

VALIDATION D'UN PROCESSUS D'IDENTIFICATION D'ÉVÉNEMENTS CANCEREUX SURVENUS DANS UNE COHORTE DE PATIENTS DIABÉTIQUES DE TYPE 2 SUIVIS AU CHU DE POITIERS

Elise Gand¹, Evelyne Liuu¹, Gautier Defossez², & Marc Paccalin¹

¹ *Centre d'investigation clinique INSERM 1402, Entrée 5 – Cour Est Jean Bernard CS 90577
86021 POITIERS CEDEX, elise.gand@chu-poitiers.fr, evelyne.liuu@chu-poitiers.fr,
marc.paccalin@chu-poitiers.fr*

² *Registre des cancers Poitou-Charentes, Université de Poitiers, gautier.defossez@univ-poitiers.fr*

Résumé. Les établissements de santé sont des sources riches et multiples de données de santé individuelles. Dans un contexte réglementaire de valorisation des données médicales disponibles, L'axe ACDC du Centre d'investigation Clinique 1402 souhaite valider un processus d'identification de survenue de cancers dans la cohorte SURDIAGENE de patients diabétiques de type 2. Le processus mis en place comporte trois phases. La première étape consiste à recueillir différentes sources d'informations disponibles pour chacun des patients au sein du système informatique de l'hôpital: les courriers d'hospitalisation, les comptes rendus d'anatomopathologie et les réunions de concertation pluridisciplinaires. Les venues comportant des actes reliés à un cancer sont également identifiées à partir des données du Programme de Médicalisation des Systèmes d'Information (PMSI). La deuxième étape est entièrement informatisée : un algorithme de décision, utilisant notamment une lecture automatisée des documents, identifie les patients potentiellement porteurs de cancer. Enfin suit une phase d'adjudication par des médecins spécialisés en oncologie afin de valider chaque cas de cancer, et ses caractéristiques notamment leur localisation. Les cas identifiés sont confrontés à ceux enregistrés par le Registre Général des Cancers Poitou-Charentes, membre du réseau FRANCIM (réseau français des registres du cancer), qui recense de façon standardisée et exhaustive les cas incidents de cancers depuis 2008, dans l'ex-région Poitou-Charentes, sous l'égide de l'Institut de veille sanitaire (InVS) et de l'institut national du Cancer (INCa). Notre processus montre une très bonne sensibilité (91%) et spécificité (98%). Nous pensons donc développer cet outil pour la recherche d'autres événements de santé et généraliser la démarche à d'autres cohortes de patients.

Mots-clés. Cohorte, Cancer, Données Censurées, Identification Automatique

Abstract. We aimed to validate cancer occurrence identification by using multisources electronic health record (EHR) data in a type 2 diabetes patient cohort (SURDIAGENE cohort). The process follows three steps. The first one consists in collecting EHR ie hospitalization letters, pathology reports and multidisciplinary consultation meetings reports. Data referred to cancer are also collected from PMSI (Programme de Médicalisation des Systèmes d'Information) system. The second stage is fully computerized. A decisional algorithm, including an automated reading of documents, identifies patients with potential history of cancer. Finally this is followed by cancers topology adjudication with practitioners specialists in oncology. Identified cases are compared to those recorded by the Poitou-Charentes Cancer Registry, member of the French network of cancer registries, which collects systematically all incident cases of cancers since 2008 in the Poitou-Charentes area. This process shows a good performance, with sensitivity (91%) and specificity (98%). We planned to extend this process for the research of other major health events.

Keywords. Cohort, Cancer, Censored Data, Automatic Identification

1. Introduction

Les sources de données disponibles sont de plus en plus riches et complexes notamment dans le milieu médical. Cela donne lieu au développement de nombreuses démarches de traitement des Big Data et au développement de l'intelligence artificielle.

Plus particulièrement, en France, une volonté forte d'ouverture de base de données a été portée par la loi de modernisation du système de santé français du 26 janvier 2016. Cette loi détermine les contours du Système National de Données de Santé (SNDS) permettant à des structures identifiées d'avoir accès aux données de santé telles que celles de l'assurance maladie (SNIIRAM), des hôpitaux (Programme de Médicalisation des Systèmes d'Information PMSI), et des causes de décès (Centre d'épidémiologie sur les causes médicales de Décès CépiDC).

L'augmentation des sources d'information devrait permettre de mieux phénotyper les patients inclus dans les études médicales. Toutefois, il est important de valider des algorithmes d'identification des maladies. C'est un travail porté nationalement notamment par la Caisse Nationale de l'Assurance Maladie (CNAM). Cependant cela ne peut pas être réalisé à l'échelle locale à partir des données disponibles uniquement dans les établissements de soins.

Dans ce contexte, nous avons souhaité développer un outil permettant d'identifier des événements survenus au cours de suivi de patients inclus dans une cohorte du Centre d'Investigation Clinique (CIC) 1402 du Centre Hospitalier Universitaire (CHU) de Poitiers.

2. Méthodes

Afin de valider la performance du processus d'identification, nous avons besoin d'une source indépendante recherchant les mêmes événements sur les mêmes patients. C'est pourquoi nous avons fait le choix de développer l'outil en recherchant les événements de cancer, dans une cohorte de diabétique de type 2 suivis au CHU de Poitiers et donc habitant à proximité. La pertinence de notre identification des cas de cancers pouvait ainsi être confrontée aux données du Registre Général des Cancers Poitou Charentes (RGCP).

2.1 Le registre général des cancers Poitou Charentes

Depuis 2008, le RGCP, membre du réseau FRANCIM (réseau français des registres des cancers), identifie de façon standardisée et exhaustive les nouveaux cas de cancers au sein de la population domiciliée dans les quatre départements de l'ancienne région Poitou-Charentes. Il a pour objectif de fournir des indicateurs épidémiologiques de référence utiles à la connaissance et à la prise en charge des cancers, de permettre de décrire la dynamique d'évolution des cancers, d'évaluer l'état de santé de la population générale et d'identifier les principaux déterminants associés à la survenue de ces maladies, à l'aggravation ou à l'importance de leur retentissement, sous le patronage de l'institut de veille sanitaire (InVS) et l'institut national du cancer (INCa).

Trente localisations de cancers, définies par la classification internationale des maladies pour l'oncologie, CIM-O, ont été considérées, seuls les cancers invasifs ont été inclus dans l'étude. De fait, les tumeurs in situ, les lésions précancéreuses, ainsi que les lésions cutanées autres que le mélanome, n'ont pas été comptabilisées.

2.2 La cohorte SURDIAGENE

La cohorte SURDIAGENE décrite dans l'article de Hadjadj (2008) est une cohorte monocentrique de patients diabétiques de type 2 recrutés entre 2002 et 2012 au CHU de Poitiers. Son objectif est d'identifier les déterminants génétiques et environnementaux des complications microvasculaires et macrovasculaires de patients diabétiques de type 2. Les principaux critères d'exclusion sont de vivre en dehors de la région de Poitiers et d'avoir une maladie rénale non liée au diabète.

Au sein de la cohorte SURDIAGENE, dix évènements dont le décès, les complications macrovasculaires et microvasculaires font déjà parti d'un processus d'adjudication réalisé tous les deux ans depuis 2007. Les évènements sont identifiés à partir des comptes rendus d'hospitalisation et des appels aux médecins traitant.

Grâce à une autorisation de la Commission Nationale de l'Informatique et des Libertés (CNIL) obtenue en 2017, nous avons pu faire le lien entre la cohorte SURDIAGENE et les patients identifiés par le RGCPC grâce à plusieurs éléments d'identito-vigilance incluant le nom, le prénom, le sexe, la date de naissance et le code de la commune de naissance de chacun des patients.

2.3 Processus d'identification des évènements cancéreux

Les cancers recherchés par notre processus d'identification et leurs caractéristiques étaient les mêmes que ceux définis par le RGCPC.

Les données sources utilisées étaient des informations disponibles dans le système informatique du CHU de Poitiers. Nous avons collecté l'ensemble des comptes rendus de Réunions de Concertation Pluridisciplinaire (RCP) réalisées par le réseau 3C, tous les comptes rendus d'examens anatomo-pathologie, l'ensemble des comptes rendus d'hospitalisations et de consultations réalisées au CHU. De plus par le biais du médecin du Département d'Information Médicale (DIM) de l'établissement nous avons récupéré toutes les venues à l'hôpital pour lesquelles un code PMSI lié au cancer a été saisi.

Nous avons également identifié une liste de 26 mots (ex : tumeur), préfixe (exemple : cancé*) et suffixe (exemple : *ome) pouvant évoquer des cancers. A partir de ces sources d'information, nous avons demandé à un programme informatique développé en interne de lire tous les documents et d'identifier ceux pour lesquels un mot en lien avec le cancer était précisé.

A la suite de cela, un comité d'adjudication composé de deux médecins spécialisés en oncologie validait les documents identifiés par le programme afin de caractériser chaque cas de cancers, notamment leurs localisations. Un cas était adjudiqué quand il y avait consensus entre les deux experts médicaux.

3. Résultats

Nous avons travaillé sur 1 468 dossiers de patients diabétiques de type 2 de la cohorte SURDIAGENE suivis de 2002 à décembre 2015. Cette population comporte 620 (42%) femmes avec un âge à l'inclusion de 65 ± 11 ans en moyenne. Le suivi médian des patients est de 7,1 ans. Nous cherchions à identifier des évènements de cancer avant l'inclusion dans la cohorte (antécédents) et au cours du suivi.

Nos documents sources étaient composés de 65 685 courriers médicaux, de 241 comptes rendus de RCP, de 2 539 bilans d'anatomo-pathologie et de 742 codes PMSI en lien avec le cancer.

Grâce à ces différentes sources d'informations, 1 289 documents de patients ont été identifiés comme portant une notion de cancer par le processus automatisé.

Après adjudication par les médecins spécialistes, nous avons adjudiqué 595 cancers chez 329 patients de la cohorte SURDIAGENE.

En moyenne les patients ayant un cancer adjudiqué avaient été identifiés par l'outil automatisé, à partir de 4 sources différentes contre 2 pour les patients non cancéreux.

Sur la période 2008-2015, couverte également par le RGCPC, nous avons identifié 183 cas incidents de cancer contre 152 cas incidents pour le registre. Notre processus d'adjudication

retrouvait 138 cas (90%) en concordance avec le registre.

Le RGCCPC identifiait sur cette même période 144 patients de notre cohorte avec au moins un cancer incident. L'adjudication mise en place par le CIC retrouvait 167 patients. Treize patients identifiés par le registre n'ont pas été retrouvés par le processus d'adjudication, même si ces dossiers avaient été identifiés par le processus automatisé. Notre processus d'adjudication des patients ayant un événement cancéreux incident a une sensibilité de 91% et une spécificité de 98%.

4. Conclusion

Dans un contexte d'ouverture des bases de données médicales, il est important de valider nos procédures d'identification des événements d'intérêt et notamment de pouvoir développer des outils permettant le travail sur les données censurées. En couplant différentes sources de données internes, le CIC 1402 est capable d'avoir une puissance d'identification comparable au Registre des cancers pour des patients suivis localement. Nous travaillons sur la généralisation de ce processus à d'autres événements et à d'autres cohortes de patients.

Bibliographie

LOI n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé - Article 193

Hadjadj et al (2008). Prognostic Value of the Insertion/Deletion Polymorphism of the ACE Gene in Type 2 Diabetic Subjects, *Diabetes Care* 31:1847–1852.

Binder-Foucard F et al (2014), Cancer incidence and mortality in France over the 1980-2012 period: solid tumors. *Revue Epidemiol Sante Publique* 62:95-108.

Le Guyader-Peyrou S et al (2016) Cancer incidence in France over the 1980-2012 period : Hematological malignancies. *Rev Epidemiol Sante Publique* 64:103-12.