Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Validation Of Cancer Diagnoses In Electronic Health Records: Results From The Information System For Research In Primary Care (SIDIAP) In Northeast Spain

This article was published in the following Dove Press journal:
Clinical Epidemiology

Martina Recalde,[1,2,*] Cyntia B Manzano-Salgado,[1,2,*] Yesika Díaz,[1] Diana Puente,[1,2] Maria Del Mar Garcia-Gil,[1] Rafael Marcos-Gragera,[3,4] Josefa Ribes-Puig,[5,6] Jaume Galceran,[7] Margarita Posso,[8] Francesc Macià,[8] Talita Duarte-Salles[1]

[1]Fundació Institut Universitari per a la Recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGoL), Barcelona, Spain; [2]Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Cerdanyola del Vallès, Spain; [3]Unitat d'Epidemiologia i Registre de Càncer de Girona (UERCG), Pla Director d'Oncologia, Institut Català d'Oncologia, Institut d'Investigació Biomèdica de Girona (IdIBGi), Universitat De Girona, Girona, Spain; [4]CIBER de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain; [5]Catalan Cancer Plan, Department of Health of Catalonia, Barcelona, Spain; [6]Department of Clinical Sciences, University of Barcelona (UB), Barcelona, Spain; [7]Registre de Càncer de Tarragona, Fundació per a la Investigació i Prevenció del Càncer (FUNCA), IISPV, Reus, Spain; [8]Cancer Prevention Unit and Cancer Registry, Department of Epidemiology and Evaluation, Hospital del Mar, Barcelona, Spain

*These authors contributed equally to this work

Correspondence: Talita Duarte-Salles
Fundació Institut Universitari per a la Recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGoL) Gran Via Corts Catalanes, 587 Àtic, Barcelona 08007, Spain
Tel +34 93 482 43 42
Email tduarte@idiapjgol.org

Background: Electronic health records are becoming an increasingly valuable resource for epidemiology but their data quality needs to be quantified. We aimed to validate twenty-five types of incident cancer cases in the Information System for Research in Primary Care (SIDIAP) in Catalonia with the population-based cancer registries of Girona and Tarragona as the gold-standard.

Methods: We calculated the sensitivity, positive predictive values (PPV), and the time-difference between the date of diagnosis entered into the SIDIAP and into the registries. We added hospital discharge cancer diagnoses to the SIDIAP to assess sensitivity changes.

Results: We identified 27,046 incident cancer diagnoses in the SIDIAP from 2009–2015 among the 949,841 residents of Girona and Tarragona. The cancer types with the highest sensitivity were breast (89%, 95% CI: 88–90%), colorectal (81%, 95% CI: 80–82%), and prostate (81%, 95% CI: 80–83%). Trachea, bronchus and lung cancers had the highest PPV (76%, 95% CI: 74%-78%) followed by stomach (72%, 95% CI: 68–75%) and pancreas (71%, 95% CI: 67–75%). Most cancer diagnoses were reported with less than three months of difference between the SIDIAP and the registries. More cases were registered first in the registries than in the SIDIAP. By adding cancer diagnoses based on hospital discharge data, sensitivity increased for all cancers, especially for gallbladder and biliary tract for which the sensitivity increased by 21%.

Conclusion: The SIDIAP includes 76% of the cancer diagnoses in the cancer registries but includes a considerable number of cases that are not in the registries. The SIDIAP reports most of the cancer diagnoses within a three-month period difference from the date of diagnosis in the cancer registries. Our results support the use of the SIDIAP cancer diagnoses for epidemiological research when cancer is the outcome of interest. We recommend adding hospital discharge data to the SIDIAP to increase data quality, particularly for less frequent cancer types.

Keywords: validation studies, cancer, electronic health records, primary health care, population-based cancer registries

Introduction

Cancer is one of the leading causes of morbidity and mortality worldwide.[1] In 2018, there were 18 million new cases and 9 million deaths.[2] In Spain, cancer is a significant burden for the National Health System: cancer is the second most frequent overall cause of death and results in more than 250,000 new invasive

submit your manuscript | www.dovepress.com
DovePress
http://doi.org/10.2147/CLEP.S225568

cancer cases every year.[3] Therefore, conducting research focused on understanding cancer epidemiology is important both at the national and international levels.

The use of databases of routinely collected electronic health records (EHRs) is becoming more common in epidemiology and clinical research. Due to their size, amount of data availability, representativeness, and long-term follow-up, EHR databases offer a great opportunity to conduct cancer research.[4] Another advantage of large health record databases is that they provide sufficient statistical power to detect uncommon outcomes such as rare cancer types.[5] However, validation processes are required to quantify the correctness of the data and to increase the reliability of large health record databases for use in subsequent observational studies.[6]

The information recorded in EHRs by primary health care professionals in Catalonia – a region in Northeast Spain with 7.5 million inhabitants (2017) – comprises the Information System for Research in Primary Care (SIDIAP) platform.[7] Since the SIDIAP aims to provide reliable information to support research in primary health care, validation studies are performed regularly.[8] A previous study assessed the validity of lung, colon and rectum, prostate, breast, and cervix uteri cancers in the SIDIAP during the period 2009–2012 with sensitivities ranging from 64% (cervix uteri) to 92% (breast).[9] However, this study compared SIDIAP cancer cases with those from the registry of a single hospital in Barcelona. Although the data collection for this hospital is rigorous for a specific area in Barcelona, this area is not representative of the general population of Catalonia. Furthermore, the hospital does not have data available for research use on hematological cancers. A study validating more cancer types and using population-based cancer registries as the gold-standard may increase the scope of the validity of cancer diagnosis in the SIDIAP as well as its use in new areas of research.

The aim of this study was to validate twenty-five types of incident cancer cases in the SIDIAP using the population-based cancer registries of Girona and Tarragona as the gold-standard and to assess the time-difference in the date of diagnosis between the SIDIAP and these cancer registries.

Methods

Data Sources

We performed a cross-sectional study in the SIDIAP during the years 2009–2015, using data from the two population-based cancer registries that exist in Catalonia, the Girona and Tarragona cancer registries, as the gold-standard. The SIDIAP includes information recorded in EHRs by health professionals during routine visits at 287 primary health care centers from the Institut Català de la Salut (ICS, Catalan Health Institute).[10,11] The SIDIAP has anonymized records for more than seven million people and is representative of the Catalan population in terms of age, sex, and geographic distribution.[11] It includes information on disease diagnoses (International Classification for Diseases, 10th revision [ICD-10]), drug prescriptions and dispensations in the primary care setting, and clinically relevant parameters (eg, weight, blood pressure, laboratory tests). It is also linked to a hospital discharge database for patients who attend ICS hospitals (30% of the SIDIAP population).[12,13] The cancer registries of Girona (created in 1994) and Tarragona (in 1980) cover 20% of the Catalan population.[14,15] They collect cancer diagnoses from public and private hospitals, anatomopathological and hematological laboratories, mortality registries, and other information sources.[16–18] Both cancer registries comply with the International Agency for Research on Cancer quality requirements.[19]

Study Population And Cancer Case Definition

In the SIDIAP, incident cancer cases were identified as the first cancer diagnosis from 2009 to 2015 among inhabitants of the provinces of Girona and Tarragona. We had the number of incident cancer cases from the cancer registries during 2005–2015 for Girona and during 2005–2013 for Tarragona available for reference. Cases registered during 2005–2008 were used to clean prevalent cases (Figure 1). The linkage between the SIDIAP and the cancer registries data was performed by a Trusted Third Party (the ICS in this study) using the unique personal identification number of patients. We obtained approval from the Clinical Research Ethics Committee of the IDIAPJGol (project code: P14/074) and the Research Ethics Committee of the Hospital Doctor Josep Trueta (project code: 2017.024).

We used ICD-10 codes and date of diagnosis to identify the following 25 cancer types in adults (aged ≥18 years): head and neck (ICD-10 codes: C00-C14), esophagus (C15), stomach (C16), colorectal (C18–21), liver (C22), gallbladder and biliary tract (C23-24), pancreas (C25), larynx (C32), trachea, bronchus, and lung (C33-34), bone and articular cartilage (C40-C41), malignant

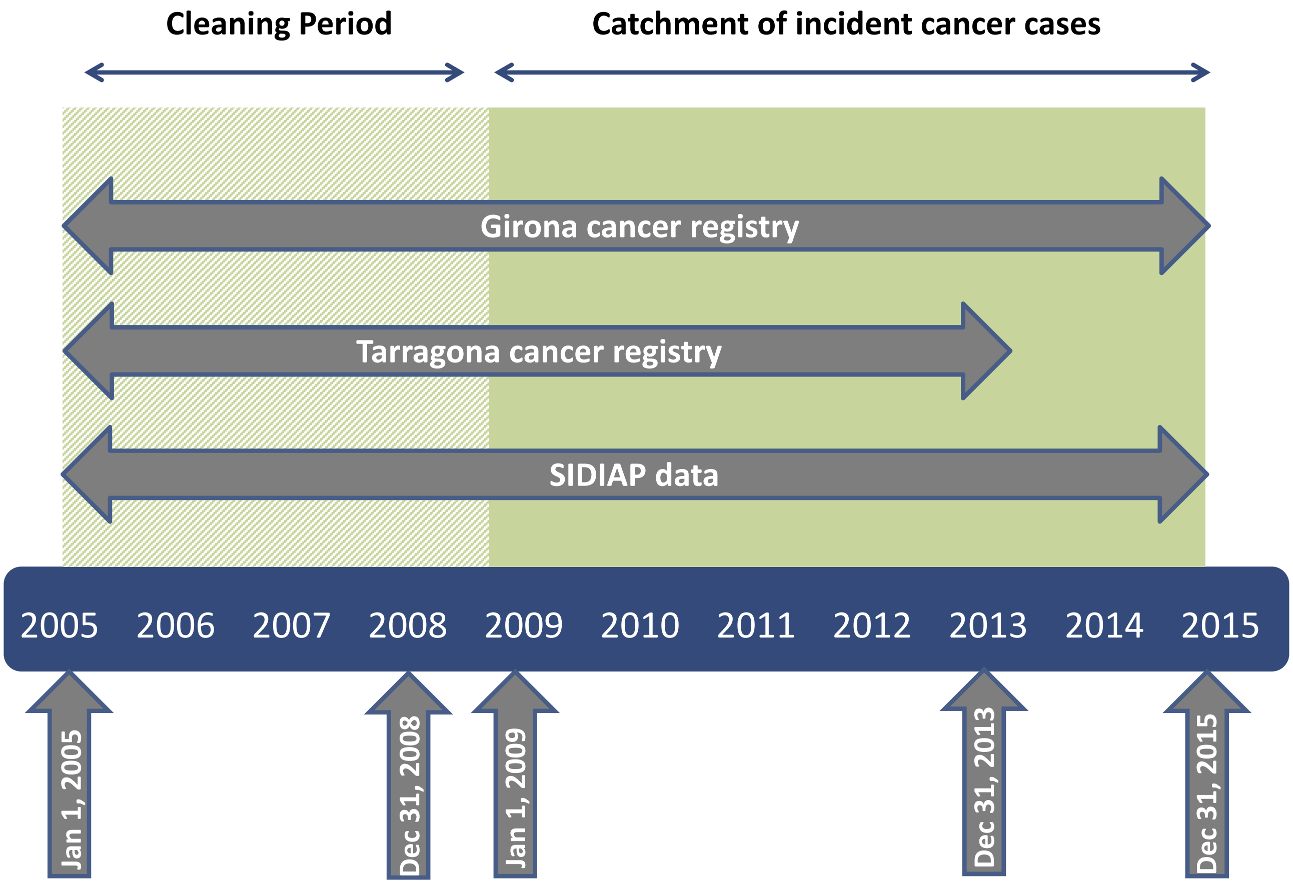

Figure 1 Time period covered by each data source with respect to the duration of the study.
Notes: Figure adapted from Margulis, A. et al. (2017). Validation of Cancer Cases Using Primary Care, Cancer Registry, and Hospitalization Data in the UK. Epidemiology, 29(2), 1.
Abbreviation: SIDIAP, Information System for Research in Primary Care.

melanoma of skin (C43), breast (C50), cervix uteri (C53), corpus uteri (C54-C55), ovary (C56), prostate (C61), testis (C62), kidney (C64), bladder (C67), brain, central nervous system, pituitary gland and pineal gland (C70-72, C75.1-C75.3), thyroid (C73), Hodgkin lymphoma (C81), non-Hodgkin lymphoma (C82-C86, C96), multiple myeloma (C90), and leukemia (C91-95).[20] We excluded other and unspecified malignant neoplasm of skin (C44). Other unspecified or very low-frequency cancers (n<100) were excluded. Diagnosis in hospital discharge data was registered using ICD-9 codes.[21] We mapped diagnosis codes to ICD–10 using available conversion codes eCIEMaps v3.1.9, which we have provided in Supplementary Table S1.

Other Variables

In the SIDIAP, we had information on the primary care center to which individuals were assigned in 2016 (Girona, Tarragona), date of diagnosis, sex (women, men), age (18–35, 36–50, 51–65, ≥66), and nationality (Spanish, non-Spanish). Socioeconomic status was assessed using the “Mortalidad en áreas pequeñas Españolas y Desigualdades Socioeconómicas y Ambientales” (MEDEA) deprivation index, which we categorized into quintiles for anonymization purposes. The 1st and the 5th quintiles represent the least and most deprived levels of the urban population in Catalonia, respectively.[22] We included a rural category since the MEDEA index was not available for people living in these areas.

Statistical Analysis

We performed a descriptive analysis of the overall number of cancer cases in SIDIAP and of the confirmed (ie, matched diagnoses between the SIDIAP and the cancer registries) vs non-confirmed cases (ie, in the SIDIAP but not in the cancer registries) by sex, age, nationality, MEDEA deprivation index, and year of diagnosis, in Girona and Tarragona, and we used a Chi-squared test to assess for significant differences.[23] We used the Catalonia Cancer Registries (CCRs, Girona and Tarragona combined) data as the gold-standard to calculate the sensitivity and the positive predictive values (PPVs) for each cancer type (an illustration of our calculations is available in Figure S1. As secondary analyses, we stratified the sensitivity and PPV analyses by province (Girona and Tarragona) to assess if there were geographical differences and by sex, nationality, age, and the MEDEA deprivation index to assess if there were differences for specific population groups. We also checked if the sensitivities improved after including cancer diagnoses from the hospital discharge database.

For the confirmed cases, we calculated the time difference (months) between the date of diagnosis registered in the SIDIAP and the date registered in the CCRs.

We used R version 3.5.0 for all the statistical analyses and considered p-values <0.05 to be statistically significant.

Results

Sociodemographic Characteristics Of SIDIAP And Confirmed Cases

In the SIDIAP, we identified 496,356 inhabitants of Girona in 2016, of which 16,211 had a cancer diagnosis between 2009 and 2015, and 453,485 inhabitants of Tarragona, of which 10,835 had a cancer diagnosis between 2009 and 2013. There were more cancer cases registered in the SIDIAP among men (55%, 56% for Girona and Tarragona, respectively), people aged 66 years or older (45%, 49%), Spanish citizens (94%, 95%), and people living in rural areas (32%, 37%) (Table 1).

We confirmed 9,296 cancer cases in Girona and 7,182 in Tarragona. Compared to non-confirmed cases, confirmed cases had a higher proportion of men in Tarragona (58% vs 52%) as well as people aged 51 to 65 in both provinces (35% vs 31% in Girona; 34% vs 27% in Tarragona) but a lower proportion of socioeconomically deprived individuals in Girona (11% vs 14%) (Table 1).

Overall Validation

Out of the 21,559 cancer cases registered in the CCRs, 16,478 (76%) were in the SIDIAP. The cancer types with the highest sensitivities in Catalonia were breast (89%, 95% CI: 88–90%), colorectal (81%, 95% CI: 80–82%), and prostate (81%, 95% CI: 80–83%) (Table 2). Almost all cancer types had sensitivities above 60% in both provinces. The exceptions were head and neck (51%, 95% CI: 47–55%) and gallbladder and biliary tract (29%, 95% CI: 23–35%) (Table 2).

Out of the 27,046 SIDIAP cancer cases present in Catalonia, 16,478 (61%) were also in the CCRs. The trachea, bronchus and lung cancers had the highest PPV (76%, 95% CI: 74–78%) followed by stomach (72%, 95% CI: 68–75%) and pancreas (71%, 95% CI: 67–75%) cancers (Table 2). On the other hand, bone and articular cartilage (23%, 95% CI: 15–31%) and cervix uteri (28%, 95% CI: 24–33%) cancers had the lowest PPVs (Table 2).

Most cancer diagnoses were reported within less than three months of difference between the SIDIAP and the registries (Figure 2). More cases were reported first in the cancer registries than in the SIDIAP. Only kidney cancer had more than twenty-five percent of cases reported first in the SIDIAP compared to the CCRs.

Secondary Analyses

Overall, Girona had higher sensitivities than Tarragona, especially for cancers of the cervix uteri (68% vs 52%, for Girona and Tarragona, respectively), Hodgkin lymphoma (69% vs 56%) and head and neck (56% vs 45%) (Supplementary Table S2). The only cancer for which Tarragona had a higher sensitivity than Girona was for bone and articular cartilage (56% vs 75%). Regarding PPVs, Tarragona had higher estimates than Girona, except for six cancer types. We observed the biggest differences for bladder (33% vs 69% %, for Girona and Tarragona, respectively), colorectal (65% vs 77%) and larynx (52% vs 63%) cancers. The cancer types for which Girona had the biggest differences in PPVs with Tarragona were gallbladder and biliary tract (56% vs 44%) and Hodgkin lymphoma (71% vs 56%) (Supplementary Table S2).

Overall, sensitivity estimates differed by age groups, and PPVs estimates differed by age, nationality and socioeconomic status. Those older than 66 years showed lower sensitivities than those aged between 36 and 65 years for most cancer types (Supplementary Table S3). Overall, PPVs were lower in those aged between 18 and 35 years than in the rest of age groups, in non-Spanish than in the Spanish population and in the most deprived compared to the least deprived MEDEA quintiles (Supplementary Table S4). Besides the abovementioned situations, we did not observe any other change in the sensitivity and PPVs according to sex, age, nationality, and socioeconomic status, with exception of certain specific cancer types (Supplementary Tables S3 and S4).

Table 1 Descriptive Characteristics Of The SIDIAP Population With A Cancer Diagnosis In Girona (2009–2015) And Tarragona (2009–2013) By Confirmation Status From The Cancer Registries

	SIDIAP Cancer Cases, n (%)							
	Girona				**Tarragona**			
	SIDIAP Cases	**Confirmed Cases**	**Non-Confirmed**[a]	**p-value**[b]	**SIDIAP Cases**	**Confirmed Cases**	**Non-Confirmed**[a]	**p-value**[b]
Characteristics	**N=16,211**	**N=9296**	**N=6915**		**N=10,835**	**N=7182**	**N=3653**	
Sex								
Women	7300 (45.0%)	4207 (45.3%)	3093 (44.7%)	0.515	4732 (43.7%)	2994 (41.7%)	1738 (47.6%)	<0.001
Men	8911 (55.0%)	5089 (54.7%)	3822 (55.3%)		6103 (56.3%)	4188 (58.3%)	1915 (52.4%)	
Age (years)[c]								
18–35	624 (3.8%)	298 (3.2%)	326 (4.7%)	<0.001	375 (3.5%)	168 (2.3%)	207 (5.7%)	<0.001
36–50	2830 (17.5%)	1634 (17.6%)	1196 (17.3%)		1636 (15.1%)	1058 (14.7%)	578 (15.8%)	
51–65	5417 (33.4%)	3280 (35.3%)	2137 (30.9%)		3471 (32.0%)	2474 (34.5%)	997 (27.3%)	
≥ 66	7340 (45.3%)	4084 (43.9%)	3256 (47.1%)		5353 (49.4%)	3482 (48.5%)	1871 (51.2%)	
Nationality								
Spanish	15,182 (93.7%)	8731 (93.9%)	6451 (93.3%)	0.110	10,328 (95.3%)	6930 (96.5%)	3398 (93.0%)	<0.001
Non-Spanish	1029 (6.3%)	565 (6.1%)	464 (6.7%)		507 (4.7%)	252 (3.5%)	255 (7.0%)	
MEDEA deprivation index[d]								
Quintile 1	1784 (11.0%)	1171 (12.6%)	613 (8.9%)	<0.001	813 (7.5%)	528 (7.3%)	285 (7.8%)	0.022
Quintile 2	1461 (9.0%)	978 (10.5%)	483 (7.0%)		1062 (9.8%)	716 (10.0%)	346 (9.5%)	
Quintile 3	2180 (13.5%)	1152 (12.4%)	1028 (14.9%)		1201 (11.1%)	784 (10.9%)	417 (11.4%)	
Quintile 4	2540 (15.7%)	1209 (13.0%)	1331 (19.2%)		1664 (15.4%)	1097 (15.3%)	567 (15.5%)	
Quintile 5	2025 (12.5%)	1029 (11.1%)	996 (14.4%)		1366 (12.6%)	936 (13.0%)	430 (11.8%)	
Rural areas	5179 (31.9%)	3205 (34.5%)	1974 (28.5%)		4001 (36.9%)	2677 (37.3%)	1324 (36.2%)	
"Missing"	1042 (6.4%)	552 (5.9%)	490 (7.1%)		728 (6.7%)	444 (6.2%)	284 (7.8%)	
Year of diagnosis								
2009	2354 (14.5%)	1151 (12.4%)	1203 (17.4%)	<0.001	2143 (19.8%)	1150 (16.0%)	993 (27.2%)	<0.001
2010	2329 (14.4%)	1342 (14.4%)	987 (14.3%)		2154 (19.9%)	1433 (20.0%)	721 (19.7%)	
2011	2310 (14.3%)	1401 (15.1%)	909 (13.1%)		2046 (18.9%)	1443 (20.1%)	603 (16.5%)	
2012	2374 (14.6%)	1438 (15.5%)	936 (13.5%)		2212 (20.4%)	1553 (21.6%)	659 (18.1%)	
2013	2365 (14.6%)	1471 (15.8%)	894 (12.9%)		2280 (21.0%)	1603 (22.3%)	677 (18.5%)	
2014	2259 (13.9%)	1383 (14.9%)	876 (12.7%)		-	-	-	
2015	2220 (13.7%)	1110 (11.9%)	1110 (16.1%)		-	-	-	

Notes: [a]Non-confirmed cases either have different diagnoses in SIDIAP and the registry or were not available in the registry. [b]Comparison of confirmed vs non-confirmed cases using Chi-squared of independence test. [c]Age in 2009. [d]Quintile 1 of the MEDEA Index represents the least deprived and quintile 5 represents the most deprived. Rural was included as a category since the index cannot be calculated for people living in rural areas.
Abbreviations: SIDIAP, Information System for Research in Primary Care.

Table 2 Validity Of The ICD-10 Codes Used To Identify Incident Cancer Diagnoses Registered In The SIDIAP Database, Catalonia[a] (2009–2015)[b]

Cancer Type (ICD-10 CM)	Cancer Cases, n			Sensitivity, % (95% CI)	PPV, % (95% CI)
	CCRs	SIDIAP	Confirmed		
Head and neck (C00-C14)	650	819	332	51.1 (47.2–54.9)	40.5 (37.2–43.9)
Esophagus (C15)	211	255	157	74.4 (68.5–80.3)	61.6 (55.6–67.5)
Stomach (C16)	673	633	455	67.6 (64.1–71.1)	71.9 (68.4–75.4)
Colorectal (C18-C21)	3743	4329	3035	81.1 (79.8–82.3)	70.1 (68.7–71.5)
Liver (C22)	561	625	364	64.9 (60.9–68.8)	58.2 (54.4–62.1)
Gallbladder & biliary tract (C23-C24)	197	107	57	28.9 (22.6–35.3)	53.3 (43.8–62.7)
Pancreas (C25)	578	590	419	72.5 (68.8–76.1)	71.0 (67.4–74.7)
Larynx (C32)	337	403	226	67.1 (62.0–72.1)	56.1 (51.2–60.9)
Trachea, bronchus & lung (C33-C34)	2152	2155	1631	75.8 (74.0–77.6)	75.7 (73.9–77.5)
Bone and articular cartilage (C40-C41)	39	106	24	61.5 (46.3–76.8)	22.6 (14.7–30.6)
Malignant melanoma of skin (C43)	550	962	417	75.8 (72.2–79.4)	43.3 (40.2–46.5)
Breast (C50)	3325	4456	2958	89.0 (87.9–90.0)	66.4 (65.0–67.8)
Cervix uteri (C53)	198	416	118	59.6 (52.8–66.4)	28.4 (24.0–32.7)
Corpus uteri (C54-C55)	576	661	424	73.6 (70.0–77.2)	64.1 (60.5–67.8)
Ovary (C56)	263	398	190	72.2 (66.8–77.7)	47.7 (42.8–52.6)
Prostate (C61)	2820	3596	2286	81.1 (79.6–82.5)	63.6 (62.0–65.1)
Testis (C62)	139	175	102	73.4 (66.0–80.7)	58.3 (51.0–65.6)
Kidney (C64)	536	730	397	74.1 (70.4–77.8)	54.4 (50.8–58.0)
Bladder (C67)	1456	2370	1108	76.1 (73.9–78.3)	46.8 (44.7–48.8)
Brain and CNS (C70-C72, C75.1-C75.3)[c]	393	544	298	75.8 (71.6–80.1)	54.8 (50.6–59.0)
Thyroid (C73)	395	432	264	66.8 (62.2–71.5)	61.1 (56.5–65.7)
Hodgkin lymphoma (C81)	144	142	92	63.9 (56.0–71.7)	64.8 (56.9–72.6)
Non-Hodgkin lymphoma (C82-C86, C96)	709	909	472	66.6 (63.1–70.0)	51.9 (48.7–55.2)
Multiple myeloma (C90)	294	362	233	79.3 (74.6–83.9)	64.4 (59.4–69.3)
Leukemia (C91-C95)	620	871	419	67.6 (63.9–71.3)	48.1 (44.8–51.4)

Notes: [a]Provinces of Girona and Tarragona. [b]Data from the Tarragona Cancer Registry was only available for 2009–2013. [c]Include pituitary gland and pineal gland tumors.
Abbreviations: CI, Confidence Interval; CNS, Central Nervous System; CCRs, Catalonia Cancer Registries; ICD-10, International Classification for Diseases, 10th revision; PPV, positive predictive values; SIDIAP, Information System for Research in Primary Care.

When adding cancer diagnoses from hospital discharge to primary care data, we observed an increase in sensitivity for all cancer types. Gallbladder and biliary tract cancer had the most substantial change in sensitivity, changing from 29% to 50% (Supplementary Table S5). We also observed changes above 10% for larynx (67% to 83%), head and neck (51% to 66%) and liver (65% to 78%) cancers (Supplementary Table S5).

Discussion

This study validated cancer diagnoses recorded in primary care using the data of the two provincial population-based cancer registries that exist in Catalonia as the gold-standard. We found that 23 out of 25 cancer types had sensitivities above 60%. PPV estimates were generally lower than the sensitivities observed in most cancer types. The number of cancer cases in the SIDIAP that were not confirmed by the cancer registries was high for some specific cancer sites. More cases were first recorded in the cancer registries rather than in the SIDIAP, though for most cancer cases, the time difference between both data sources did not exceed three months. Including cancer diagnoses from hospital discharge data considerably improved the reliability of the data for specific cancer types.

We observed a high sensitivity for the majority of cancer types. Breast, colorectal and prostate cancers had the highest sensitivities, which are some of the most incident tumors and thoroughly screened cancers in systematic programs (breast and colorectal) and strongly sought by opportunistic screening (prostate) in Catalonia.[24,25] Furthermore, these cancers take part in the rapid diagnostic circuit program run in Catalonia, which could also contribute to an increase in the accuracy of diagnosis in primary care.[26] Previous studies conducted in the United Kingdom (UK) that compared primary care data with

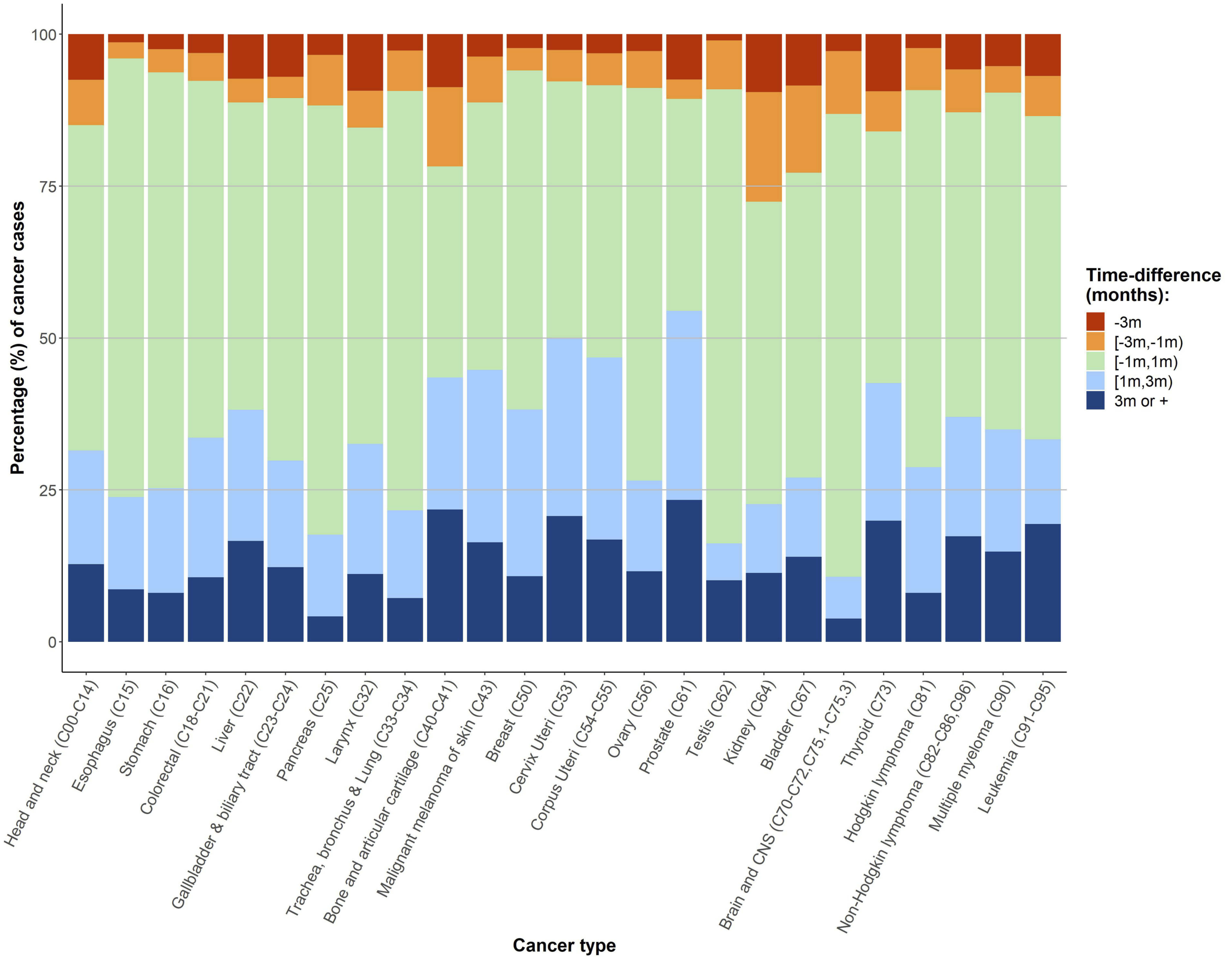

Figure 2 Time-difference (months) in the date of cancer diagnosis recorded in the SIDIAP and the population-based Catalonia Cancer Registries[a] (2009–2015)[b].

Notes: [a]Population-based cancer registries from the provinces of Girona and Tarragona. [b]Data from the Tarragona Cancer Registry was only available for 2009–2013. Negative values indicate SIDIAP diagnosis before the registries' diagnosis date. Brain and CNS include pituitary gland and pineal gland tumors.

Abbreviations: CNS, Central Nervous System; m, months; SIDIAP, Information System for Research in Primary Care.

hospital and cancer registry data also reported high sensitivities for breast, prostate, and colorectal cancers, highlighting that these cancers are usually managed by general practitioners.[27,28] In Catalonia, a previous study comparing SIDIAP cases with those registered in a hospital cancer registry in Barcelona, also reported high sensitivities for breast, colorectal and prostate cancers.[9] High sensitivities are important to enhance study inclusiveness and to be able to ascertain common exposures.[29] A high sensitivity paired with a high specificity (which is important for classifying outcomes) facilitates both the study of cancer as an outcome as well as the identification of the cases' common exposures. In our study, the lowest sensitivities were found for cancers that are less frequent and that are more commonly managed in hospitals, such as gallbladder and biliary tract or bone and articular cartilage.[9,24,30,31] We are not aware of any previous national or international studies validating the primary care diagnosis of these cancer types using external sources. Thus, our results indicate that using SIDIAP cancer diagnoses for research when cancer is the outcome of interest is reliable for most common cancer types in Catalonia but may be insufficient for less frequent types.

PPV estimates were generally lower than the sensitivities observed in most cancer types. The number of cancer cases in the SIDIAP that were not confirmed by the cancer registries was high for some specific cancer sites. A previous study validating only colorectal, lung, gastro-esophageal and urological cancer diagnosis in primary care in the UK reported higher PVV estimates than in our study, ranging from 92% to 98%.[28] This study hypothesized that some of the reasons behind non-confirmed cases might be a disagreement in the type of cancer diagnosed in each data source, or the possibility of suspicious symptoms being registered as cancer diagnoses in primary care.[28] In agreement with this hypothesis, we found that approximately 10% of the non-confirmed cases by the cancer registries were due to disagreement in the type of cancer diagnosis between the data sources. The low PPV for cervix uteri cancer (included in the rapid diagnostic circuit in Catalonia) could be due to detected suspicious symptoms recorded as cancers in SIDIAP; however, we did not have the information needed to prove this hypothesis. Another factor that can influence PPVs is the prevalence of the cancer type which could partially explain the low PPVs of bone and articular cartilage (106 cases registered in the SIDIAP) and gallbladder and biliary tract (107 cases registered). High PPVs are important when we want to identify a cohort of people that only includes people with the condition of interest but do not need to be representative of all cases.[29] Therefore, the SIDIAP does not appear to be an appropriate database to create a cohort of cancer patients, except for certain cancer types (eg, trachea, bronchus and lung, stomach, pancreas or colorectal cancers). More research needs to be conducted to understand the reasons behind non-confirmed cancer cases in SIDIAP.

Most cancer diagnoses were reported within less than three months of difference between the SIDIAP and the registries, and generally, the cancer registries reported the cases earlier than the SIDIAP. Our results are in line with two previous studies in the UK which assessed the time difference between the date of cancer diagnoses registered in the cancer registries and primary care databases. One study reported a median time difference in the date of diagnosis of 11 days (range 6–30 days) between a UK primary care database and the Cancer Registry in England for colorectal, lung, gastro-esophageal and urological cancers.[28] The other study, also using information from the same UK primary care database and cancer registry but combining 11 cancer types, reported that 63% of cancer diagnoses were recorded with one month of difference between the data sources and 24% within one to three months of difference. However, the authors did not specify which source registered the diagnosis first.[32] Although the time difference between the data sources was not substantial in our study, investigators should be aware of it when addressing time-related research questions in the SIDIAP, such as those in the cancer survival field.

In our study, the inclusion of hospital discharge data to SIDIAP cancer diagnoses improved the sensitivity estimates for most cancer sites, with substantial improvements observed particularly for less frequent cancer types. The use of multiple data sources is highly recommended when using EHRs for epidemiological research since the advantages of each database can overcome the limitations of the others.[4,33] Specifically, the need to link primary care databases to those from hospitals and cancer registries to correctly identify certain cancer types has been proposed in the UK.[27] Therefore, considering both SIDIAP and hospital discharge databases can improve the reliability in the results of future research. This may be especially important for larynx, head and neck and liver cancers. For gallbladder and biliary tract cancer, despite the sizeable improvement in sensitivity after adding hospital discharge to SIDIAP cancer diagnoses, the final sensitivity estimate (50%) seems insufficient to perform future studies using this cancer type as an outcome. If data is available, future studies may consider restricting their analyses to confirmed cases only to avoid misclassifications and attain data robustness.

The main strengths of this study are first, the use of the SIDIAP database, which provides a large and representative sample of the Catalonian population and increases external validity.[11] Second, the use of two population-based cancer registries as the gold-standard allowed us to validate numerous cancer types. Third, we were able to calculate the sensitivity of the SIDIAP cancer diagnoses, a type of measure that is often not reported in cancer validation studies. However, our study has limitations. First, since the SIDIAP is a primary care database, certain cancer types are harder to be detected at this level; nevertheless, we assessed the inclusion of hospital discharge information to account for this limitation. Second, textual information in medical records could be of value to distinguish cancer suspicions from actual diagnoses in the SIDIAP, but this information was not available in this study. Third, for this study we were only able to add cancer diagnoses from hospital discharge from the ICS hospitals, therefore we cannot confirm whether including information from all Catalan hospitals would permit better identification of cases for the same cancer types we found. Finally, our population of reference was the population of individuals assigned to a primary care center in Girona and Tarragona provinces in 2016 and, thus, we could not account for changes in patient address during the whole study period.

Conclusion

The SIDIAP includes 76% of the cancer diagnoses present in the cancer registries of Catalonia but also includes a considerable number of cases that are not in the registries.

Overall, the SIDIAP reports cancer cases later than the registries but the time difference in the date of diagnosis between the databases is usually less than three months. Our results support the use of SIDIAP cancer diagnoses for national and international epidemiological research when cancer is used as an outcome, especially for the most frequent cancer types. The inclusion of cancer diagnoses from hospital discharge data is recommended to improve the reliability of certain cancer types such as head and neck, liver, larynx, and leukemia. However, our results do not support the use of SIDIAP data for all cancer sites when the purpose of the study is to identify a cohort of cancer patients. Further research is needed to understand the cancer cases recorded in the SIDIAP that were not confirmed by the cancer registries.

Abbreviations

CI, Confidence Interval; CCRs, Catalonia Cancer Registries; CNS, Central Nervous System; EHRs, Electronic Health Records; ICD-9,International Classification for Diseases, 9th revision; ICD-10, International Classification for Diseases, 10th revision; ICS, Institut Català de la Salut; MEDEA, Mortalidad en áreas pequeñas Españolas y Desigualdades Socioeconómicas y Ambientales; PPV, Positive Predictive Value; SIDIAP, Information System for Research in Primary Care; UK, United Kingdom.

Ethics Approval

The Clinical Research Ethics Committee of the IDIAPJGol (project code: P14/074) and the Research Ethics Committee of the Hospital Doctor Josep Trueta (project code: 2017.024) approved this study.

Data Availability

In accordance with current European and national law, the data used in this study is only available for the researchers participating in this project. Thus, we are not allowed to distribute or make publicly available the data to other parties. However, researchers from public institutions can request data from the SIDIAP and other sources (eg, Cancer Registries) if they comply with certain requirements. Further information is available online (https://www.sidiap.org/index.php/menu-solicitudes-en/application-proccedure) or by contacting Anna Moleras amoleras@idiapjgol.org).

Acknowledgements

We thank the Registre del conjunt mínim de bases de dades (CMBD), for providing data on Hospital Discharges. We also thank Leonardo Mendez-Boo and Eduardo Hermosilla for their support on the data linkage processes and Matthew Bennett for revising the English language composition.

Funding

Funding [for grant number: 2017/1630] was obtained from Wereld Kanker Onderzoek Fonds (WKOF), as part of the World Cancer Research Fund International grant program. TDS is funded by the Department of Health of the Generalitat de Catalunya, awarded on the 2016 call under the Strategic Plan for Research and Innovation in Health (PERIS) 2016–2020, modality incorporation of scientists and technologists, with reference SLT002/16/00308.

Disclosure

The authors report no conflicts of interest in this work.

References

1. International Agency for Research on Cancer. *World Cancer Report 2014*. Lyon; 2014.
2. Global Cancer Observatory. Cancer tomorrow. International Agency for Research on Cancer. https://gco.iarc.fr/tomorrow/home. Published 2018. Accessed March 10, 2019.
3. Galceran J, Ameijide A, Carulla M, et al. Cancer incidence in Spain, 2015. *Clin Transl Oncol*. 2017;19(7):799–825. doi:10.1007/s12094-016-1607-9
4. Cook JA, Collins GS. The rise of big clinical databases. *Br J Surg*. 2015;102(2):93–101. doi:10.1002/bjs.9723
5. Haynes K, Forde KA, Schinnar R, Wong P, Strom BL, Lewis JD. Cancer incidence in the health improvement network. *Pharmacoepidemiol Drug Saf*. 2009;18:730–736. doi:10.1002/pds.v18:8
6. Nissen F, Quint JK, Morales DR, Douglas IJ. How to validate a diagnosis recorded in electronic health records. *Breathe*. 2019;15(1):64–68. doi:10.1183/20734735.0344-2018
7. Instituto Nacional de Estadística. Estadísticas Territoriales: cataluña. http://www.ine.es/FichasWeb/RegComunidades.do?fichas=49&busc_comu=&botonFichas=Ir+a+la+tabla+de+resultados. 2018. Accessed October 10, 2018.
8. Ramos R, Balló E, Marrugat J, et al. Validity for use in research on vascular diseases of the SIDIAP (Information System for the Development of Research in Primary Care): the EMMA study. *Rev Española Cardiol (English Ed*. 2012;65(1):29–37. doi:10.1016/j.recesp.2011.07.017
9. Garcia-Gil M, Elorza J-M, Banque M, et al. Linking of primary care records to census data to study the association between socioeconomic status and cancer incidence in Southern Europe: a nation-wide ecological study. *PLOS ONE*. 2014;9(10):e109706. doi:10.1371/journal.pone.0109706
10. Bolíbar B, Fina Avilés F, Morros R, et al. Base de datos SIDIAP: la historia clínica informatizada de Atención Primaria como fuente de información para la investigación epidemiológica. *Med Clin (Barc)*. 2012;138(14):617–621. doi:10.1016/j.medcli.2012.01.020
11. García-Gil MDM, Hermosilla E, Prieto-Alhambra D, et al. Construction and validation of a scoring system for selection of high quality data in a Spanish population primary care database (SIDIAP). *Inf Prim Care*. 2012;20(2):1.

12. Generalitat de Catalunya. Conjunt mínim bàsic de dades (CMBD). https://catsalut.gencat.cat/ca/proveidors-professionals/registres-catalegs/registres/cmbd/index.html#googtrans(ca%7Ces. 2017. Accessed March 5, 2019.
13. Instituto Nacional de Estadística. Hospital morbidity Survey: general methodology. http://www.ine.es/en/daco/daco42/sanitarias/notaemh_en.htm. 2019. Accessed March 6, 2019.
14. Josep MB, Izquierdo A, Vilardell L, et al. Cancer Incidence in Girona (2008-2012). In: Bray F, Colombet M, Mery L, et al., editors. *Cancer Incidence in Five Continents*. Vol. XI (electronic version). Lyon: International Agency for Research on Cancer; 2017. Available from: http://ci5.iarc.fr. Accessed March 10, 2018.
15. Galceran J, Carulla M, Ameijide A, et al. Cancer incidence in Tarragona (2008–2012). In: Bray F, Colombet M, Mery L, et al., editors. *Cancer Incidence in Five Continents*. Vol. XI (electronic version). Lyon: International Agency for Research on Cancer; 2017. Available from: http://ci5.iarc.fr. Accessed March 10, 2018.
16. Unitat d'Epidemiologia i Registre de Càncer de Girona. *El Càncer a Girona 2010–12: Projeccions De La Incidència 2017*. Girona: Institut Català d'Oncologia; 2016.
17. Navarro C, Martos C, Ardanaz E, et al. Population-based cancer registries in Spain and their role in cancer control. *Ann Oncol*. 2010;21(Supplement3):iii3–iii13. doi:10.1093/annonc/mdq094
18. International Agency for Research on Cancer. Indices of data Quality: All sites except non-melanoma skin (C00-96 exc. C44). *Cancer Incidence in Five Continents Volume XI*. http://ci5.iarc.fr/CI5-XI/PDF/INDICES/21.pdf. 2017. Accessed March 10, 2018.
19. International Agency for Research on Cancer. Chapter 5: Data Comparability and Quality. Cancer Incidence in Five Continents Volume XI. http://ci5.iarc.fr/CI5-XI/Pages/Chapter5.aspx. 2017. Accessed March 10, 2018.
20. World Health Organization. *ICD-10 : International Statistical Classification of Diseases and Related Health Problems: Tenth Revision*. 2nd ed. Geneva: World Health Organization; 2004.
21. World Health Organization. Ninth revision of the International Classification of Diseases. 1976. doi:10.1084/jem.143.4.741
22. Domínguez-Berjón MF, Borrell C, Cano-Serral G, et al. Construcción de un índice de privación a partir de datos censales en grandes ciudades españolas (Proyecto MEDEA). *Gac Sanit*. 2008;22(3):179–187. doi:10.1157/13123961
23. Subirana I, Vila J, Sanz H, Lucas G, Penafiel J, Gimenez D. Building bivariate tables: the comparegroups package for R. *J Stat Softw*. 2014;57(12):1–16. doi:10.18637/jss.v057.i12
24. Pla director d'Oncologia de Catalunya. *El Càncer a Catalunya*. Barcelona; 2016.
25. Instituto Catalán de Oncología. Prevención del cáncer. http://ico.gencat.cat/es/el-cancer/programes_de_deteccio_precoc/. Accessed April 17, 2019.
26. Generalitat de Catalunya. Circuit de diagnòstic ràpid. http://canalsalut.gencat.cat/ca/salut-a-z/c/cancer/recursos-per-a-professionals/diagnostic/circuit-de-diagnostic-rapid/index.html. 2018. Accessed April 16, 2019.
27. Margulis AV, Fortuny J, Kaye JA, et al. Validation of cancer cases using primary care, cancer registry, and hospitalization data in the United Kingdom. *Epidemiology*. 2018;29(2):308–313. doi:10.1097/EDE.0000000000000786
28. Dregan A, Moller H, Murray-Thomas T, Gulliford MC. Validity of cancer diagnosis in a primary care database compared with linked cancer registrations in England. Population-based cohort study. *Cancer Epidemiol*. 2012;36(5):425–429. doi:10.1016/j.canep.2012.05.013
29. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. *J Clin Epidemiol*. 2012;65(3):343–349.e2. doi:10.1016/j.jclinepi.2011.09.002
30. American Cancer Society. Can gallbladder cancer be found early? https://www.cancer.org/cancer/gallbladder-cancer/detection-diagnosis-staging/detection.html. 2018. Accessed March 5, 2019.
31. American Cancer Society. Can bone cancer be found early? https://www.cancer.org/cancer/bone-cancer/detection-diagnosis-staging/detection.html. 2018. Accessed March 5, 2019.
32. Boggon R, Van Staa T, Chapman M, Gallagher AM, Hammad TA, Richards MA. Cancer recording and mortality in the general practice research database and linked cancer registries. *Pharmacoepidemiol Drug Saf*. 2013;22:168–175. doi:10.1002/pds.v22.2
33. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the general practice research database: a systematic review. *Br J Clin Pharmacol*. 2010;69(1):4–14. doi:10.1111/bcp.2010.69.issue-1