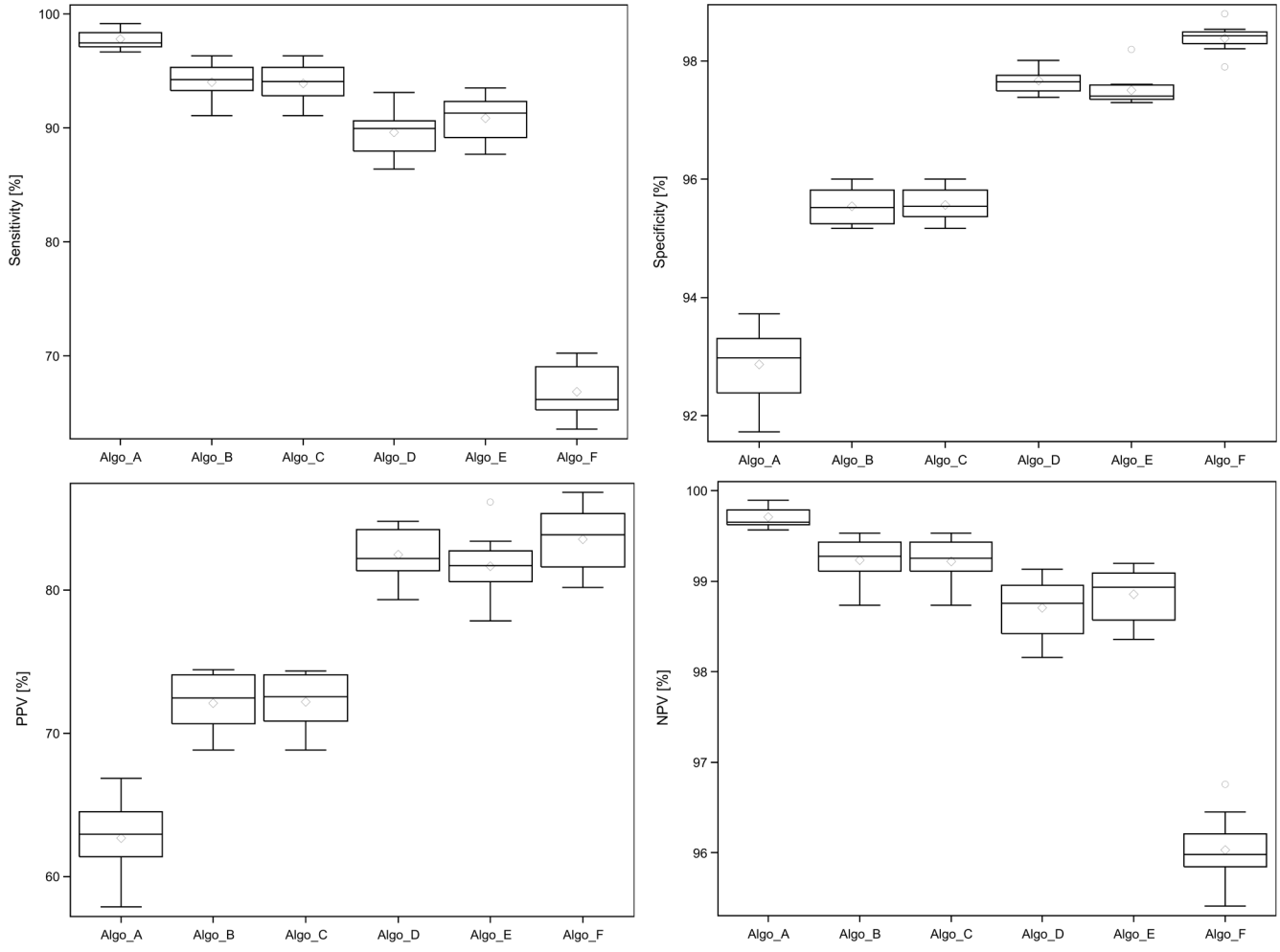

Figure 2 Boxplots for the results of the 10 disjunct equally sized test samples for sensitivity, specificity, negative predictive value (NPV) and positive predictive value (PPV) for the classification of six algorithm version (Algo_A to Algo_F) compared with the official cause of death.

Table 3 Measures of accordance between the individual official cause of death (CoD) as the 'gold standard' and the CoD classification based on the breast cancer death algorithm (E), by age group

Age group (years)	Breast cancer deaths (N) as classified by algorithm version (E)	Validity measures for accordance (%) (95% CIs)			
		Sensitivity	Specificity	Negative predictive value	Positive predictive value
40 to <50	156	96.2 (91.4 to 98.4)	96.5 (94.9 to 97.5)	99.4 (98.5 to 99.7)	81.4 (74.6 to 86.7)
50 to <65	1169	92.5 (90.8 to 94.0)	97.4 (97.0 to 97.7)	98.9 (98.6 to 99.1)	83.7 (81.5 to 85.8)
65 to <70	611	91.2 (88.5 to 93.3)	97.7 (97.2 to 98.1)	98.9 (98.5 to 99.2)	82.8 (79.6 to 85.6)
70 to <80	788	87.0 (84.3 to 89.2)	97.6 (97.3 to 98.0)	98.7 (98.5 to 99.0)	77.8 (74.8 to 80.6)

Among the versions with the highest values for specificity and PPV (D with 97.6% and 82.2%, E with 97.4% and 81.7%, F with 98.4% and 83.8%), we selected algorithm E for the further evaluation as it offered the highest value for sensitivity (91.3% vs 89.9% and 66.2%).

For algorithm E, sensitivity decreased with increasing age (from 96.2% in age group 40–50 years to 87.0% in age group 70–80 years) and the PPV was lowest in the highest age group (77.8% in age group 70–80 years vs 81.4%, 83.7% and 82.8% in the other age groups) while only marginal changes with age occurred for the specificity (between 96.5% and 97.7%) and the NPV (between 98.7% and 99.4%) (table 3).

Comparison of algorithm-based and official BC mortality rates at population level

The study samples exemplarily used for the calculation of the algorithm-based BC mortality rates in 2007, 2010 and 2012 comprised n=7 257 975, n=7 540 664 and n=7 825 758 females, respectively. It included n=47 763, n=55 013 and n=60 506 deceased women of which n=2709, n=2959 and n=3141, respectively, were classified as BC deaths by algorithm E. The BC mortality rates based on algorithm E agreed well with the data from the ZfKD in all age categories except for the highest age group (≥85 years) where the algorithm-based rates were 25%–30% lower than the national rates from the ZfKD (figure 3). This difference was significant in all study years.

DISCUSSION

In this study, we demonstrated the feasibility of a health claims data-based algorithm for the identification of BC deaths. Using German SHI claims data, we adapted the algorithm presented by Gagnon *et al*[3] for Canadian administrative hospital data by also including outpatient diagnoses and adding further criteria to optimise the performance of the algorithm. We validated the different algorithm versions by direct comparison with the official CoD on the individual level as well as by indirect comparison with official data on the population level.

Based on a sample of more than 22 000 deceased women from our data source, the most promising algorithm comprised data on the presence of BC, metastases, diagnoses indicating terminal illness and the absence of other tumours with a high case fatality rate as potentially competing CoDs. These individual criteria were applied within the quarter (terminal illness) or the year prior to death. The finally selected algorithm showed high

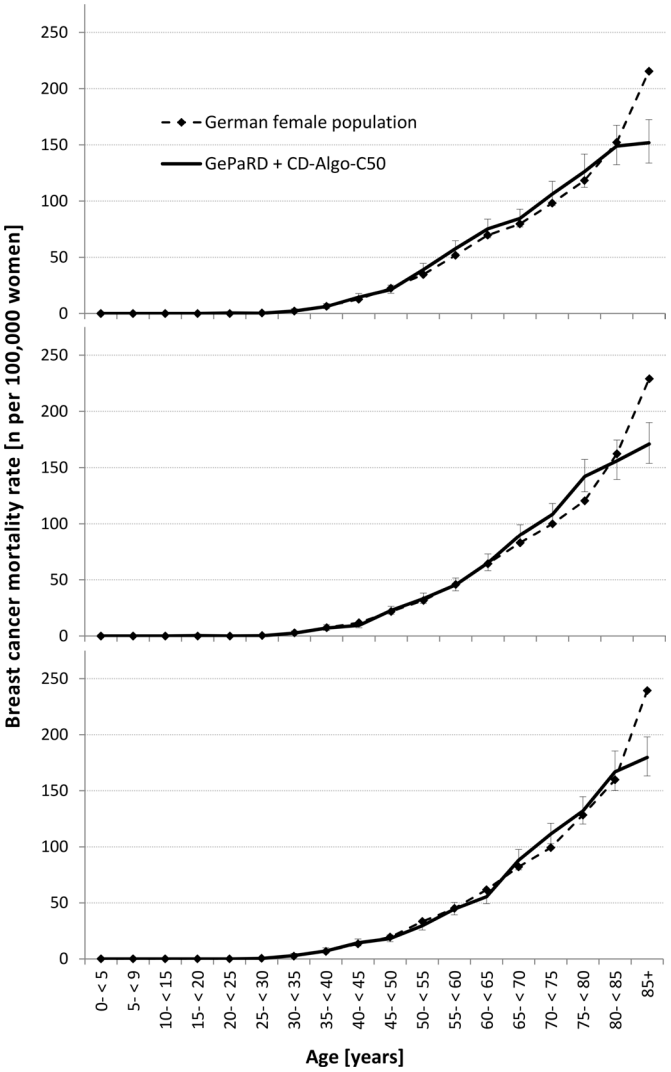

Figure 3 Official age-specific breast cancer mortality rates for German women in 2007, 2010 and 2012 as well as corresponding rates and 95% confidence limits (vertical bars) calculated using the female population included in the German Pharmacoepidemiological Research Database (GePaRD) in these years, respectively, and the classification of deceased women resulting from the application of the cause of death algorithm version (E) for breast cancer.

specificity and NPVs and represents a compromise in terms of a reasonably high sensitivity and a PPV that is not too low.

On the population level, the application of this algorithm resulted in a good agreement of algorithm-based age-specific BC mortality rates in the GePaRD sample with general population rates. However, the predictive performance (expressed as sensitivity and PPV) of the algorithm was better for deaths below the age of 70 than for older ages. Especially for the age group of 85 years and older, we saw a marked underestimation of the official rate by the algorithm-based rates.

One reason for this finding could be that the intensity of the search for metastases might be lower in very old patients with cancer compared with younger ones as elderly patients are less likely to receive additional cancer treatment.[9 10] Consequently, as the algorithm requires identifying metastasis diagnoses for the classification of BC deaths, the rate of false negative decisions by the algorithm would be higher among the elderly leading to a more pronounced underestimation of the BC mortality rate in this age group. Indeed, among cases with an official CoD 'BC' which were misclassified by the algorithm, the rate of those misclassified due to missing metastasis diagnoses increased with age (results not shown), which in part explains the slightly lower PPV for the age group 70 years and older compared with the younger age groups. As the sample enriched with the individual official CoD excluded subjects aged 80 years or older, we were not able to examine reasons for the underestimation on the individual level in more detail.

We explored the validity of several algorithm versions, which differed from the approach presented by Gagnon *et al*.[3] For this, we used an approach with 10 test samples to explore the variance of the estimates of the measures of accordance. Certain changes between two algorithm versions (version B vs C and version D vs E) led only to minor changes of the measures of accordance. In extension to version B, version C required the presence of at least one condition indicating terminal illness.[3] As the corresponding list of diagnoses is rather long, the inclusion of ambulant diagnoses in addition to inpatient diagnoses (in contrast to Gagnon *et al*) might have increased the probability of such diagnoses being present in an individual which could have weakened the discrimination of this criterion. In contrast to algorithm version D, version E did not exclude all individuals with any inpatient cancer main discharge diagnoses of cancer (criterion Cr5) but only those with such diagnoses for specific tumours (criterion Cr6). This restriction was intended to reduce false negatives by focusing on tumours with a higher fatality rate. However, this characteristic could be associated with a higher probability of such diagnoses occurring as a main discharge diagnosis compared with diagnoses with a lower fatality rate that would reduce the differences in discrimination between criterion Cr5 and criterion Cr6. The setting of our algorithm C was the version most similar to the algorithm used by these authors, however, with the major difference that we also considered outpatient diagnoses. Compared with their results (sensitivity 95%, specificity 89%, PPV 98%, NPV 77%) which were based on a sample of only 119 deceased women, our algorithm C showed a higher specificity and NPV, a similar sensitivity but a lower PPV. Adding further criteria lowered the rate of false positives but led to a decline in sensitivity (our algorithm versions D–F).

Of note, the comparison of the validity measures obtained in the two studies is constrained by the fact that, contrary to our approach of using the official CoD for validation, Gagnon *et al* relied on the decision of a palliative care specialist who reviewed the entire hospital medical chart to determine whether a study subject had died of BC or not. Furthermore, Gagnon *et al*[3] derived their algorithm from a sample of women with BC in need of end-of-life cancer care, while our approach intended to identify the outcome 'BC death' among deceased women in an epidemiological follow-up study. Therefore, the two studies may have prioritised different criteria to select the 'best-fitting' algorithm. It should further be noted that Canadian study setting, characterised by only a small proportion (16%) of women not dying of BC, that is, a high a priori probability of BC death, facilitated the generation of validity measures with a combination of high specificity, high PPV and low NPV. By contrast, our study sample comprised a proportion of 89% of women not dying of BC, that is, a low a priori probability of BC death, which more likely leads to results with combinations of high sensitivity, low PPV and high NPV. Another notable difference in the study setting was that we did not restrict our sample to women dying in hospital (about 50% of all cases included in our study). Therefore, comparisons between the study of Gagnon *et al* and our study should only be cautiously invoked. However, it is generally accepted that a high PPV is important to minimise bias in relative effect estimates, that is, a high PPV will be important when the algorithm is used for outcome definition in future studies on risk factors of BC death or the evaluation of screening measures.

Limitations

Some general limitations have to be considered when interpreting our results. First, this study shows that it appears principally feasible to adapt published algorithms to the specific characteristics and requirements of another data setting such as a claims database in another country. Of note, 87% of the German population are insured by SHIs all of which routinely collect claims data on the same structured data basis as GePaRD and the algorithm presented here is most likely also applicable to data from other German SHIs. In theory, the algorithm could even be transferred to other countries because the three-digit diagnoses used (figure 1) are the same in the international version of ICD-10. However, the purpose of collecting health data and coding practices may be different and this, in turn, could influence the informative quality (specificity) of the diagnoses included. Further,

different profoundness and details in the medical data at hand (eg, tumour staging, disease severity, laboratory measures) may necessitate modifications of the algorithm and subsequently also additional evaluation.

Second, our reference, the official CoD, is not in perfect agreement with the true underlying CoD. There is an acknowledged degree of uncertainty in the abstraction of the CoD from death certificates with some inter-rater variability.[11 12] For example, the medical examiner may occasionally not be aware of the full medical history omitting relevant concomitant disease data.[13 14] In contrast, the algorithm in our study was based on all medical diagnoses recorded in the year before death. Thus, using the official CoD as a gold standard for the validation of the algorithm was a pragmatic choice in this study. Therefore, achieving complete agreement between the official and the algorithm-based CoD 'BC' was not to be expected.

Third, the record linkage procedure applied to add the official CoD to the GePaRD data was based on a probabilistic method using pseudonymised data. The linkage method exhibited very low error rates in an independent validation study.[15] Additionally, the matches resulting from the probabilistic linkage used in this study were verified by a second, but different linkage method.[7] Mismatches in the linkage procedure with subsequent misclassification, which would result in an underestimation of the performance of the algorithm versions examined in this study, are thus expected to be low. Further, only 5.28% of the records of the original linkage sample had no successful link with CR-NRW data. The cases that could not be linked may be explained by data errors at the CR-NRW or in the core data of the SHIs concerning the personal identifiers used for the linkage.[7] These data errors (mostly entry errors or various spellings of complex names) are unlikely to be associated with the CoD. The proportion of 10.9% with an official CoD 'BC' among those with a successful link corresponds closely to national data of the Federal Statistical Office[16] in the observed age range. Thus, substantial biases introduced by the linkage procedure appear unlikely.

Fourth, although the age-specific BC mortality rates produced with algorithm E and the official BC mortality rates provided by the ZfKD showed hardly any differences one needs to keep in mind differences between the underlying study populations. While the latter are representing the general population of Germany, three of the four SHIs contributing data to GePaRD comprise relatively high proportions of insured persons with a better educational and economic position.[17] Social position is generally associated with morbidity and mortality risk,[18] and meta-analyses indicate increased BC incidence rates in women with higher socioeconomic status,[19 20] while the evidence regarding differences in BC mortality is heterogeneous.[19] Therefore, it is not clear whether the selection of SHIs represented in the database could have had an effect on the algorithm-based BC mortality rate. Additionally, the official mortality rates are based on the monocausal documentation of the underlying CoD. Thus, if two different cancer diseases are documented by the medical examiner on the death certificate, the official CoD is coded as 'malignant neoplasms of independent (primary) multiple sites' (ICD-10-GM: C97). Of note, for only 0.6% of our sample (n=67) was C97 the official CoD but n=10 of them were classified as BC deaths by algorithm E. This, however, limits the comparability of the algorithm-based mortality rates with official data only to a small extent.

Fifth, not all of the criteria defined for the different algorithm versions were defined independently from the data used for the evaluation of the algorithm performance. Three of eight criteria were defined after additional considerations following a descriptive analysis of the official causes of death of those selected as false positives results with another algorithm version which was based on a priori set criteria. Although we did not apply typical machine learning for the algorithm versions in our study and we did not directly analyse the diagnosis entries used by the algorithm to classify deceased women into 'BC deaths' and 'non-BC deaths', our procedure might have led to some overoptimistic results for two of the six tested algorithm versions. On the other hand, the boxplots of the results of the test samples showed only small variances for the calculated accordance measures of the test samples that indicates rather a robust algorithm.

CONCLUSION

Our study showed that the algorithm-based classification of BC as CoD, as adapted from a recently published Canadian algorithm, is feasible with German claims-based data (with the possible exemption of the age group 85+ years). This may indicate that such algorithms could also be adapted to healthcare databases in other countries. If individual linkage with official CoD is not possible, the algorithm is a useful tool that can be applied in epidemiological studies using BC mortality as an endpoint, for example, in the context of monitoring and evaluating routine MSPs. Given that in Germany, about 90% of the total female population are members of an SHI the algorithm could be used to investigate BC deaths in the vast majority of the German female population. However, given the variability observed with different algorithm versions, the transferability of the adapted algorithm to other data sources will certainly require further specific adaptations and additional validation.

Author affiliations
[1]Clinical Epidemiology, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany
[2]Health Services Research, IGES Institut GmbH, Berlin, Germany
[3]High-Profile Research Area Health Sciences, University of Bremen, Bremen, Germany
[4]Institute of Epidemiology and Social Medicine, Westfälische Wilhelms-Universität Münster, Münster, Germany
[5]State Cancer Registry North Rhine Westphalia, Münster, Germany
[6]SOCIUM Research Center on Inequality and Social Policy, University of Bremen, Bremen, Germany

Acknowledgements We would like to thank the AOK Bremen/Bremerhaven, the hkk Krankenkasse, the Techniker Krankenkasse (TK) and the DAK-Gesundheit, which provided data for this study. We also thank the Techniker Krankenkasse (TK) and the DAK-Gesundheit for their intensive support. The study presented here was conducted in the framework of a feasibility study for the evaluation of breast cancer-related mortality in the German mammography screening programme, which was funded by the Federal Office for Radiation Protection (Bundesamt für Strahlenschutz), the Federal Ministry for Environment, Nature Conservation, Building and Nuclear Safety (BMUB), the Federal Ministry of Health (BMG) and the Kooperationsgemeinschaft Mammographie under grant no. UFOPLAN 3610S40002 and 3614S40002. The publication of this article was funded by the Open Access Fund of the Leibniz Association.

Contributors IL, CO, UH, HWH and HZ conceived the study and planned the design. IL and UH obtained the permission from SHIs and administrative authorities to use the claims data for this study. IL and HWH managed the record linkage between claims data and epidemiological cancer registry. IL performed the statistical analyses. IL and CO wrote the first draft of the manuscript. UH, HWH, JC and HZ provided supervision and critical review of the manuscript. All authors contributed to and have approved the final manuscript.

Funding The present study was funded by the Federal Office for Radiation Protection (Bundesamt für Strahlenschutz) of the Federal Ministry for Environment, Nature Conservation and Nuclear Safety (BMU), the Federal Ministry of Health (BMG) and the Kooperationsgemeinschaft Mammographie under grant no UFOPLAN 3610S40002 and 3614S40002.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The utilisation of SHI data for scientific research is regulated by the Code of Social Law in Germany (SGB X). All four involved SHIs, the Federal Social Insurance Authority (as the responsible authority of the three nationwide-operating SHIs) and the regional Senator for Science, Health and Consumer Protection (for the one regional SHI) approved the use of the data for this study.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement In accordance with German data protection regulations, access to the data of the German Pharmacoepidemiological Database must not be given to third parties. Furthermore, as we are not the owners of the data we are not legally entitled to grant access to the data.

REFERENCES

1. Malek D, Kääb-Sanyal V. Implementation of the German Mammography Screening Program (German MSP) and First Results for Initial Examinations, 2005-2009. *Breast Care* 2016;11:183–7.
2. Fuhs A, Bartholomäus S, Heidinger O, *et al*. Evaluation of effects of the Mammography-Screening-Program on breast cancer mortality: feasibility study fort he record linkage of different data sources in North Rhine-Westphalia [Evaluation der Auswirkungen des Mammographie-Screening-Programms auf die Brustkrebsmortalität: Machbarkeitsstudie zur Verknüpfung verschiedener Datenquellen in Nordrhein-Westfalen]. *Bundesgesundheitsbl* 2014;57:60–7.
3. Gagnon B, Mayo NE, Laurin C, *et al*. Identification in administrative databases of women dying of breast cancer. *J Clin Oncol* 2006;24:856–62.
4. Ohlmeier C, Langner I, Hillebrand K, *et al*. Mortality in the German Pharmacoepidemiological Research Database (GePaRD) compared to national data in Germany: results from a validation study. *BMC Public Health* 2015;15:570.
5. Garbe E, Mikolajczyk RT, Banaschewski T, *et al*. Drug treatment patterns of attention-deficit/hyperactivity disorder in children and adolescents in Germany: results from a large population-based cohort study. *J Child Adolesc Psychopharmacol* 2012;22:452–8.
6. Pigeot I, Ahrens W. Establishment of a pharmacoepidemiological database in Germany: methodological potential, scientific value and practical limitations. *Pharmacoepidemiol Drug Saf* 2008;17:215–23.
7. Langner I, Krieg V, Heidinger O, *et al*. [Enrichment of Claims Data with Official Causes of Death Using a Record Linkage with the Epidemiological Cancer Registry of North Rhine-Westphalia: Feasibility Study and Comparison of Procedures]. *Gesundheitswesen* 2018.
8. Newcombe RG, Altman DG, *et al*Proportions and their differences. In: Altman DG, Machin D, Bryant TN, Gardner MJ, . eds. *Statistics with confidence*. 2nd ed. Bristol: JW Arrowsmith Ltd, 2001:45–56.
9. Chawla N, Yabroff KR, Mariotto A, *et al*. Limited validity of diagnosis codes in Medicare claims for identifying cancer metastases and inferring stage. *Ann Epidemiol* 2014;24:666–72.
10. Warren JL, Mariotto A, Melbert D, *et al*. Sensitivity of Medicare Claims to Identify Cancer Recurrence in Elderly Colorectal and Breast Cancer Patients. *Med Care* 2016;54:e47–e54.
11. Jahn I, Jöckel KH, Bocter N, *et al*. Studie zur Verbesserung der Validität und Reliabilität der amtlichen Todesursachenstatistik. *Baden-Baden: Nomos* 1995:230.
12. Giersiepen K, Greiser E. [Coding of cause of death for mortality statistics-a comparison with results of coding by various statistical offices of West Germany and West Berlin]. *Offentl Gesundheitswes* 1989;51:40–7.
13. Schubert-Fritschle G, Eckel R, Eisenmenger W, *et al*. Quality of indications on death certificates- Is the cause of death statistic concerning cancer better than its reputation? [Qualität der Angaben von Todesbescheinigungen - Ist die Todesursachenstatistik zu Krebserkrankungen besser als ihr Ruf?]. *Dtsch Arztebl* 2002;99:50–5.
14. Schröder AS, Wilmes S, Sehner S, *et al*. Post-mortem external examination: competence, education and accuracy of general practitioners in a metropolitan area. *Int J Legal Med* 2017;131:1701–6.
15. 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.
16. Federal Statistical Office. Mortality Statistics. 2019 https://www-genesis.destatis.de/genesis/online/data;sid=7A797275F348011F6C4D45595BCE75CD.GO_1_3?operation=abruftabelleAbrufen&selectionname=23211-0004&levelindex=0&levelid=1550570635876&index=4 (Accessed 19 Feb 2019).
17. Hoffmann F, Icks A. [Structural differences between health insurance funds and their impact on health services research: results from the Bertelsmann Health-Care Monitor]. *Gesundheitswesen* 2012;74:291–7.
18. Mackenbach JP, Kunst AE, Cavelaars AE, *et al*. Socioeconomic inequalities in morbidity and mortality in western Europe. The EU Working Group on Socioeconomic Inequalities in Health. *Lancet* 1997;349:1655–9.
19. Akinyemiju TF, Genkinger JM, Farhat M, *et al*. Residential environment and breast cancer incidence and mortality: a systematic review and meta-analysis. *BMC Cancer* 2015;15:191.
20. Lundqvist A, Andersson E, Ahlberg I, *et al*. Socioeconomic inequalities in breast cancer incidence and mortality in Europe-a systematic review and meta-analysis. *Eur J Public Health* 2016;26:804–13.