Déterminer le statut vital des patients domiciliés en France métropolitaine et dans les Départements d'Outre Mer en chaînant données hospitalières et données de mortalité de l'INSEE anonymisées

I. FOURNEL[1], M. SCHWARZINGER[2], E. BENZENINE[1], C. BINQUET[1], B. RIANDEY[3], C. HILL[4], C. QUANTIN[1]

[1] Département d'information médicale. CHU Dijon.

[2] Service d'information médicale, Direction des études médico-économiques, Institut Gustave Roussy.

[3] Institut National d'Etudes Démographiques.

[4] Département de biostatistiques et d'épidémiologie, Institut Gustave Roussy.

1. Introduction

Élément clé de l'évaluation des soins médicaux, de la recherche clinique et des études épidémiologiques, le statut vital des patients peut être déterminé par différentes méthodes. Parmi celles-ci, on peut citer la mise en relation de fichiers, ou chaînage, qui consiste à associer les observations se rapportant à un même individu, mais provenant de deux fichiers différents. Largement utilisé à l'étranger dans de nombreuses études épidémiologiques pour déterminer le statut vital des patients, le chaînage est peu utilisé en France pour atteindre cet objectif. Pourtant, le statut vital peut être obtenu, en France, à partir du fichier annuel de mortalité de l'Institut National des Statistiques et des Études Économiques (INSEE) qui répertorie l'ensemble des décès des personnes domiciliées en France métropolitaine et dans les Départements d'Outre Mer (DOM). Il est ainsi possible de mettre en relation les fichiers hospitaliers et les fichiers de mortalité de l'INSEE. Cependant, le croisement de ces fichiers nécessite le respect des législations française et européenne relatives au traitement des données à caractère personnel. La Commission Nationale de l'Informatique et des Libertés (CNIL) a estimé que l'utilisation d'informations pouvant être chaînées nécessite des procédures d'anonymisation reconnues et évaluées [1].

L'objectif de ce travail était d'évaluer la performance de la détermination du statut vital des patients par croisement entre des données hospitalières et les données de mortalité de l'INSEE par une méthode de chaînage, après avoir rendu ces informations anonymes.

2. Populations et méthodes

2.1. Population

2.1.1. Critères d'inclusion

L'ensemble des patients domiciliés en France métropolitaine ou dans les DOM et hospitalisés pour la première fois entre 1998 et 2000 à l'Institut Gustave Roussy (IGR) pour une tumeur de malignité certaine ou possible ont été inclus.

2.1.2. Bases de données utilisées

Les fichiers de mortalité annuels de l'INSEE ont été utilisés pour les années 1998 à 2004, et les données hospitalières ont été obtenues à partir du fichier d'hospitalisation de l'IGR.

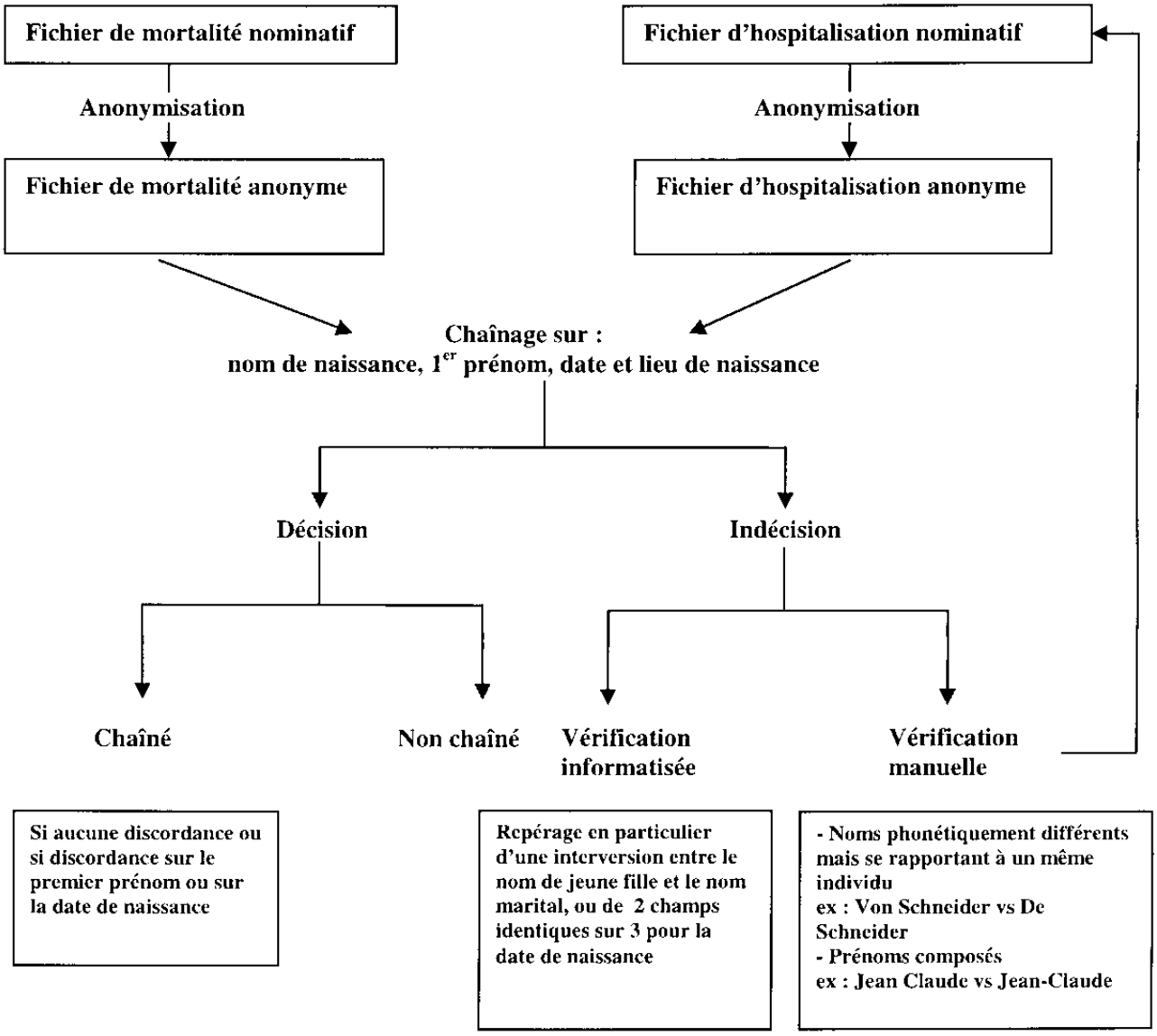
Le statut vital réel des patients au 31 décembre 2004 a été déterminé à partir des informations disponibles à l'IGR ou, dans le cas où l'IGR ne disposait d'aucune information sur le statut vital, par l'interrogation du fichier du Répertoire National d'Identification des Personnes Physiques (RNIPP).

2.2. Méthode d'anonymisation et de chaînage (figure 1)

2.2.1. Anonymisation

Un traitement phonétique adapté à la langue française a été appliqué d'emblée sur le nom de naissance, le nom marital et les prénoms pour limiter l'impact des erreurs dues à la saisie [2].

FIGURE 1 : DESCRIPTION DE LA MÉTHODE D'ANONYMISATION ET DE CHAÎNAGE.



Chacune de ces variables d'identification des fichiers d'hospitalisation de l'IGR et des fichiers de mortalité de l'INSEE a ensuite été anonymisée séparément avec le logiciel ANONYMAT, agréé par la CNIL (avis de la CNIL du 18 décembre 1996, donné à la demande

d'avis numéro 455 451). Cette technique d'anonymisation entraîne la transformation irréversible des variables d'identification d'un individu, pour obtenir un code strictement anonyme, mais toujours le même pour une chaîne de caractères donnée.

2.2.2. Le chaînage probabiliste

Le fichier d'hospitalisation de l'IGR et le fichier de mortalité de l'INSEE ont ensuite été chaînés selon la méthode probabiliste de Jaro [3]. Le chaînage portait sur les variables suivantes : le nom de naissance, le premier prénom et la date de naissance anonymisés, ainsi que le code INSEE de la commune de naissance en clair. Les codes INSEE du lieu de naissance avaient été préalablement harmonisés, afin de tenir compte des changements de codes survenus au cours du temps. De plus, il a été nécessaire de stratifier sur le lieu de naissance. En effet, le code INSEE de la commune de naissance n'a pas le même pouvoir discriminant selon que le patient est né en France (il correspond alors à la commune de naissance) ou à l'étranger (il correspond alors au pays de naissance).

La méthode probabiliste de Jaro permet de chaîner les observations provenant de différentes sources en tenant compte de la capacité discriminante de chacune des variables d'identification utilisées. En effet, cette méthode permet d'estimer un poids pour chacune d'entre elles, ce poids étant d'autant plus élevé que la variable est discriminante. Un poids composé, obtenu par la somme des poids de chaque variable, est attribué à chaque paire d'enregistrements. La décision de chaîner ou non deux observations dépend de la valeur de ce poids composé, et donc de la teneur de l'ensemble des variables. Les paires d'enregistrements sont alors classées en trois ensembles : à chaîner, à ne pas chaîner ou en indécision [3].

2.2.3. Règles de décision

L'algorithme décisionnel a été divisé en deux étapes : la première était complètement automatisée, et la seconde comportait des étapes de validation manuelle. Ces étapes de validation manuelle consistaient à rechercher des erreurs non récupérables par le traitement orthographique, comme les erreurs orthographiques sur le nom, ou des prénoms composés. Elles supposaient donc de retourner aux données en clair, ce qui peut être réalisé en accord avec la directive européenne dans le centre où le patient est traité, car chaque centre source de l'information est en droit de conserver la correspondance entre le numéro d'anonymat et l'identité du patient. Le coordinateur de l'étude peut donc demander la vérification ou la correction des données correspondant à un numéro d'anonymat particulier. Le centre, source de l'information, renvoie alors l'ensemble des enregistrements corrigés, après une nouvelle anonymisation [4].

3. Résultats

3.1. Description de la population

Un total de 10 089 patients (dont 8 592 nés en France) correspondaient aux critères d'inclusion. Les patients nés en France comportaient une proportion de femmes plus importante que les patients nés à l'étranger (52,9% vs 50,0%, p=0,03), et étaient plus jeunes (52,8 ans vs 55,8 ans, p<0,0001).

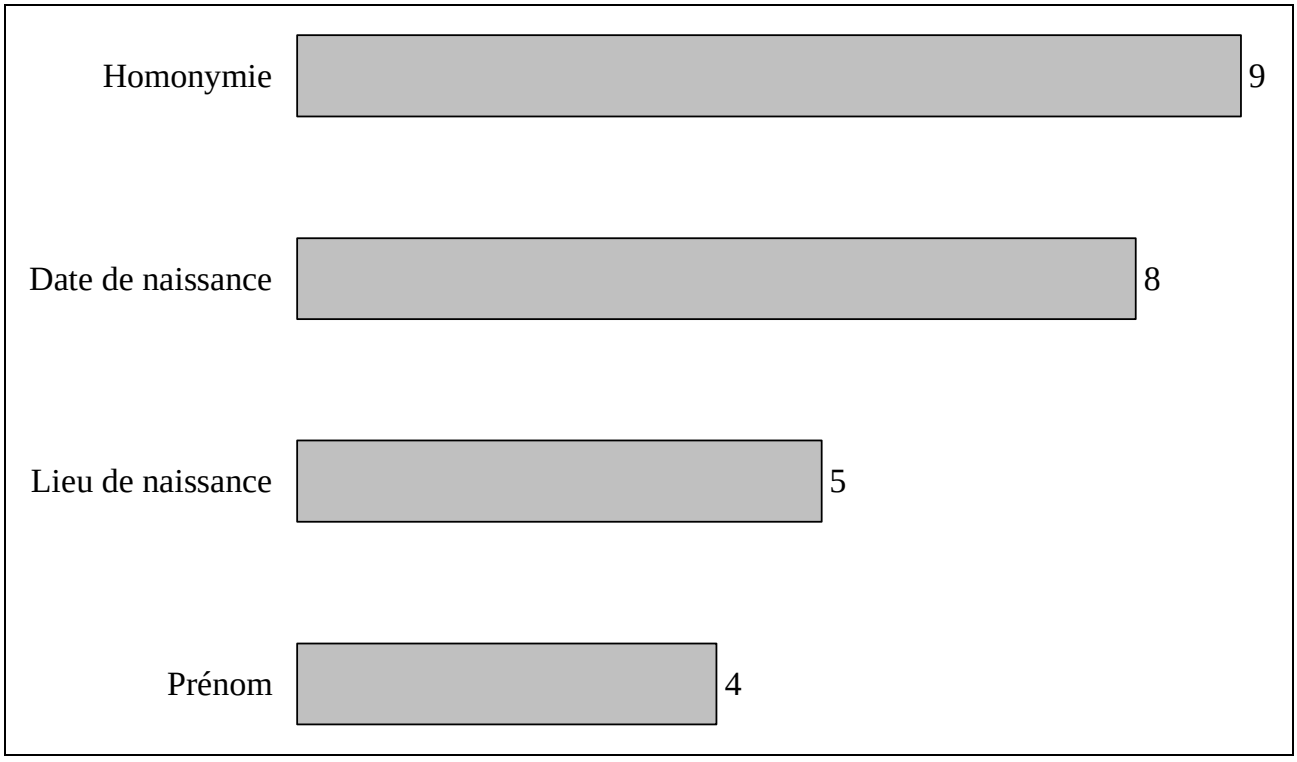
La fréquence de la mortalité était 48,5% au 31 décembre 2004 (48,7% pour les patients nés en France et 47,0% pour les patients nés à l'étranger, p=0,21).

3.2. Indicateurs de performance du chaînage

3.2.1. Globalement

Les résultats du chaînage étaient très satisfaisants pour l'ensemble des patients inclus. La proportion de bien classés était de 97,2%, la sensibilité (probabilité que le patient IGR soit retrouvé dans la base INSEE s'il était décédé) de 94,8% et la spécificité (probabilité que le patient IGR ne soit pas retrouvé dans la base INSEE s'il était en vie) de 99,5%. Seuls 256 patients étaient déclarés vivants à tort par notre méthode, et 26 étaient déclarés décédés à tort (figure 2).

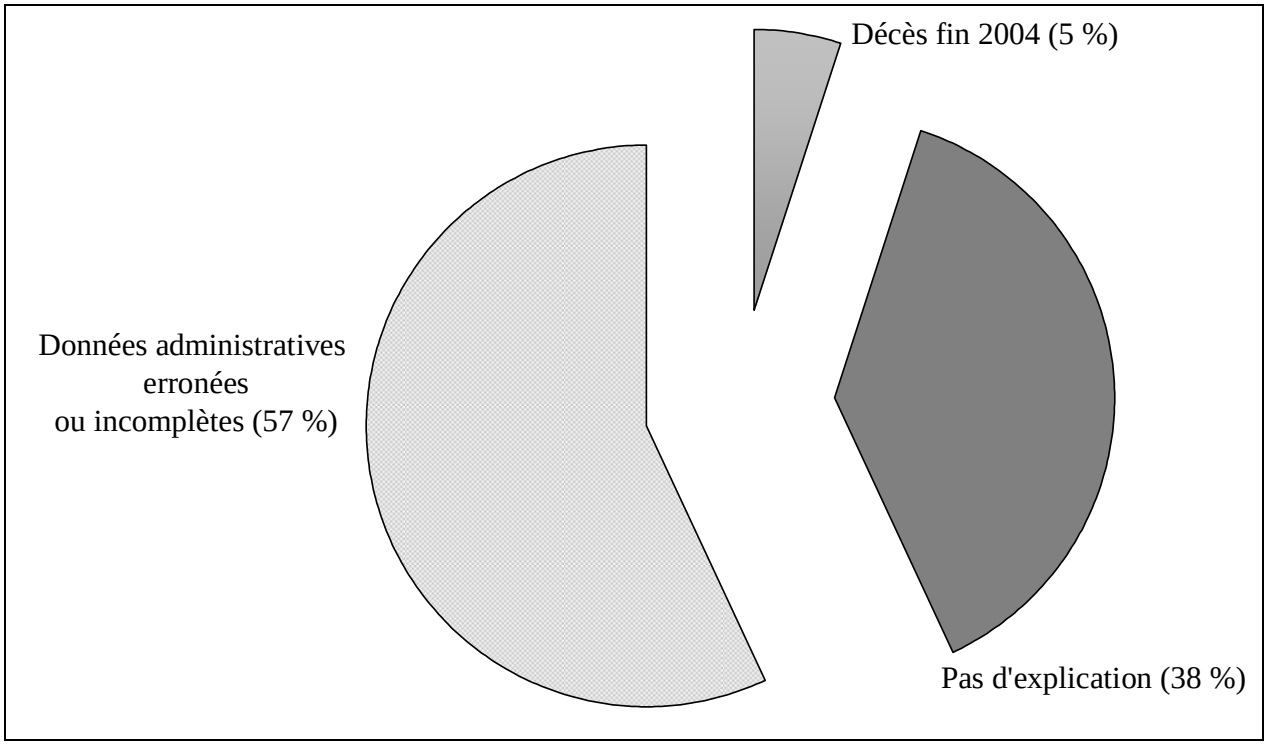
FIGURE 2 : RAISONS DE DISCORDANCE POUR LES 26 PATIENTS DÉCLARÉS DÉCÉDÉS À TORT PAR NOTRE MÉTHODE DE CHAÎNAGE.



Les principales raisons de discordance (figure 3) entre les résultats obtenus avec notre méthode et le statut vital de référence concernaient majoritairement des erreurs sur les données administratives (erreurs orthographiques sur le nom, prénoms composés, nom de jeune fille non renseigné, code du lieu de naissance erroné ou date de naissance incomplète). Il semblerait également que les décès survenus en décembre 2004 n'étaient pas encore tous intégrés dans la base INSEE.

L'incorporation d'étapes de validation manuelle au processus automatique permettait d'accroître la sensibilité de la méthode. Cette procédure de validation manuelle des noms et des prénoms a permis de récupérer 120 des 256 décès non retrouvés initialement par notre stratégie de chaînage automatisé au prix de seulement 4 décès supplémentaires pour des patients ayant des noms très proches entre les deux bases. La sensibilité passait alors à 97,2% et la spécificité à 99,4%, et la proportion de bien classés à 98,4%.

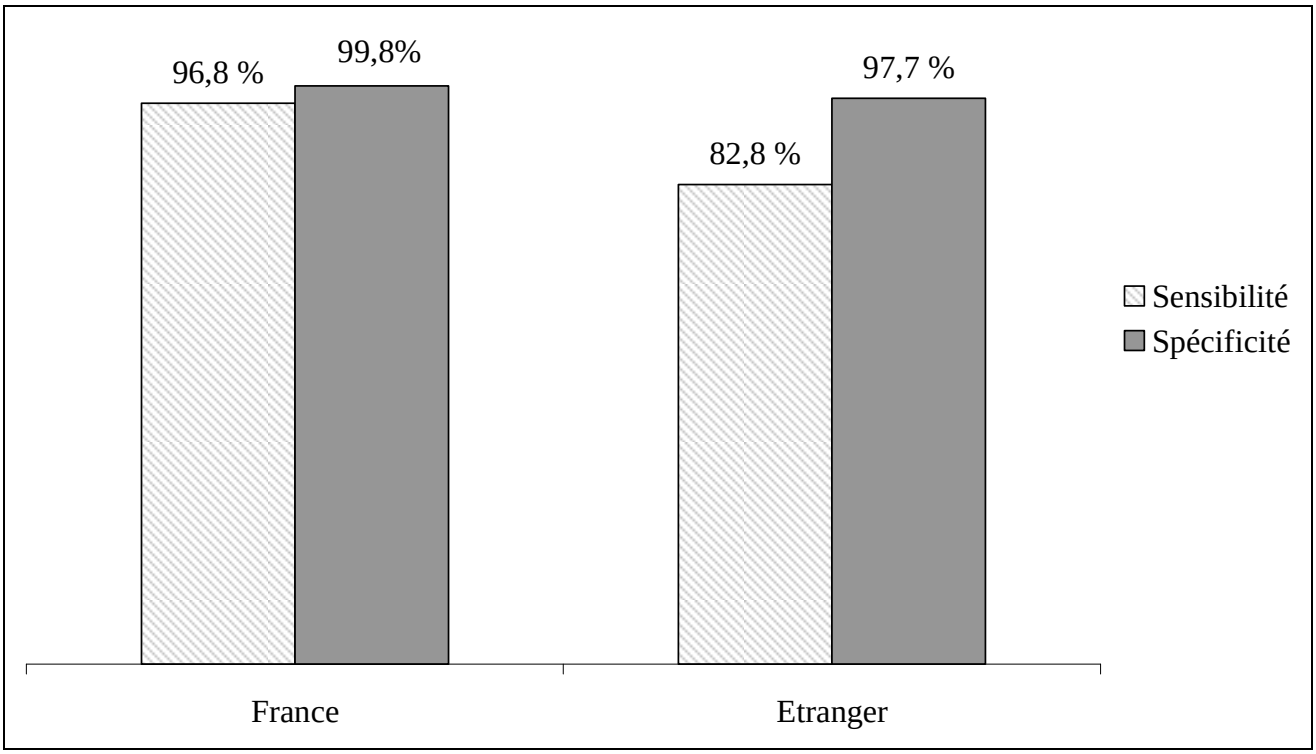
FIGURE 3 : RAISONS DE DISCORDANCE POUR LES 256 PATIENTS DÉCLARÉS VIVANT À TORT PAR NOTRE MÉTHODE DE CHAÎNAGE EN AUTOMATIQUE.



3.2.2. Selon le lieu de naissance (figure 4)

La méthode était très performante, même en l'absence de validation manuelle, pour les patients nés en France, avec une sensibilité de 96,8% et une spécificité de 99,8%, et un peu moins bonne pour les patients nés à l'étranger, pour lesquels la sensibilité est à 82,8% et la spécificité à 97,7%.

FIGURE 4 : PERFORMANCES DU CHAÎNAGE SELON LE LIEU DE NAISSANCE



3.2.3. Selon le sexe et le type de cancer

La sensibilité du chaînage était meilleure chez les hommes (97,9%) que chez les femmes (95,0%), et la spécificité est proche de 100% quel que soit le sexe (figure 5). Les cancers pour lesquels les performances du chaînage étaient les meilleures sont les cancers de la prostate et du testicule, pour lesquels la spécificité est 100%, et la sensibilité de 97,6% et 100% respectivement (tableau 1).

FIGURE 5 : PERFORMANCES DU CHAÎNAGE SELON LE SEXE

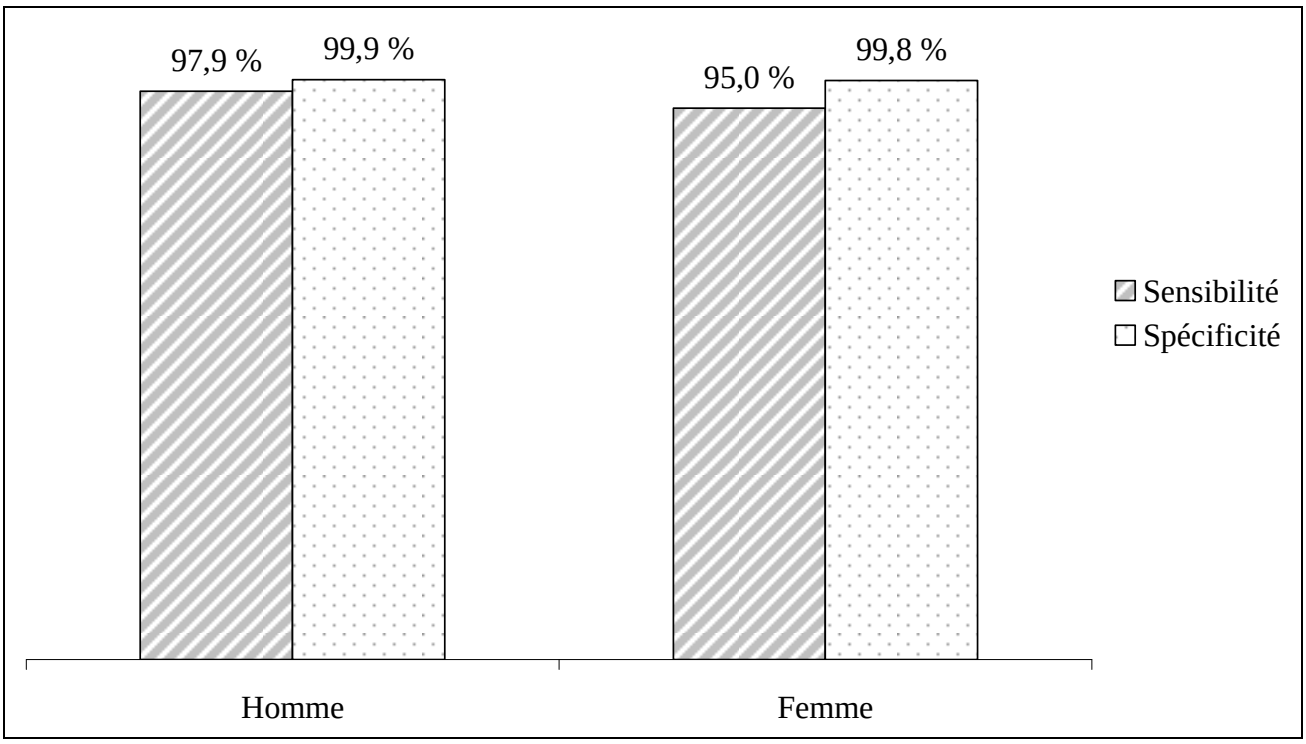


TABLEAU 1 : PERFORMANCES DU CHAÎNAGE SELON LE TYPE DE CANCER

	Effectif	Mortalité (%)	Sensibilité (%)	Spécificité (%)
Colo-rectal	476	67,4	97,5	99,4
Thyroïde	204	17,2	91,4	100,0
Sein	1 774	19,1	94,1	99,8
Larynx	272	54,4	96,6	100,0
Prostate	298	42,3	97,6	100,0
Rein	476	67,4	97,5	99,4
Poumon	500	90,0	97,8	100,0
Os et cartilage articulaire	152	42,1	93,8	100,0
Testicule	149	14,8	100,0	100,0
Lèvre, bouche, pharynx	906	69,1	97,6	99,6
Utérus	503	34,2	94,2	100,0

4. Discussion

Cette approche de détermination du statut vital par une méthode de chaînage probabiliste présente un intérêt certain lorsque les bases de données sont importantes. Bien que l'interrogation du RNIPP soit bien adaptée aux grandes bases de données [5], c'est une méthode dont le coût dépend du nombre d'individus (91,47 euros hors taxe par fichier avec un

coût supplémentaire variant de 0,35 à 0,83 euros hors taxe par individu selon son lieu de naissance), alors que la base de mortalité de l'INSEE a un coût fixe de 4 000 euros pour une année donnée, quel que soit le nombre d'individus. De plus, le coût de la base de mortalité de l'INSEE est mutualisé entre les différents centres de lutte contre le cancer, revenant ainsi seulement à 200 euros par an et par centre. L'utilisation de la base de mortalité de l'INSEE permet donc de réduire les coûts et d'éviter que des études épidémiologiques soient empêchées pour des raisons financières. Enfin, pour faciliter ces recherches par croisement avec les données de cette base, il serait souhaitable de pouvoir disposer dans les données de l'étude concernée non seulement des noms et prénoms, mais aussi de la date et de la commune de naissance des individus dont on cherche à connaître le statut vital. Les méthodes traditionnelles de recherche du statut vital, quant à elles, nécessitent du temps et sont davantage adaptées à une recherche du statut vital pour un nombre restreint d'individus. Elles nécessitent en outre de disposer des coordonnées (adresse ou numéro de téléphone) précises et à jour des patients afin de pouvoir les contacter, ce qui n'est pas toujours le cas dans les études rétrospectives. Cependant, le chaînage entre données hospitalières et données de mortalité de l'INSEE ne s'applique qu'aux patients domiciliés en France métropolitaine et dans les DOM, puisque seuls ces patients sont présents dans le fichier de mortalité de l'INSEE.

L'autre avantage majeur de notre approche est de montrer que, malgré les difficultés réglementaires, il est possible de croiser des données hospitalières et des données de mortalité sur des bases de données importantes tout en préservant la confidentialité des informations médicales. En effet, le chaînage sur données anonymisées donne des résultats très satisfaisants notamment pour les patients nés en France. Pour les patients nés à l'étranger, le code lieu de naissance correspond alors à un pays, et non pas à une commune, et de ce fait, était beaucoup moins discriminant que pour les patients nés en France.

Dans notre étude, les performances du chaînage étaient meilleures chez les hommes (sensibilité de 97,9% et spécificité de 99,9%). Ceci peut probablement s'expliquer par le fait que le nom de jeune fille soit inconstamment renseigné, ou accolé au nom marital. Les performances du chaînage semblent ainsi être davantage liées à la qualité des identifiants qu'à la fréquence de la mortalité. En effet, les cancers pour lesquels la sensibilité et la spécificité sont les plus élevées sont des cancers exclusivement masculins, alors que la fréquence de mortalité est très différente.

Dans tous les cas, la possibilité de croiser de façon fiable les données hospitalières et les données de mortalité suppose que des identifiants compatibles soient utilisés dans les deux sources. Notre travail montre l'importance de la qualité de ces identifiants, et conforte les résultats obtenus par l'étude de faisabilité d'une étude pilote pour mesurer la mortalité post-hospitalière à 30 jours [6]. La difficulté soulevée par cette étude est relative à la qualité de l'identifiant FOIN 1 (Fonction d'Occultation d'Identifiants Nominatifs) du PMSI (Programme de Médicalisation des Systèmes d'Information), fondé sur le Numéro d'Inscription au Répertoire national d'identification des personnes physiques (NIR) de l'ouvrant droit et non celui du patient, identifiant disponible dans les seuls fichiers de l'assurance maladie. D'autre part, la procédure FOIN transforme par hachage les trois identifiants (numéro de sécurité sociale, date de naissance et sexe du patient) en un seul numéro, contrairement à la méthode appliquée dans notre étude, pour laquelle chaque champ était anonymisé séparément [7]. Par conséquent, une erreur sur un seul des identifiants de la procédure FOIN (sexe par exemple) peut conduire à rejeter le chaînage des données d'un même patient. De ce fait, un nombre non négligeable d'appariements sont perdus dans cette étude du fait d'erreurs potentielles de saisie [6]. Il nous paraît donc urgent que soit mise en œuvre la carte vitale 2 fondé sur le NIR haché du patient, dans la logique de l'identifiant qui se met en place pour le Dossier Médical Personnel, dont nous détaillons l'historique et les principes dans une autre communication [8].

5. Conclusion

Ce travail montre que l'on peut améliorer considérablement l'information sur le statut vital en ajoutant aux informations hospitalières les informations de mortalité nationale, tout en respectant les règles de la confidentialité.

Cette méthode permet d'envisager la réalisation d'études épidémiologiques à large échelle à un coût humain et financier plus abordable, lorsque les données hospitalières sont de qualité. Avant d'être appliquées au croisement d'une autre base de données avec la base de l'INSEE, il faudrait préalablement valider, sur un petit échantillon, les règles de décision présentées dans cette étude. Si l'algorithme apparaît satisfaisant, il pourra être appliqué à l'ensemble de la base. Dans le cas contraire, il sera nécessaire de redéfinir d'autres règles adaptées aux poids obtenus par l'application du modèle probabiliste de Jaro à la nouvelle base étudiée.

Cette étude, comme l'étude pilote de l'hôpital Paul Brousse [6], montre qu'il existe encore un fossé entre les études épidémiologiques hospitalières et les études de mortalité, pour des raisons de qualité et de compatibilité des identifiants mais aussi des raisons financières et qu'il serait souhaitable que la passerelle entre les deux devienne aisée dans les prochaines années.

BIBLIOGRAPHIE

[1] Délibération n°97-008 du 4 février 1997 portant adoption d'une recommandation sur le traitement des données de santé à caractère personnel. Journal officiel du 12 avril 1997

[2] BOUZELAT H. Anonymat et chaînage de fichiers médicaux en vue d'études épidémiologiques. Thèse de Docteur d'Université spécialiste en Informatique Médicale. Université de Bourgogne. 1998 : p 97

[3] JARO MA. Probabilistic linkage of large public health data files. Stat med 1995 Mar 15-Apr 15 ; 14(5-7) : 491-8

[4] QUANTIN C, BOUZELAT H, ALLAERT FA, BENHAMICHE AM, FAIVRE J, DUSSERRE L. How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure. Int J Med Inform. 1998 Mar;49 (1) : 117-22

[5] BOSSARD N, VELTEN M, REMONTET L, BELOT A, MAAROUF N, BOUVIER AM, GUIZARD AV, TRETARRE B, LAUNOY G, COLONNA M, DANZON A, MOLINIE F, TROUSSARD X, BOURDON-RAVERDY N, CARLI PM, JAFFRÉ A, BESSAGUET C, SAULEAU E, SCHVARTZ C, ARVEUX P, MAYNADIE M, GROSCLAUDE P, ESTEVE J, FAIVRE J. Survival of cancer patients in France : a population-base study from the Association of the French Cancer Registries (FRANCIM) European Journal of Cancer 43, 2007 : 149-60

[6] VALLET O, VILLEMINOT S, GASQUET I, FALISSARD B. Direction de la recherche, des études, de l'évaluation et des statistiques. DREES. Élaboration d'un outil de mesure de la mortalité post hospitalière. Série Études 2004 ; 40

[7] QUANTIN C, GOUYON B, ALLAERT FA, COHEN O. Méthodologie pour le chaînage de données sensibles tout en respectant l'anonymat: application au suivi des informations médicales. Journal de la SFdS 2005 ;146 (3) :15-26

[8] QUANTIN C, RIANDEY B, COHEN O. Traitement épidémiologique et démographique du Dossier médical personnel. L'identifiant santé. XIV^{ème} colloque national de démographie. Démographie et santé. Bordeaux, France, mai 2007

Remerciements

Pour la détermination du statut vital de référence, les auteurs remercient Josianne Chavanne (Centre de Ressources Informatiques appliquées à l'épidémiologie, aux sciences sociales et à la santé publique) et Astrid Galtrand (IGR).