Cancer Epidemiology 98 (2025) 102906

Contents lists available at ScienceDirect

Cancer Epidemiology

journal homepage: www.elsevier.com/locate/canep

Algorithm development for identifying breast cancer incident cases and epidemiological updates: A cohort study based on multiple secondary sources

Andrea Faragalli [a,1], Marica Iommi [a,*,1], Donatella Sarti [b,c], Chiara Peconi [b,c], Marco Pompili [d], Emilia Prospero [b,c], Flavia Carle [a,d], Rosaria Gesuita [a,e]

[a] Center of Epidemiology, Biostatistics and Medical Information Technology, Department of Biomedical Sciences and Public Health, Università Politecnica delle Marche, Ancona, Italy
[b] Department of Biomedical Sciences and Public Health, Section of Hygiene Preventive Medicine, and Public Health, Università Politecnica delle Marche, Ancona, Italy
[c] Registro Tumori Regionale Marche, Regional Health Agency Marche Region, Ancona, Italy
[d] Regional Health Agency Marche Region, Ancona, Italy
[e] IRCCS INRCA, Ancona, Italy

ARTICLE INFO

Keywords:
Breast cancer
Epidemiology
Cancer Surveillance
Healthcare Utilization Databases
Incidence Trends
Algorithm Validation

ABSTRACT

Purpose: This study aimed to develop and validate an algorithm for identifying incident breast cancer (BC) cases using Healthcare Utilization Databases (HUDs) and to assess BC incidence trends in the Marche Region, Italy, from 2010 to 2021.

Methods: This population-based longitudinal study included women aged ≥ 18 years residing in Marche. The HUDs Algorithm was developed to identify new BC cases using hospital discharge, outpatient, and beneficiary databases, and it was validated against the Cancer Registry by evaluating agreement, sensitivity, and positive predictive value (PPV). Age-standardized BC incidence rates were estimated. A Poisson regression model was used to assess trends, including comparisons between pre/post COVID-19 pandemic periods.

Results: Validation results showed a sensitivity of 81.2 % and PPV of 85.0 %. A total of 18,158 incident BC cases were identified, with a mean incidence rate of 224.7 per 100,000 person-years (95 % CI: 221.5–228.0). No significant increase in BC incidence was observed over time, but a marked decline occurred in 2020–2021, likely due to COVID-19-related disruptions.

Conclusions: HUDs can be a valuable complementary data source, providing additional information useful for timely epidemiological surveillance and supporting rapid public health responses in cases where Cancer Registry data are delayed. Further refinements and integration with other data could enhance the accuracy of the HUDs Algorithm.

1. Introduction

Breast cancer (BC) remains a major oncological challenge, affecting women of all ages worldwide, with incidence rates steadily increasing across various populations [1]. Italy counted approximately 55,900 new diagnoses in women in 2023, and 15,500 BC-related deaths were estimated in 2022 [2]. According to the Italian Association of Cancer Registries (AIRTum), BC incidence has risen by 0.3 % in recent decades [3]. This upward trend not only emphasizes the disease's impact on individuals' health and well-being but also highlights the growing burden on healthcare systems, especially considering that survival rates are currently very high [2]. The downside of this high survival rate is the high risk of recurrence: previous research highlights the long-term nature of recurrence in women with breast cancer, with a 20-year recurrence risk ranging from 13 % to 41 %, depending on tumor size and lymph node involvement [4,5].

* Correspondence to: Via Tronto 10/A, Ancona 60126, Italy.
E-mail addresses: a.faragalli@staff.univpm.it (A. Faragalli), m.iommi@staff.univpm.it (M. Iommi), sarti.donatella05@gmail.com (D. Sarti), peconi.chiara@gmail.com (C. Peconi), marco.pompili@regione.marche.it (M. Pompili), e.prospero@staff.univpm.it (E. Prospero), f.carle@staff.univpm.it (F. Carle), r.gesuita@staff.univpm.it (R. Gesuita).
[1] Joint first authors: Andrea Faragalli and Marica Iommi contributed equally and share first authorship.

https://doi.org/10.1016/j.canep.2025.102906
Received 9 June 2025; Received in revised form 1 August 2025; Accepted 12 August 2025
Available online 19 August 2025

Traditionally, cancer registries have played a crucial role in gathering comprehensive, continuous, and systematic data on neoplasms, allowing for detailed analyses of cancer incidence and prevalence, geographic distribution, and trends over time [6–8]. However, the labour-intensive and time-consuming processes of data collection and verification often result in significant delays, with data becoming available months or even years after diagnosis, diminishing the cancer registries' usefulness in timely cancer prevention and control efforts [9].

To overcome these limitations and enhance the responsiveness of cancer surveillance systems, there is growing interest in leveraging health administrative data, such as Healthcare Utilization Databases (HUD). In recent decades, Italy's healthcare information system has significantly advanced its databases by improving their coverage and accuracy [10]. Tumour registries have benefited from this digital transformation, which has simplified the identification of new cases and reduced dependence on manual tasks [11]. HUD data sources, including Hospital Discharge Records and the Outpatient Care Database, offer a more immediate view of healthcare interactions within the population, potentially allowing for quicker identification of new cancer cases.

The objectives of this study are twofold: to develop and validate a novel algorithm to identify new cases of BC based on Healthcare Utilization Databases, by comparing it with the gold-standard represented by the Marche Region's Cancer Registry; and to estimate breast cancer incidence rates and trends among women over an extended period (2010–2021) by combining data from both the Cancer Registry and Healthcare Utilization Databases.

2. Materials and methods

2.1. Study design and population

This is a population-based longitudinal study comprising women aged 18 years and older, residing in the Marche Region and beneficiaries of the National Health Service (NHS), from 2010 to 2021. Marche is a region of Central Italy with 1498,236 inhabitants as of January the 1st, 2021, of whom 659,665 are adult women. Given that breast cancer in males is extremely rare, we conducted the study considering only the condition in the female population.

2.2. Data sources

The HUDs cover the entire population residing in the Marche Region who benefit from the Regional Health Service; all residents are included in the Regional Beneficiary database and can be traced in other HUDs if they need healthcare.

Three HUDs of the Health Regional System were employed to develop the identification algorithm: (1) the Regional Beneficiaries database (RBD) provides information on the beneficiary's date of birth, sex, start and end dates of regional healthcare and date of death, if any; (2) the Hospital Discharge Records (HDR) report information on admission and discharge dates, primary diagnosis and intervention and up to five secondary diagnoses and interventions (coded using International Classification of Diseases, 9th Revision, Clinical Modification, ICD-9-CM); (3) Outpatient Care Database (OCD) which reports outpatient specialistic visits and outpatient exams reimbursed by the NHS. Such databases were established by national law for administrative and health expenditure monitoring and control and provide useful information on healthcare service utilization, recording all the episodes of care for each beneficiary. Moreover, HUDs can be linked by the beneficiary code as a primary key, providing real-world data that can be analysed for epidemiological purposes [10]. This was an observational study based on secondary sources, in which patients were not directly involved.

2.3. Cancer registry

The Cancer Registry of the Marche Region (CR), established by a regional law in 2013 [12], records all new cases of cancer since January 1st, 2010, collecting information on basic demographic data (sex, age, date and place of birth, residence), and of characteristics of the individual tumour (site, morphology, behaviour, tumour stage, grading, markers, receptor status, biological indicators); at the time of this analysis (2023), the CR data were updated to 2018 (consolidated up to 2017). New cases are detected using multiple sources, including HDR, death certificates, histological reports, and, if necessary, medical records. Data quality and completeness are assessed using the Joint Research Centre–European Network of Cancer Registries Quality Check Software (JRC-ENCR QCS), in line with the international standards of cancer registration [13]. The quality indicators include: 1 % of cases diagnosed based on death certificate only, 97 % of cases confirmed by histological or cytological examination, and 2 % diagnosed through clinical or instrumental methods.

2.4. Identification Algorithm for BC Cases based on HUD (HUDs Algorithm)

The target population is represented by adult women beneficiaries residing in the Marche Region in 2017. In the Italian healthcare context, most newly diagnosed breast cancer cases result in hospital admission, especially when biopsy or mastectomy are required. Therefore, hospital discharge records represent a reliable starting point for case identification. The identification algorithm of new cases of BC was based on HUDs linked by a deterministic record linkage procedure and consisted in the following steps:

1. *Index Admission:* identification from HDR of the cases hospitalized in 2017 with a primary or secondary diagnosis of invasive breast cancer (ICD-9-CM: 174.0–174.9) or breast carcinoma in situ (ICD-9-CM: 233.0). If multiple hospitalizations occurred, only the first one was considered (index admission);
2. *Women beneficiaries*: for each BC case identified in step 1, sex, age and residence were determined through the linkage with RBD. Females aged $\geq$ 18 years and resident at the time of index admission were included;
3. *Diagnostic Procedures:* identification of mammography, breast ultrasound, or breast biopsy procedures performed within twelve months preceding the index admission from both HDR (ICD-9-CM: 87.37, 88.73, 85.11, 85.12) and OCD (Regional code: 87.37.1, 87.37.2, 88.73.1, 88.73.2, 91.46.5, 91.47.1, 85.11.0, 85.11.1);
4. *Index Date:* setting the index date as the date of the diagnostic procedure closest to the index admission. If no procedure was recorded, the date of the index admission was used as the index date. The date of the procedure closest to the index admission was selected to improve temporal alignment between diagnostic investigation and hospital-recorded diagnosis;
5. *Exclusion Criteria:* Women were excluded if they had: (a) at least one previous hospital discharge with a diagnosis of breast cancer (ICD-9-CM: 174.0–174.9, 233.0) in primary or secondary position between 01/01/2011 and the index date; (b) at least one hospital discharge with a diagnosis of history of breast cancer (ICD-9-CM: V10.3) in the period between 01/01/2011 and the index date included; (c) a diagnosis of non-specific breast neoplasm (ICD-9-CM: 238.3, 239.3) in the period between 01/01/2011 and the index date included; or (d) evidence of metastatic disease at initial presentation (ICD-9-CM: 197–199) in the period between 01/01/2011 and the index date included. These exclusions were applied to ensure accurate identification of newly diagnosed, non-metastatic breast cancer cases.

2.5. Statistical analysis

The HUD Algorithm was validated through a comprehensive comparison between the cases identified by the algorithm and those recorded in the Cancer Registry (CR) in 2017. To assess the algorithm's accuracy, the percentage of agreement (PA), the sensitivity, and positive predictive value (PPV) were calculated, each with their 95 % confidence intervals (95 % CI). PA was defined as the ratio of cases simultaneously identified by both the CR and the HUD Algorithm to the total cases identified by either the CR or the HUD Algorithm. Sensitivity was defined as the ratio of cases identified simultaneously by both the CR and the HUD Algorithm to the total cases identified by the CR. PPV was defined as the ratio of cases identified simultaneously by both the CR and the HUD Algorithm to the total cases identified by the HUD Algorithm.

The analysis of incidence was conducted over the twelve-year period (2010–2021), retrieving new BC cases from the Marche Region CR for the period 2010–2017 (since CR data were consolidated up to 2017) and through the HUD Algorithm for the period 2018–2021. The overall and annual BC incidence rates per 100,000 person-years (py) and their 95 % CIs were estimated by dividing the number of new cases over the total adult female population residing in the Marche Region (source: Italian National Statistics Institute, ISTAT, semi-sum of population present at the beginning and end of the calendar year). Age-standardize incidence rates are also provided using the European standard population, considering the EU-27 +EFTA (27 countries of European Union and European Free Trade Association) average populations based on the 2011–20 projections (data provided by Eurostat) [14], and the 2019 Global Burden Disease world female standard population (data provided by Global Health Data Exchange) [15]. The use of both standard populations allows for comparisons at both European and global level.

Poisson regression was used to estimate the age-adjusted incidence trend. Observed rates for the years 2018 and 2019 were compared with those predicted by the model based on CR data for the period 2010–2017, in order to assess whether the incidence estimates based on HUDs were in line with the CR period trend.

A second evaluation concerned the two years of the COVID-19 pandemic, assuming a different incidence trend conditioned by the emergency situation. Therefore, the observed rates for the years 2020 and 2021 were compared with those predicted by the model based on both sources for the period 2010–2019.

Finally, a sensitivity analysis was conducted to compare the incidence rates across different age classes observed using only the 2010–2017 Cancer Registry data with those observed using the HUD Algorithm for 2018–2019 (pandemic years were purposely excluded from this analysis). The comparison was performed using the Incidence Rate Ratio (IRR, $IR_{2010–2017}$ as reference category) and its 95 % CI to determine whether the HUD Algorithm provided consistent estimates of incidence rates across age groups.

All analyses were performed using the R statistical software and the significance level was set at 5 %.

3. Results

3.1. Validation process

The identification process of the HUD algorithm in 2017 is reported in the flow-chart in Fig. 1. Briefly, 1645 women were identified by HUDs as being discharged for the first time with a diagnosis of breast cancer. Of these, 68 were excluded due to a prior admission with a diagnosis of a history of breast cancer, 23 for having a prior admission for an unspecific breast cancer, and 65 for a prior admission for metastatic disease, either in the index admission or during the period from 01/01/2011 to the index date.

The validation was performed using data from the year 2017, comparing the 1489 new BC cases identified by the HUD Algorithm with the 1559 new cases reported in CR. Table 1 summarises the intersections between the HUD Algorithm and the CR: the incident cases correctly

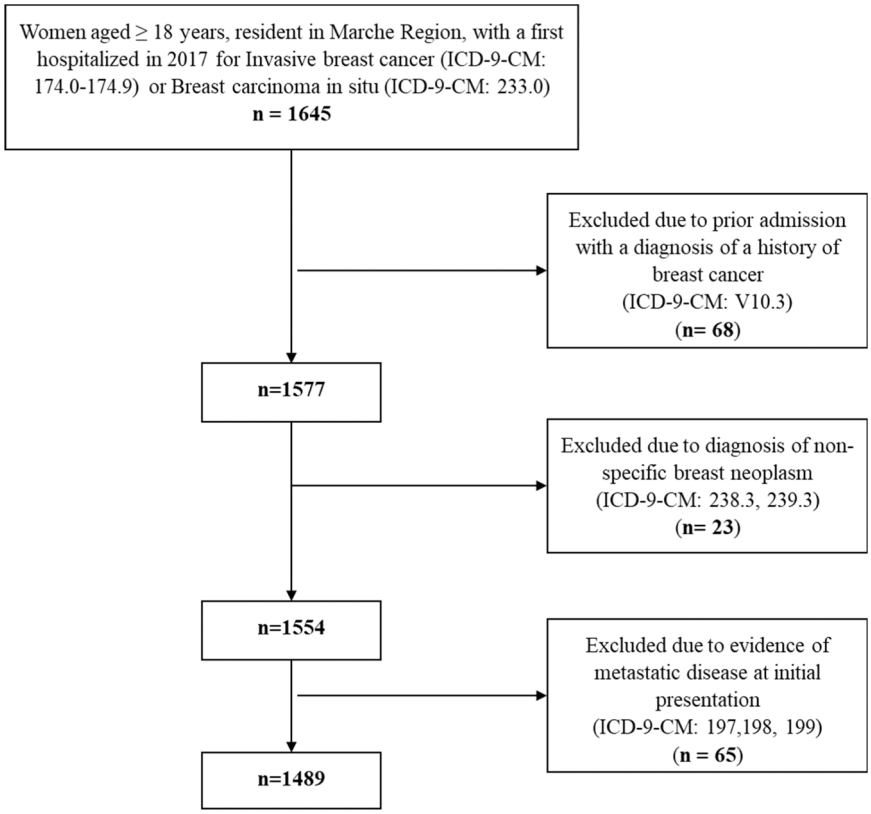

Fig. 1. Flow-chart of the identification process of the Algorithm for BC Cases based on HUD (HUDs Algorithm).

Table 1
Comparison of incident cases (IC) of Breast Cancer identified by the algorithm based on regional Healthcare Utilization Databases (HUDs) and the Cancer Registry of the Marche Region (CR).

	1st evaluation	2nd evaluation
IC of 2017 in the CR n = 1559 (gold standard)		
IC of 2017 identified by the HUDs Algorithm (n_{HUD})	1489	1489
Cases present in the CR, IC in 2017[(a)]	1252	1266
Cases present in the CR, IC in a different year[(b)]	122	277
IC in 2017 by the HUDs Algorithm, recorded before 2017 by the CR	*36*	*159*
IC in 2017 by the HUDs Algorithm, recorded after 2017 by the CR	-	*32*
IC in 2017 by the CR, detected before 2017 by the HUDs Algorithm	*51*	*51*
IC in 2017 by the CR, detected after 2017 by the HUDs Algorithm	*35*	*35*
IC of 2017 in the CR, did not meet the inclusion criteria[(c)]	113	113
IC of 2017 identified only by the HUDs Algorithm[(d)]	201	33[#]
IC of 2017 in the CR not detect by the HUDs Alghoritm[(e)]	108	104[*]
Percentage of agreement (95 % CI) [100*a/(a+b+c+d+e)]	69.7 % (67.5–71.8)	70.6 % (68.4–72.7)
Sensitivity (95 % CI) [100*a/n]	80.3 % (78.2–82.2)	81.2 % (79.2–83.1)
Positive Predictive Value (95 % CI) [100*a/n_{HUD}]	84.1 % (82.1–85.9)	85.0 % (83.1–86.8)

IC: incidence cases; CR: Cancer Registry of the Marche Region; 95 %CI: 95 % Confidence Interval;
#From 201 to 33: 14 cases confirmed as incident in 2017 by CT, 32 were incident cases of CT in the following year, 122 were relapsed cases.
[*] From 108 to 104: 3 non-residents at the time of diagnosis and 1 case incident in 2002.

identified by the HUD Algorithm were 1252, yielding a percentage of agreement (PA) of 69.7 % (95 % CI 67.5–71.8), sensitivity of 80.3 % (95 % CI 78.2–82.2), and positive predictive value (PPV) of 84.1 % (95 % CI 82.1–85.9). For 122 women, the year of diagnosis did not match between the two sources. Additionally, 113 women were identified as incident cases by the CR but excluded by the HUD Algorithm based on eligibility criteria, 108 were registered in the CR but not traced by the algorithm, and 201 were identified exclusively by the HUD Algorithm.

An in-depth examination of cases identified only by the HUD Algorithm or only by CR was carried out. This second evaluation revealed that among the 201 cases identified solely by the HUD Algorithm, 14 had been correctly identified and were, therefore, included in the CR as new breast cancers in 2017; 32 were incident cases in the following year (misclassification of the incidence date by the HUD Algorithm); 122 were cases of breast cancer relapse (misclassification of incidence cases by the HUD Algorithm). Among the 108 cases in the CR alone, four were not detected by the HUD Algorithm as 3 of them were non-residents at the time of diagnosis and one was an incident case in 2002 as reported in the CR. Considering the results of the second evaluation, the algorithm correctly identified 1266 incident cases, resulting in a PA of 70.6 % (95 % CI 68.4–72.7), sensitivity of 81.2 % (95 % CI 79.2–83.1), and PPV of 85.0 % (95% CI 83.1–86.8).

For the cases present only in the CR (n = 104), it was found that in 33 cases, the hospital discharge related to the BC diagnosis had an erroneously compiled unique identification code, and in 22 cases, an ICD-9-CM code different from 174.* or 233.0 was erroneously reported. Finally, 49 women were not tracked by HUD databases because the case was confirmed as BC by the CR only through histology or through a death certificate.

3.2. BC incidence rate

Between 2010 and 2021, there were 18158 new BC diagnoses in the adult female resident population of the Marche Region, resulting in a mean period incidence rate of 224.7 cases per 100000 person-years (/100000 py) (95 % CI 221.5–228.0).

Table 2 shows BC incidence cases, person-years, and rates by year of diagnosis: the highest incidence rate was observed in 2016 (239.18/100000 py, 95 % CI 227.7–251.1) while the lowest were observed during the two pandemic years (in 2020: 189.91/100000 py, 95 % CI 179.6–200.7; in 2021: 210.32/100000 py, 95 % CI 199.4–221.7).

3.3. Trend analysis

The annual age-adjusted breast cancer incidence trend for 2010–2017, based on the cases recorded in the CR, was stable over time ($\widehat{\beta} = 0.002,\ 95\%CI: -0.006; 0.009, p = 0.696$).

The observed rates for the biennium 2018–2019, calculated using cases detected by the HUDs Algorithm, did not significantly differ from those predicted (2018: $p = 0.326$; 2019: $p = 0.557$) by the trend estimation model (Fig. 2).

During the pandemic years, 2020 and 2021, the observed incidence rates were significantly lower (Figs. 3), 189.9/100000 py (95 % CI 179.6–200.7) and 210.3/100000 py (95 % CI 199.4–221.7), respectively, compared to the corresponding rates estimated by the model based on data from 2010 to 2019, 237.7/100000 py (95 % CI 226.1–249.8) and 238.5/100000 py (95 % CI 226.8–250.6), respectively (both years: $p < 0.001$).

3.4. Sensitivity analysis

The sensitivity analysis revealed differences in incidence rates across different age classes according to the data source (Fig. 4). For the age group 50–59 years (IRR=0.89, 95 % CI: 0.82–0.97) and 90 + years (IRR=0.68, 95 % CI: 0.53–0.88) a significant underestimation in the incidence rates by the HUDs Algorithm was observed; a significant overestimation was observed in the age class 70–79 years (IRR=1.15, 95 % CI: 1.06–1.25); finally, the rates estimated using HUDs Algorithm were closely aligned with those obtained by CR data in the other age groups.

4. Discussion

This study provides a comprehensive understanding of the possibility of using healthcare utilisation databases to assess the incidence of breast cancer and evaluate the trend over time. HUDs are secondary data sources that have been widely used in epidemiology for several decades because they contain useful information for the assessment of population health conditions [16,17]. To use HUDs in epidemiology, it is necessary to combine information from different databases using appropriate linking procedures, taking into account the objective, study design, and outcome measures. In particular, when analysing epidemiological measures such as incidence, look-back periods must always be considered [18]. In our study, the linkage of data from multiple HUDs is a winning strategy in detecting new breast cancer cases from a well-defined and circumscribed population; we found a high probability of being detected by the HUD algorithm when present in the CR (sensitivity above 80 %), and a low frequency of false positives (n = 223) with a PPV of 85 %. In addition, we found a fair degree of agreement between the HUD Algorithm and the CR. These results are consistent with those observed in three Italian studies based on local HUDs: Baldi et al. [19] and Yuen et al. [20] performed a population-based study and validated the identification process of incident cases using their respective cancer registries; Abraha et al. [21] conducted a multicentre study evaluating only a sample of new breast

Table 2
New cases of breast cancer and incidence rate per 100000 py (95 % CI) by year of diagnosis and over the entire 2010–2021 period. Cases were retrieved from Cancer Registry between 2010 and 2017 and from Healthcare Utilization Databases between 2018 and 2021 using the identification HUDs algorithm.

Data source	Year of diagnosis	New cases	Person-year*	IR	(95 % CI)	SIR Europe	SIR GBD
Cancer Registry	2010	1483	678,024	218.7	(207.7–230.1)	161.4	116.0
	2011	1518	679,659	223.3	(212.3–234.9)	164.3	119.4
	2012	1517	680,376	223.0	(211.9–234.5)	164.1	119.9
	2013	1626	680,596	238.9	(227.4–250.8)	173.2	125.9
	2014	1492	679,753	219.5	(208.5–230.9)	156.2	113.0
	2015	1600	677,576	236.1	(224.7–248.0)	167.2	120.3
	2016	1614	674,799	239.2	(227.7–251.1)	168.5	121.6
	2017	1559	672,044	232.0	(220.6–243.8)	161.4	115.8
Healthcare Utilization Databases Algorithm	2018	1540	669,534	230.0	(218.7–241.8)	159.0	112.9
	2019	1568	667,392	234.9	(223.5–246.9)	162.3	116.3
	2020	1259	662,959	189.9	(179.6–200.7)	130.1	93.0
	2021	1382	657,094	210.3	(199.4–221.7)	143.8	103.9
Period 2010 - 2021		18,158	8079,803	224.7	(221.5–228.0)	159.3	114.8

* Female population ≥ 18 years residing in the Marche Region (source: Italian National Statistics Institute, ISTAT, semi-sum of population present at the beginning and end of the calendar year) **IR**: Incidence rate; **95 % CI**: 95 % Confidence interval; **SIR Europe**: European standard population considering the EU-27 +EFTA average populations based on the 2011–20 projections (Eurostat); **SIR GBD**: Global Burden Disease world female standard population (Global Health Data Exchange). All rates are reported per 100,000 person-year

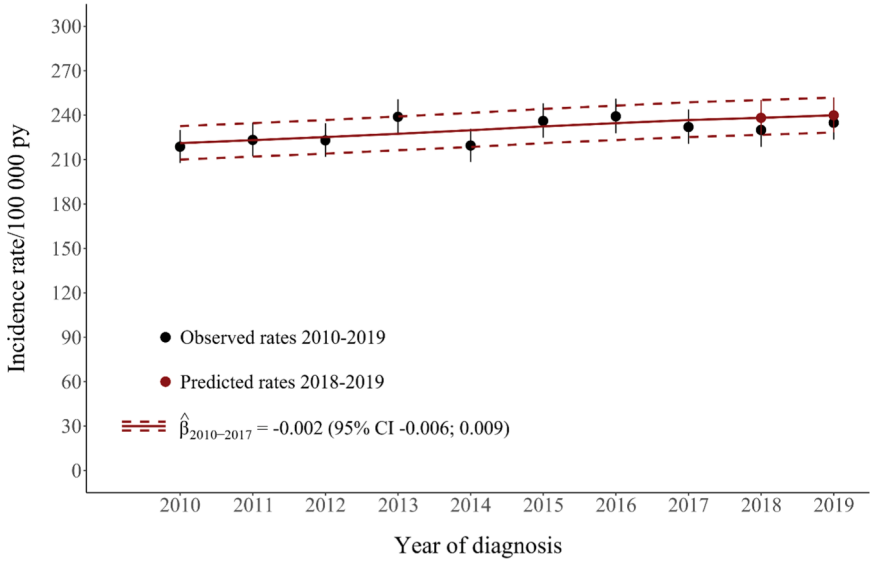

Fig. 2. Trend analysis of breast cancer incidence rates. Circles and vertical black lines: observed rates and 95 % confidence interval (CI) per year, 2010–2019. Solid red line: incidence trend estimated by the Poisson regression model adjusted by age classes during the period 2010–2017 (Cancer Registry data). Dotted red lines: 95 % CIs of the estimated incidence trend. Circles and vertical red lines: predicted rates for 2018–2019 estimated by the regression model. py: person-year.

cancer diagnoses and the validation process was based on medical charts. Furthermore, in all three studies, the identification process was based only on hospital discharge records that reported a breast cancer code exclusively in primary diagnosis; Baldi et al. and Yuen et al. limited the identification to women for whom a surgical procedure was simultaneously recorded in the same discharge record. The HUDs Algorithm proposed here broadens this definition by selecting new cases based on both primary and secondary diagnoses, does not limit identification to women who have undergone surgery, and incorporates outpatient diagnostic procedure to more accurately determine the date of breast cancer onset.

An overall incidence rate of 224.7 per 100000 person-years was estimated in the Marche Region between 2010 and 2021. In the two years preceding the COVID-19 pandemic, the Marche region recorded approximately 30 more new cases per 100000 person-years than the estimated Italian incidence rate (232.5 vs 203.0, respectively), as reported in AIRTum publications [22]. This difference is unlikely to be due to the use of the HUD algorithm, as reported in Table 2, for estimating breast cancer incidence, as the incidence rate based on CR data in Marche (230.3) was also considerably higher than the national estimate (183.6) for the period 2011–2017. This discrepancy can be partly explained by differences in data sources. The Italian data are estimated to provide relevant information rather than a precise number, which is susceptible to error, as acknowledged in AIRTum publications; conversely, the Marche Region's registry reports data with histological confirmation. Additionally, AIRTum data are based on accredited cancer registries covering 70 % of the Italian population, with just one province from the Marche Region included.

Comparing the age-standardised incidence rates (European Standard Population) with those of Italy, Marche Region recorded a higher incidence rate (164.4) than Northern Italy (161.9), Central Italy (141.7) and Southern Italy (124.9) in the period 2010–2015 [22]. The Marche Region also recorded a higher rate than the U.S. mean age-adjusted rates of new cases for the period 2010–2021, which was 132.6 per 100000 person-years [23].

In the two pandemic years, although an inverse result emerged, with the Italian incidence rate higher than that of the Marche Region, 212.8 versus 200.1 respectively, the gap narrowed considerably, suggesting a

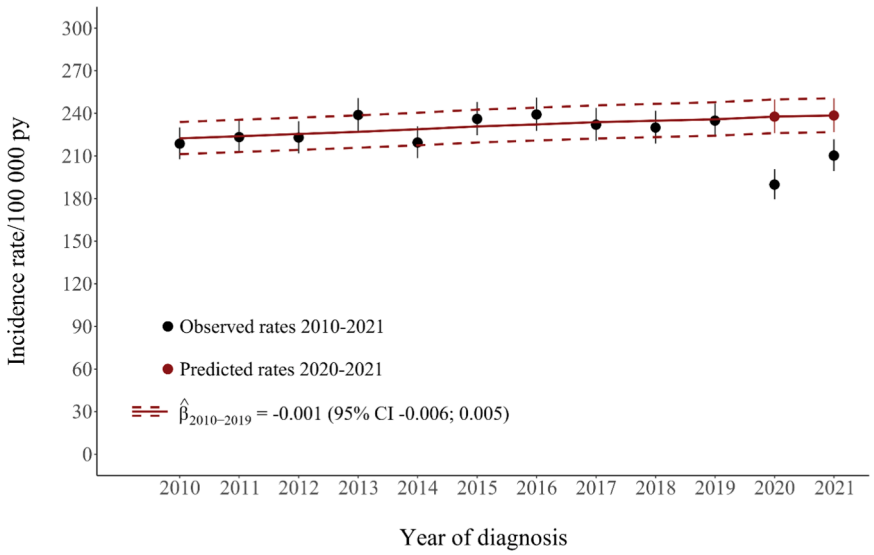

Fig. 3. Trend analysis of breast cancer incidence rates. Circles and vertical black lines: observed rates and 95 % confidence interval (CI) per year, 2010–2021. Solid red line: incidence trend estimated by the Poisson regression model adjusted by age classes during the period 2010–2019 (Cancer Registry data 2010–2017, Algorithm 2018–2019). Dotted red lines: 95 % CIs of the estimated incidence trend. Circles and vertical red lines: predicted rates for 2020–2021 estimated by the regression model. py: person-year.

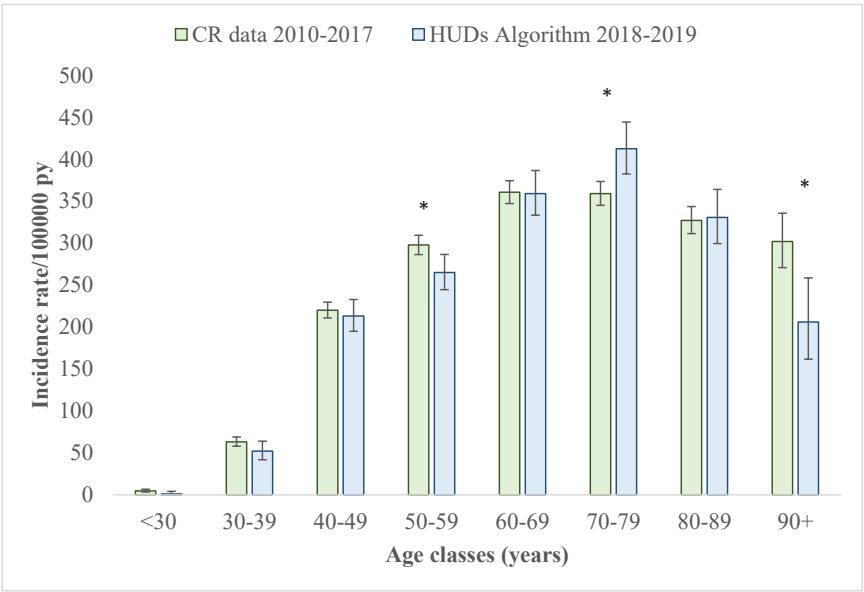

Fig. 4. Incidence rates of breast cancer across different age classes according to the data source (CR=Cancer Registry, HUDs=Healthcare Utilization Databases Algorithm). Error bars indicate 95 % Confidence Interval. The asterisks indicate IRR 95 % confidence intervals that do not contain 1.

possible alignment of the regional incidence with the national one.

In our study no increasing trend over the study period was detected. This contrasts with findings from Italy [3] and worldwide [1] analyses, though the Italian study examined an earlier decade (2003–2014) and reported only a slight increase. In the global study, which covered a much longer period (1990–2019), the global trend was steadily increasing but of modest magnitude; however, the Central Europe Estimated Annual Percentage Change indicated a significant decline over time (-0.17, 95 % CI −0.27;-0.07).

Furthermore, during the pandemic years the observed incidence rates were significantly lower than the rates predicted by the model based on 2010–2019 data. This suggests that the COVID-19 pandemic has had a substantial impact on new BC diagnoses due to the overwhelming burden of COVID-19 pandemic on the healthcare system, which disrupted screening programs, postponed/cancelled diagnostic assessments [24–26]. Therefore, the observed reduction in BC incidence rates during the pandemic is likely due to underdiagnosis rather than an actual decrease in occurrence.

The strengths of the study include the use of the regional Cancer Registry as the gold standard for assessing the accuracy of the HUD algorithm and the inclusion of a large, unselected population covering the entire region. However, HUDs are secondary data sources designed primarily for administrative and cost-containment purposes rather than epidemiological research, which may affect the quality of diagnosis coding and lead to potential misclassification. One limitation of the HUD algorithm is the possible misclassification of recurrent breast cancer cases as incident cases, since our look-back period covers only the six years preceding the index date. Additionally, disease onset may be detected with a delay if the diagnostic assessment (i.e., mammography, breast ultrasound, or breast biopsy procedures) is conducted in a private

setting, and finally cases managed exclusively in outpatient settings may be missed by the HUD algorithm, which relies on diagnoses recorded in Hospital Discharge Records.

The HUDs algorithm allows to adequately estimate the development of new cases of breast cancer in the general population, providing useful epidemiological indications ready to be used in public health policies. However, in the age group 70–79, the HUDs algorithm overestimates the incidence rate, failing to exclude long-term recurrences. This bias can be controlled with the availability of HUDs covering more than 20 years. In addition, the underestimation observed in the age group 90 and over could result from the role of death certificates used by the CR for case identification. The integration of the HUDs with the regional mortality registry, which is not yet active in the Marche region, could further improve the accuracy of the HUDs algorithm.

In conclusion, by integrating healthcare utilisation databases with cancer registry data, this study demonstrated how these methodologies can provide a more dynamic and comprehensive approach to the timely epidemiological monitoring of breast cancer. While cancer registries remain the gold standard for accuracy, they typically involve delays due to the time required for case verification and data consolidation. In contrast, HUDs data can offer near real-time insights that are particularly valuable during public health emergencies. For example, during the COVID-19 pandemic, access to timely hospital discharge data made it possible to rapidly assess the impact of diagnostic delays or disruptions in care pathways, enabling quicker public health responses and more targeted resource allocation

HUDs can be a useful surrogate for cancer registry data when timely estimates of incidence trends or monitoring of healthcare disruptions (e.g., during a pandemic) are needed. However, they are less suitable for analyses requiring detailed clinical data, such as tumour staging or molecular subtypes, where cancer registry data remain essential.

Ethics statement

This observational study fulfils the Italian regulations of ethics committees, which require only standard written informed consent at the time of hospital admission.

Ethical review and approval were waived for this study. We did not mention ethical safeguards simply because not pertinent in our study. All data were anonymized and handled in a manner that protected the privacy and confidentiality of individuals represented in the datasets.

According to Article 9 of the General Data Protection Regulation (European Union Regulation 2016/679), pseudonymized administrative data can be used without specific written informed consent when patient information is collected for healthcare management, quality evaluation, and improvement. All procedures adhered to the 1964 Helsinki Declaration and its subsequent amendments.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Flavia Carle: Writing – review & editing. **Rosaria Gesuita:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization. **Marica Iommi:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Donatella Sarti:** Writing – review & editing, Data curation. **Andrea Faragalli:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Emilia Prospero:** Writing – review & editing. **Chiara Peconi:** Writing – review & editing, Data curation. **Marco Pompili:** Writing – review & editing, Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Restrictions apply to the availability of these data. The datasets generated and/or analysed during the current study are property of a third party that is the Regional Health Agency of Marche (ARSMarche) and, although they are anonymized, datasets are not publicly available due to the current regulation on privacy. The description of the administrative databases is available from the website ARSMarche/Flussi.

Other researchers can obtain access to the data through a formal request based on a research project to the Regional Health Agency of Marche.

References

[1] Y. Xu, M. Gong, Y. Wang, Y. Yang, S. Liu, Q. Zeng, Global trends and forecasts of breast cancer incidence and deaths, Sci. Data 10 (1) (2023) 334.
[2] 2023, Associazione Italiana di Oncologia Medica (AIOM) AIRTA, Società Italiana di Anatomia Patologica e di Citologia Diagnostica (SIAPEC-IAP), Fondazione AIOM, sorveglianze PASSI e PASSI d'Argento (PdA) dell'ISS e Osservatorio Nazionale Screening (ONS): I numeri del cancro in Italia 2023. In. Intermedia Editore.
[3] Mediagraf SpA, 2019.
[4] H. Pan, R. Gray, J. Braybrooke, C. Davies, C. Taylor, P. McGale, R. Peto, K. I. Pritchard, J. Bergh, M. Dowsett, et al., 20-Year risks of Breast-Cancer recurrence after stopping endocrine therapy at 5 years, N. Engl. J. Med. 377 (19) (2017) 1836–1846.
[5] R.N. Pedersen, B.O. Esen, Mellemkjaer L. Christiansen, P. Ejlertsen, B. Lash, T. L. Norgaard, M. Cronin-Fenton, D: **the incidence of breast cancer recurrence 10-32 years after primary diagnosis**, J. Natl. Cancer Inst. 114 (3) (2022) 391–399.
[6] O.M. Jensen, H.H. Storm, Cancer registration: principles and methods. Reporting of results, IARC Sci. Publ. 95 (1991) 108–125.
[7] D.M. Parkin, The role of cancer registries in cancer control, Int. J. Clin. Oncol. 13 (2) (2008) 102–111.
[8] Silva IdS, Chapter 17. The role of cancer registries (edn). Cancer Epidemiology: Principles and Methods, International Agency for Research on Cancer, 1999.
[9] A.M. Jabour, B.E. Dixon, J.F. Jones, D.A. Haggstrom, Toward timely data for cancer research: assessment and reengineering of the cancer reporting process, JMIR Cancer 4 (1) (2018) e4.
[10] E. Skrami, F. Carle, S. Villani, P. Borrelli, A. Zambon, G. Corrao, P. Trerotoli, V. Guardabasso, R. Gesuita, Availability of Real-World data in Italy: a tool to navigate regional healthcare utilization databases, Int. J. Environ. Res Public Health 17 (1) (2019).
[11] S. Ferretti, S. Guzzinati, P. Zambon, G. Manneschi, E. Crocetti, F. Falcini, S. Giorgetti, C. Cirilli, M. Pirani, L. Mangone, et al., [Cancer incidence estimation by hospital discharge flow as compared with cancer registries data], Epidemiol. Prev. 33 (4-5) (2009) 147–153.
[12] Giunta Regionale della Regione Marche. Costituzione del Registro Tumori Regionale. Deliberazione della Giunta Regionale n 1629 del 2/12/2013.
[13] F. Giusti, C. Martos, S. Adriani, M. Flego, R.N. Carvalho, M. Bettio, E. Ben, The joint research Centre-European network of cancer registries quality check software (JRC-ENCR QCS), Front Oncol. 13 (2023) 1250195.
[15] Global Burden of Disease Study 2019 (GBD 2019) Population Estimates 1950-2019.
[14] Eurostat EU: Revision of the European standard population: Report of the Eurostat's task force. In. Edited by Union POotE. Luxembourg; 2013.
[16] G. Corrao, G. Mancia, Generating evidence from computerized healthcare utilization databases, Hypertension 65 (3) (2015) 490–498.
[17] N. Gavrielov-Yusim, M. Friger, Use of administrative medical databases in population-based research, J. Epidemiol. Community Health 68 (3) (2014) 283–287.
[18] C. Mazzali, P. Duca, Use of administrative data in healthcare research, Intern Emerg. Med 10 (4) (2015) 517–524.
[19] I. Baldi, P. Vicari, D. Di Cuonzo, R. Zanetti, E. Pagano, R. Rosato, C. Sacerdote, N. Segnan, F. Merletti, G. Ciccone, A high positive predictive value algorithm using hospital administrative data identified incident cancer cases, J. Clin. Epidemiol. 61 (4) (2008) 373–379.
[20] E. Yuen, D. Louis, L. Cisbani, C. Rabinowitz, R. De Palma, V. Maio, M. Leoni, R. Grilli, Using administrative data to identify and stage breast cancer cases: implications for assessing quality of care, Tumori 97 (4) (2011) 428–435.
[21] I. Abraha, D. Serraino, A. Montedori, M. Fusco, G. Giovannini, P. Casucci, F. Cozzolino, M. Orso, A. Granata, M. De Giorgi, et al., Sensitivity and specificity of breast cancer ICD-9-CM codes in three Italian administrative healthcare databases: a diagnostic accuracy study, BMJ Open 8 (7) (2018) e020627.

[22] I numeri del cancro, https://www.registri-tumori.it/cms/pagine/i-numeri-del-cancro [accessed 06 March 2025].
[23] Cancer Stat Facts: Female Breast Cancer, https://seer.cancer.gov/statfacts/html/breast.html [accessed 06 March 2025].
[24] F. Battisti, P. Falini, G. Gorini, P. Sassoli de Bianchi, P. Armaroli, P. Giubilato, P. Giorgi Rossi, M. Zorzi, J. Battagello, C. Senore, et al., Cancer screening programmes in Italy during the COVID-19 pandemic: an update of a nationwide survey on activity volumes and delayed diagnoses, Ann. Ist. Super. Sanita 58 (1) (2022) 16–24.
[25] L. Fortunato, G. d'Amati, M. Taffurelli, C. Tinterri, L. Marotti, L. Cataliotti, Severe impact of Covid-19 pandemic on breast cancer care in Italy: a senonetwork national survey, Clin. Breast Cancer 21 (3) (2021) e165–e167.
[26] J.S. Ng, D.G. Hamilton, Assessing the impact of the COVID-19 pandemic on breast cancer screening and diagnosis rates: a rapid review and meta-analysis, J. Med. Screen 29 (4) (2022) 209–218.