*Research Article*

# A Suitable Approach to Estimate Cancer Incidence in Area without Cancer Registry

**Nicolas Mitton,[1] Marc Colonna,[1, 2] Béatrice Trombert,[3] Frédéric Olive,[4] Frédéric Gomez,[5] Jean Iwaz,[6, 7, 8] Stéphanie Polazzi,[9] Anne-Marie Schott-Petelaz,[9] Zoé Uhry,[10] Nadine Bossard,[6, 7, 8] and Laurent Remontet[6, 7, 8]**

[1] *Registre des Cancers de l'Isère, 38240 Meylan, France*
[2] *Réseau des Registres de Cancers FRANCIM, 31000 Toulouse, France*
[3] *Service de Santé Publique et d'Information Médicale, CHU de Saint-Etienne, 42100 Saint Etienne, France*
[4] *Département d'Information Médicale, CHU de Grenoble, 38043 Grenoble, France*
[5] *Département d'Information Médicale, Centre Léon Bérard, 69373 Lyon, France*
[6] *Hospices Civils de Lyon, Service de Biostatistique, 69003 Lyon, France*
[7] *Université de Lyon, Université Lyon I, 69100 Villeurbanne, France*
[8] *CNRS, UMR 5558, Laboratoire Biostatistique Santé, 69495 Pierre-Bénite, France*
[9] *Pole Information Médicale Evaluation Recherche, Hospices Civils de Lyon, Université de Lyon,
   EA 4129 Santé Individu Société, 69373 Lyon, France*
[10]*Institut de Veille Sanitaire (InVS), Département des maladies chroniques et des traumatismes, 94415 Saint-Maurice, France*

Correspondence should be addressed to Nicolas Mitton, nmitton.registre@orange.fr

*Objective*. Use of cancer cases from registries and PMSI claims database to estimate Département-specific incidence of four major cancers. *Methods*. Case extraction used principal diagnosis then surgery codes. PMSI cases/registry cases ratios for 2004 were modelled then Département-specific incidence for 2007 estimated using these ratios and 2007 PMSI cases. *Results*. For 2007, only colon-rectum and breast cancer estimations were satisfactorily validated for infranational incidence not ovary and kidney cancers. For breast, the estimated national incidence was 50,578 cases and the incidence rate 98.6 cases per 100,000 person per year. For colon-rectum, incidence was 21,172 in men versus 18,327 in women and the incidence rate 38 per 100,000 versus 24.8. For ovary, the estimated incidence was 4,637 and the rate 8.6 per 100,000. For kidney, incidence was 6,775 in men versus 3,273 in women and the rate 13.3 per 100.000 versus 5.2. *Conclusion*. Incidence estimation using PMSI patient identifiers proved encouraging though still dependent on the assumption of uniform cancer treatments and coding.

## 1. Introduction

In being responsible for about 350,000 new cases and 150,000 deaths yearly in France, cancer is a major health problem and its surveillance the utmost public health concern. Regarding surveillance, FRANCIM, the French network of cancer registries, is responsible for exhaustive collections of cancer cases in 10 to 14 French Départements (depending on the cancer type) corresponding to 15% to 20% of the French population. However, estimating epidemiological indicators at the scale of whole Départements over the country is necessary not only to reveal etiological factors and geographical or social discrepancies but also to plan the needs in terms of medical resources (prevention, treatment, and surveillance).

Over the previous ten years or so, FRANCIM and the Department of Biostatistics of the Hospices Civils de Lyon have been providing national estimations of cancer incidence [1, 2]. Their usual approach to produce these estimations is to use registry incidence data together with CépiDC mortality data (Centre d'Epidémiologie sur les Causes médicales de Décès) [3]. The principle is to calculate a mean ratio

between incidence and mortality in an area covered by cancer registries then use that ratio with national mortality data to derive an estimation of national incidence. While the mean ratio estimated in a registry area can be reasonably considered as representative of the ratio for the whole country, the same is not true at the level of a single Département because this ratio may be highly variable between Départements and because identical incidence values do not necessarily lead to identical mortalities. Indeed, a great number of factors are able to affect patient survival and generate heterogeneity of the ratio between Départements: differences in patient management (diagnostic or therapeutic procedures), prevention or screening policies, or compliance of the population with these policies. Therefore, because incidence and mortality cannot be used to provide Département-specific incidence estimations, a new approach should be sought for.

One interesting source of data with a nationwide coverage has been recently used together with registry data to estimate Département-specific cancer incidence: the hospital database of the Programme de Médicalisation des Systèmes d'Information Médicale (PMSI) [4–6]. This medicoadministrative database is held to help manage health institutions and provide budget estimates.

A previous work [7] has discussed the problem of using hospital stays from PMSI data to estimate Département-specific incidence of breast cancer. However, the constant improvement of the quality of patient identification in PMSI data using a single-patient identifier makes it possible now to use patient-specific rather than stay-specific data

In the present paper, our objective was to estimate Département-specific incidence of colon-rectum, breast, kidney, and ovary cancers for 2007 using mean ratios of PMSI-extracted cases to registry-extracted (incident) cases.

## 2. Materials and Methods

*2.1. PMSI Database, Hospital Stay-Specific Data, and Patient-Specific Data.* The French Agence Technique de l'Information sur l'Hospitalisation (ATIH) made available its data on all short stays in all health institutions over France for 2002–2007. In our analyses, we kept the variables related to personal characteristics (sex, age, and code of the residence area), hospital stay (stay number and principal diagnosis according to the International Classification of Diseases (ICD-10) [8], and medical procedures according to the Catalogue des Actes Médicaux (CdAM) until 2004 [9] and to the new Classification Commune des Actes Médicaux (CCAM) [10] from 2004 to 2007, plus an anonymous alphanumerical patient identifier [11, 12] to allow chaining of hospital stays of the same patient in successive institutions. That identifier is systematically generated by procedure FOIN (Fonction d'Occultation des Informations Nominatives) in all French health institutions since 2001. Hospital stays with no available patient identifier were excluded from the present analyses.

Two algorithms were independently used to extract hospital stays to analyze. The first one extracted all stays with cancer as principal diagnosis. The corresponding CIM 10 codes were C18 to C21 for colon-rectum cancer, C50

for breast cancer, C56 and C57.0 to C57.4 for ovary cancer, and C64 to C66 plus C68 for kidney cancer. The second algorithm extracted stays with cancer as principal diagnosis *and* with surgical procedures for cancer. Using CdAM and CCAM codes, the latter extraction considered 95 procedures for colon-rectum cancer, 31 for breast cancer, 114 for ovary cancer, and 44 for kidney cancer.

After each type of extraction, hospital stays were ordered by their serial numbers to spot the first stay of each patient, then only these stays were kept for analysis. These stays were then counted over each Département by age group: ten groups for colon-rectum cancer (15–44, 45–49, 50–54, ..., 80–84, and ≥85 yrs), eleven groups for kidney cancer in women (15–39, 40–44, 45–49, ..., 80–84, and ≥85 yrs), and thirteen groups for each of breast cancer, ovary cancer, and kidney cancer in men (15–29, 30–34, 35–39, ..., 80–84, ≥85 yrs).

*2.2. Registry Incident Cancer Cases.* FRANCIM network made available the data on incident cancer cases registered in 2004 (the most recent and checked data set when the present study was initiated).

Cancer sites were determined according to the International Classification of Diseases for Oncology, third version (ICD-O-3) [8] and corresponded to invasive tumors. These codes were: C18 to C21 for colon-rectum cancers, C50 for breast cancer, C56 and C57.0 to C57.4 for ovary cancer (excluding morphological codes 8442/3, 8451/3, 8461/3, 8462/3, 8472/3, and 8473/3), and C64 to C66 plus C68 for kidney cancer.

The cancer registries used were those of eleven Départements: Calvados, Côte d'Or, Doubs, Hérault, Isère, Loire-Atlantique, Bas-Rhin, Haut-Rhin, Saône-et-Loire, Somme, and Tarn. Incidents cases of cancers were counted by the same age group as for hospital stays.

*2.3. Modeling the Ratio of PMSI Cases to Incident Cases.* Our approach was to model, in function of age, the PMSI cases/incident cases ratio; that is, the ratio of the number of patients with hospital stays for cancer present in the PMSI database to the number of incident cases present in the registries. This ratio was obtained from Départements where the two sources of information exist (i.e., Départements with a registry). It was then applied to PMSI data of Départements without registry, in order to estimate cancer incidence in these areas. This ratio was obtained from registry and PMSI data of year 2004; it was then applied to PMSI data for 2007 to derive Département-specific incidence for 2007.

The method adopted to model that ratio was detailed by Remontet et al. [7]. It is a calibration method where incidence, as obtained from cancer registries, is considered as reference or "true" value whereas PMSI cases allow only an approximation of this value. More precisely, the ratio is modelled as a function of age (effects smoothed using cubic regression spline) and Département (considered as random effect).

Further, to estimate the PMSI cases/incident cases ratio, a data quality criterion was required: that the chaining rate be greater or equal to 95%. This chaining rate was defined as
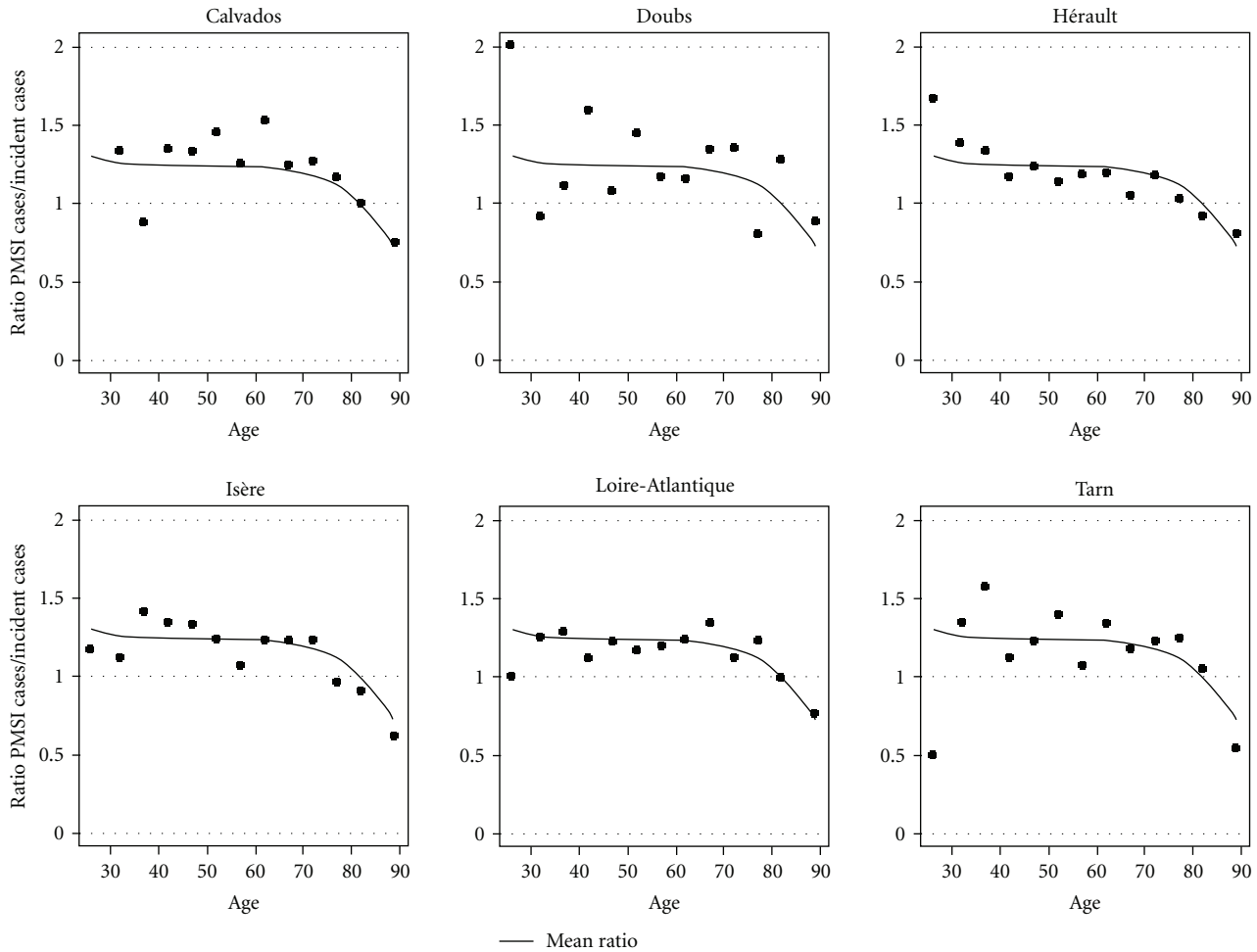
FIGURE 1: Breast cancer-PMSI cases/incidence case ratio by age classes by Département in year 2004.

the ratio of the number of stays with personal identifier to the total number of stays.

The analysis was carried out for the four cancer sites in women but only for colon-rectum cancer and kidney cancer in men. Whenever the number of cancer cases was small, it was not possible to take into account between-Département variability in areas with registries. We had then to sum up data from several Départements and calculate the PMSI cases/incident cases ratio per age group over the entire zone.

The overall national incidence was calculated by summing all Département estimations. For validation, we compared these national incidence values to FRANCIM values previously obtained by modeling the incidence/mortality ratio [1, 2, 13].

*2.4. Validation of the Estimations.* Validation was carried out through three steps. In step 1, the PMSI cases/incident cases ratio was calculated over all ages using each algorithm per cancer site-sex combination and Département. Indeed, Département-specific estimations stemming from this approach are invalid unless the ratio is homogeneous between Départements (always >1 or always <1). In step 2, the observed ratio for a given age group and Département

was graphically compared to the modelled mean ratio over all Départements with registries (Figure 1). However, to be applicable to all Départements, the modelled mean ratio should not suffer a "Département effect". In the presence of this effect, the observed ratios per age groups tend to be systematically higher or lower than the mean ratio whereas in its absence, the observed ratios are distributed around the mean ratio. For cancer sites with high incidence (breast, colon-rectum), step 3 was a cross-validation [7]. The incidence in a given Département with registry is estimated from PMSI data together with the PMSI cases/incident cases ratio obtained by a model from which the data of that Département were excluded. A comparison between the number of observed cases and the number of predicted cases yields a Prediction Error (PE) [7]. Under hypothesis $H_0$ of a correct prediction, the PE obeys a $\chi^2$ rule whose degree of freedom is equal to the number of age groups. A 5% $\alpha$-risk was adopted to set the critical value for rejecting $H_0$. For cancer with low incidence (ovary, kidney), cross-validation could not be used, because it was difficult to determine the statistical distribution of the PE and so only graphical validation was done. In addition, a comparison between the total numbers of observed and predicted cases was carried

TABLE 1: Registry data and PMSI data used for estimations of the ratio in 2004. Validation steps of prediction of incidence using each algorithm.

| | Registry data | | PMSI data | | | Validation steps | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cancer site | Number of Départements | Observed cases | Algorithm | Extracted cases | Number of Départements with ratio <1 | Département effect | Cross-Validation | |
| | | | | | | | PE[1] | $\chi^2$[2] |
| *Colon-Rectum* | | | | | | | | |
| Men | 8 | 2,009 | 1 | 2,660 | 0 | No | 0 | 0 |
| | | | 2 | 1,850 | 7 | Yes | 0 | 2 |
| Women | 8 | 1,625 | 1 | 2,042 | 0 | No | 0 | 0 |
| | | | 2 | 1,373 | 7 | Yes | 2 | 1 |
| *Breast* | | | | | | | | |
| Women | 6 | 3,946 | 1 | 4,624 | 0 | No | 0 | 0 |
| | | | 2 | 4,037 | 2 | No | 2 | 0 |
| *Ovary* | | | | | | | | |
| Women | 6 | 390 | 1 | 529 | 0 | No | NA[3] | NA |
| | | | 2 | 383 | 3 | No | NA | NA |
| *Kidney* | | | | | | | NA | NA |
| Men | 8 | 657 | 1 | 786 | 0 | No | NA | NA |
| | | | 2 | 573 | 7 | No | NA | NA |
| Women | 5 | 231 | 1 | 284 | 0 | No | NA | NA |
| | | | 2 | 202 | 3 | No | NA | NA |

[1] Prediction error: number of registries presenting a statistically significant difference between observed and predicted cases by age group.
[2] $\chi^2$: Number of registries presenting a statistically significant difference between total observed and total predicted cases predicted.
[3] Not applicable.

out ($\chi^2$ with one degree of freedom). A relative error (RE) was calculated as the difference between the observed and predicted cases divided by the number of observed cases.

The results were mapped as standardized incidence ratios (SIR) using software MAPINFO (version 7.0).

## 3. Results

In 2004, the PMSI database was including 20,721,587 hospital stays of which 718,044 stays had cancer as principal diagnosis. The chaining rate of those stays was 96.4%. In 2007, there were 21,201,102 stays of which 721,823 had cancer as principal diagnosis and the chaining rate was 99%.

To illustrate our validation procedure the analyses relative to breast cancer were plainly detailed whereas those relative to the other cancer sites were less detailed.

*3.1. Breast Cancer.* Six registries were selected to estimate the incidence of breast cancer (Table 1). Four Départements with registries were excluded because, in 2004, their chaining rate was too low; it ranged between 49.7% and 94.9%.

Table 1 shows the total number of cases by cancer site and case-extraction algorithm as well as the three steps of estimate validation: homogeneity of the PMSI

cases/incident cases ratio through Départements with ratio <1, the Département effect, and the cross-validation.

Regarding the homogeneity of the ratio, the results show that algorithm 2 was inadequate. Indeed, whereas with algorithm 1 the number of PMSI cases was always higher than the incident cases in all Départements, with algorithm 2, the former number was sometimes higher (2 Départements) and sometimes lower (4 Départements) than the latter.

Regarding the Département effect, Figure 1 presents, for each Département with registries and for algorithm 1, the observed ratio by age group as well as the modelled mean ratio over all Départements with registries. The absence of heterogeneity in that ratio between Départements is illustrated by the fact that there was no Département in which the ratio per age group was systematically higher or lower than the modelled mean ratio (though Département Calvados tend to be systematically higher than the modelled mean ratio). It can be therefore concluded that there is no Département effect for breast cancer with algorithm 1 (the variance of the random effect was small: 0.023).

Regarding the cross-validation, Table 2 presents the detailed results for breast cancer. With algorithm 1, the differences between the observed and the predicted number of cases per age group (PEs) were small whereas with algorithm 2, two Départements displayed large differences

Table 2: Cross-validation of breast cancer estimations for each algorithm (PMSI data and registry data correspond to year 2004).

| Département | Observed cases | | Algorithm 1 | | | | | Algorithm 2 | | | |
| | | PMSI cases | Predicted cases | PE [1] | $\chi^2$ [2] | RE [3] | PMSI cases | Predicted cases | PE [1] | $\chi^2$ [2] | RE [3] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Calvados | 449 | 566 | 487 | 8 | 3.0 | +8.5 | 493 | 480 | **27.4** | 2.0 | + 6.8 |
| Doubs | 409 | 495 | 426 | 12.8 | 0.7 | +4.2 | 401 | 398 | **34.3** | 0.3 | −2.7 |
| Hérault | 919 | 1,031 | 872 | 7.1 | 2.5 | −5.1 | 900 | 880 | 12 | 1.8 | −4.3 |
| Isère | 901 | 1,039 | 873 | 9 | 0.9 | −3.1 | 901 | 869 | 8.9 | 1.2 | −3.5 |
| Loire-Atlantique | 963 | 1,138 | 959 | 5 | 0.0 | −0.4 | 1,036 | 1,019 | 10.9 | 3.1 | +5.8 |
| Tarn | 305 | 355 | 306 | 7.3 | 0.0 | +0.3 | 306 | 304 | 10.1 | 0.0 | −0.5 |
| Total | 3,946 | 4,624 | 3,923 | | | | 4,037 | 3,950 | | | |

[1] PE: prediction error, under the hypothesis that prediction is correct, PE follows a $\chi^2$ law with 10 degrees of freedom; that is, with a 5% $\alpha$-risk, if PE >18.3 then the prediction is not correct for that Département.
[2] $\chi^2$ with one degree of freedom, threshold value 3.84.
[3] RE: relative error; that is, predicted cases-observed cases/observed.

Table 3: National estimations of the number of cancer cases and of world-standardized incidence rates (per 100,000 persons) in 2007 based on the modeling of the ratio of PMSI cases to incidence cases using algorithm 1. Comparison with FRANCIM national estimates based on the modeling of the ratio incidence/mortality.

| Cancer site | National estimates based on the PMSI cases/Incident cases ratio | | FRANCIM estimates | |
| | Number of cases | Standardized rates[1] | Number of cases | Standardized rates[1] |
|---|---|---|---|---|
| *Colon-rectum* | | | | |
| Men | 21,172 | 38.0 | 20,453 | 37.4 |
| Women | 18,327 | 24.8 | 18,052 | 24.4 |
| *Breast* | | | | |
| Women | 50,578 | 98.6 | 52,492 | 103.4 |
| *Ovary* | | | | |
| Women | 4,637 | 8.6 | 4,402 | 7.9 |
| *Kidney* | | | | |
| Men | 6,775 | 13.3 | 6,328 | 13 |
| Women | 3,273 | 5.2 | 3,201 | 5.3 |

[1] standardized on world population age structure.

especially concerning the last age group. Furthermore, the difference between the total observed and the total predicted cases ($\chi^2$) was small with algorithm 1: thus, this algorithm 1 may be reliably used for breast cancer estimates.

Table 3 shows the national estimations obtained by adding the estimations obtained from all Départements by algorithm 1 as well as the national estimations elaborated by FRANCIM [14] through the use of the incidence/mortality ratio [1, 2, 13]. Comparing these two estimations was one way to validate the results of the present analysis.

The national incidence of breast cancer for 2007 was estimated at 50,578 cases. The World age-standardized incidence rate (WASR) was 98.6 cases per 100,000 and varied between Départements from 71.4 to 127.1 (Table 4).

### 3.2. Colon-Rectum Cancer.
To estimate Département-specific incidence of colon-rectum cancer, we used data from eight registries. Data from three Départements with registries were excluded because the chaining rate in 2004 was too poor (72.6% to 94.9%).

Irrespective of sex, the PMSI cases/incident cases ratios obtained with algorithm 1 were homogeneous (the variance

of the random effect was small: 0.024), which was not the case with algorithm 2 for which there was a Département effect in both gender. In addition, contrary to algorithm 1, cross-validation invalidated the estimations made with algorithm 2 in two Départements. Thus, we present only estimations made with algorithm 1. The national incidence for 2007 was estimated at 21,172 cases in men and 18,327 cases in women. The estimated national WASR was 38 cases per 100,000 in men and 24.8 in women. At the national level, our estimations were in high agreement with those of FRANCIM (40.8 cases per 100,000 in men and 24.8 cases per 100,000 in women).

Among Départements, the WASR ranged from 23.1 to 48.9 in men and from 17.4 to 31.3 in women.

### 3.3. Ovary Cancer.
The estimations of Département-specific incidence used six cancer registries. Three Départements with registries were excluded because the chaining rate was too poor (80% to 93.5%).
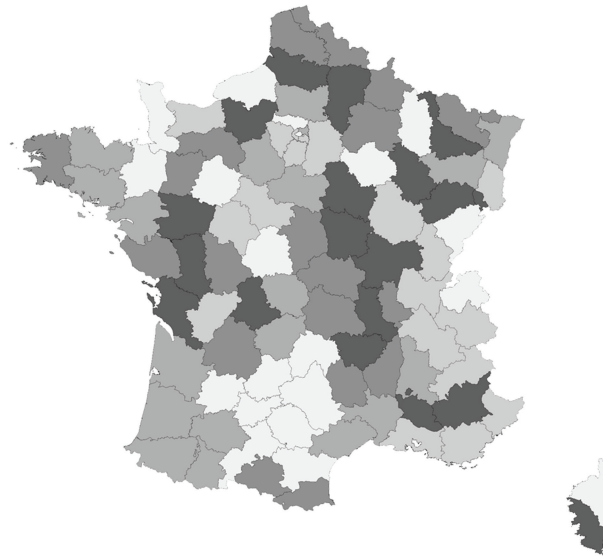
The PMSI cases/incident cases ratio carried out with algorithm 1 yielded homogeneous ratio between Départements, and no Département effect was observed. This was

TABLE 4: World standardized incidence rates in all French Départements (per 100,000 persons) in year 2007.

| Département | Code | Colon-rectum | | Breast |
| --- | --- | --- | --- | --- |
| | | Male | Female | |
| Ain | 01 | 35.1 | 25.2 | 82 |
| Aisne | 02 | 44.1 | 26.1 | 103.2 |
| Allier | 03 | 38.9 | 27.4 | 81.8 |
| Alpes-de-Haute-Provence | 04 | 43.5 | 24.3 | 98.1 |
| Hautes-Alpes | 05 | 35.9 | 17.4 | 94.3 |
| Alpes-Maritimes | 06 | 37 | 26.1 | 98.6 |
| Ardèche | 07 | 44.3 | 26 | 84.9 |
| Ardennes | 08 | 43.6 | 19.5 | 95.7 |
| Ariège | 09 | 40.5 | 25.8 | 80.7 |
| Aube | 10 | 28.8 | 23.3 | 84.1 |
| Aude | 11 | 35.2 | 24.3 | 95.6 |
| Aveyron | 12 | 30.3 | 23 | 80.3 |
| Bouches-du-Rhône | 13 | 37.7 | 26.5 | 111.6 |
| Calvados | 14 | 36.6 | 20.3 | 92.5 |
| Cantal | 15 | 27.3 | 20.5 | 83.3 |
| Charente | 16 | 38.2 | 24.8 | 93.5 |
| Charente-Maritime | 17 | 43.8 | 27 | 92.5 |
| Cher | 18 | 40.2 | 27.7 | 94.9 |
| Corrèze | 19 | 32.9 | 22.9 | 84.7 |
| Côte-d'Or | 21 | 35.7 | 22.7 | 84.9 |
| Côtes-d'Armor | 22 | 40.3 | 28.7 | 95.1 |
| Creuse | 23 | 41 | 27 | 83.2 |
| Dordogne | 24 | 40.8 | 22.8 | 91.6 |
| Doubs | 25 | 34.5 | 24.7 | 89.6 |
| Drôme | 26 | 37.3 | 28.3 | 97.9 |
| Eure | 27 | 43.4 | 25.2 | 97.2 |
| Eure-et-Loir | 28 | 38.2 | 25.9 | 101.1 |
| Finistère | 29 | 39.3 | 24 | 94.4 |
| Corse-du-Sud | 2A | 48.9 | 29.2 | 79.5 |
| Haute-Corse | 2B | 23.1 | 22.5 | 107.1 |
| Gard | 30 | 40.8 | 23.6 | 109 |
| Haute-Garonne | 31 | 34.4 | 22.5 | 96.8 |
| Gers | 32 | 34.6 | 26 | 79.7 |
| Gironde | 33 | 37.9 | 28 | 93.6 |
| Hérault | 34 | 38.6 | 22.7 | 94 |
| Ille-et-Vilaine | 35 | 30.8 | 19.6 | 90.3 |
| Indre | 36 | 34.6 | 27.6 | 86.8 |
| Indre-et-Loire | 37 | 35 | 27.5 | 108.8 |
| Isère | 38 | 35.9 | 23.4 | 96.3 |
| Jura | 39 | 36.3 | 19.5 | 81.8 |
| Landes | 40 | 37.5 | 24.4 | 92.7 |
| Loir-et-Cher | 41 | 37.9 | 27.6 | 104.9 |
| Loire | 42 | 40.7 | 27.1 | 97.5 |
| Haute-Loire | 43 | 43 | 25.3 | 77.9 |
| Loire-Atlantique | 44 | 39.8 | 23.7 | 108.4 |
| Loiret | 45 | 40 | 22.7 | 96.7 |
| Lot | 46 | 35.3 | 22.3 | 83.3 |
| Lot-et-Garonne | 47 | 32.3 | 24.2 | 71.4 |
| Lozère | 48 | 39.8 | 20.9 | 98.1 |

TABLE 4: Continued.

| Département | Code | Colon-rectum | | Breast |
|---|---|---|---|---|
| | | Male | Female | |
| Maine-et-Loire | 49 | 42.8 | 25.7 | 127.1 |
| Manche | 50 | 32.6 | 18.6 | 86.8 |
| Marne | 51 | 38.9 | 27.2 | 106.6 |
| Haute-Marne | 52 | 41.2 | 18.4 | 99.4 |
| Mayenne | 53 | 42.2 | 20.4 | 91.4 |
| Meurthe-et-Moselle | 54 | 41.7 | 26.5 | 109.9 |
| Meuse | 55 | 34.8 | 31.3 | 101.7 |
| Morbihan | 56 | 38.1 | 23.3 | 92 |
| Moselle | 57 | 41.6 | 24.4 | 101.8 |
| Nièvre | 58 | 48.6 | 22.8 | 80.6 |
| Nord | 59 | 41.6 | 26.2 | 102 |
| Oise | 60 | 38.8 | 26.3 | 93.7 |
| Orne | 61 | 41.5 | 27.4 | 112.6 |
| Pas-de-Calais | 62 | 40.5 | 27.7 | 97.6 |
| Puy-de-Dôme | 63 | 42.2 | 24.7 | 87.7 |
| Pyrénées-Atlantiques | 64 | 37.4 | 22.9 | 91.8 |
| Hautes-Pyrénées | 65 | 39.6 | 23.8 | 83.8 |
| Pyrénées-Orientales | 66 | 40.9 | 24 | 96.3 |
| Bas-Rhin | 67 | 38.1 | 23.5 | 97.7 |
| Haut-Rhin | 68 | 35.3 | 25.3 | 86.8 |
| Rhône | 69 | 39.9 | 24.6 | 99.3 |
| Haute-Saône | 70 | 42 | 28.5 | 100.5 |
| Saône-et-Loire | 71 | 43.4 | 25.3 | 97.9 |
| Sarthe | 72 | 31.2 | 19.8 | 103.9 |
| Savoie | 73 | 35.3 | 23.5 | 88.8 |
| Haute-Savoie | 74 | 29 | 17.8 | 96.3 |
| Paris | 75 | 35.8 | 25.5 | 117.6 |
| Seine-Maritime | 76 | 34.9 | 24.3 | 103.5 |
| Seine-et-Marne | 77 | 37.5 | 30.2 | 101.6 |
| Yvelines | 78 | 34.9 | 26.1 | 107.1 |
| Deux-Sèvres | 79 | 44.1 | 23.4 | 91.3 |
| Somme | 80 | 44.6 | 23.6 | 104.7 |
| Tarn | 81 | 28.7 | 21.2 | 87.5 |
| Tarn-et-Garonne | 82 | 34.1 | 23.6 | 108.8 |
| Var | 83 | 36.3 | 24.6 | 109.1 |
| Vaucluse | 84 | 45.6 | 25.7 | 101.6 |
| Vendée | 85 | 40.2 | 25.3 | 103.6 |
| Vienne | 86 | 39.7 | 20.1 | 93.7 |
| Haute-Vienne | 87 | 39.6 | 24.8 | 89.3 |
| Vosges | 88 | 37.7 | 25.3 | 89.6 |
| Yonne | 89 | 43.2 | 25.4 | 99.5 |
| Territoire-de-Belfort | 90 | 43.6 | 26.6 | 88 |
| Essonne | 91 | 35.1 | 26.3 | 101.3 |
| Hauts-de-Seine | 92 | 35.6 | 25.5 | 100.8 |
| Seine-Saint-Denis | 93 | 37.7 | 26.3 | 100 |
| Val-de-Marne | 94 | 35.6 | 24.6 | 104.2 |
| Val-d'Oise | 95 | 34.1 | 23.2 | 103.9 |
| **FRANCE** | | **38** | **24.8** | **98.6** |

SIR subgroups

☐ 0.633–0.92              ▨ 1.035–1.114
▨ 0.92–0.991             ▨ 1.114–1.289
▨ 0.991–1.035

FIGURE 2: Map of SIR of colon-rectum cancer in men.



SIR subgroups

☐ 0.754–0.928            ▨ 1.022–1.075
▨ 0.928–0.978           ▨ 1.075–1.236
▨ 0.978–1.022
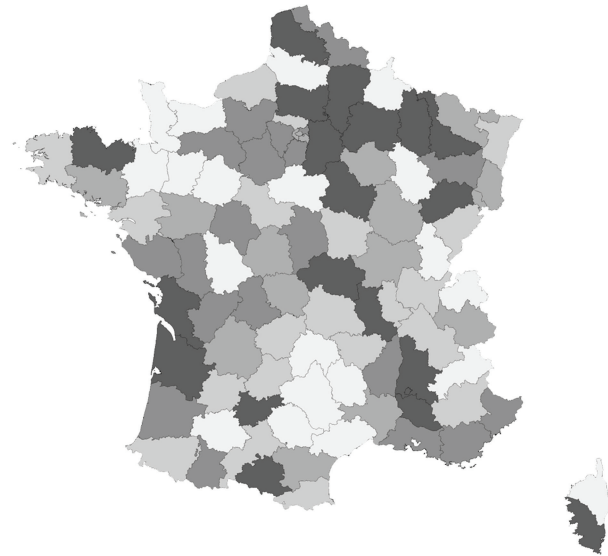
FIGURE 3: Map of SIR of colon-rectum cancer in women.

not the case with algorithm 2 for which heterogeneity was observed. However, as already mentioned in section Methods, cross-validation could not be carried out to validate estimate from algorithm 1 and thus, because of difficult formal validations of the Département-specific estimations, we only present national estimations based on this algorithm 1. The national incidence was 4,637 cases, which corresponds to a WASR of 8.6 cases per 100,000.

*3.4. Kidney Cancer.* The quality of data chaining in Départements with registries differed according to sex. Thus, eight Département registries were considered for men (one registry excluded because of a 92.5% chaining rate) and five registries for women (four registries excluded because of chaining rates ranging between 89.7% and 94.4%).

Here too, only a graphical validation of the estimations could be carried out and, because of difficult validations of Département-specific estimations, only national estimations are given.

As for ovary cancer and irrespective of sex, algorithm 1 performed better than algorithm 2. In 2007, the national incidence of kidney cancer was estimated at 6,775 in men and 3,273 in women; the national WASR was 13.3 per 100,000 in men but much lower (5.2 per 100,000) in women.

*3.5. SIR Maps.* Département-specific estimations for colon-rectum and breast cancer are shown in Table 4 and SIR maps of these cancers are shown in Figures 2, 3, 4. These maps were not constructed for ovary and kidney cancers

because of difficult formal validations of Département-specific incidence.

Overall, no clear geographical gradient could be seen. However, for colon-rectum cancer in men, the southwest was a low incidence area. This area was much larger in women. One well-marked low-incidence area for breast cancer was the southwest quadrant of France.

## 4. Discussion

To estimate the incidence of the four cancers in each Département, case extraction from PMSI database used two algorithms. algorithm 1 targeted all hospitalized cancer patients; that is, those whose principal diagnosis is cancer because of positive laboratory tests, metastasis staging, cancer-related procedures, or sudden potentially fatal progression (exacerbation or relapse). Thus, some prevalent cases were included along with incident cases. This is common in incidence estimations based on hospital data [15–17] and was confirmed here: the number found in PMSI data was higher than incident cases found in the registries (truer with algorithm 1 than with algorithm 2). Nevertheless, the PMSI cases/incident cases ratio obtained with algorithm 1 seemed somehow stable between Départements, which allowed estimations of Département-specific incidence. In contrast, algorithm 2 that used initial surgical procedures was more selective; it extracted a closer number of PMSI patients to the number of incident cases than algorithm 1. Thus, the PMSI cases/incident cases ratio was more heterogeneous between Départements (at different degrees according to the

SIR subgroups

☐ 0.7–0.897          ▨ 1.011–1.043
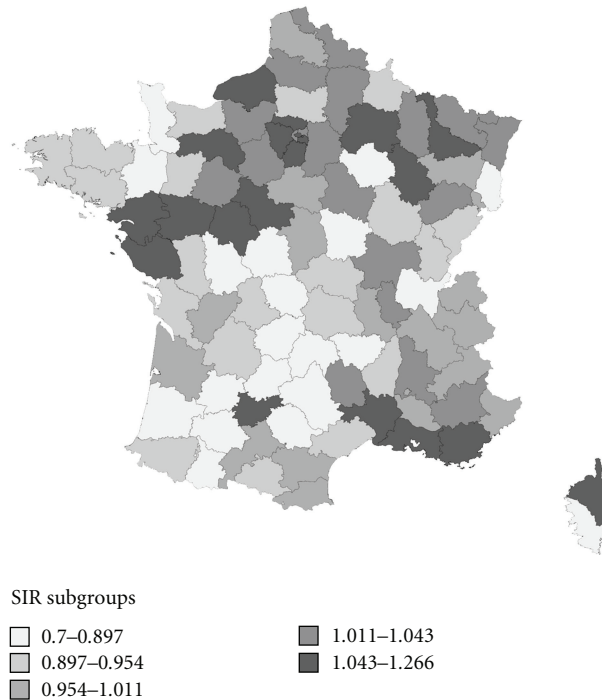▨ 0.897–0.954        ■ 1.043–1.266
▨ 0.954–1.011

FIGURE 4: Map of SIR of breast cancer.

cancer site). Besides, cross-validation revealed that incidence estimations in some Départements with registries were not valid. Indeed, with algorithm 2, the critical value of the prediction error was crossed for colon-rectum cancer in both sexes and for breast cancer in women. In sum, the simpler and less selective algorithm 1 was more adequate than algorithm 2 to estimate Département-specific incidence.

At the national level, our estimations were in agreement with FRANCIM projections [13], except for breast cancer (our estimations were lower). This was expected because estimations using incidence/mortality ratios do not take into account the recent trend towards a slow decline of breast cancer incidence in France [18] and in other countries [19, 20]. Our estimation made in 2007 at 50,578 new yearly cases in France seemed thus more realistic than the 52,492 cases stemming from FRANCIM projections. This comparison is interesting because it shows that in average, over all Départements, our approach leads to reliable estimations. Finally, our graphical validation cannot constitute a formal validation because, in borderline situations, one cannot claim a "Département effect". The graphical method would be only as a tool to detect important departures from the model assumptions.

PMSI data concern only hospitalized patients; thus, our method does not apply to cancers such as skin melanomas or basocellular cancers, which are usually treated early without hospitalization. Besides, the method supposed identical treatment choices in all Départements; which motivated the choice of the four cancer sites [21]. In two successive articles on colon and rectum cancers [22, 23], Phelip et al. have shown that surgical resection was performed in 90% of cases without significant geographical variation between

Départements. A high variability in treatment choices leads to problems with case-extraction from the PMSI database. For example, in aged men, prostate cancer can be treated by surgery, radiotherapy, or even hormone therapy alone (without hospitalization) and, in bladder cancer, the surgical treatment depends on the stage. If surgery is avoided, the lack of "surgery for cancer" in the PMSI database leads to missingness of PMSI cancer cases.

In addition, our method lays fundamentally on the hypothesis that within a given age group, the PMSI cases/incident cases ratio is constant between Départements whereas several factors may affect that rate. First, chaining of hospital stays is essential. Indeed, in a previous work [7], an insufficient chaining led to consider the number of stays rather than the number of patients. Consequently, the between-Département variability in the mean number of stays per patient—essentially due to very different hospitalization policies and coding practices—prevented correct estimations. In 2002, the chaining rate was quite low (92%) but improved up to 2004 (nearly 96%) because of the implementation of a "Tarification A l'Activité" (a *prospective payment system*). This rate further improved nationwide up to 2007. This high quality will soon become a norm. The use of a single-patient identifier allows keeping a single record per patient whatever the number of hospital stays and insures a better homogeneity of PMSI data.

Variability may also stem from various coding habits in different health institutions because of ignorance or misinterpretation of coding rules. A wide interinstitution variability of the PMSI cases/incident cases ratio may lead to a wide between-Département variability, especially between Départements with few health institutions.

Besides, the PMSI cases/incident cases ratio reflects the fact that for a given incidence level, the prevalence may vary between Départements. Indeed, different rates of cancer-specific survivals between Départements affect hospital prevalence; thus, the number of PMSI cases. In fact, survival in a given Département may vary with the presence/absence of systematic screening, the existence or not of a reference health institution [24], and the educational [25] and socioeconomic levels of the population [26]. If survival is high because of complete cures, hospital prevalence and PMSI cases will decrease, which will underestimate incidence, but if survival is high but associated with more procedures, hospital prevalence and PMSI cases will increase, which will overestimate Département-specific incidence. The impact of different survival rates between Départements on hospital prevalence is complex to seize and undoubtedly dependent on the cancer site under study.

Another implicit hypothesis in our method is a constant PMSI cases/incident cases ratio from 2004 to 2007. This is plausible because data quality of both sources over that short period was deemed constant despite improvements in cancer therapies that would have changed the prevalence of cancer. Another effect of time is the change in the national standards of coding PMSI data. For example, coding palliative care as "cancer" has been replaced by a specific ICD-10 code (Z51.5). More recently, in 2009, the rules for the choice of the principal diagnosis have changed; the impact of that change should

be evaluated. Nevertheless, that impact would be limited within the context of the four cancers under study here.

Improvements of the present method are possible. A better followup of the same patients over several years would exclude a number of prevalent cases. The ratio would vary less between Départements. Ideally, if all prevalent cases were excluded, the ratio would be interpreted as the proportion of hospitalized incident cases and would be no more affected by different prevalence in different Départements. Another improvement would be to add data from the health insurance (Affections Longue Durée (ALD30) database of Caisses d'Assurance Maladie); a feasibility study of that possibility is underway.

## 5. Conclusion

Using an adequate method, it seems now possible to estimate Département-specific incidence of some cancers for a given year. A validation procedure should accompany these estimations. Nevertheless, this validation is only partial because Département-specific estimations will still suffer the basic assumption of similar coding practices in all hospitals.

## Abbreviations

WASR: World age-standardized incidence rate
SIR:     Standardized incidence ratio.

## Acknowledgments

## References

[1] L. Remontet, J. Estève, A. M. Bouvier et al., "Cancer incidence and mortality in France over the period 1978–2000," *Revue d'Epidemiologie et de Sante Publique*, vol. 51, no. 1 I, pp. 3–30, 2003.

[2] A. Belot, P. Grosclaude, N. Bossard et al., "Cancer incidence and mortality in France over the period 1980–2005," *Revue d'Epidemiologie et de Sante Publique*, vol. 56, no. 3, pp. 159–175, 2008.

[3] Centre d'Epidémiologie sur les Causes Médicales de Décès, A DEFINIR, 2009.

[4] C. M. Couris, C. Colin, M. Rabilloud, A. M. Schott, and R. Ecochard, "Method of correction to assess the number of hospitalized incident breast cancer cases based on claims databases," *Journal of Clinical Epidemiology*, vol. 55, no. 4, pp. 386–391, 2002.

[5] Z. Uhry, M. Colonna, L. Remontet et al., "Estimating infra-national and national thyroid cancer incidence in France from cancer registries data and national hospital discharge database," *European Journal of Epidemiology*, vol. 22, no. 9, pp. 607–614, 2007.

[6] Z. Uhry, L. Remontet, P. Grosclaude, M. Velten, and M. Colonna, "Estimating infra-national incidence of colorectal cancer in France from hospital discharge database," *Revue d'Epidemiologie et de Sante Publique*, vol. 57, no. 5, pp. 329–336, 2009.

[7] L. Remontet, N. Mitton, C. M. Couris et al., "Is it possible to estimate the incidence of breast cancer from medico-administrative databases?" *European Journal of Epidemiology*, vol. 23, no. 10, pp. 681–688, 2008.

[8] A. Fritz, C. Percy, A. Jack et al., *International Classification of Disease for Oncology*, 2000.

[9] ATIH, *Catalogue des Actes Médicaux*, ATIH, 2007.

[10] ATIH, *Classification Commune des Actes Médicaux*, ATIH, 2004.

[11] Ministère de l'emploi et de la solidarité, "Circulaire n°106 du 22 février 2001 relative au chaînage des séjours en établissements de santé dans le cadre du programme de médicalisation des systèmes d'information (PMSI)," *Bulletin Officiel*, vol. 13, 2001.

[12] G. Trouessin and F. A. Allaert, "FOIN: a nominative information occultation function," *Studies in Health Technology and Informatics*, vol. 43, pp. 196–200, 1997.

[13] L. Remontet, A. Belot, and N. Bossard, "Tendances de l'incidence et de la mortalité par cancer en France et projections pour l'année en cours : méthodes d'estimation et rythme de production," *Bulletin Epidémilogique Hebdomadaire*, vol. 38, pp. 405–408, 2009.

[14] Hospices civils de Lyon, Institut de veille sanitaire, and Institut national du cancer, "Projections de l'incidence et de la mortalité par cancer en France en 2009," Tech. Rep., 2009.

[15] M. E. Brackley, M. J. Penning, and M. L. Lesperance, "In the absence of cancer registry data, is it sensible to assess incidence using hospital separation records?" *International Journal for Equity in Health*, vol. 5, article no. 12, 2006.

[16] C. M. Couris, C. Forêt-Dodelin, M. Rabilloud et al., "Sensitivity and specificity of two methods used to identify incident breast cancer in specialized units using claims databases," *Revue d'Epidemiologie et de Sante Publique*, vol. 52, no. 2, pp. 151–160, 2004.

[17] J. L. Freeman, D. Zhang, D. H. Freeman, and J. S. Goodwin, "An approach to identifying incident breast cancer cases using Medicare claims data," *Journal of Clinical Epidemiology*, vol. 53, no. 6, pp. 605–614, 2000.

[18] B. Séradour, H. Allemand, A. Weill, and P. Ricordeau, "Changes by age in breast cancer incidence, mammography screening and hormone therapy use in France from 2000 to 2006," *Bulletin du Cancer*, vol. 96, no. 4, pp. E1–E6, 2009.

[19] A. K. Hausauer, T. H. M. Keegan, E. T. Chang, S. L. Glaser, H. Howe, and C. A. Clarke, "Recent trends in breast cancer incidence in US white women by county-level urban/rural and poverty status," *BMC Medicine*, vol. 7, article 31, 2009.

[20] A. G. Glass, J. V. Lacey, J. D. Carreon, and R. N. Hoover, "Breast cancer incidence, 1980–2006: combined roles of menopausal hormone therapy, screening mammography, and estrogen receptor status," *Journal of the National Cancer Institute*, vol. 99, no. 15, pp. 1152–1161, 2007.

[21] F. Olive, F. Gomez, A. M. Schott et al., "Analyse critique des données du PMSI pour l'épidémiologie des cancers: une approche longitudinale devient possible," Revue d'Epidémiologie et de Santé Publique .

[22] J. M. Phelip, P. Grosclaude, G. Launoy et al., "Are there regional differences in the management of colon cancer in France?" *European Journal of Cancer Prevention*, vol. 14, no. 1, pp. 31–37, 2005.

[23] J. M. Phelip, G. Launoy, M. Colonna et al., "Regional variations in management of rectal cancer in France," *Gastroenterologie Clinique et Biologique*, vol. 28, no. 4, pp. 378–383, 2004.

[24] X. Q. Yu, D. L. O'Connell, R. W. Gibberd, and B. K. Armstrong, "Assessing the impact of socio-economic status on cancer survival in New South Wales, Australia 1996–2001," *Cancer Causes and Control*, vol. 19, no. 10, pp. 1383–1390, 2008.

[25] O. Dejardin, L. Remontet, A. M. Bouvier et al., "Socioeconomic and geographic determinants of survival of patients with digestive cancer in France," *British Journal of Cancer*, vol. 95, no. 7, pp. 944–949, 2006.

[26] Ø. Kravdal, "Does place matter for cancer survival in Norway? A multilevel analysis of the importance of hospital affiliation and municipality socio-economic resources," *Health and Place*, vol. 12, no. 4, pp. 527–537, 2006.