Available online at
ScienceDirect
www.sciencedirect.com
Elsevier Masson France
EM|consulte
www.em-consulte.com
Revue d'Épidémiologie et de Santé Publique 64 (2016) 145–152
Revue d'Épidémiologie et de Santé Publique
Epidemiology and Public Health

Original article

Estimation of breast, prostate, and colorectal cancer incidence using a French administrative database (general sample of health insurance beneficiaries)

Estimation de l'incidence des cancers du sein, de la prostate et du côlon-rectum à partir des bases de données médico-administratives (échantillon généraliste des bénéficiaires de l'assurance maladie)

S. Doat [a,b,c], S. Samson [d], A. Fagot-Campagna [d], P. Tuppin [d], F. Menegaux [a,b,*]

[a] *Inserm, center for research in epidemiology and population health (CESP), U1018, gene, environment cancer epidemiology team, 94807 Villejuif cedex, France*
[b] *UMRS 1018, Paris-Sud University, 94807 Villejuif cedex, France*
[c] *Hepato-gastroenterology department, unit of gastrointestinal tumor screening and treatment, Pitié-Salpêtrière hospital, Paris Public Hospital Authority (AP–HP), 75013 Paris, France*
[d] *National Health Insurance Organization for Salaried Workers (CNAM), 75986 Paris cedex 20, France*

Received 13 March 2015; accepted 15 December 2015
Available online 26 May 2016

Abstract

Aim. – The aim of this study was to compare incidence of breast, prostate, and colorectal cancer incidence estimated from a French administrative database with the incidences estimated from the cancer registry data.

Materials and methods. – A cohort of 426,410 people included in the general sample of health insurance beneficiaries (EGB) database as of January 1, 2007, was constituted. Several algorithms were developed to estimate cancer incidence between 2008 and 2012 using principal diagnosis (PD) of hospital discharge data (medical information systems program [PMSI]) and/or long-term disease (LTD) and together with a procedure necessary for histological diagnosis and indicating initial disease management. The incidence rates obtained were compared with those from the registry data using the standardized incidence ratio (SIR).

Results. – The algorithm taking into account LTD and PD in the PMSI and the mandatory presence of a marker procedure provided estimates close to those from the registry data for breast cancer (SIR: 1.12 [1.07–1.18]) and colorectal cancer (SIR: 0.94 [0.88–1.02] in men and SIR: 0.93 [0.86–1.01] in women). For prostate cancer, taking into account specific procedures and drugs in addition to LTD and PD in the PMSI enhanced the estimation of incidence (SIR: 1.03 [0.98–1.08]).

Conclusion. – The PMSI together with reimbursement data (LTD, procedures, drugs) provided estimates of breast, prostate, and colorectal cancer incidence, at a national level, comparable to those from the cancer registry data.

Keywords: Cancer; Incidence; Algorithms; Medico-administrative databases

Résumé

Objectif. – L'objectif de cette étude était de comparer l'incidence des cancers du sein, de la prostate et du côlon-rectum obtenue à partir d'algorithmes appliqués à un échantillon de la base de données médico-administratives de l'assurance maladie, à celle estimée à partir des données de registres du cancer.

Matériel et méthodes. – Une cohorte de 426 410 personnes résidant en France métropolitaine, affiliées au régime général, présentes dans l'échantillon généraliste des bénéficiaires (EGB) de l'assurance maladie au 1er janvier 2007 a été constituée. Plusieurs algorithmes ont été

* Corresponding author. Inserm U1018, CESP, équipe épidémiologie des cancers, gènes et environnement, 16, avenue Paul-Vaillant-Couturier, 94807 Villejuif cedex, France.
E-mail address: florence.menegaux@inserm.fr (F. Menegaux).

http://dx.doi.org/10.1016/j.respe.2015.12.020
0398-7620/

développés à partir des données hospitalières et de remboursement pour estimer l'incidence des cancers entre 2008 et 2012. Ils utilisaient le code CIM-10 du cancer d'intérêt codé en diagnostic principal dans le programme de médicalisation des systèmes d'information (PMSI) et/ou dans les affections de longue durée (ALD), ainsi qu'un acte nécessaire au diagnostic histologique et révélateur de la prise en charge initiale. Concernant le cancer de la prostate, la présence d'actes ou de médicaments spécifiques à la prise en charge de ce cancer était également prise en compte. Les taux d'incidence obtenus ont été comparés à ceux issus des données de registres à l'aide d'un calcul du ratio d'incidence standardisé (SIR).

Résultats. – L'algorithme prenant en compte non seulement l'ALD mais aussi le diagnostic principal dans le PMSI et la présence obligatoire d'un acte traceur permettait d'obtenir des estimations proches de celles issues des données de registres pour le cancer du sein (SIR : 1,12 [1,07–1,18]) et le cancer colorectal chez l'homme (SIR : 0,94 [0,88–1,02]) et chez la femme (0,93 [0,86–1,01]). Pour le cancer de la prostate, la prise en compte en sus des données du PMSI et d'ALD, des actes et médicaments spécifiques permettait une meilleure approche de l'incidence et parvenait à un SIR de 1,03 [0,98–1,08].

Conclusion. – Les données de l'EGB de l'assurance maladie permettent, à l'aide du couplage des données hospitalières et de remboursement, d'obtenir des estimations de taux d'incidence des cancers du sein, de la prostate et du côlon-rectum, à un niveau national, comparables à celles issues des données de registres.

Mots clés : Cancer ; Incidence ; Algorithmes ; Bases de données médico-administratives

1. Introduction

In France, national annual cancer incidence rates are estimated by extrapolation from the data in the departmental registries of the Francim network, which covers about 20% of mainland France, and the national mortality data of the Epidemiological Center for the Medical Causes of Death (CépiDC–Inserm) [1]. Departmental data are extrapolated to national data using mortality as a correlate of the incidence in each department [2]. A time interval of about 3 years for the publication of the incidence observed in the zones covered by the registries is currently necessary to ensure data validation. An annual projection of the incidence and mortality for the current year is implemented and is based on the time course scenarios for the most recent years [1]. This methodology currently constitutes the gold standard for estimation of the national incidence of cancers in France. For a finer estimation, on the regional or departmental scale, of zones not covered by a registry, there must not be any regional disparities in the incidence/mortality ratio (specific survival, screening policy), which is not often the case [3,4].

For some 10 years, with a view to overcoming the problems related to the time to data provision, which induce low sensitivity to changes in incidence and to the extrapolation on the departmental or regional scale in zones not covered by the registries, alternative methodologies based on the use of administrative and, particularly, hospital discharge databases have been developed worldwide [3–6]. However, as shown in a review of studies using UK primary care databases to identify incident cancers, there is a lack of transparency and heterogeneity in the methodologies used [7], and further studies are needed for validation against external data sources to improve reproducibility of methodologies. In France, the coverage of these databases is national and the data are available within about 1 year. Several algorithms for the detection of incident cases of cancer have been developed. Most of the algorithms are based on hospital data only, derived from the hospital discharge database (medical information systems program [PMSI]). They combine the code of the cancer of interest as the principal diagnosis (PD) in the PMSI (International Classification of Diseases, revision 10 code [ICD-10]) with a procedure necessary for histological diagnosis or characteristic of initial disease management (common classification of medical procedures code [CCAM]). The CCAM codes used are, for example, breast biopsy, mastectomy, or tumorectomy for breast cancer, or colonoscopy, colonic resection, or colonic prosthesis for colorectal cancer. This procedure helps lower the false-positives related to prevalent cases [6,8,9]. Another medico-administrative database that is increasingly used in France is the database of the French national health insurance system (via the general sample of beneficiaries [EGB]), which contains not only the PMSI data, but also the data on outpatient reimbursement and information on the diagnosis, such as the statement of 100% reimbursement for cancer treatment (long-term disease [LTD]). In this context, the objective of this study was to test, using the data of the national health insurance system's EGB, several algorithms for national incidence estimation combining hospital discharge data and reimbursement data for the three most frequent cancers in France – breast, colorectal, and prostate cancer – and to compare these estimates with the incidence estimates generated by the departmental registry data extrapolated to the national level.

2. Material and methods

2.1. Database

The EGB of the health insurance system is obtained by random 1/97 sampling of the population with health insurance coverage [10]. For each individual, the database includes data on any hospitalizations in short-stay establishments since 2005, derived from the hospital discharge database (PMSI). The International Classification of Diseases, revision 10 (ICD-10) is used to code the principal (PD), associated (AD), or linked (LD) diagnosis of the hospitalization. The hospital medical and surgical procedures are also coded using the CCAM for the procedures conducted during hospitalization. The EGB also

contains, for a given beneficiary, information on the outpatient prescriptions and procedures reimbursed to patients with health insurance coverage, such as the drugs prescribed in open-care settings for which reimbursement has been made since 2003. No clinical information with regard to the results of the consultations, prescriptions, or examinations is collected. Nonetheless, there is information on LTD, coded using the ICD-10. Recognition by a health insurance advisory physician, pursuant to the attending physician's request, enables 100% reimbursement. The individual data are linked from one year to the next and from one table to the next using a single anonymous identifier that in no circumstance enables the source patient to be identified. This database can be easily and rapidly accessed by every authorized research team in France with a personal login after a short training session. Twenty years of data are accessible for each individual. In contrast, the national database from the health insurance system has no more than 3 years of medical data and specific extraction needs special authorization and is much longer to obtain. This makes the EGB very convenient for epidemiological studies on frequent diseases, especially pharmaco-epidemiological studies.

2.2. Study population

From the EGB, a fixed cohort of 426,410 subjects was constituted. To ensure comparability with the data of the Francim network registries, the subjects were required to reside in mainland France. To prevent nonexhaustive data for the period for the other minority regimens, the subjects had to be covered by the general regimen, present in the EGB on January 1, 2007, and alive on January 1, 2008 (Fig. 1). From the 426,410 subjects (203,764 men and 222,646 women), three subcohorts to study each cancer of interest were constituted.

The subcohort of interest for the estimation of the incidence of breast cancer, after exclusion of all men and women with a PD, LD, or AD of invasive breast cancer (C50) between 2005 and 2007 during a hospitalization or with LTD status for breast cancer before 2008, consisted of 218,458 women. With regard to colorectal cancer, after elimination of the patients with a primary, linked, or AD of invasive colorectal cancer (C18, C19, C20) between 2005 and 2007 or with LTD status for colorectal cancer before 2008, there remained 424,813 subjects in the subcohort. Lastly, for prostate cancer, there was a subcohort of 201,019 men after exclusion of men hospitalized for prostate cancer as a PD, LD, or AD between 2005 and 2007 and those with LTD status for prostate cancer before 2008 or having received a specific drug (GnRH analogs, androgen inhibitors, estrogens, estramustine, etc.) or having undergone a procedure specific to the treatment of prostate cancer (pulpectomy, vesiculoprostatectomy, specific brachytherapy) before 2008.

2.3. Algorithms for identification of the incident cases

The occurrence of breast, colorectal, or prostate cancer was determined for each of the subcohorts between January 1, 2008, and December 31, 2012, using several algorithms. It should be noted that the incidence date was the first date on which the ICD-10 code for cancer was found in the hospital discharge data or reimbursement data.

For breast and colorectal cancer, algorithm 1 used the ICD-10 codes of interest for the PD from the PMSI. Algorithm 2 used the ICD-10 codes of interest for LTD. Algorithm 3 used the presence of one or the other datum. In addition to the previous algorithm, algorithm 4 also required implementation of a mandatory procedure for the initial diagnosis of breast or

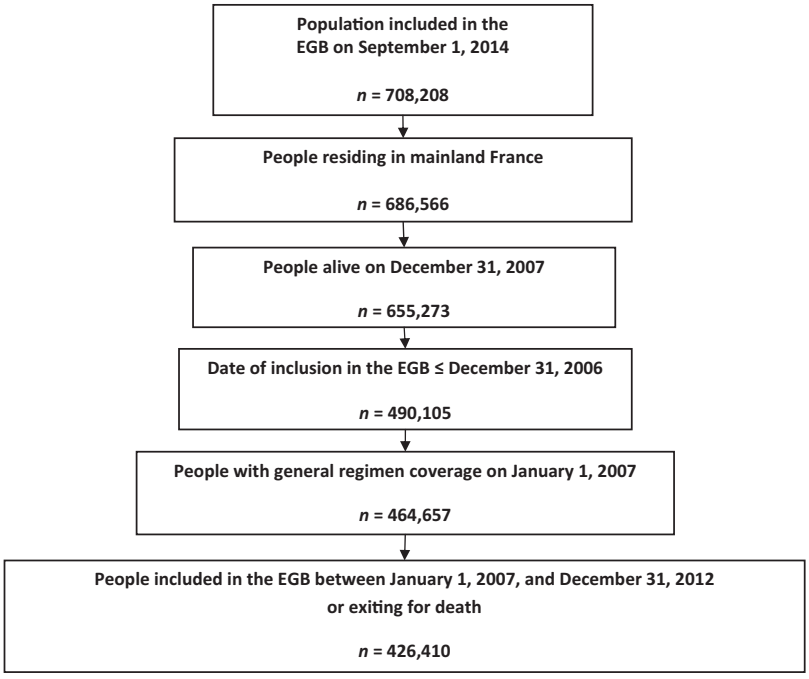

Fig. 1. Flowchart for constitution of the overall cohort derived from the general sample of beneficiaries (EGB).

Table 1
Expected and observed incidence of breast cancer between 2008 and 2012 according to algorithms used.

Breast cancer algorithm	No. of cancers observed	Raw rate (*n*/100,000)	No. of cancers expected	National rate[a] (*n*/100,000)	SIR [95% CI]
Algorithm 1 (PMSI PD)	1042	97.4	1589	148.5	0.66 [0.62–0.70]
Algorithm 2 (LTD)	1893	177.3	1586	148.5	1.19 [1.14–1.25]
Algorithm 3 (PMSI PD or LTD)	1972	184.7	1586	148.5	1.24 [1.19–1.30]
Algorithm 4 (PMSI PD or LTD and procedure)	1783	166.9	1586	148.5	1.12 [1.07–1.18]

No.: number; SIR: standardized incidence ratio; PMSI PD: medical information systems program, principal diagnosis; LTD: long-term disease.
[a] Estimated by the InVS for 2012 [12].

colonic cancer in the 3 months preceding or following the incidence date.

With regard to prostate cancer, algorithm 1 was supplemented by the addition of possible LDs of prostate cancer (C61) with a primary diagnosis of chemotherapy, radiotherapy, or palliative care (Z452, Z510, Z511, Z512, Z514, Z515, Z518) (algorithm 1b). Algorithms 2 and 3 were the same as those used for the other cancer sites. The algorithm excluding patients not having undergone a procedure indispensable for initial diagnosis of the cancer in question was not used since it was considered irrelevant in the case of prostate cancer. Prostate cancer may simply be monitored or be treated in an outpatient setting, and therefore not be the subject of a procedure or hospitalization. In contrast, algorithm 4 took into account the potential presence of a specific procedure with or without hospitalization or LTD status for prostate cancer. To algorithm 4, algorithm 5 added the possibility of prescription of drugs specific to management of prostate cancer, as published recently [11].

2.4. Statistical analysis

For each cancer site and each algorithm, the number of incident cases detected between 2008 and 2012, the raw incidence rate (calculated using the number of person-years) the expected number of cases (calculated from the national rate published by the French Public Health Surveillance Institute [InVS] [12] using the Francim network registry data), and the standardized incidence rate (SIR) ratio between the rates observed in the databases and those expected on the basis of the registries were reported. To calculate the number of person-years in each cohort, the date of the end of follow-up was defined as the first date between date of death, date of cancer of interest (which differed between the source used for cancer identification in each algorithm), or December 31, 2012.

The sources from which the data enabling incident cancer case identification were specified for each cancer site as a function of the various age groups.

All the analyses were performed using the SAS Enterprise Guide software package, version 4.3.

3. Results

3.1. Breast cancer

Among the 218,458 women followed in the cohort, the median age was 43 years (one-third below 30 years old, one-third 30–50 years old, and approximately 40% 50 years old and above). Relative to the estimates extrapolated from the registries, algorithm 1 identified fewer cases of breast cancer: 1042 cases were detected with a raw rate of 97.4/100,000 people and an SIR of 0.66 [0.62–0.70] (Table 1).

In contrast, algorithm 2, using LTD status, identified 1893 cases, i.e., a raw rate of 177.3/100,000 and an SIR of 1.19 [1.14–1.25]. Algorithm 3 identified even more cases: 1972 incident cases with a raw rate of 184.7/100,000 and an SIR of 1.24 [1.19–1.30]. Adding a marker procedure for initial disease management or histological proof to algorithm 4 limited the prevalent cases that may still have been present. The number of incident cases was 1783, i.e., a raw rate of 166.9/100,000 with an SIR of 1.12 [1.07–1.18]. It should be noted that among the patients who did not have a diagnostic procedure, the median age was 69 years vs. 58 years for the patients who underwent a procedure. Using algorithm 4 for the breast cancer cohort, the median age at breast cancer diagnosis was 60 years.

3.2. Colorectal cancer

Among the 424,813 persons followed in the colorectal cohort, 202,966 were men (48%) and 221,847 were women (52%). The median age was 41 years old, 39 years in men and 42 years in women. The distribution according to age was 33% below 30 years old, one-third 30–50 years old, and around 37% 50 years old and above.

Using algorithm 1 for the 2008–2012 period, 692 cases of colorectal cancer in men and 562 cases in women were detected, i.e., raw rates of 69.7/100,000 and 51.7/100,000, respectively, and SIRs of 0.93 [0.86–1.00] and 0.90 [0.82–0.97], respectively. The estimate was thus lower than that derived from the registries (Table 2).

In contrast to what was observed for breast cancer, a greater number of incident cases were not detected with algorithm 2 using LTD status vs. the registries. Many fewer cases in both men and women were detected, with 580 incident cases for men and 476 incident cases for women. The raw rates were 58.4/100,000 and 43.8/100,000, respectively, with SIRs of 0.78 [0.72–0.84] and 0.76 [0.69–0.83], respectively. Including the two databases in algorithm 3 enabled approximation of the incidence rate estimated by the registries with an SIR of 1.06 for men and women. Addition of a procedure necessary for histological diagnosis resulted in an SIR of 0.94 [0.88–1.02] for men and 0.93 [0.86–1.01] for women. The patients who did not undergo a procedure had a median age of 72 years (vs. 69 years

Table 2
Expected and observed incidence of colorectal cancer between 2008 and 2012 according to algorithms used.

Colorectal cancer algorithms	No. of cancers observed	Raw rate (*n*/100,000)	No. of cancers expected	National rate[a] (*n*/100,000)	SIR [95% CI]
Algorithm 1 (PMSI PD)					
Men	692	69.7	746	75.2	0.93 [0.86–1.00]
Women	562	51.7	627	57.7	0.90 [0.82–0.97]
Algorithm 2 (LTD)					
Men	580	58.4	746	75.2	0.78 [0.72–0.84]
Women	476	43.8	627	57.7	0.76 [0.69–0.83]
Algorithm 3 (PMSI PD or LTD)					
Men	794	80.0	746	75.2	1.06 [0.99–1.14]
Women	667	61.4	627	57.7	1.06 [0.99–1.15]
Algorithm 4 (PMSI PD or LTD and procedure)					
Men	704	71.0	746	75.2	0.94 [0.88–1.02]
Women	585	53.8	627	57.7	0.93 [0.86–1.01]

No.: number; SIR: standardized incidence ratio; PMSI PD: medical information systems program, principal diagnosis; LTD: long-term disease.
[a] Estimated by the InVS for 2012 [12]

for the patients with a marker procedure). The proportion of women was slightly higher in that population (48% vs. 45%). For the colorectal cancer cohort and using algorithm 4, the median age at colorectal cancer diagnosis was 71 years (69 years in men and 74 years in women).

3.3. Prostate cancer

Among the 201,019 men followed in the prostate cohort, the median age was 39 years. The distribution according to age was 35% below 30 years old, 31% 30–50 years old, and around 34%, 50 years old and above.

Markedly fewer incident cases of prostate cancer were detected with algorithm 1, relative to the incidence estimated by the registry data, with 723 observed cases, i.e., a raw rate of 73.5/100,000 and an SIR of about 0.42 [0.39–0.45] for the PD (Table 3). The estimate remained stable taking into account the LD of prostate cancer when the PD was a chemotherapy, radiotherapy, or palliative care code as in algorithm 1b (774 observed cases, i.e., a raw rate of 78.7/100,000 and an SIR of 0.45 [0.42–0.48]). Algorithm 2 improved the number of cases detected: 1419, i.e., 144.6/100,000 and an SIR of 0.82 [0.78–0.87]. Algorithm 3, which included both databases, improved sensitivity with 1562 observed cases, i.e., 159.2/100,000 and an SIR of 0.90 [0.86–0.95]. Inclusion of specific procedures enabled detection of 23 additional cases with a raw rate of 161.6/100,000 and an SIR of 0.92 [0.87–0.96]. Algorithm 5, which included drugs specific to the treatment of prostate cancer in addition to algorithm 4, enabled detection of 1778 incident cases, i.e., a raw rate of 181.3/100,000 and an SIR of 1.03 [0.98–1.08]. For the prostate cancer cohort and using algorithm 5, the median age at prostate cancer diagnosis was 69 years.

3.4. Sources of information enabling detection of incident cases of cancer

Table 4 shows the sources of the cancer information as a function of disease site and age group for algorithm 4 for breast cancer and colorectal cancer, and for algorithm 5 for prostate cancer.

With regard to the 1783 incident cases of breast cancer, the use of PD from the PMSI only detected slightly more than half of the breast cancer cases. LTD status contributed to case detection in 96% of the cases, particularly for the younger cases (92–100% of cases). The concordance rate for the two methods was only 52%.

In contrast, with regard to colorectal cancer diagnosis, the use of PD information in the PMSI detected more cases than LTD status (89% vs. 72%), particularly among elderly subjects

Table 3
Expected and observed incidence of prostate cancer between 2008 and 2012 according to algorithms used.

Prostate cancer algorithms	No. of cancers observed	Raw rate (*n*/100,000)	No. of cancers expected	National rate[a] (*n*/100,000)	SIR [95% CI]
Algorithm 1 (PMSI PD)	723	73.5	1730	176.0	0.42 [0.39–0.45]
Algorithm 1b (PMSI PD + LD)	774	78.7	1730	176.0	0.45 [0.42–0.48]
Algorithm 2 (LTD)	1419	144.6	1727	176.0	0.82 [0.78–0.87]
Algorithm 3 (PMSI PD + LD or LTD)	1562	159.2	1727	176.0	0.90 [0.86–0.95]
Algorithm 4 (PMSI PD + LD or LTD or procedure)	1585	161.6	1727	176.0	0.92 [0.87–0.96]
Algorithm 5 (PMSI PD + LD or LTD or procedure or drug)	1778	181.3	1726	176.0	1.03 [0.98–1.08]

No.: number; SIR: standardized incidence ratio; PMSI PD: medical information systems program, principal diagnosis; PMSI LD: medical information systems program, linked diagnosis; LTD: long-term disease.
[a] Estimated by the InVS for 2009 [12].

Table 4
Sources of information enabling detection of incident cases of cancer as a function of disease site and age group.

Site/age	PMSI PD (with or without LTD) *n* (%)	LTD (with or without PMSI) *n* (%)	PMSI PD and LTD (with or without another marker) *n* (%)	Drug (with or without another marker) *n* (%)	Procedure (with or without another marker) *n* (%)
Breast cancer (algorithm 4)					
0–29 years	6 (46)	12 (92)	5 (38)		
30–39 years	64 (48)	133 (100)	64 (48)		
40–49 years	210 (54)	385 (98)	204 (52)		
50–59 years	230 (54)	416 (97)	218 (51)		
60–69 years	222 (56)	386 (97)	210 (53)		
70–79 years	178 (64)	262 (94)	160 (57)		
≥ 80 years	81 (58)	124 (89)	65 (46)		
All	991 (56)	1718 (96)	926 (52)		
Colorectal cancer (algorithm 4)					
0–29 years	3 (75)	3 (75)	2 (50)		
30–39 years	15 (75)	20 (100)	15 (75)		
40–49 years	59 (92)	56 (87.5)	51 (80)		
50–59 years	212 (90)	191 (81)	167 (71)		
60–69 years	289 (88)	239 (73)	199 (60)		
70–79 years	345 (89)	276 (71)	233 (60)		
≥ 80 years	222 (89.5)	145 (58.5)	119 (48)		
All	1145 (89)	930 (72)	786 (61)		
Prostate cancer (algorithm 5)					
0–29 years	0	0	0	29 (97)	1 (3)
30–39 years	0	0	0	11 (100)	0
40–49 years	15 (30)	30 (60)	15 (30)	27 (54)	18 (36)
50–59 years	150 (46)	290 (89)	136 (42)	77 (23.5)	190 (58)
60–69 years	328 (56)	525 (89)	288 (49)	160 (27)	348 (59)
70–79 years	121 (39)	446 (82)	163 (30)	288 (53)	86 (16)
≥ 80 years	69 (30)	128 (56)	29 (13)	163 (71.5)	2 (0.9)
All	774 (43.5)	1419 (80)	631 (35.5)	755 (42.5)	645 (36)

PMSI PD: medical information systems program, principal diagnosis; LTD: long-term disease.

for whom LTDs were underreported (59% of patients aged 80 years or older with LTD status for colorectal cancer among the cases). There was no between-gender difference with regard to the sources of information.

For prostate cancer, the use of the PD from the PMSI contributed little to any age group and detected less than 45% of the cases identified. LTD underreporting in subjects aged 80 years or more was observed (56% vs. 80% overall). In over 70% of the cases, those in the extreme age groups (0–39 years and 80 years or older) were identified on the basis of a drug specific to prostate cancer. This was far from the case for the other age groups.

4. Discussion

In this study, in addition to hospital data, reimbursement data enabled a rapid and relatively easy improvement in estimation of the incidence rates for breast, prostate, and colorectal cancer at the national level, with a minimum of 3 years of data history to detect prevalent cases. The standardized incidence rates obtained with algorithm 4, taking into account the PD during hospitalization or LTD status together with the mandatory presence of a marker procedure, were close to those obtained from the breast and colorectal cancer registry data. With regard to prostate cancer, in addition to the PD from the PMSI or LTD status, the potential presence of a specific procedure or drug improved sensitivity (algorithm 5).

The medico-administrative databases are currently being increasingly used to estimate the incidence of cancers, particularly breast and colorectal cancer [8,9,13]. Due to the almost systematic need for hospitalization and a diagnostic procedure, these diagnoses are readily recorded in the databases, which have the advantage of being exhaustive and national in scope. Since 2004, the tariff system based upon the hospital discharge database (PMSI) has improved the quality of the information derived from the ICD-10 codes and the CCAM. The temporal sequencing enables exclusion of cases identified in previous years (prevalent cases) and linkage with reimbursement data enables addition of information on LTD status as well as open-care procedures and drugs. To the authors' knowledge, few studies have specifically addressed algorithms derived from hospital data in combination with health insurance reimbursement data [11,14]. The algorithms herein were developed using the EGB data. The results in this study (cases detected in medical and administrative databases) were compared to the estimates derived from the registries by calculation of the SIR. Initially, the comparability of the registry population (general population derived from INSEE data [15]) with the population derived from the EGB with general regimen coverage was verified. With 48% men and 52%

women and the same age pyramid in the two populations, the comparability was excellent.

National and international publications estimating the incidence of cancers from medico-administrative databases are few and use relatively old data with a short follow-up [5,6,8,9,13]. With regard to breast cancer, the highest sensitivity found in the literature was 0.77 with a positive predictive value (PPV) of 0.93 in an Italian study, which was already old. The study used hospital data from 2000. The PD combined with a marker procedure and the previous year's cases were excluded [6]. The French studies using the same methodology had a sensitivity of 0.64–0.75 [8,13] with a 30% lower rate of false-positives when a marker procedure was also used [8]. The incidence rates obtained with LTD status alone had equivalent sensitivity but a higher PPV (0.76) [13]. For colorectal cancer, the PD from the PMSI yielded greater sensitivity than the PD combined with the LTD data (0.83 vs. 0.55) but with an inversely proportional PPV, which improved when a marker procedure was used (PPV: 0.73) [9]. In the present study, the C21 code corresponding to anal cancer was not taken into account because it is a distinct histology from colorectal cancer with distinct risk factors and physiopathology. This disease represents a small number of patients and did not impact the comparison with registry data when taken into account with a similar SIR with each algorithm. Prostate cancer would appear to be very poorly detected by the hospital data due to frequent outpatient treatment, as was shown by a study on Spanish data [5]. In the present study, algorithm 1 using the PD from the PMSI alone identified fewer incident cases of prostate cancer compared to the estimates extrapolated from the registry data (SIR: 0.42–0.45), fewer breast cancer cases (SIR: 0.66) and, to a lesser extent, fewer colorectal cancer cases (SIR: 0.93 in men and 0.90 in women). The reverse trend was found with LTD status alone, which enabled greater detection of breast cancer cases (SIR: 1.19), detection of similar prostate cancer populations (SIR: 0.82), and detection of fewer cases of colorectal cancer (SIR: 0.78 in men and 0.76 in women). This finding was confirmed by studying the variables necessary for determination of incidence as a function of site: the LTD status enabled detection of 96% of the breast cancer cases vs. 56% for the PMSI PD (concordance: 52%). For prostate cancer, the LTD data enabled detection of 80% vs. 44% for the PMSI data (concordance: 36%). This situation was reversed for colorectal cancer, which was detected in 89% of cases with the PD from the PMSI data vs. 72% with the LTD data (concordance: 61%). This was probably related to the specificities of cancer management with almost obligatory hospitalization for colonic cancer (colonoscopy under general anesthesia or discovery during a complication), frequent outpatient treatment of prostate cancer, which was better detected by the LTD data, and mixed treatment for breast cancer. Pooling the two types of information thus enables the inadequacies of each type to be overcome, providing an enhanced approximation of the incidence of prostate cancer and that of colorectal cancer in men and women (SIR: 1.03 and SIR: 1.06 and 1.06, respectively). More cases of breast cancer were identified than with the registries (SIR: 1.24). The surplus of identified cases is probably related to false incident cases (persistence of prevalent cases, coding error) and had already been reported in the only study incorporating PMSI and LTD data of which the present authors are aware [16]. For the disease site in question, the surplus was partially corrected using a procedure necessary for histological diagnosis and marking the initial management (SIR: 1.12). The SIR fell to 0.94 and 0.93 in men and women, respectively, for colorectal cancer. For prostate cancer, the marker procedures are not pertinent since management may be entirely in an outpatient setting or consist in simple monitoring. The enhanced sensitivity resulting from addition of procedures and drugs specific to the management of prostate cancer yielded an SIR of 1.03.

For all the algorithms, in order to decrease the false-positives related to prevalent cases, LTD status and PD, LD, or AD in the 3 years preceding the diagnosis were excluded. This enabled longer follow-up than most of the published studies on this subject, which had a follow-up of 1 or at most 2 years [5,6,11].

The fact that prostate cancer drugs more frequently give rise to detection of prostate cancer in young or even very young subjects probably relates to a classification bias related to use of the drugs in another indication (e.g., precocious puberty). In contrast, prostate cancer drugs are prescribed in the majority of subjects aged 80 years or older (71% of cases), who may only receive one drug related to prostate cancer without hospitalization or any associated procedure. This reflects the differences in the treatment of prostate cancer as a function of patient age.

One of the limitations of this study is related to the fact that the sensitivity and specificity of the algorithms cannot be estimated since it is impossible to return to the files and validate cases using the histological data. Moreover, this classification bias may persist during the coding of the hospital data, sometimes with a delay in obtaining histological proof of cancer or in the use of drugs specific to prostate cancer for other reasons. However, the indications remain marginal (precocious puberty, misuse). An information bias may also be related to underreporting of LTD by the attending physician, who may, erroneously, not formulate an LTD application when the patient already has LTD status for multiple diseases or when the patient has good complementary insurance coverage, particularly when the disease does not require expensive treatment. Algorithm 4, using marker procedures for the initial treatment and frequently necessary for histological proof of breast and colorectal cancer, may induce selection bias by excluding elderly subjects for whom the procedures are not performed due to their fragile condition. The strength of algorithm 4 resides in the rapid availability of data derived from several sources and exhaustively collected based on an economic logic. This results in enhanced sensitivity to changes in incidence trends compared to the cancer incidence estimates published from registry data. This was observed by a study published in 2009, which showed a reduction in the incidence of breast cancer based on the health insurance system data before the registries were able to confirm it [17].

In conclusion, this study proposes a new approach, based on medical and administrative data derived from the EGB, to estimate national incidence rates of several cancers. The

algorithm providing the closest SIR to those derived from the registry data estimates incorporates the PMSI with the PD and/or LTD data obligatorily associated with a marker procedure for breast cancer and colorectal cancer, while excluding the cases identified in the PMSI and LTD in the previous 3 years insofar as the database permits. With regard to prostate cancer, which may only necessitate outpatient follow-up or simple monitoring, a more complex algorithm incorporating the PMSI data with PD and LD, LTD, and procedures and drugs specific to the cancer, while excluding the patients for whom the foregoing occurred in the previous 3 years, provides estimates similar to those deriving from the registry data.

Disclosure of interest

The authors declare that they have no competing interest.

Acknowledgments

We acknowledge the gene, environment cancer epidemiology team of the center for research in epidemiology and population health (CESP), Inserm U1018, the Paris-Sud university, UMRS 1018, the hepato-gastroenterology department, unit of gastrointestinal tumor screening and treatment, Pitié-Salpêtrière hospital, Paris public hospital authority (AP–HP) and the National health insurance organization for salaried workers (CNAM).

References

[1] Binder-Foucard F, Bossard N, Delafosse P, Belot A, Woronoff A-S, Remontet L. Cancer incidence and mortality in France over the 1980–2012 period: solid tumors. Rev Epidemiol Sante Publique 2014;62(2):95–108.

[2] Colonna M, Bossard N, Mitton N, Remontet L, Belot A, Delafosse P, et al. Éléments d'interprétation des estimations régionales de l'incidence du cancer en France sur la période 1980–2005. Rev Epidemiol Sante Publique 2008;56(6):434–40.

[3] Olive F, Gomez F, Schott A-M, Remontet L, Bossard N, Mitton N, et al. [Critical analysis of French DRG based information system (PMSI) databases for the epidemiology of cancer: a longitudinal approach becomes possible]. Rev Epidemiol Sante Publique 2011;59(1):53–8.

[4] Mitton N, Colonna M, Trombert B, Olive F, Gomez F, Iwaz J, et al. A suitable approach to estimate cancer incidence in area without cancer registry. J Cancer Epidemiol 2011;2011:418968.

[5] Bernal-Delgado EE, Martos C, Martínez N, Chirlaque MD, Márquez M, Navarro C, et al. Is hospital discharge administrative data an appropriate source of information for cancer registries purposes? Some insights from four Spanish registries. BMC Health Serv Res 2010;10:9.

[6] Baldi I, Vicari P, Di Cuonzo D, Zanetti R, Pagano E, Rosato R, et al. A high positive predictive value algorithm using hospital administrative data identified incident cancer cases. J Clin Epidemiol 2008;61(4):373–9.

[7] Rañopa M, Douglas I, van Staa T, Smeeth L, Klungel O, Reynolds R, et al. The identification of incident cancers in UK primary care databases: a systematic review. Pharmacoepidemiol Drug Saf 2015;24(1):11–8.

[8] Couris CM, Polazzi S, Olive F, Remontet L, Bossard N, Gomez F, et al. Breast cancer incidence using administrative data: correction with sensitivity and specificity. J Clin Epidemiol 2009;62(6):660–6.

[9] Quantin C, Benzenine E, Hägi M, Auverlot B, Abrahamowicz M, Cottenet J, et al. Estimation of national colorectal-cancer incidence using claims databases. J Cancer Epidemiol 2012;2012:298369.

[10] Tuppin P, de Roquefeuil L, Weill A, Ricordeau P, Merlière Y. French national health insurance information system and the permanent beneficiaries sample. Rev Epidemiol Sante Publique 2010;58(4):286–90.

[11] Tuppin P, Samson S, Fagot-Campagna A, Lukacs B, Alla F, Paccaud F, et al. Prostate cancer outcomes in France: treatments, adverse effects and two-year mortality. BMC Urol 2014;14(1):48.

[12] Binder-Foucard F, Belot A, Delafosse P, Remontet L, Woronoff A, Bossard N. Estimation nationale de l'incidence et de la mortalité par cancer en France entre 1980 et 2012. Partie 1 – tumeurs solides. Saint-Maurice (France): Institut de veille sanitaire; 2013.

[13] Grosclaude P, Dentan C, Trétarre B, Velten M, Fournier E, Molinié F. Utilité des bases de données médico-administratives pour le suivi épidémiologique des cancers. Comparaison avec les données des registres au niveau individuel. Bull Epidemiolo Hebd 2012;5–6:63–7.

[14] Neumann A, Weill A, Ricordeau P, Fagot JP, Alla F, Allemand H. Pioglitazone and risk of bladder cancer among diabetic patients in France: a population-based cohort study. Diabetologia 2012;55(7):1953–62.

[15] Insee – Population – Bilan démographique 2010 – La population française atteint 65 millions d'habitants [Online]. Insee Premiere 2011;332:1–4. [Available on:http://www.insee.fr/fr/themes/document.asp?ref_id=ip1332 (cited October 14, 2014)].

[16] Kudjawu Y, de Maria F, Decool E, Altana M, Harlin J, Weill A. Croisement de deux bases de données médico-administratives : méthodologie et étude descriptive pour une application à la surveillance épidémiologique des cancers en France. Bull Epidemiol Hebd 2013;49–58.

[17] Séradour B, Allemand H, Weill A, Ricordeau P. Changes by age in breast cancer incidence, mammography screening and hormone therapy use in France from 2000 to 2006. Bull Cancer 2009;96(4):E1–6.