Hindawi Publishing Corporation
Journal of Cancer Epidemiology
Volume 2012, Article ID 298369, 7 pages
doi:10.1155/2012/298369

Research Article

Estimation of National Colorectal-Cancer Incidence Using Claims Databases

C. Quantin,[1,2] E. Benzenine,[1] M. Hägi,[1] B. Auverlot,[1] M. Abrahamowicz,[3] J. Cottenet,[1] E. Fournier,[4] C. Binquet,[1,2] D. Compain,[1] E. Monnet,[5] A. M. Bouvier,[2,6] and A. Danzon[4]

[1] *Service de Biostatistique et d'Informatique Médicale (DIM), Centre Hospitalier Universitaire, BP 77908, 21079 Dijon Cedex, France*
[2] *INSERM U866, Université de Bourgogne, 21000 Dijon, France*
[3] *Department of Epidemiology and Biostatistics, McGill University, Montreal, QC, Canada H3A 1A2*
[4] *Registre des Tumeurs du Doubs EA 3181, Université de Franche-Comté, 25000 Besançon, France*
[5] *Service d'Hépatologie et de Soins Intensifs Digestifs, Hôpital Jean-Minjoz, 25000 Besançon, France*
[6] *Registre Bourguignon des Cancers Digestifs, Faculte de Medecine, EPI-INSERM 0106, BP 87900, 21000 Dijon, France*

Correspondence should be addressed to C. Quantin, catherine.quantin@chu-dijon.fr

Received 21 February 2012; Revised 19 April 2012; Accepted 4 May 2012

Academic Editor: Hermann Brenner

Background. The aim of the study was to assess the accuracy of the colorectal-cancer incidence estimated from administrative data. *Methods.* We selected potential incident colorectal-cancer cases in 2004-2005 French administrative data, using two alternative algorithms. The first was based only on diagnostic and procedure codes, whereas the second considered the past history of the patient. Results of both methods were assessed against two corresponding local cancer registries, acting as "gold standards." We then constructed a multivariable regression model to estimate the corrected total number of incident colorectal-cancer cases from the whole national administrative database. *Results.* The first algorithm provided an estimated local incidence very close to that given by the regional registries (646 versus 645 incident cases) and had good sensitivity and positive predictive values (about 75% for both). The second algorithm overestimated the incidence by about 50% and had a poor positive predictive value of about 60%. The estimation of national incidence obtained by the first algorithm differed from that observed in 14 registries by only 2.34%. *Conclusion.* This study shows the usefulness of administrative databases for countries with no national cancer registry and suggests a method for correcting the estimates provided by these data.

1. Introduction

Cancer registries provide reliable statistical material, but they usually collect information only from specific geographic areas, thus cover only part of the population of a country. To estimate nationwide cancer incidence, the most commonly used method worldwide is to extrapolate the incidence/mortality ratio recorded in population-based registries to the total number of cases where cancer is reported as the underlying cause of death on death certificates at the national level. This method is clearly more efficient for cancers with a high mortality rate and requires the incidence/mortality ratio to be consistent across the country. In addition, the quality of cancer mortality data obtained from death certificates varies greatly according to the cancer site. Indeed, for some sites like the digestive system, several studies have shown that there can be more cases recorded in population-based registries than are reported in death certificates [1–3]. Likewise, if the patient dies of another cause, the cancer is most often not mentioned in the death certificate [4, 5]. In such cases, the death rate for cancer in a given site will probably be underreported, thus leading to an underestimation of the incidence of that cancer. The opposite is true for other cancers, especially in cases when the metastasis rather than the primary cancer was recorded as the cause of death. Given the above, estimations of colorectal-cancer incidence at the national level should preferentially be based on morbidity data rather than on mortality data

and should rely on a larger data source than cancer registries. Concerning the latter, administrative claims databases are widely regarded as a valuable source of data. Previous surveys studied information that dated back more than 12 years [6–14] and the quality of administrative data has improved since then. In light of the above, in this study, we aimed to define and compare two algorithms constructed to identify new cases of colorectal-cancer in the nationwide DRG-system based French administrative database, and to use these in a model validation study.

2. Materials and Methods

We first defined two different algorithms to identify new cases of colorectal-cancer in the national administrative database. Secondly, we applied these algorithms to a subset of administrative data concerning patients for whom cancer information was available from other sources, namely the rest of the medical record and two local population-based cancer registries. We then assessed our algorithms by comparing their results to the baseline data of the two registries, and when differences occurred, we explored the corresponding medical records to understand the discrepancy. Finally, once the causes of the discrepancy had been identified, they were incorporated into two separate multivariable logistic regression models. These models finally helped us to correct estimates of colorectal-cancer incidence at the national level obtained by applying our algorithms to the entire administrative database. The national estimate obtained was also compared with the data of all available registries.

French cancer registries are managed in accordance with the recommendations of both the International Agency for Research on Cancer and the European Network of Cancer Registries. In this study, we approached two colorectal-cancer registries that identify and record all new cases of inpatients diagnosed with invasive tumours within two geographical districts, "Côte d'Or" and "Doubs". We also approached all public and private hospitals of these districts (18 hospitals; Côte d'Or: 11; Doubs: 7) and asked them to provide their relevant data. There were no refusals. As the data hosted in the registries come directly from all relevant sources of information (public and private pathology and cytology laboratories, patients' medical files for both outpatients and inpatients, death certificates, and data from the National Health Service for patients whose costs are completely reimbursed) [15, 16], and as these data are regularly checked and validated, we assumed that they were far more reliable than any estimate and thus used them as the reference.

The national administrative database gathers information regarding inpatients and is based on the so-called DRG system. This kind of system is widely used around the world, but the French model has the specific feature of covering the entire population of the country. As all of the reimbursements of healthcare expenditure to health facilities are exclusively based on this system, the major strength of this database is that data are exhaustive. The diagnoses are coded according to the 10th edition of the International Classification of Diseases (ICD10). The procedures are coded according the CCAM codes, the French equivalent of the HCPCS or CPT codes, which include both medical and surgical procedure codes.

2.1. Identification of Incident Cases in the Administrative Claims Database. An incident case is above all a case, which means that the diagnosis of cancer had to be retrievable from the patient's information. It also had to be a new case, and two ways to check for this are commonly described in the literature [6–8, 12, 17–23]. The first is based on the need to retrieve a procedure specific to the first occurrence of the disease. The second is based on the absence of a previous diagnosis for that cancer in the administrative data over a certain period of time, which would ideally be the patient's lifetime.

In our study, we chose to use the two approaches simultaneously by developing two corresponding algorithms.

Algorithm 1 is mainly based on diagnosis and procedure codes, without taking into account the timing of the events. It defines incident cases as inpatients with both a principal diagnosis of colorectal-cancer (ICD 10 code C18 to C20) and a specific colorectal-cancer procedure mainly associated with initial treatment, recorded for 2004 and 2005. These specific codes were as follows: "endoscopic examination of the colon or rectum," "partial or total exeresis of the colon or rectum (primary or secondary surgery)," "excision, exeresis or destruction of polyps or tumours in the colon or rectum," "colostomy repair or closure," "secondary restoration of continuity" and "implantation of a colon endoprosthesis." Chemotherapy and radiotherapy may also have been used as the initial treatment or in the case of recurrences, but when they are used as an initial treatment, they are almost always adjuvant to the surgery [24]. That is why we chose to include only surgery-related codes when we created the list of specific codes. When several admissions occurred for the same patient during the same year, only the first hospital stay was considered as reflecting an incident case. Algorithm 2 is almost exclusively based on diagnosis codes (same codes as those used for Algorithm 1) but the past history of the patient is also considered. It defines incident cases as inpatients with a principal or associated diagnosis of colorectal-cancer recorded for 2004 and 2005, with no other record over the previous five-year period, which was as far back as we could go.

By comparing the results of applying the two algorithms to local data, we aimed to determine which of the two definitions of "incident case" would be most likely to give accurate results.

As the subset of administrative data examined came from the hospitals of Côte d'Or and Doubs, those of their inhabitants admitted to hospital in another district may not have been included in our paper. In order to detect such cases and to prevent underestimates, Algorithm 1 was applied twice to the entire national database; the first time to detect migrant inhabitants of Côte d'Or and the second time for those of Doubs.

2.2. Assessment of the Identification. In compliance with confidentiality policies, data must be rendered anonymous prior to treatments. In practice, administrative data are

rendered anonymous before they are passed on by hospitals, and thus we applied the same anonymization procedures to registry data in order to make them linkable. For this purpose, we used our ANONYMAT software [25] based on hash-coding techniques. This software was also used to perform the linkage between cases identified as incident in administrative data and validated incident cases in registries.

As previously mentioned, the information recorded in the two registries was considered the gold standard, and any case identified as incident in administrative data by either algorithm but not identified as such in the registries was considered a false positive. Conversely, a case recorded as incident in the registries but not retrieved as such by either algorithm was considered a false negative.

For each algorithm, the sensitivity and the positive predictive values (PPV) were accordingly estimated.

To determine in detail the causes of the inaccuracy of the algorithms in identifying incident colorectal-cancers in the administrative data, an exploratory analysis of false negatives and false positives was conducted using the same methodology as in a previous study for breast cancer. False negatives were studied by going back to registry information while false positives were investigated through the medical records.

2.3. Computation of the Total Number of Incident Colorectal Cancer Cases from the National Administrative Database. As the validation study showed that Algorithm 1 performed better, we chose to use it to estimate the total number of incident colorectal-cancer cases at the national level.

We tried to correct the number of cases selected by Algorithm 1 in the national administrative data by taking into account that the quality of administrative data may vary with a patient's characteristics and geographical area. Indeed, there may be differences between the two districts and the entire country for the distribution of covariates associated with the probability of a person having an incident cancer.

For this purpose, two separate multivariable logistic regression models were used to estimate how the probability of a false negative and a false positive depended on the patient's characteristics. Then, each model was assessed using data from the Côte d'Or and Doubs dataset for which the "true" incidence status was known from the registries.

The first regression model was estimated using data on all cases identified as "positive" based on the administrative data (i.e., retrieved from the administrative database). Among these "positive" subjects, the binary response variable was assigned the value of "1" or "0" depending on whether a given case represented in fact a "false positive" or a "true positive" (i.e., was actually, resp., truly negative or truly positive, according to the registries). Similarly, the second regression model was estimated using data on all cases identified as "negative" in the administrative data (i.e. not retrieved by the administrative database query). Among these "negative" subjects, the binary response variable was assigned the value of "1" or "0" depending on whether a given case represented in fact a "false negative" or a "true negative" (i.e., was actually, resp, truly positive or truly negative, according to the registries). In the model for false positives, the independent variables included "age," "gender," and "geographical area," and "hospital type." In the model of false negatives, the independent variables included age and gender. Indeed, the variables "geographical area" and "hospital type" could not be used as, by definition, there was no admission for negatives cases (not retrieved by the administrative database query). The estimated parameters of the model were then applied to the inpatients selected as not incident by Algorithm 1. Specifically, for each of these we calculated the estimated probability that a given inhabitant actually had an incident cancer, as a function of the individual's aforementioned covariates. The total number of incident cases missed by the national administrative database (false negatives) was estimated by summing up all the individual probabilities.

The variance of 95% of the estimated total number of false negatives depends on the variance and the covariance of the regression coefficients of the logistic model used to estimate the probabilities of false negative results. Therefore, the 95% confidence interval (CI) for the total number of false negatives was estimated on 500 simulations. In each simulation, the entire vector of logistic regression coefficients was randomly sampled from the multivariate normal distribution in which both the mean values and the variance-covariance matrix corresponded to the estimates from the original model. For each simulation, the probability that an inpatient with no hospitalization selected by Algorithm 1 in the national database had an incident cancer was recalculated using the corresponding, randomly sampled vector of regression coefficients, and the resulting estimate of the total number of "false negatives" was obtained as the sum of these probabilities. Finally, the 95% CI for the total number of false negatives was obtained as the interval between the 2.5th and the 97.5th percentile of the distribution of the 500 estimates, each corresponding to one simulation.

A similar procedure was used for false positives, using all patients identified as incident cases by Algorithm 1 in the administrative data. In this second model, the independent variables included along with age and gender geographical area (rural versus urban) and hospital type (public versus private).

The total number of incident colorectal-cancer cases at the national level was estimated by (i) adding the number of patients selected by Algorithm 1 in the national administrative database to (ii) the estimated number of false negatives, and then (iii) subtracting the estimated number of false positives, computed as defined above. The 95% CI for the estimated total number of incident cases was obtained by summing the estimated variances of the last two components of the estimate. Because the proportions of false negatives or positives were very small relative to the national population, the dependence between the three components was negligible, which justifies summing up their respective variances.

To validate this model, colorectal-cancer incidence obtained by applying it to the national database was then compared with the data of 14 registries, which together cover 10.5 million inhabitants or 16.7% of the French population.

The SAS macro that implemented the above procedure is available from the first author upon request.

Table 1: Number of incident cases in Côte d'Or and Doubs estimated by Algorithms 1 and 2.

		Registry	Estimated number of incident cases	
			Administrative data	
			Algorithm 1	Algorithm 2
Côte d'Or	2004	332	313 (94.3%)	457 (137.7%)
	2005	313	333 (106.4%)	465 (148.6%)
Doubs	2005	273	265 (98.2%)	—

*Percentage with regard to the total number of incident cases in the registry.

Table 2: Algorithm 1 results by district and diagnostic year: sensitivity and positive predictive value of administrative data for identifying incident colorectal-cancer cases versus cancer registries used as the gold standard.

District	Year	Incident cases identified by Algorithm 1	Administrative data/registry discordances		Sensitivity (%) (95% CI)	PPV (%) (95% CI)
			False positives	False negatives		
Côte d'Or	2004	313	69	88	73.5 (68.7–78.2)	77.9 (73.3–82.5)
	2005	333	88	68	78.3 (73.7–82.9)	73.6 (68.9–73.3)
Doubs	2005	268	70	75	72.5 (67.2–77.8)	73.9 (68.6–79.2)

3. Results

The Côte d'Or digestive cancer registry identified 332 new colorectal-cancer cases in 2004 and 313 in 2005. The Doubs tumour registry identified 273 new colorectal-cancer cases in 2005. Whatever the year analysed and the district, the estimate using Algorithm 1 was close to the number of incident cases collected by the cancer registries, whereas the estimate using Algorithm 2 overestimated the number of incident cases by almost 50% in 2005 (Table 1).

Tables 2 and 3 show the results of the sensitivity and PPV calculations for administrative data for Algorithm 1 and 2, respectively. Whereas for Algorithm 1, the sensitivity and PPV were very similar (around 75%), for Algorithm 2 the high sensitivity (87.5% in 2005) was counterbalanced by a low PPV (58.9% in 2005).

Concerning patients admitted to hospitals outside their district of residence, Algorithm 1 identified 17 among 354 inpatients (4.8%) from Côte d'Or and 5 among 276 (1.8%) from Doubs.

The results of the explanatory analysis of patients misclassified by the two algorithms were very similar to those obtained previously for breast cancer [16]. Regarding false positives, most were prevalent cases (66%) and the others were mainly related to errors in information collection, namely three-quarters of diagnosis coding errors and one-quarter of erroneous post codes. Among the prevalent cases, the majority (96%) predated our anteriority period of 5 years, whereas the remaining 4% were due to a time gap between diagnosis (year y) and hospitalisation (year $y + 1$), as already mentioned in other studies [26].

False negatives mainly concerned patients who did not receive care during the year of the diagnosis due to a time gap between diagnosis and hospitalisation and patients who were never hospitalised for their cancer. Coding errors also explained a part of the false positives.

The results of the logistic regression (AUC = 0.604) are given in Table 4.

Among the four independent variables of the model of false positives: "age," "gender," "geographical area" and "hospital type," the latter three had no significant effect on the appearance of false positives. However, old age and male gender seem to affect the proportion of false negatives.

By applying the models, there were an estimated 10884.02 false positives (95% confidence interval: 9542.37, 12616.12) and 8885.07 false negatives (95% confidence interval: 7687.90, 10554.48).

Finally, the national estimation of colorectal-cancer incidence in France in 2005 was $41121 - 10884 + 8885 = 39122$, (95% confidence interval: 37020, 41224). The comparison between these results and registry data, when available, is shown in Table 5. The final discrepancy was only 2.34%.

4. Discussion

Algorithm 1 provided an estimated incidence close to those given by registries. Indeed, for the period of 2004 and 2005, the summed number of incident cases detected by this algorithm in Côte d'Or was 646 (268 for 2005 in Doubs), while the true numbers observed by registry were 645 and 273, respectively. The sensitivity and the positive predictive values of Algorithm 1 are also quite good (about 75% for both).

The fact that the number of false positives is greater when previous years (Algorithm 2) are taken into account seems quite surprising, at least at the first glance. Indeed, one would have expected that this method would be better at detecting prevalent cases and, thus would have given a more precise estimate. However, the validation study conducted on the corresponding medical records showed not only that most (66%) of the false positives were prevalent cases but,

Table 3: Algorithm 2 results in Côte d'Or by diagnostic year: sensitivity and positive predictive value of administrative data for identifying incident colorectal-cancer cases versus cancer registries used as the gold standard.

Year	Incident cases identified by Algorithm 2	Administrative data/registry discordances		Sensitivity (%) (95%CI)	PPV (%) (95%CI)
		False positives	False negatives		
2004	457	180	55	83.4 (79.4–87.4)	60.6 (56.1–65.1)
2005	465	191	39	87.5 (83.8–91.2)	58.9 (54.4–63.4)

Table 4: Coefficients and standard errors of the predictive models estimated from the regional data (Côte d'Or and Doubs).

Parameter			Beta	SE	Khi2	*P* value
False positive model	Intercept		0.7191	0.4999	2.0697	0.1502
	Age		−0.0253	0.00719	12.3965	0.0004
False negative model	Intercept		−9.6632	0.1530	3989.3101	<.0001
	Age	≥75	2.4619	0.1689	212.4791	<.0001
	Gender	Male	0.3718	0.1690	4.8400	0.0278

above all, that the vast majority of these prevalent cases (96%) predated our anteriority period of 5 years. In other words, most of "false positives" were already prevalent in 1999, which was as far back as we could go. Under these circumstances, Algorithm 1, which was exclusively based on diagnosis and procedure codes and did *not* take into account the timing of the events, was not affected by this issue and performed better than Algorithm 2, which overestimated the number of incident cases by almost 50%.

Another way to explain the discrepancy between the two algorithms is that, although the sensitivity of Algorithm 1 was lower than that of Algorithm 2, its PPV was higher, leading to balanced false negatives and false positives that cancelled each other out. Indeed, the decisive date recorded in the registries for incident cases was the date of diagnosis, whereas the only date that was relevant for our purposes in the administrative data was the date of admission. In colorectal-cancer, admission for treatment can occur sometime after the histologically confirmed diagnosis. However, false negatives, missed by Algorithm 1 because of a diagnosis date in year "Y" (registry data), but treated in "$Y + 1$" (administrative data) are balanced by the false positives treated in "Y" (administrative data) but diagnosed, in "$Y-1$" (registry data).

Concerning the national estimate, Algorithm 1 overestimated cancer incidence by only 2.34% compared with the summed data of the 14 registries, after correction of the results by our models. These good results can be contrasted with the underestimation of incident cases observed in a previous study for colorectal-cancer (642 rather than the 799 incident cases recorded in a registry) [6].

The discrepancy in the treatment (surgery versus chemotherapy and/or radiotherapy) would mainly affect the performance of Algorithm 1, as it would miss patients not treated with surgery. However, these cases are relatively rare, as more than 90% of cancer patients are treated with surgery and/or endoscopic resection (both included in Algorithm 1). Questions could be raised about colonoscopy because if this examination was not performed under general anaesthesia, there would have been no admission and the patient would therefore have been missed by the algorithm. However, endoscopic resection without general anaesthesia is tending to disappear. Though it was still the case for about 5–7% of the patients in 2004, nowadays, almost all patients receive general anaesthesia and can thus be detected by the algorithm.

The impact of old age and male gender on the proportion of false negatives could be explained by the fact that older patients are less willing to accept surgical treatment, as it involves a quite burdensome hospitalization, and that the rejection of any aggressive therapy is classically more common among men. The clinical pathway and care sequences may also have had an impact on the proportions of false positives and false negatives. Unfortunately, it was not feasible to analyse this hypothesis during the present study as the relevant information was not recorded in the studied data. However, we are currently working on a study of the patients' pathways using the French health insurance claims database, but due to a technical limitation (data anonymization of the insurance claim database), will not be possible to link insurance data with registry data, and the impact of the patients' pathway will be assessed using other appropriate methods.

In France in 2004, the global endowment system was replaced by a system in which remuneration is calculated on the basis of Price per Activity. Since then, the quality of national administrative data has greatly improved and one could expect that future studies on the same subject but carried out on recent administrative data will not generate the same results. However, there is a delay of about 3 or 4 years before registry data become available. Under these circumstances, we have no choice but to work on 9-year-old data in order to have an anteriority period of 5 years, and future studies as mentioned above will not be feasible for many years.

5. Conclusion

This study shows the usefulness of administrative databases and suggests a method to correct the estimates of cancer

Table 5: Colorectal-cancer incidence comparison between results of predictive model and registry data.

District	Registry incidence (1)	Incidence estimated by Algorithm 1 (2)	Incidence estimated by the model (3)	(1) − (2)	(1) − (3)
Bas-Rhin	625	681	604	8.96%	−3.36%
Haut-Rhin	448	388	359	−13.39%	−19.87%
Calvados	336	336	340	0.00%	1.19%
Manche	304	331	322	8.88%	5.92%
Côte d'Or	350	351	334	0.29%	−4.57%
Saône et Loire	402	452	424	12.44%	5.47%
Finistère	765	802	726	4.84%	−5.10%
Doubs	283	258	258	−8.83%	−8.83%
Hérault	655	714	675	9.01%	3.05%
Tarn	304	334	311	9.87%	2.30%
Loire-Atlantique	688	804	757	16.86%	10.03%
Vendée	367	448	421	22.07%	14.71%
Somme	309	382	358	23.62%	15.86%
Isère	564	722	680	28.01%	20.57%
Total	6400	7003	6569	9.42%	2.34%

incidence provided by these data. Detecting incident cases using a mix of diagnosis and procedure codes specific to new cases of cancer appears to be an efficient and reliable way to estimate incidence rates from one year's worth of data in the absence of long-term patient history. Furthermore, even when a patient's history is retrievable, our results showed that this detection method still performs better than one based on the timing of the events.

This method may also be useful for many countries in which claims data are gathered and where no national cancer registries exist. In addition, as administrative data are generally available quickly (less than six months in France), a system derived from our method could operate in almost real-time, while processing registry data currently takes much longer. For instance, such a system could be implemented to automatically estimate the number of new cases of cancer in the population of a specific geographical area in order to optimize the organization of health care in that area.

Of course, although the risk of underestimating the incidence of low-mortality cancers, such as colorectal-cancer, primarily motivated our decision to rely on morbidity data, the method presented here is suitable for high-mortality cancers as well.

However, incidence is not the only key statistic, and beyond estimating incidence, our method is of little use. Indeed, Algorithm 1 proposed in this study is useful for counting incident cases only because the false negatives and false positives tend to have similar frequencies and, thus, to cancel each other out. Some of individual patients identified through our method may not necessarily have the cancer, and some actual cancer patients may escape detection. Therefore, Algorithm 1 is unable to accurately identify cases and cannot be used in longitudinal studies. In addition, administrative data do not provide any information concerning the tumor stage, grade, or localization. Therefore, registries remain essential to study prognostic factors and to compare cancer care management in different facilities.

Abbreviations

ICD-10: International classification of disease, tenth revision
PPV: positive predictive value
DRG: Diagnosis-related group
CCAM: French classification of procedures (classification commune des actes médicaux)
HCPS: Healthcare common procedure coding system
CPT: Current procedural terminology.

Acknowledgments

This work was supported by the "Institut de Recherches et d'Expertises en Sécurité Publique (IRESP)." The authors wish to thank, for their help, Patrick Arveux (CGFL), Claude Klepping (Clinique de Chenôve), Sylvie Grosjean (Hôpital de Semur-en-Auxois et de Saulieu), Michel Roux (Hôpital de Beaune), Jean-Claude Naudin (Hôpital Chatillon-Montbard), Pascaline Bataillon-Charles (Clinique de Fontaine et de Sainte Marthe), Stéphanie Gathion (Clinique Drevon), Claude Petit-Marnier (Clinique de Talant), Annie Billod-Girard (Centre Hospitalier de Belfort-Montbeliard), Jacques-Henri Bauer (Polyclinique de Franche-Comté, Clinique des Portes du Jura), Jean-François Viel (Centre Hospitalier Universitaire de Besançon), Vincent Provitolo (Clinique St. Vincent), Thierry Dispot (Clinique de Laennec), and Jean Rudloft (Centre Hospitalier de Pontarlier). We are grateful to Anne-Marie Bouvier (Registre des Cancers Digestifs de Côte d'Or) and Arlette Danzon

(Registre des Tumeurs du Doubs) as well as the teams of the two registries. The authors wish to thank Philip Bastable for comments on the paper. They also wish to thank the Francim Network for its help.

References

[1] F. Ederer, M. S. Geisser, S. J. Mongin, T. R. Church, and J. S. Mandel, "Colorectal cancer deaths as determined by expert committee and from death certificate: a comparison. The Minnesota study," *Journal of Clinical Epidemiology*, vol. 52, no. 5, pp. 447–452, 1999.

[2] R. R. German, A. K. Fink, M. Heron et al., "The accuracy of cancer mortality statistics based on death certificates in the United States," *Cancer Epidemiology*, vol. 35, no. 2, pp. 126–131, 2011.

[3] C. L. Percy, B. A. Miller, and L. A. Gloeckler Ries, "Effect of changes in cancer classification and the accurcy of cancer death certificates on trends in cancer mortality," *Annals of the New York Academy of Sciences*, vol. 609, pp. 87–97, 1990.

[4] 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.

[5] J. Powell, "Cancer registration: principles and methods. Data sources and reporting," *IARC Scientific Publications*, no. 95, pp. 29–42, 1991.

[6] I. Baldi, P. Vicari, D. Di Cuonzo et al., "A high positive predictive value algorithm using hospital administrative data identified incident cancer cases," *Journal of Clinical Epidemiology*, vol. 61, no. 4, pp. 373–379, 2008.

[7] N. Carré, Z. Uhry, M. Velten et al., "Predictive value and sensibility of hospital discharge system (PMSI) compared to cancer registries for thyroïd cancer (1999–2000)," *Revue d'Epidemiologie et de Sante Publique*, vol. 54, no. 4, pp. 367–376, 2006.

[8] 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.

[9] C. M. Couris, S. Polazzi, F. Olive et al., "Breast cancer incidence using administrative data: correction with sensitivity and specificity," *Journal of Clinical Epidemiology*, vol. 62, no. 6, pp. 660–666, 2009.

[10] O. Ganry, A. Taleb, J. Peng, N. Raverdy, and A. Dubreuil, "Evaluation of an algorithm to identify incident breast cancer cases using DRGs data," *European Journal of Cancer Prevention*, vol. 12, no. 4, pp. 295–299, 2003.

[11] L. Penberthy, D. McClish, A. Pugh, W. Smith, C. Manning, and S. Retchin, "Using hospital discharge files to enhance cancer surveillance," *American Journal of Epidemiology*, vol. 158, no. 1, pp. 27–34, 2003.

[12] 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.

[13] "From case mix data bases to health geography," in *Proceedings of the 19th International PCS/E Working Conference*, B. Trombert, C. Martin, and P. Vercherin, Eds., Washington, DC, USA, 2003.

[14] 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.

[15] M. Goldberg, E. Jougla, M. Fassa, R. Padieu, and C. Quantin, "The French health information system," *Journal of the International Association for Official Statistics*, vol. 28, no. 1-2, pp. 31–41, 2012.

[16] C. Quantin, E. Benzenine, M. Fassa et al., "Evaluation of the interest of using discharge abstract databases to estimate breast cancer incidence in two French departments," *Journal of the International Association for Official Statistics*, vol. 28, no. 1-2, pp. 73–85, 2012.

[17] A. M. McBean, J. D. Babish, and J. L. Warren, "Determination of lung cancer incidence in the elderly using Medicare claims data," *American Journal of Epidemiology*, vol. 137, no. 2, pp. 226–234, 1993.

[18] D. K. McClish, L. Penberthy, M. Whittemore et al., "Ability of medicare claims data and cancer registries to identify cancer cases and treatment," *American Journal of Epidemiology*, vol. 145, no. 3, pp. 227–233, 1997.

[19] Z. Hafdi-Nejjari, C. M. Couris, A. M. Schot et al., "Role of hospital claims databases from care units for estimating thyroid cancer incidence in the Rhône-Alpes region of France," *Revue d'Epidemiologie et de Sante Publique*, vol. 54, no. 5, pp. 391–398, 2006.

[20] S. M. Koroukian, G. S. Cooper, and A. A. Rimm, "Ability of Medicaid claims data to identify incident cases of breast cancer in the Ohio Medicaid population," *Health Services Research*, vol. 38, no. 3, pp. 947–960, 2003.

[21] K. M. Leung, A. G. Hasan, K. S. Rees, R. G. Parker, and A. P. Legorreta, "Patients with newly diagnosed carcinoma of the breast: validation of a claim-based identification algorithm," *Journal of Clinical Epidemiology*, vol. 52, no. 1, pp. 57–64, 1999.

[22] C. M. Couris, A. Seigneurin, S. Bouzbid et al., "French claims data as a source of information to describe cancer incidence: predictive values of two identification methods of incident prostate cancers," *Journal of Medical Systems*, vol. 30, no. 6, pp. 459–463, 2006.

[23] C. W. Ko, J. A. Dominitz, P. Green, W. Kreuter, and L. M. Baldwin, "Accuracy of Medicare claims for identifying findings and procedures performed during colonoscopy," *Gastrointestinal Endoscopy*, vol. 73, no. 3, pp. 447–453.e1, 2011.

[24] SNFGE, "National Thesaurus of digestive oncology," 2008, http://www.snfge.asso.fr/01-bibliotheque/0g-thesaurus-cancerologie/publication5/sommaire-thesaurus.asp.

[25] C. Quantin, H. Bouzelat, F. A. Allaert, A. M. Benhamiche, J. Faivre, and L. Dusserre, "Automatic record hash coding and linkage for epidemiological follow-up data confidentiality," *Methods of Information in Medicine*, vol. 37, no. 3, pp. 271–277, 1998.

[26] F. Olive, F. Gomez, A. M. Schott et al., "Critical analysis of French DRG based information system (PMSI) databases for the epidemiology of cancer: a longitudinal approach becomes possible," *Revue d'Epidemiologie et de Sante Publique*, vol. 59, no. 1, pp. 53–58, 2011.