ÉVALUATION DU PMSI COMME MOYEN D'IDENTIFICATION DES CAS INCIDENTS DE CANCER COLORECTAL

Catherine Quantin *et al.*

S.F.S.P. | *Santé Publique*

2014/1 - Vol. 26
pages 55 à 63

ISSN 0995-3914

Évaluation du PMSI comme moyen d'identification des cas incidents de cancer colorectal

Evaluation of medical information systems as a mean of identification of incident cases of colorectal cancer

Catherine Quantin[1, 2], Éric Benzenine[1], Mathieu Hägi[1], Bertrand Auverlot[1], Michal Abrahamowicz[3], Jonathan Cottenet[1], Évelyne Fournier[4], Christine Binquet[1, 2], Delphine Compain[1], Élisabeth Monnet[5], Anne-Marie Bouvier[2, 6], Arlette Danzon[4]

Résumé

Contexte : pour estimer l'incidence nationale d'un cancer, les registres sont une source fiable de données mais celles-ci ne sont pas toujours disponibles sur tous les territoires. Nous avons voulu évaluer l'intérêt du programme de médicalisation des systèmes d'information (PMSI) pour l'identification des cas incidents de cancers colorectaux.

Méthode : afin de retrouver ces cas incidents dans la base PMSI, nous avons élaboré deux algorithmes. Le premier se base sur les codes diagnostiques et actes médicaux, le second uniquement sur les diagnostics et leur absence au cours des cinq dernières années. Les résultats obtenus sur deux départements ont été confrontés à ceux de deux registres, constituant la référence. Nous avons ensuite élaboré deux modèles de régression logistique multi-variée permettant de corriger le nombre de cas incidents estimé au niveau national par l'algorithme retenu après évaluation des résultats.

Résultats : le premier algorithme a fourni des résultats très proches de ceux observés au niveau des registres (646 *vs* 645 cas), avec une bonne sensibilité et valeur prédictive positive de 75 %. Le second surestime l'incidence (≈ 50 %), avec une valeur prédictive positive de 60 % et n'a donc pas été retenu pour l'estimation nationale. En appliquant le premier algorithme à la base nationale du PMSI MCO (médecine, chirurgie, obstétrique), et après modélisation, l'incidence estimée ne diffère que de 2,34 % par rapport à celle observée par l'ensemble de 14 registres. L'estimation au niveau national est de 39 122[37 020 ; 41 224] cas pour l'année 2005 et est cohérente avec celle publiée par le réseau national des registres, Francim (37 413).

Conclusion : cette étude montre l'utilité des données PMSI pour l'estimation de l'incidence nationale de certains cancers, en l'absence de registres du cancer. Une correction des résultats bruts reste cependant nécessaire, et la méthode ici proposée permet d'y parvenir.

Mots-clés : Systèmes informatisés de dossiers médicaux ; Tumeurs colorectales ; Incidence ; Recherche en efficacité comparative ; Couplage des dossiers médicaux.

Summary

Background: *Cancer registries are a reliable source of data to estimate national cancer incidence rates, but they are not always available in all regions. This study assessed the value of medical information systems (PMSI) to identify incident cases of colorectal cancer.*

Methods: *Two algorithms were elaborated to identify these incident cases in the PMSI database. The first algorithm was based on diagnosis and medical procedure codes and the second algorithm was based exclusively on diagnoses and the absence of diagnoses over the last five years. The results obtained for two departments were compared with those derived from two cancer registries, constituting the reference data. We then elaborated two multivariate logistic regression models to correct the national number of incident cases estimated by the algorithm adopted after evaluation of the results.*

Results: *The first algorithm provided results that were very close to those derived from the regional registries (646 vs 645 cases) with a good sensitivity and positive predictive value of 75%. The second algorithm overestimated the incidence by about 50% with a positive predictive value of 60% and was therefore not adopted for the national estimation. By applying the first algorithm to the national PMSI MCO database (medicine, surgery, obstetrics), and after modelling, the estimated incidence differed by only 2.34% compared to that observed by all 14 registries. The national estimation of cancer incidence was 39,122 [37,020; 41,224] cases for 2005, which is consistent with the figure published by the Francim national registry network (37,413).*

Conclusion: *This study demonstrates the value of PMSI data for estimation of national incidence rates for certain cancers in the absence of cancer registries. However, raw data must be corrected and can be achieved by the method proposed here.*

Keywords : *Medical Records Systems, Computerized; Colorectal Neoplasms; Incidence; Comparative Effectiveness Research; Medical Record Linkage.*

[1] CHRU Dijon – Service de Biostatistique et d'Informatique Médicale (DIM) – BP 77908 – 21079 Dijon – France.
[2] Inserm, U866 – Université de Bourgogne – 21000 Dijon – France.
[3] Department of Epidemiology and Biostatistics – McGill University – Montreal – Quebec – Canada H3A 1A2.
[4] Registre des tumeurs du Doubs EA 3181 – Université 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, 21000 Dijon – France.

Correspondance : C. Quantin
catherine.quantin@chu-dijon.fr

Réception : 31/05/2013 – *Acceptation :* 02/01/2014

Introduction

Les données de morbidité fournies par les registres et les données de mortalité sont très utilisées pour établir des estimations d'incidence du cancer au niveau national. Ainsi les registres de population fournissent, entre autres, des informations sur l'incidence du cancer et sur ses variations temporelles avec la mise à jour régulière de statistiques internationales. Toutefois, l'extrapolation des résultats obtenus au niveau national est parfois limitée par l'absence de couverture de l'ensemble de la population française par les registres.

Plusieurs auteurs se sont intéressés à la question de l'utilisation des données du Programme de médicalisation des systèmes d'information (PMSI) pour approcher la mesure des cas incidents ou les repérer [1-10]. En effet, l'intérêt du PMSI est de fournir des bases de données structurées et codées de manière standardisée et exploitables au niveau national. Un résumé est produit à la sortie de chaque séjour à l'hôpital. Ce résumé rassemble les informations recueillies au cours des passages dans les différentes unités d'hospitalisation. Il comporte des données médicales : dont les diagnostics principaux et l'ensemble des diagnostics associés, ainsi que la liste des actes pratiqués. Les diagnostics sont codés selon la dixième révision de la classification internationale des maladies (Cim-10), et les actes selon la classification commune des actes médicaux (CCAM) ou le catalogue des actes médicaux (CDAM) en fonction de la période étudiée. L'utilisation de ces données et leur contrôle par l'Assurance maladie pour l'allocation budgétaire des établissements de santé constituent un incitatif important en faveur d'une bonne exhaustivité (de l'ordre de 100 %) et qualité.

Par ailleurs, la population française des patients atteints de cancer est bien représentée dans les bases de données PMSI, notamment dans sa partie MCO (médecine, chirurgie, obstétrique). En effet la majorité des moyens thérapeutiques utilisés contre le cancer nécessitent une hospitalisation. Si certaines personnes sont susceptibles d'échapper au PMSI parce qu'elles sont suivies uniquement en consultation, cela concerne très peu le cancer colorectal. Ainsi de nombreux registres considèrent le PMSI parmi leurs sources de notifications

L'objectif de ce travail est d'évaluer l'intérêt du PMSI MCO comme moyen d'identification des cas incidents de cancer. Pour ce faire, la confrontation des données du PMSI à celles de deux registres, constituant la référence, a été réalisée à l'échelle de deux départements, la Côte d'Or et le Doubs, selon une méthode que nous avons développée [4, 11, 12].

Méthodes

Nous avons défini deux algorithmes différents (cf. infra méthode d'identification des cas) afin d'identifier les nouveaux cas de cancers dans la base PMSI. Puis nous les avons appliqués aux données PMSI des patients domiciliés en Côte d'Or (respectivement dans le Doubs) et pris en charge dans un des établissements de ce département. Les 18 établissements hospitaliers, publics comme privés, de ces deux départements (onze en Côte d'Or et sept dans le Doubs), ont accepté de nous fournir les données PMSI des années 1999 à 2005. Après avoir comparé les résultats des deux algorithmes à ceux fournis par les registres, nous avons étudié les cas discordants par retour au dossier d'hospitalisation du patient.

Nous avons mis en œuvre une recherche des facteurs expliquant les discordances, à l'aide une régression logistique. Nous avons ensuite appliqué les paramètres obtenus aux données nationales PMSI afin d'obtenir une estimation nationale corrigée du nombre total de cas incidents de cancer colorectal.

L'efficacité globale de cette méthode a finalement été vérifiée en confrontant nos résultats obtenus aux nombres de cas observés sur les 14 départements couverts par les registres, puis à l'estimation nationale réalisée par Francim.

Méthode d'identification des cas incidents dans la base PMSI

Pour repérer un « cas incident », il convient que le diagnostic de cancer soit retrouvé dans l'information relative au patient. Pour vérifier qu'il s'agit d'un « nouveau » cas, deux approches sont fréquemment retrouvées dans la littérature [1, 3, 7, 9, 13-18]. La première se base sur la nécessité de retrouver un acte spécifique, propre à la prise en charge d'une première occurrence de la maladie. La seconde se base sur l'absence de diagnostic précédent au cours d'une période antérieure. Afin de déterminer laquelle de ces deux approches serait la plus adaptée à notre problématique, nous avons choisi, de développer pour chacune d'elle un algorithme de sélection des cas.

Le premier algorithme est essentiellement construit sur les codes des diagnostics (Cim-10) et des actes (CCAM et CDAM). Il considère comme cas incident tout patient hospitalisé présentant à la fois un diagnostic principal (DP) de cancer colorectal (Cim-10 C18 à C20) et un code d'acte en relation avec le traitement initial : « endoscopie

colorectale », « exérèse partielle ou totale du colon ou du rectum », « excision, exérèse ou destruction de polypes ou de tumeurs du colon ou du rectum », « réfection ou fermeture de colostomie », « rétablissement secondaire de la continuité » et « pose d'une endoprothèse du côlon ». Nous avons choisi d'utiliser des codes essentiellement chirurgicaux et d'écarter la chimiothérapie et la radiothérapie, qui ne sont pas uniquement utilisées en cas de traitement initial. D'autre part, lorsque c'est le cas, elles sont quasiment toujours associées à la chirurgie [19]. Quand plusieurs hospitalisations survenaient pour un même patient au cours d'une même année, seule la première était sélectionnée afin de ne repérer que des cas incidents.

Le second algorithme se base sur les mêmes codes de diagnostics que l'algorithme 1 mais, c'est sur l'histoire du patient que cet algorithme repose pour écarter les cas prévalents. Seuls ont été retenus les patient pour lesquels un diagnostic principal ou associé (DAS) de cancer colorectal (codés C18 à C20) était retrouvé dans un résumé du PMSI en 2004-2005, sans qu'aucun autre ne le soit sur une période précédente de cinq ans, ce qui était le maximum de recul dont nous disposions. Les données antérieures à 2005 n'étant pas disponibles dans tous les établissements du Doubs, seul l'algorithme 1 a pu être appliqué dans ce département, et uniquement sur l'année 2005.

Afin d'évaluer les fuites, c'est-à-dire d'obtenir des données relatives aux habitants de la Côte d'Or et du Doubs, hospitalisés dans un autre département, nous avons appliqué l'algorithme 1 à toutes les hospitalisations de la base nationale PMSI MCO en nous limitant dans un premier temps aux patients domiciliés en Côte d'Or et dans un second temps à ceux domiciliés dans le Doubs.

Validation de la méthode d'identification

Cette validation s'est déroulée en confrontant les résultats obtenus à partir de l'application de nos algorithmes (1 et 2) sur les données PMSI aux données de références mesurées dans les registres, par chaînage des deux bases. Nous avons développé cette méthodologie de confrontation des données PMSI aux données des registres par anonymisation et chaînage [12]. En accord avec la Cnil (Autorisation n° 1220801 du 27/03/2007), nous avons appliqué aux données des registres, la même méthode d'anonymisation (logiciel ANONYMAT [12], basé sur des techniques de hachage) qu'aux données PMSI des établissements hospitaliers.

Tout cas repéré dans le PMSI comme potentiellement incident par un algorithme mais non retrouvé comme tel dans les données des registres a été considéré comme un faux positif (Fp) dudit algorithme (cf. figure 1, partie confrontation PMSI/Registre au niveau des départements). À l'inverse, un cas avéré par les registres mais non retrouvé par un algorithme a été classé comme faux négatif (Fn). Nous avons pu ainsi déterminer la sensibilité et la valeur prédictive positive (Vpp) de chaque algorithme. Afin de déterminer les causes d'erreur de chaque algorithme, une étude approfondie des cas discordants a été menée, grâce aux données des registres pour l'étude des faux négatifs et à un retour aux dossiers des patients pour les faux positifs.

Estimation nationale de l'incidence

L'étude de validation ayant mis en exergue la meilleure performance du premier algorithme par rapport au second (cf. résultats), c'est celui-ci que nous avons choisi d'utiliser pour l'estimation nationale. Le postulat de départ était que la qualité des données du PMSI pouvait varier non seulement en fonction des caractéristiques du patient, mais également en fonction de la zone géographique. Cette variation de qualité devait être prise en compte pour corriger l'estimation fournie par l'algorithme 1. Pour le vérifier, nous avons mis en œuvre deux modèles de régression logistique multi-variée, l'un permettant d'estimer la part des Fp dans les résultats, l'autre celle des Fn. Ces deux modèles ont tout d'abord été appliqués aux données collectées dans les deux départements, afin de rechercher les variables pouvant être associées à une augmentation de la proportion des Fp ou des Fn.

Le premier modèle, utilisé pour sélectionner les variables associées aux Fp, a été construit à partir des cas détectés comme potentiellement incidents dans la base PMSI, avec une variable à expliquer dichotomique : '0' représentant les cas validés par les registres (vrais positifs, Vp) et '1' les cas faux positifs. Les variables indépendantes testées étaient l'âge (considérée comme une variable quantitative continue), le sexe, la zone géographique (rurale *vs* urbaine) et le type d'établissement (public *vs* privé).

Dans le second modèle, recherchant les variables associées aux faux négatifs, et appliqué aux cas non retrouvés dans la base PMSI, la variable à expliquer valait '0' pour les vrais négatifs (absents également des registres) et '1' pour les faux négatifs. L'ensemble des vrais négatifs a été construit à partir des données Insee (Côte d'Or et Doubs) auxquelles ont été retranchés les Vp, Fp et Fn. Dans ce modèle, seuls l'âge (considérée comme une variable dichotomique : inférieur ou supérieur à 75 ans) et le sexe ont été utilisés, puisque s'agissant de cas négatifs, aucune trace d'admission ne pouvait être retrouvée dans les données du

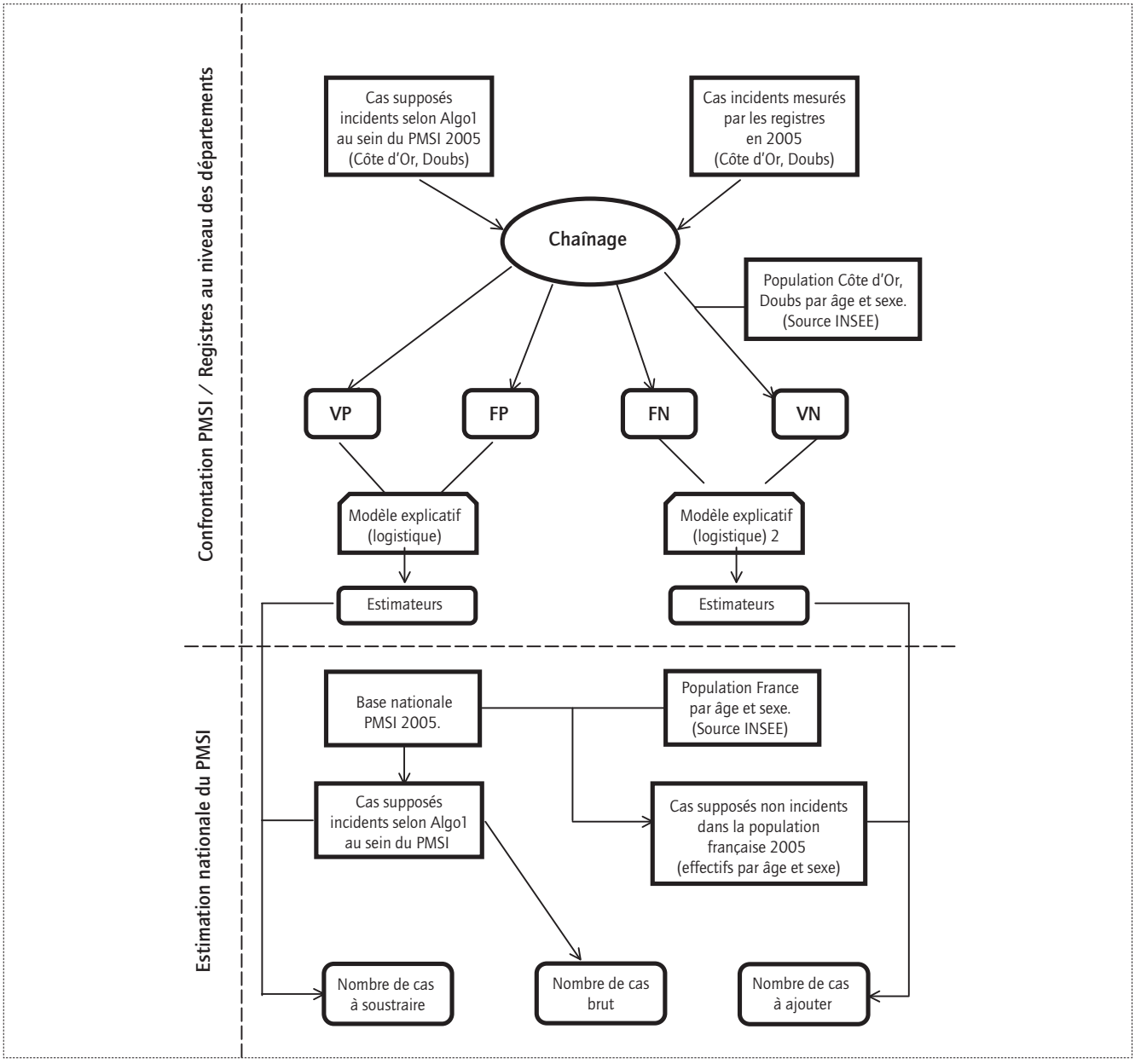

Figure 1 : Principes de la méthode utilisée

PMSI, et donc les variables liées à l'établissement étaient non disponibles. Pour chacun de ces cas négatifs, nous avons calculé la probabilité que le patient soit en réalité un cas incident, en fonction de chacune des variables du modèle. En additionnant toutes ces probabilités individuelles, nous avons ainsi estimé le nombre total de cas incidents non identifiés par l'algorithme 1 (faux négatifs).

Pour estimer l'intervalle de confiance de ce nombre, 500 simulations (selon la loi normale multi-variée) ont été réalisées à partir du vecteur des paramètres et de la matrice de variance covariance obtenue par le modèle de régression logistique. Les bornes de l'intervalle de confiance ont été définies à partir des valeurs correspondant aux 2,5 et 97,5 percentiles de cette distribution. La même procédure a été utilisée pour les faux positifs, en étudiant cette fois les cas identifiés comme incidents par l'algorithme 1.

L'estimation du nombre total de cas de cancers colorectaux incidents a été obtenue au niveau national en

additionnant (i) le nombre total de cas sélectionnés par l'algorithme 1 à (ii) l'estimation du nombre total de faux négatifs fournie par le second modèle, puis en y soustrayant (iii) l'estimation du nombre total de faux positifs fournie par le premier modèle (cf. figure 1, partie évaluation nationale). L'intervalle de confiance à 95 % de cette estimation a été obtenu en additionnant les variances estimées de (ii) et (iii). Afin de valider cette méthodologie, notre estimation des cas incidents *via* la base PMSI nationale a été confrontée aux données de 14 registres, qui couvrent 10,5 millions de personnes, soit près de 16,7 % de la population française.

La macro SAS utilisée dans la procédure décrite ci-dessus est disponible auprès du premier auteur par simple demande.

Résultats

Comparaison des mesures d'incidences entre le PMSI et les registres

Le registre de la Côte d'Or recensait 332 nouveaux cas de cancers colorectaux pour 2004 et 313 pour 2005. Celui du Doubs en identifiait 273 pour 2005. Quelle que soit l'année prise en compte, les estimations fournies par l'algorithme 1 étaient toutes très proches des mesures des registres, alors que l'algorithme 2 surestimait invariablement les résultats, jusqu'à près de 50 % pour 2005 (tableau I). Les tableaux II et III présentent les valeurs de sensibilité et de Vpp calculées pour chaque algorithme. Alors que la sensibilité et la Vpp de l'algorithme 1 étaient similaires (autour de 75 %), une sensibilité plus élevée était obtenue pour l'algorithme 2 (87,5 % pour 2005) au détriment de la Vpp (58,9 % pour 2005).

Concernant la détection des patients admis dans un autre département que celui d'origine (fuites), l'algorithme 1 a identifié 17 migrants sur les 354 patients de la Côte d'Or (4,8 %) et cinq sur les 276 du Doubs (1,8 %).

Les résultats de l'étude de validation sont cohérents avec ceux obtenus lors d'une étude précédente sur le cancer du sein [11].

Concernant les faux positifs, la majorité d'entre eux s'est avérée correspondre à des cas prévalents (66 %). Le reste s'explique par des erreurs dans la collecte de l'information : il s'agissait pour les trois-quarts d'erreurs dans le code du diagnostic et pour le quart restant de codes postaux erronés. Parmi les cas prévalents, 96 % avaient été diagnostiqués plus de cinq ans auparavant et n'avaient donc pas pu être éliminés par la recherche de diagnostics dans les résumés PMSI antérieurs (avec un recul de cinq ans). Les 4 % restants étaient dus à un décalage entre l'année du diagnostic (N-1), et l'année de l'hospitalisation (N), phénomène par ailleurs décrit dans d'autres études [20].

Concernant les faux négatifs, il s'agissait essentiellement de patients n'ayant pas été hospitalisés au cours de l'année du diagnostic (N), (par exemple l'année civile suivante (N+1) pour les cas diagnostiqués en fin d'année). Plus rarement, ces faux négatifs concernaient des patients n'ayant jamais été hospitalisés pour leur cancer, ou à des erreurs de codage.

Les résultats de la régression logistique (AUC = 0,604) sont les suivants : parmi les variables incluses dans les modèles, seul l'âge était associé significativement à une diminution du taux de faux positifs, alors que l'âge avancé et le fait d'être un homme étaient associés à une augmentation du taux de faux négatifs.

Estimations départementales et nationale corrigées

Pour ce qui concerne l'estimation nationale, après application des deux modèles correctifs des résultats de l'algorithme 1, le nombre total de faux positifs a été estimé à 10 884 [9 542 ; 12 616] et celui de faux négatifs à 8 885 [7 687 ; 10 554], ce qui conduit à une estimation finale du nombre de cas incidents de cancers colorectaux de (41 121 + 8 885 - 10 884) = 39 122 [37 020 ; 41 224]. Cette estimation est cohérente avec celle publiée par Francim (37 413). D'autre part, le résultat de la modélisation s'est avéré différer de 2,34 % par rapport aux données agrégées des 14 registres (tableau IV) [4, 7, 10, 21-23].

Tableau I : Nombre de cas incidents de la Côte d'Or et du Doubs estimés par les algorithmes 1 & 2 *vs* données observées par les registres

		Nombre estimé de cas incidents		
		Registre	PMSI[a]	
			Algorithme 1	Algorithme 2
Côte d'Or	2004	332	313 (94,3 %)[b]	457 (137,7 %)
	2005	313	333 (106,4 %)	465 (148,6 %)
Doubs	2005	273	265 (98,2 %)	–

[a] : programme de médicalisation des systèmes d'information.
[b] : pourcentage par rapport au nombre de cas enregistrés dans le registre.

Tableau II : Résultats de l'algorithme 1, par département et par année de diagnostic : sensibilité et valeur prédictive positive des données du PMSI, données des registres prises comme référence

Département	Année	Cas identifiés comme incidents par Algorithme 1	Discordance PMSI[a]/Registre		Sensibilité (%) [IC 95 %]	Vpp[b] (%) [IC 95 %]
			Faux positifs	Faux négatifs		
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]

[a] : programme de médicalisation des systèmes d'information.
[b] : valeur prédictive positive.

Tableau III : Résultats de l'algorithme 2 en Côte d'Or, par année de diagnostic : sensibilité et valeur prédictive positive des données du PMSI, données des registres prises comme référence

Année	Cas identifiés comme incidents par Algorithme 2	Discordance PMSI[a]/Registre		Sensibilité (%) [IC 95 %]	Vpp[b] (%) [IC 95 %]
		Faux positifs	Faux négatifs		
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]

[a] : programme de médicalisation des systèmes d'information.
[b] : valeur prédictive positive.

Discussion

Au final, l'algorithme 1, basé sur les diagnostics en DP et les actes du PMSI MCO, et présentant de bonnes sensibilités et Vpp (proches de 75 %), s'est montré plus efficace que l'algorithme 2. Il a conduit à une meilleure estimation du nombre de cas incidents, voisin du nombre mesuré par les registres. L'algorithme 2, basé sur les diagnostics (DP et DAS) et considérant les hospitalisations antérieures pour qualifier un cas de prévalent ou d'incident, s'est montré plus sensible. Cependant, sa Vpp étant moins bonne, il a conduit à une surestimation de l'ordre de 50 % du nombre de cas incidents. Nous espérions que les cas prévalents soient mieux détectés par cet algorithme. L'antériorité considérée était de cinq ans, car il ne nous a pas été possible d'obtenir les données PMSI antérieures à 2009. Cependant, le retour au dossier médical a montré que la plupart des cas prévalents (96 %) l'étaient depuis bien plus longtemps.

Une autre explication de la supériorité de l'algorithme 1 est une compensation mutuelle des faux négatifs et des faux positifs. Cette compensation provient en partie du fait que les faux négatifs (cas manqués à cause d'un diagnostic année N et d'une admission pour traitement année N+1) sont contrebalancés par les faux positifs (cas faussement identifiés par le biais d'une admission année N, mais en fait diagnostiqués l'année N-1). En effet, l'hospitalisation pour traitement d'un cancer colorectal peut intervenir après la confirmation histologique de la maladie notamment si celle-ci intervient en fin d'année civile, amenant à démarrer la prise en charge dans une année civile différente.

Au final, après correction par nos modèles de régression logistique multi-variée, l'algorithme 1 a donc fourni une meilleure estimation du nombre réel de cas incidents de cancers colorectaux, ne différant que de 2,34 % par rapport aux données observées par un ensemble de 14 registres départementaux. L'estimation nationale est de 39 122 [37 020 ;41 224] et présente un écart de 4,5 % avec celle de Francim (37 413), qui est incluse dans notre intervalle de confiance. La modélisation a permis de réduire de plus

Tableau IV : Incidences départementales des cancers colorectaux : comparaison entre les données des registres et la méthode de l'algorithme 1, avant et après correction

Département	Incidence mesurée dans registre (1)	Incidence estimée à partir du PMSI[a] par algorithme 1 (2)	Incidence estimée après correction des résultats de l'algorithme 1 par les modèles 1 et 2 (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	6 400	7 003	6 569	9,42	2,34

[a] : programme de médicalisation des systèmes d'information.

de moitié l'écart qui était initialement de 9,9 % entre l'estimation fournie par l'algorithme 1 (41 121 cas), sans correction par la modélisation, et celle proposée par le réseau Francim (regroupement national des registres). Toutefois, l'écart reste supérieur à celui obtenu par Mitton *et al.* [21] sur l'année 2007 (2,6 %). Cette différence pourrait s'expliquer en partie par l'amélioration au fil des années de la qualité et de l'exhaustivité du codage du PMSI MCO, suite à la mise en place de la tarification à l'activité en 2004, avec augmentation progressive du taux d'application de 10 % en 2004 à 100 % en 2008. L'amélioration du codage avec les années et les différences en termes de valorisation des séjours entre les pays pourraient aussi expliquer partiellement (en plus du choix de l'algorithme) pourquoi les résultats obtenus en France semblent meilleurs que ceux publiés, plus tôt en Italie où 642 cas de cancer colorectal [1], avaient été identifiés, alors que les registres en enregistraient 799, pour l'année 2001.

Le choix d'écarter la chimiothérapie et la radiothérapie, et donc de ne baser l'algorithme 1 que sur des procédures chirurgicales, pourrait être discuté. Cependant, les cas de patients exclusivement traités par chimio et/ou radiothérapie lors d'un premier épisode de la maladie sont relativement rares, puisque, par exemple, plus de 90 % des cas incidents bénéficient d'une résection soit chirurgicale, soit endoscopique, toutes deux incluses dans la liste des procédures de l'algorithme 1. De même, il pourrait être objecté que, dans le cas d'une endoscopie réalisée en externe, le patient ne soit pas détectable par notre méthode. Bien que cela soit encore le cas pour 5 à 7 % de ces gestes en 2004, la pratique de coloscopies réalisées sans anesthésie générale tend à disparaître. La plupart des patients bénéficient actuellement d'une anesthésie générale et sont donc hospitalisés, bénéficiant d'un résumé PMSI les rendant alors détectables par notre algorithme.

L'impact observé de l'âge avancé et du sexe masculin sur l'apparition de faux négatifs peut s'expliquer, d'une part, par la moindre acceptance ou faisabilité du traitement chez les personnes âgées, du fait de l'hospitalisation lourde que cela implique et, d'autre part, par le fait que, classiquement, les hommes acceptent moins facilement une intervention lourde que les femmes. Il serait très intéressant de pouvoir étudier

l'ensemble de la prise en charge du patient, en médecine de ville comme à l'hôpital, ce qui permettait de réduire les biais, grâce notamment aux données du Sniiram (Système national d'information inter-régimes de l'Assurance maladie).

La méthode que nous avons utilisée pour le chaînage entre les données PMSI et les données des registres est une méthode largement éprouvée, que nous avons développée et appliquée au chaînage des données PMSI avec celles des registres des cancers digestifs de Côte d'Or dès 1996 [12]. Nous avons eu l'occasion d'utiliser cette méthode pour d'autres pathologies telles que le cancer du sein dans plusieurs départements [4, 11] et les accidents vasculaires cérébraux [24]. Cette méthode nécessite un croisement individuel des données du PMSI et des registres, qui repose sur l'application d'une méthode de chaînage probabiliste garantissant la qualité de ce croisement. Toutefois, notre étude ne s'appuie que sur les données de 2 départements. Compte tenu de la variabilité des résultats fournis par la confrontation PMSI-registre dans différents départements français [4, 7, 15], il aurait été intéressant de pouvoir s'appuyer, pour l'extrapolation à la France entière, sur les données des registres d'autres départements.

Cette étude repose sur des données déjà anciennes, comme beaucoup d'autres travaux reposant sur la confrontation des données médico-administratives avec des données de registre pour d'autres localisations cancéreuses, du fait de la nécessité de disposer de données de registres validées (procédure longue) et du temps nécessaire à l'obtention des autorisations et à l'accès aux données dans les établissements et à leur validation [21, 25-29].

Le changement des règles de codage du diagnostic principal en 2009 ne devrait pas avoir d'impact dans le contexte de notre étude. En effet, lors de la prise en charge initiale d'un cancer, celui-ci reste noté le plus souvent en DP. De plus, comme les données du PMSI ont gagné en qualité depuis 2004, date de l'instauration de la tarification à l'activité, nous pensons que de futures études, menées sur des données plus récentes et, si possible, à une échelle plus large, devraient arriver à des conclusions peut-être encore plus favorables à l'utilisation des données PMSI pour identifier les cas incidents de cancers colorectaux.

Conclusion

Cette étude montre l'intérêt des données du PMSI MCO pour l'épidémiologie des cancers colorectaux, en proposant une méthode permettant d'obtenir une estimation du nombre de cas incidents. L'utilisation, comme critères de sélection, de codes de diagnostics et d'actes, s'est avérée plus efficace que de se baser sur l'élimination des cas prévalents. Toutefois, il convient d'être extrêmement vigilant sur les utilisations possibles du PMSI en dehors du calcul de l'incidence ou de la prévalence des maladies. En effet, si le premier algorithme utilisé fournit une bonne estimation du nombre total de cas incidents, il faut noter que les faux négatifs et les faux positifs, de par leur fréquence similaire, se compensent mutuellement. De plus, aucune information concernant le stade, le type ou la localisation histologique de la tumeur, ne peut être retrouvée dans les données du PMSI, ce qui souligne l'intérêt des registres, qui reste un outil incontournable, notamment pour l'étude des facteurs pronostiques et la comparaison des prises en charge.

Aucun conflit d'intérêt déclaré

Remerciements

Ce travail a reçu le support de l'Institut de recherches et d'expertises en santé publique (Iresp). Les auteurs souhaitent remercier pour leur aide : 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-Montbéliard), 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), Jean Rudloft (Centre hospitalier de Pontarlier), Anne-Marie Bouvier et son équipe (Registre des cancers digestifs de Côte d'Or) et Arlette Danzon et son équipe (Registre des tumeurs du Doubs) . Les auteurs remercient également le réseau Francim ainsi que Jean-Paul Beyeme-Ondoua, pour les échanges fructueux lors de la conception des algorithmes.

Une version anglaise de cet article a été publié dans le Journal of Cancer Epidemiology (Volume 2012, Article ID 298369, 7 pages).

Références

1. 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.

2. Beyeme-Ondoua JP. Evaluation of the quality of surveillance data for colorectal cancer from the national PMSI database in 2003. Sante Publique. 2007;19(6):471-80.
3. Carre N, Uhry Z, Velten M, Tretarre B, Schvartz C, Molinie F, *et al.* [Predictive value and sensibility of hospital discharge system (PMSI) compared to cancer registries for thyroid cancer (1999-2000)]. Rev Epidemiol Sante Publique. 2006;54(4):367-76.
4. 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.
5. Ganry O, Taleb A, Peng J, Raverdy N, Dubreuil A. Evaluation of an algorithm to identify incident breast cancer cases using DRGs data. Eur J Cancer Prev. 2003;12(4):295-9.
6. Penberthy L, McClish D, Pugh A, Smith W, Manning C, Retchin S. Using hospital discharge files to enhance cancer surveillance. Am J Epidemiol. 2003;158(1):27-34.
7. Remontet L, Mitton N, Couris CM, Iwaz J, Gomez F, Olive F, *et al.* Is it possible to estimate the incidence of breast cancer from medico-administrative databases? Eur J Epidemiol. 2008;23(10):681-8.
8. Trombert B, Martin C, Vercherin P, editors. From case mix data bases to health geography. Proceedings of the 19th International PCS/E Working Conference; 2003; Washington.
9. Couris CM, Foret-Dodelin C, Rabilloud M, Colin C, Bobin JY, Dargent D, *et al.* [Sensitivity and specificity of two methods used to identify incident breast cancer in specialized units using claims databases]. Rev Epidemiol Sante Publique. 2004;52(2):151-60.
10. Uhry Z, Colonna M, Remontet L, Grosclaude P, Carre N, Couris CM, *et al.* Estimating infra-national and national thyroid cancer incidence in France from cancer registries data and national hospital discharge database. Eur J Epidemiol. 2007;22(9):607-14.
11. Quantin C, Benzenine E, Fassa M, Hagi M, Fournier E, Gentil J, *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 (SJIAOS). 2012;28:73-85.
12. Quantin C, Bouzelat H, Allaert FA, Benhamiche AM, Faivre J, Dusserre L. Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. Methods Inf Med. 1998;37(3):271-7.
13. McBean AM, Babish JD, Warren JL. Determination of lung cancer incidence in the elderly using Medicare claims data. Am J Epidemiol. 1993;137(2):226-34.
14. McClish DK, Penberthy L, Whittemore M, Newschaffer C, Woolard D, Desch CE, *et al.* Ability of Medicare claims data and cancer registries to identify cancer cases and treatment. Am J Epidemiol. 1997;145(3):227-33.
15. Hafdi-Nejjari Z, Couris CM, Schott AM, Perrot L, Bourgoin F, Borson-Chazot F, *et al.* [Role of hospital claims databases from care units for estimating thyroid cancer incidence in the Rhone-Alpes region of France]. Rev Epidemiol Sante Publique. 2006;54(5):391-8.
16. Koroukian SM, Cooper GS, Rimm AA. Ability of Medicaid claims data to identify incident cases of breast cancer in the Ohio Medicaid population. Health Serv Res. 2003;38(3):947-60.
17. Leung KM, Hasan AG, Rees KS, Parker RG, Legorreta AP. Patients with newly diagnosed carcinoma of the breast: validation of a claim-based identification algorithm. J Clin Epidemiol. 1999;52(1):57-64.
18. Couris CM, Seigneurin A, Bouzbid S, Rabilloud M, Perrin P, Martin X, *et al.* French claims data as a source of information to describe cancer incidence: predictive values of two identification methods of incident prostate cancers. J Med Syst. 2006;30(6):459-63.
19. SNFGE. National Thesaurus of digestive oncology [internet]. 2008. [Cited 4 September, 2013]. Available from: http://www.snfge.asso.fr/01-bibliotheque/0g-thesaurus-cancerologie/publication5/sommaire-thesaurus.asp.
20. Olive F, Gomez S, Schott AM, 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;59:53-8.
21. 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.
22. Ferlay J, Parkin DM, Steliarova-Foucher E. Estimates of cancer incidence and mortality in Europe in 2008. Eur J Cancer. [Multicenter Study Research Support, Non-U.S. Gov't]. 2010;46(4):765-81.
23. Uhry Z, Belot A, Colonna M, Bossard N, Grosclaude P, Remontet L, editors. Comparaison de trois méthodes d'estimation de l'incidence nationale pour 22 cancers en France. V[e] congrès international d'épidémiologie ADELF-EPITER; 2012; Bruxelles.
24. Aboa-Eboule C, Mengue D, Benzenine E, Hommel M, Giroud M, Bejot Y, *et al.* How accurate is the reporting of stroke in hospital discharge data? A pilot validation study using a population-based stroke registry as control. J Neurol. 2012;260(2):605-13.
25. Colonna M, Mitton N, Schott AM, Remontet L, Olive F, Gomez F, *et al.* Joint use of epidemiological and hospital medico-administrative data to estimate prevalence. Application to French data on breast cancer. Cancer epidemiology. [Research Support, Non-U.S. Gov't]. 2012 Apr;36(2):116-21.
26. Trombert Paviot B, Gomez F, Olive F, Polazzi S, Remontet L, Bossard N, *et al.* Identifying prevalent cases of breast cancer in the French case-mix databases. Methods of information in medicine. [Research Support, Non-U.S. Gov't]. 2011;50(2):124-30.
27. Uhry Z, Remontet L, Colonna M, Belot A, Grosclaude P, Mitton N, *et al.* Cancer incidence estimation at a district level without a national registry: a validation study for 24 cancer sites using French health insurance and registry data. Cancer epidemiology. [Research Support, Non-U.S. Gov't Validation Studies]. 2013 Apr;37(2):99-114.
28. Uhry Z, Belot A, Colonna M, Bossard N, Rogel A, Iwaz J, *et al.* National cancer incidence is estimated using the incidence/mortality ratio in countries with local incidence data: is this estimation correct? Cancer epidemiology. 2013 Jun;37(3):270-7.
29. Coureau G, Baldi I, Saves M, Jaffre A, Barat C, Gruber A, *et al.* [Performance evaluation of hospital claims database for the identification of incident central nervous system tumors compared with a cancer registry in Gironde, France, 2004]. Revue d'Epidemiologie et de Sante Publique. [Comparative Study]. 2012 Aug;60(4):295-304.