THÈSE

Pour l'obtention du grade de
DOCTEUR DE L'UNIVERSITÉ DE POITIERS
UFR de médecine et de pharmacie
Centre d'investigation clinique - CIC (Poitiers)
(Diplôme National - Arrêté du 25 mai 2016)

École doctorale : Sciences Biologiques et Santé (Limoges)
Secteur de recherche : Recherche clinique, Innovation technologique et thérapeutique, Santé Publique

Présentée par :
Gautier Defossez

Le système d'information multi-sources du Registre général des cancers de Poitou-Charentes. Conception, développement et applications à l'ère des données massives en santé

Directeur(s) de Thèse :
Pierre Ingrand, François Guilhot-Gaudeffroy

Soutenue le 01 juillet 2021 devant le jury

Jury :

Président	**Marc Cuggia**	Professeur et praticien hospitalier, Université de Rennes
Rapporteur	**Catherine Quantin**	Professeur et praticien hospitalier, Université de Bourgogne
Rapporteur	**Pascale Grosclaude**	Praticien hospitalier, Institut Claudius Regaud, Toulouse
Membre	**Pierre Ingrand**	Professeur et praticien hospitalier, Université de Poitiers
Membre	**François Guilhot-Gaudeffroy**	Professeur et praticien hospitalier, Université de Poitiers
Membre	**Virginie Migeot**	Professeur et praticien hospitalier, Université de Poitiers

Pour citer cette thèse :
Gautier Defossez. *Le système d'information multi-sources du Registre général des cancers de Poitou-Charentes. Conception, développement et applications à l'ère des données massives en santé* [En ligne]. Thèse Recherche clinique, Innovation technologique et thérapeutique, Santé Publique. Poitiers : Université de Poitiers, 2021. Disponible sur Internet <http://theses.univ-poitiers.fr>

THÈSE

Pour l'obtention du grade de
DOCTEUR DE L'UNIVERSITÉ DE POITIERS
UFR de Médecine et de Pharmacie
Centre d'investigation clinique CIC 1402 INSERM (Poitiers)
(Diplôme National - Arrêté du 25 mai 2016)
École doctorale : Sciences Biologiques et Santé (SBS 615)
Secteur de recherche : Recherche clinique, Innovation technologique et thérapeutique, Santé Publique

Présentée par :

Gautier DEFOSSEZ

Le système d'information multi-sources du Registre général des cancers de Poitou-Charentes. Conception, développement et applications à l'ère des données massives en santé.

Directeur de Thèse :

Pierre INGRAND

Co-Directeur de Thèse :

François GUILHOT

Soutenance le 1er Juillet 2021 devant le jury

Pr Catherine QUANTIN, PU-PH, Rapporteur, CHU de Dijon Bourgogne
Dr Pascale GROSCLAUDE, PH, Rapporteur, Institut Claudius Regaud Toulouse
Pr Marc CUGGIA, PU-PH, Examinateur, Université de Rennes
Pr Virginie MIGEOT, PU-PH, Examinateur, Université de Poitiers
Pr Pierre INGRAND, PU-PH, Directeur de thèse, Université de Poitiers
Pr François GUILHOT, PU-PH, Co-directeur de thèse, Université de Poitiers

Résumé

Les registres du cancer sont au plan international l'outil de référence pour produire une vision exhaustive (non biaisée) du poids, de la dynamique et de la gravité du cancer dans la population générale. Leur travail de classification et de codage des diagnostics selon des normes internationales confère aux données finales une qualité spécifique et une comparabilité dans le temps et dans l'espace qui les rendent incontournables pour décrire l'évolution et la prise en charge du cancer dans un environnement non contrôlé. Leur travail repose sur un processus d'enquête rigoureux dont la complexité est largement dépendante des capacités à accéder et à rassembler efficacement toutes les données utiles concernant un même individu. Créé en 2007, le Registre Général des Cancers de Poitou-Charentes (RGCPC) est un registre de génération récente, débuté à une période propice à la mise en œuvre d'une réflexion sur l'optimisation du processus d'enregistrement. Porté par l'informatisation des données médicales et l'interopérabilité croissante des systèmes d'information, le RGCPC a développé et expérimenté sur 10 ans un système d'information multi-sources associant des méthodes innovantes de traitement et de représentation de l'information fondées sur la réutilisation de données standardisées produites pour d'autres finalités.

Dans une première partie, ce travail présente les principes fondateurs et l'implémentation d'un système capable de rassembler des volumes élevés de données, hautement qualifiantes et structurées, et rendues interopérables sur le plan sémantique pour faire l'objet d'approches algorithmiques. Les données sont collectées pluri annuellement auprès de 110 partenaires représentant sept sources de données (cliniques, biologiques et médico-administratives). Deux algorithmes assistent l'opérateur du registre en dématérialisant une grande partie des tâches préalables à l'enregistrement des tumeurs. Un premier algorithme crée les tumeurs et leurs caractéristiques (publication), puis un 2ème algorithme modélise le parcours de soin de chaque individu selon une séquence ordonnée d'évènements horodatés consultable au sein d'une interface sécurisée (publication). Des approches de machine learning sont testées pour contourner l'éventuelle absence de codification des prélèvements anatomopathologiques (publication).

La deuxième partie s'intéresse au large champ de recherche et d'évaluation rendu possible par la disponibilité de ce système d'information intégré. Des appariements avec d'autres données de santé ont été testés, dans le cadre d'autorisations réglementaires, pour enrichir la contextualisation et la connaissance des parcours de soins, et reconnaître le rôle stratégique des registres du cancer pour l'évaluation en « vie réelle » des pratiques de soins et des services de santé (preuve de concept) : dépistage, diagnostic moléculaire, traitement du cancer, pharmaco épidémiologie (quatre publications principales). L'appariement des données du RGCPC à celles du registre REIN (insuffisance rénale chronique terminale) a constitué un cas d'usage veillant à expérimenter un prototype de plateforme dédiée au partage collaboratif des données massives en santé (publication).

La dernière partie de ce travail propose une discussion ouverte sur la pertinence des solutions proposées face aux exigences de qualité, de coût et de transférabilité, puis dresse les perspectives et retombées attendues pour la surveillance, l'évaluation et la recherche à l'ère des données massives en santé.

Mots-clés : Système d'information, algorithme, registre, cancer, efficience, données massives

Abstract

Population-based cancer registries (PBCRs) are the best international option tool to provide a comprehensive (unbiased) picture of the weight, incidence and severity of cancer in the general population. Their work in classifying and coding diagnoses according to international rules gives to the final data a specific quality and comparability in time and space, thus building a decisive knowledge database for describing the evolution of cancers and their management in an uncontrolled environment. Cancer registration is based on a thorough investigative process, for which the complexity is largely related to the ability to access all the relevant data concerning the same individual and to gather them efficiently. Created in 2007, the General Cancer Registry of Poitou-Charentes (RGCPC) is a recent generation of cancer registry, started at a conducive time to devote a reflection about how to optimize the registration process. Driven by the computerization of medical data and the increasing interoperability of information systems, the RGCPC has experimented over 10 years a multi-source information system combining innovative methods of information processing and representation, based on the reuse of standardized data usually produced for other purposes.

In a first section, this work presents the founding principles and the implementation of a system capable of gathering large amounts of data, highly qualified and structured, with semantic alignment to subscribe to algorithmic approaches. Data are collected on multiannual basis from 110 partners representing seven data sources (clinical, biological and medical administrative data). Two algorithms assist the cancer registrar by dematerializing the manual tasks usually carried out prior to tumor registration. A first algorithm generate automatically the tumors and its various components (publication), and a second algorithm represent the care pathway of each individual as an ordered sequence of time-stamped events that can be access within a secure interface (publication). Supervised machine learning techniques are experimented to get around the possible lack of codification of pathology reports (publication).

The second section focuses on the wide field of research and evaluation achieved through the availability of this integrated information system. Data linkage with other datasets were tested, within the framework of regulatory authorizations, to enhance the contextualization and knowledge of care pathways, and thus to support the strategic role of PBCRs for real-life evaluation of care practices and health services research (proof of concept): screening, molecular diagnosis, cancer treatment, pharmacoepidemiology (four main publications). Data from the RGCPC were linked with those from the REIN registry (chronic end-stage renal failure) as a use case for experimenting a prototype platform dedicated to the collaborative sharing of massive health data (publication).

The last section of this work proposes an open discussion on the relevance of the proposed solutions to the requirements of quality, cost and transferability, and then sets out the prospects and expected benefits in the field of surveillance, evaluation and research in the era of big data.

Keywords: Information system, algorithm, registry, cancer, efficiency, big data

Remerciements

Au Professeur Catherine QUANTIN, qui me fait l'honneur d'être rapporteuse de cette thèse. Merci de l'intérêt porté à ce travail et d'avoir accepté d'en être rapporteuse. Ce travail trouve une origine prononcée dans l'intérêt porté précocement à l'utilisation de l'information médicale et des données PMSI, et qui résonne pour moi immédiatement avec les journées ADELF-EMOIS qui ne vous sont pas inconnues. Merci beaucoup pour votre disponibilité.

Au Docteur Pascale GROSCLAUDE. Merci infiniment Pascale d'avoir accepté d'être rapporteuse de cette thèse. Ce travail a beaucoup de sens à pouvoir bénéficier de ton expertise, au vu de ton expérience, de ton implication et de ta passion perceptible pour les registres. Plus généralement, merci d'avoir toujours été disponible et de continuer à prendre le temps nécessaire pour des suggestions, commentaires et autres remarques judicieuses à chacune de mes sollicitations.

Au Professeur Marc CUGGIA, pour avoir accepté de participer à mon jury de thèse. Merci infiniment Marc pour les encouragements et la confiance témoignée, pour ces collaborations fructueuses passées et à venir et pour cette ambiance d'échange et de recherche toujours très agréable.

Au Professeur Virginie MIGEOT. Merci Virginie d'avoir accepté de participer à mon jury de thèse. C'est un plaisir de t'avoir parmi les examinateurs et de pouvoir bénéficier à nouveau de ton expérience et de tes conseils. Inexorablement, cela me renvoie quelques années en arrière et me rappelle les bons moments partagés, ceux-là mêmes qui m'ont permis d'acquérir un certain nombre de REPERES utiles ! Merci pour ta disponibilité.

Au Professeur Pierre INGRAND, qui m'a fait le plaisir et l'honneur d'être directeur de cette thèse. Merci infiniment pour l'encadrement de ce travail et pour ce soutien et cette patience incomparables. Plus encore, merci d'avoir toujours eu confiance en moi et de m'avoir accompagné pendant toutes ces années, pour ce temps et ces relectures toujours méticuleuses et d'une grande aide. Cette thèse est le fruit d'une longue collaboration et qui vient marquer de belles années passées. Je vous en suis très reconnaissant.

Au Professeur François GUILHOT, qui m'a fait l'honneur d'être co-directeur de cette thèse et de m'avoir accueilli dans son laboratoire. Merci Pr GUILHOT pour l'intérêt porté à ce travail et votre confiance, pour ces moments d'échange et de recherche très appréciables et pour votre accompagnement dans ces travaux. Merci de m'avoir notamment invité à découvrir les remarquables locaux de l'Académie Nationale de Médecine et y défendre ce qui m'anime depuis plusieurs années.

Au Professeur René ROBERT et au Professeur Pierre-Jean SAULNIER, de m'avoir accueilli dans votre laboratoire et d'avoir été présents et disponibles les moments souhaités.

A tous les nombreux partenaires et collaborateurs du registre qui, par leur pleine implication, jouent un rôle essentiel dans l'exécution du registre et la surveillance épidémiologique qui en découle. Soyez en pleinement remerciés.

A nos tutelles et financeurs et à tous les généreux donateurs de la ligue contre le cancer (Charente, Charente-Maritime, Deux-Sèvres et Vienne), sans qui, ces travaux n'auraient pu s'immiscer et se pérenniser.

A tous ceux qui ont contribué, de près ou de loin, à ce travail et à son aboutissement, pleinement convaincu que cette thèse est loin d'être un travail solitaire: Nicolas M, Isabelle, Soizic, Florian, Tiffanie, Aurélie, Valérie M, Nicolas P, Chaïma, Wendy, Delphine, Valérie J, Thomas, Mathieu, Violaine, Amine, Alexandre Q, Alexandre R, Sébastien, Louis. Merci à vous tous pour votre confiance, votre soutien, votre gentillesse et votre implication, actuelle ou passée.

A Vianney Jouhet, pour ces moments d'une intensité rare et d'une productivité hors normes, pour toutes ces lignes de programmes et ces moments partagés, tous ces moments qui ont solidement inspiré cette thèse. Au plaisir de te revoir bientôt.

A Pierre-Jean, Stéphanie, Elodie, Sophie, Michèle pour leur accueil au CIC.

A ma famille, à mes amis pour leur soutien indéfectible, et cette question récurrente bien qu'angoissante en période de doutes « Alors tu la soutiens quand cette thèse ? », merci à vous qui m'avez permis de ne jamais dévier de mon objectif et de transformer les moments laborieux en bons souvenirs.

A mon épouse et mes 2 (grands) loulous … merci à tous les trois pour votre patience, votre amour et votre soutien permanent. Vous m'avez été d'une GRANDE aide, et êtes sans aucun doute ce qu'il y a de plus précieux.

Papa, cette thèse est pour toi.

Table des matières

Liste des abréviations

ACP : Anatomie et Cytologie Pathologiques

ADICAP : Association pour le Développement Informatique en Anatomo-Pathologie

AFAQAP : Association Française d'Assurance Qualité en Anatomie et Cytologie Pathologiques

ALD : Affection Longue Durée

AMA : Assurance Maladie

ARS : Agence Régionale de Santé

ATIH : Agence Technique de l'Information Hospitalière

BIO : Laboratoires d'hématologie biologique

CCAM : Classification Commune des Actes Médicaux

CépiDc : Centre d'épidémiologie sur les causes médicales de décès

CER : Comité d'Evaluation des Registres

CDC: Centers for Disease Control and Prevention

CIM-10: Classification Internationale des Maladies ($10^{ème}$ révision)

CIM-O3 : Classification Internationale des Maladies en Oncologie ($3^{ème}$ édition)

CIRC : Centre International de Recherche sur le Cancer

CNR : Comité National des Registres

CRAP : Compte rendu d'anatomie et de cytologie pathologiques

CRISAP : Centre de Regroupement Informatique et Statistique en Anatomie Pathologique

CRLCC : Centre Régional de Lutte Contre le Cancer

CSR : Comité Stratégique des Registres

DAS : Diagnostic Associé Significatif

DCC : Dossier Communicant de Cancérologie

DO : Dépistage organisé

DP : Diagnostic Principal

DR : Diagnostic relié

EPC : Enquête Permanente Cancer

FRANCIM : Réseau français des registres du cancer

HCL: Hospices Civils de Lyon

HAS: Haute Autorité de Santé

HDH : Health Data Hub

INCa: Institut National du Cancers

INSERM: Institut National de la Santé et de la Recherche Médicale

INSEE: Institut National de la Statistique et des Etudes Economiques

NAACCR: North American Association of Central Cancer Registries

NCI : National Cancer Institute

NPCR: National Program of Cancer Registries

OCDE : Organisation de Coopération et de Développement Economiques

OMS: Organisation Mondiale de la Santé

PBCR : Population-based cancer registries

PMSI : Programme de Médicalisation des Systèmes d'Information

PTP : Programme de Travail Partenarial

RCP : Réunion de Concertation Pluridisciplinaire

RGCPC: Registre Général des Cancers de Poitou-Charentes

RUM : Résumé d'Unité Médicale

RSA : Résumé de Sortie Anonymisé

SEER: Surveillance, Epidemiology and End Results

SI-RGCPC : Système d'Information du Registre Général des Cancers de Poitou-Charentes

SNDS : Système National des Données de Santé

SNIIRAM : Système National d'Information Inter-Régimes de l'Assurance Maladie

SpF: Santé publique France

1. Introduction

1.1. L'observation des cancers

Pendant des siècles, les seules informations sur la survenue du cancer provenaient des dossiers médicaux et des rapports d'autopsie, et à l'échelle de la population, des données relatives aux causes de décès. En général, on peut s'attendre à des taux d'incidence et de mortalité associée au cancer quasiment identiques si l'issue de la maladie est systématiquement fatale, comme c'était malheureusement le cas jusqu'au milieu du 20ème siècle. Mais depuis l'introduction de traitements efficaces, les taux de mortalité ne reflètent plus les taux d'incidence. C'est pourquoi il est devenu impératif de consigner chaque nouveau cas de cancer afin d'obtenir un tableau clair du fardeau représenté par la maladie dans une population [1,2].

Dans leur introduction au premier volume de la série Cancer Incidence in Five Continents (CIV), Doll et al. discutent de l'intérêt des comparaisons de la fréquence des maladies, entre différents endroits et au fil du temps, dans le développement de la connaissance des causes du cancer [3]. Ils concluent que parmi les statistiques disponibles pour l'étude du cancer, les données les plus précieuses sont, sans aucun doute, les taux obtenus en enregistrant la survenue de chaque cas de cancer sur une période et un territoire déterminés, selon une approche fondée sur des données probantes et recueillies en population générale (*i. e.* non sélectionnée).

1.2. Les registres du cancer

Cette production d'indicateurs épidémiologiques est la fonction fondamentale des registres du cancer basés sur la population (ou registres de population ; population-based cancer registries (PBCR) en anglais), devenus au plan international l'outil de référence en ce qui concerne les informations sur le nombre de nouveaux cas de cancer (incidence), le nombre de personnes vivant avec le cancer (prévalence), ainsi que la létalité (mortalité) et la gravité (survie) du cancer dans les populations qu'ils couvrent [4]. Depuis les premiers registres établis en Europe et en Amérique du Nord à partir des années 1930, leur utilisation pour fournir des informations sur d'autres aspects du cancer et sur le contrôle de la maladie s'est progressivement développée, là où il n'existait aucune autre source d'information susceptible de fournir une vision non biaisée du poids du cancer dans la population. Au-delà de leurs missions de surveillance et d'observation, ils constituent aujourd'hui une base de connaissances indispensable pour la conception et l'évaluation des plans de lutte contre le cancer : évaluation de l'efficacité des interventions de prévention primaire (à partir de l'étude des taux d'incidence observés après l'introduction des programmes), évaluation et suivi des programmes de détection précoce et de dépistage (à partir de l'étude des variations temporelles en termes d'incidence

pour les cancers dont le dépistage permet de prévenir le maladie invasive comme le cancer du col de l'utérus, ou en termes de mortalité et de pronostic (stades d'extension) pour les programmes permettant de détecter de façon précoce les cancers invasifs comme les cancers du sein ou du côlon), évaluation du traitement du cancer (à partir de la mesure de la survie au niveau de la population).
L'augmentation constante du nombre de registres du cancer atteste aujourd'hui de leur valeur dans la recherche et le contrôle du cancer. Le nombre de registres et de populations couvertes a été multiplié par dix entre 1966 et 2013 [5,6]. Toutefois, en dépit d'important progrès, les registres du cancer ne couvrent en 2018 encore que 21% de la population mondiale (Annexe 1), avec une grande disparité entre les pays [7,8] : allant d'une couverture complète de l'Amérique du Nord et de l'Océanie, à près de 60% de la population Européenne [8], et jusqu'à atteindre une couverture aux alentours de 10% pour l'Afrique, l'Amérique du Sud ou encore l'Asie. Les registres capables de fournir des données d'incidence de qualité et comparables (publiées dans CIV) sont cependant moins nombreux, notamment en Afrique avec seulement 23% des registres inclus (*vs.* 89% pour l'Europe et 97% pour l'Amérique du Nord) [7]. La cartographie de l'enregistrement du cancer en Europe fait apparaître également des disparités importantes en termes de couverture de la population par les registres du cancer, la qualité des données et la production de données [9]. Leur présence et leur fonctionnement à l'échelle d'un pays se trouvent conditionnés finalement par le contexte politique, économique et de santé publique du pays, et par l'organisation du système de santé dans lequel ils opèrent et ses dimensions juridiques. Cette hétérogénéité a favorisé une certaine variabilité dans les pratiques et l'émergence de différents modèles de registres.

1.3. Quels modèles en France et à l'étranger ?

En pratique, le travail d'un registre de population est grandement facilité lorsqu'il accède efficacement à toutes les données utiles concernant un même individu. Le registre doit s'attacher à avoir accès à un maximum de sources de données (notifications) et à mettre en œuvre des procédures efficaces de traitement de ces données afin de fournir avec exactitude (qualité, exhaustivité, comparabilité, ponctualité) [10,11] les informations minimales et obligatoires pour l'analyse de l'incidence et de la survie par cancer [4,12]. Ces données « d'informations de base » sont consignées en application de standards internationaux de classification et de codage (normalisation du diagnostic) de sorte que les données soient comparables d'un endroit à l'autre à l'échelle internationale. Ainsi lorsque les cas ont été identifiés, les activités du registre du cancer sont universelles. Vouloir traiter quelques centaines de nouveaux cas par an ou une dizaine de milliers revient pour les registres du cancer à exécuter les mêmes tâches opérationnelles - seules les méthodes de recueil et de traitement de l'information en amont diffèrent.

Traditionnellement, les méthodes de recueil de l'information sont classées en 2 catégories : active et passive. Le recueil actif (collecte des données à la source) implique que le personnel du registre se déplace dans les services (hôpitaux, laboratoires) et collecte les informations nécessaires sur des formulaires spéciaux ou obtienne des copies des documents utiles. Dans le cas du recueil passif, c'est le personnel des services de santé qui renseigne les formulaires de notification et les transmet au registre, ou lui envoie les copies des résultats d'analyses et de prélèvements, rapports de sortie, etc… à partir desquels les données nécessaires peuvent être extraites. En pratique, un mélange de ces deux systèmes est fréquemment utilisé comme, par exemple, la réception passive des copies des compte rendus anatomopathologiques mentionnant un cancer, complétée par des visites actives dans les hôpitaux pour la consultation des dossiers médicaux.

- **En Europe du Nord**

Précurseurs dans le domaine des registres de morbidité, le fonctionnement dans les pays de l'Europe du Nord (Norvège, Suède, Finlande, Islande, Danemark, Iles Féroé, Groenland, qui totalisent une population de près de 20 millions d'habitants) repose sur un recueil passif des cas (*i. e.* tous les médecins ont l'obligation légale de fournir des informations sur chaque cas ou suspicion de cancer au registre). De plus, le numéro national d'identité des patients figure depuis les années 1980 dans les registres des cancers (ce n'est pas encore le cas en France) et permet de collecter à propos d'un même individu des informations d'origines diverses d'une rare valeur (données hospitalières, données des laboratoires de pathologie et d'analyses médicales, certificats de décès, registres de naissance, registres de professions …), facilitant ainsi grandement les pratiques d'enregistrement et satisfaisant l'intelligence curieuse des chercheurs et les capacités à mener une recherche souvent puissante et d'envergure [13–19].

- **Aux Etats-Unis**

Deux agences fédérales se coordonnent afin de disposer d'une infrastructure nationale de surveillance du cancer : le National Cancer Institute (NCI) et les Centers for Disease Control and Prevention (CDC). Le « Surveillance, Epidemiology, and End Results (SEER) » Program[1], sous l'égide du NCI, est le système de surveillance faisant autorité sur l'incidence et la survie du cancer aux Etats-Unis. Il a été établi en 1973 (National Cancer Act) et couvre aujourd'hui environ 34% de la population à partir de 16 registres du cancer stratégiquement situés sur l'ensemble des États-Unis. Le NCI's SEER Program enregistre environ 550 000 nouveaux cas tous les ans. Le « National Program of Cancer Registries » (NPCR)[2] quant à lui, sous l'égide du CDC, a été amendé par le Congrès dès 1992 (Cancer

[1] SEER Program : https://seer.cancer.gov/about/overview.html
[2] NPCR: https://www.cdc.gov/cancer/npcr/about.htm

Registries Amendment Act) et a planifié le déploiement progressif de registres centraux du cancer dans tous les états. Le NPCR couvre aujourd'hui 97% de la population des Etats-Unis et enregistre un peu plus de 1,7 million de nouveaux cas tous les ans. Ensemble, le « NCI's SEER Program » et le « CDC's NPCR » recueillent des données pour l'ensemble de la population américaine, et de façon uniforme selon les standards édités par la NAACCR (North American Association of Central Cancer Registries) de sorte que les deux programmes fédéraux soient comparables [20,21]. A la différence des registres d'Europe du Nord, la notification des cas est établie par du personnel qualifié (*highly trained cancer registrars*) directement implanté dans les hôpitaux et cliniques, et qui transmettent leurs données au registre central du cancer (au niveau de l'état donc). En parallèle, les établissements médicaux tels que les hôpitaux, les cabinets médicaux et les laboratoires de pathologie communiquent également leurs informations sur les cas de cancer au registre central du cancer. Chaque année, les registres centraux du cancer transmettent par voie électronique les informations démographiques et cliniques sur l'incidence du cancer au « Department of Health and Human Services » du CDC, chargé de fournir les statistiques fédérales officielles [3].

- **Au Canada et en Italie**

Certains pays, en l'absence d'obligation légale de déclaration des diagnostics de cancer, proposent depuis plusieurs années des algorithmes visant à remplacer le processus manuel de notification généralement effectué par les professionnels de santé ou le personnel des registres [14,22–25].

Ces algorithmes ont été introduits pour la première fois au début des années 1970 par le registre des cancers de l'Ontario (Canada). Les diagnostics de cancer sont attribués par programme en appliquant un ensemble de règles décisionnelles (*case resolution system*) visant à produire et à enregistrer le "meilleur" diagnostic à partir des données disponibles préalablement reliées à un individu [22,26,27].

Au début des années 1990, le registre des tumeurs de la région de Vénétie (Italie) a exploré la possibilité d'utiliser les données électroniques de routine des hôpitaux, des services de pathologie et des certificats de décès, codées selon la CIM-9 (certificats de sortie d'hôpital et de décès) ou SNOMED (dossiers de pathologie) [25,28,29]. En appliquant un système décisionnel binaire de concordance/discordance, un cas incident potentiel était accepté avec un diagnostic consolidé du cancer, ou rejeté. Les cas rejetés par le programme (environ 45%) étaient résolus manuellement par le personnel du registre. Cette méthodologie (*Open Registry*) a été adoptée plus tard par le registre des cancers de Varèse (Italie), puis par le registre des cancers d'Irlande du Nord (NICR), et partiellement par le registre des cancers de la Tamise (Londres), avec des résultats similaires [23,24].

[3] U.S. Cancer Statistics : https://www.cdc.gov/cancer/uscs/about/index.htm

Ces registres ont finalement adopté leurs propres règles décisionnelles pour notifier et/ou autoriser ou non l'enregistrement automatique des nouveaux cas de cancers, et proposent ainsi de réduire la part manuelle et les délais de production. Le registre des cancers de Varèse, qui propose un enregistrement automatique pour 59% des cas, relève une réduction notable de la charge de travail et une réduction globale des coûts estimée à 40 % [24]. En revanche, ces approches algorithmiques soulèvent aussi certaines limites et prérequis à leur application :

- Les ressources humaines nécessaires dépendent de la simplicité du degré d'intégration des données dans le système d'information, et en particulier du degré de normalisation des fichiers sources (*e. g.* à Varèse le principal problème était la présence de cinq systèmes d'archivage électronique différents pour les données de pathologie, imposant d'écrire cinq programmes ad hoc pour extraire les fichiers et les traduire en une forme normalisée pour le registre) ;
- Des informations de faible qualité augmentent la proportion des cas discordants, et, de fait, sont responsables d'une moindre efficience du système ;
- Le personnel est toujours tenu de vérifier manuellement les enregistrements incertains afin de préserver la qualité des données. L'enregistrement de certains items complémentaires comme l'extension tumorale, certains paramètres biologiques ou cliniques peut s'avérer dès lors délicat lorsque ces informations ne sont présentes que dans des documents non structurés ;
- Ces méthodes algorithmiques impliquent un traitement séquentiel des données, qui impose d'être en possession de la totalité des informations au moment du traitement, et de tester les jeux de données au rythme des exports définis (au risque de produire des chiffres d'incidence sous-estimés et la mise en œuvre secondaire d'un rattrapage des cas).

Au-delà il demeure une certaine méconnaissance générale des processus algorithmiques mises en œuvre et de la productivité qui en résulte. Il s'impose alors d'examiner soigneusement l'ensemble de ces critères pour opérer un traitement optimal des données.

➢ En France

En France, le fonctionnement des registres de cancer est assez différent des modèles étrangers précédemment décrits. Le système de surveillance épidémiologique des cancers repose sur un réseau quadripartite animé par Santé publique France (SpF), l'Institut National du Cancer (INCa), les registres du cancer du réseau FRANCIM et le service de biostatistique - bioinformatique des Hospices Civils de Lyon (HCL). La création des différents registres s'est échelonnée au cours du temps, témoignant à la fois d'initiatives de recherche et des besoins mis en avant à l'époque par les politiques nationales de santé publique. C'est à partir de 1986, devant la multiplication progressive des registres en France, que l'Inserm et le Ministère de la Santé ont créé le Comité National des Registres (CNR).

Scindé depuis 2013 en 2 comités, le Comité Stratégique des Registres (CSR) a pour rôle de coordonner et orienter la politique des registres, et le Comité d'Evaluation des Registres (CER) d'évaluer leur fonctionnement (exhaustivité, adéquation entre les moyens envisagés et les finalités exposées, travaux entrepris dans le domaine de la recherche). Les registres de cancer sont 28 registres labellisés en 2020 recouvrant entre 21 et 24% de la population de France métropolitaine selon la localisation étudiée (Figure 1, Annexe 2).

Il n'y a pas actuellement un modèle de fonctionnement unique à tous les registres : certains s'appuient sur un recueil de données informatisé, tandis que d'autres, créés il y a plus longtemps, reposent implicitement sur un processus « manuel ». L'ADN commun à la plupart des registres français demeure l'instauration d'un retour au dossier médical systématique pour enregistrer chaque nouveau cas et assoir des données d'excellente qualité. Ils demeurent toutefois des organisations complexes, qui ont besoin de ressources humaines et d'infrastructures pour collecter les données sur les cas de cancer à partir d'un large éventail de données médicales et civiles. Ainsi malgré leur évidente utilité, leur rôle est parfois négligé et la complexité de leur fonctionnement interroge par la lourdeur de ses circuits, jusqu'à remettre parfois en question leur place dans le système de surveillance sanitaire actuelle, notamment à l'ère des données massives de santé (ou big data) devenues peu à peu une réalité en France [30]. L'informatisation des données médicales et l'interopérabilité des sources d'information constituent pour les registres, sous condition de leur accessibilité, une opportunité de moderniser leur fonctionnement et de répondre à de nouvelles missions.

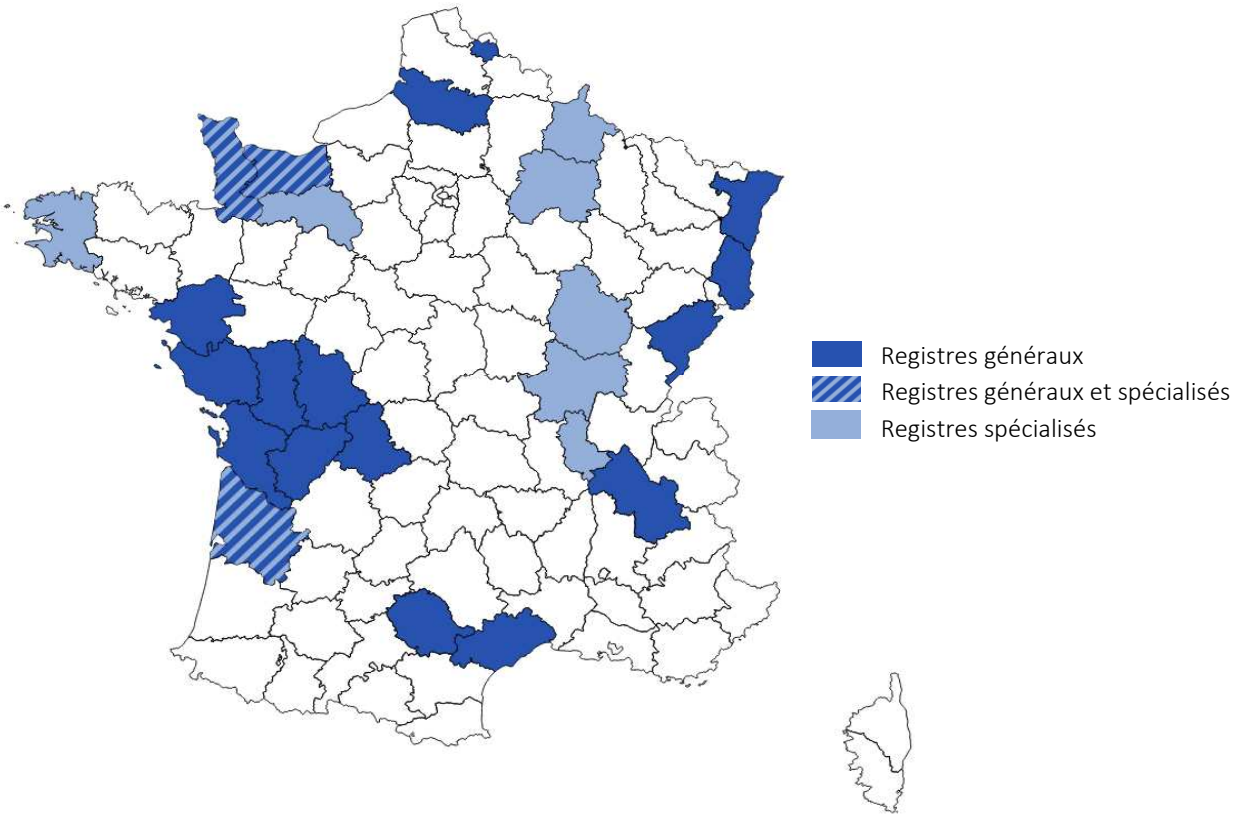


Figure 1 : Carte de l'ensemble des registres du cancer de France métropolitaine appartenant au réseau FRANCIM au 1er janvier 2021 (hors DOM-TOM et hors registre national des tumeurs de l'enfant).

1.4. Historique du registre général des cancers de Poitou-Charentes

L'optimisation de la validation de l'enregistrement a été un enjeu important au moment de la création du registre général des cancers de Poitou-Charentes (RGCPC). Porté par l'Agence Régionale de Santé Poitou-Charentes (aujourd'hui ARS Nouvelle-Aquitaine) et appuyé par la volonté des acteurs locaux, le RGCPC a été créé en 2007 avec comme ambition de se positionner d'emblée à l'échelon décisionnel de l'organisation des soins, *i. e.* la région (ex Poitou-Charentes, 1,8 million d'habitants pour une cible de 12 000 nouveaux cas annuels), et de jouer ainsi son rôle d'expert auprès de l'ARS dans la planification et l'évaluation des programmes de lutte contre le cancer.

Dans le même temps, l'augmentation de la couverture des registres des cancers était limitée par le coût important et pérenne qu'impliquait leur création. L'évaluation du premier plan cancer 2003-2007 [31] soulevait des insuffisances dans le système d'observation des cancers, en pointant les difficultés à mettre en œuvre un système d'épidémiologie national multi-sources d'une part, et celles à pouvoir étendre la couverture nationale par les registres du cancer à des zones urbaines très peuplées d'autre part. Des avancées substantielles avaient été relevées avec la création de deux des trois

nouveaux registres urbains prévues dans le plan, mais sur des zones cependant plus réduites : le registre de la Gironde (le plan prévoyait la région Aquitaine) et le registre de la zone de proximité de Lille (le plan prévoyait la région du Nord). Le registre du Val-de-Marne (le plan prévoyait l'Ile-de-France) s'est heurté quant à lui à des difficultés de collecte et de traitement de l'information face à la multiplicité des sources à solliciter, mais ce qui a obligé en contrepartie à développer une réflexion de la part des acteurs sur la mise en œuvre d'une automatisation du recueil et du traitement des données. Si ce dernier registre (*i. e.* Val-de-Marne) n'a finalement pas abouti, il a démontré l'intérêt d'une automatisation généralisée à toute la chaîne de production des données, nécessaire pour limiter les coûts en routine et raccourcir les délais de productions.

Dans une telle situation, l'enjeu était de moderniser les systèmes d'information des registres, en lien avec les avancées technologiques et l'interopérabilité croissante des systèmes d'information, afin d'en optimiser le fonctionnement et d'en accroître les domaines d'application et possibilités d'utilisation des données. Dans ce contexte, des procédures opératoires innovantes de recueil et de traitement de l'information, fondées sur la réutilisation secondaire de données médicales standardisées, ont été développées et expérimentées localement.

1.5. Objectifs

L'objectif de ce travail est de présenter la conception et les principales composantes du système d'information du registre général des cancers de Poitou-Charentes (SI-RGCPC), et d'en discuter les applications, perspectives et retombées attendues à l'ère des données massives en santé.

Ce travail est présenté en trois parties :

1. La première partie vise à décrire le cadre conceptuel et les méthodes de traitement de l'information mises en œuvre : définition des partenariats et complémentarité des sources d'information, description du processus d'automatisation du traitement des données (*i. e.* développements algorithmiques) préalable à l'enregistrement des cas (identitovigilance, notification automatisée des tumeurs, représentation du parcours de soins).
2. La seconde partie vise à illustrer l'intérêt d'une démarche intégrative de données (data linkage) au sein du registre des cancers à des fins d'évaluation et de recherche (health services research). Plusieurs applications et objectifs opérationnels sont abordés et déclinés selon la réutilisation i) de données collectées en routine par le RGCPC ou ii) de données recueillies spécifiquement en réponse à un objectif de recherche.
3. La troisième et dernière partie de ce travail a pour objectif de circonscrire la transférabilité et la généralisation de la méthode au regard de la réutilisation secondaire de données médicales standardisées et les gains d'efficience que cette informatisation procure, et d'en discuter les retombées attendues dans le système de surveillance actuel et les évolutions réglementaires.

1.6. Articles scientifiques

Les articles déduits de ces travaux originaux sont directement accessibles dans le corps de texte du document lorsque la contribution à l'obtention, à l'analyse et à l'interprétation des résultats a été substantielle et m'implique dans la majeure partie du travail scientifique : i) Deux articles en tant que 1er auteur sur le développement algorithmique du parcours de soins (méthode) et son application pratique enrichie des données des structures de gestion du dépistage organisé ; ii) Six articles en tant que 2ème auteur pour une participation active à 2 articles de méthodes sur la catégorisation automatique des comptes rendus d'anatomopathologie et la conception de l'algorithme de notification des tumeurs, et 4 articles sur la réutilisation de données externes visant à enrichir le parcours de soins (diagnostic moléculaire, traitement du cancer et pharmaco épidémiologie) et à expérimenter un prototype de plateforme dédiée au partage de données massives en santé. Au-delà, trois articles significatifs sont référencés compte tenu de mon implication et/ou la mise en œuvre conceptuelle de ces travaux sur le parcours de soins (3 articles sur les hémopathies malignes en tant que 2ème auteur).

2. Le système d'information multi-sources du registre général des cancers de Poitou-Charentes (SI-RGCPC)

Le registre général des cancers de Poitou-Charentes (RGCPC) a développé et expérimenté localement un système d'information multi-sources[4] basé sur l'exploitation intégrée de données de 110 structures partenaires. Il y associe des développements algorithmiques qui visent à renforcer le processus d'enregistrement et de surveillance épidémiologique des cancers.

2.1. Rappels : Définition, missions

Le RGCPC surveille une population d'environ 1 800 000 personnes répartis sur 4 départements : la Charente, la Charente-Maritime, les Deux-Sèvres et la Vienne. Il inclut depuis le 1er janvier 2008, conformément aux recommandations nationales et internationales, tout cas incident de :

1. tumeur maligne invasive (hémopathies malignes et tumeurs solides en-dehors des carcinomes baso-cellulaires et spino-cellulaires de la peau),
2. tumeur maligne in situ,
3. tumeur borderline des ovaires,
4. tumeur bénigne ou d'évolution imprévisible du cerveau ou de la vessie,
5. ainsi que les adénomes colorectaux en dysplasie de haut grade (catégorie V4.1 de la classification de Vienne).

C'est un registre de population qui, à la différence des registres hospitaliers qui enregistrent tous les cas pris en charge dans un hôpital donné sans tenir compte de la population de référence, recense tous les cas survenus au sein de la population domiciliée de la zone registre. Ceci implique dès lors que le registre soit capable de repérer les cas de cancers survenant dans la population surveillée et traités hors de la zone géographique, mais de pouvoir également exclure la population venant d'une autre zone qui viendrait se faire soigner dans la zone du registre.

Le registre répond à un double objectif de surveillance et de recherche. Sa priorité est donc de collecter et de diffuser des informations de qualité décrivant tous les cas de cancer diagnostiqués chez les résidents du Poitou-Charentes. Au-delà de son activité de veille sanitaire, il constitue également un outil de référence pour la définition et la constitution des populations d'étude dans les projets de recherche.

[4] **Logiciel SINOET (Système d'Information pour la Notification et l'Enregistrement des Tumeurs) v1.0**, déposé en Juillet 2016 à l'Agence pour la Protection des Programmes

2.2. Sources de données

Aucun registre de population, que ce soit en France ou à l'international, ne peut fonctionner sans un mécanisme préalable d'identification des cas. Si tant est qu'une déclaration obligatoire existe dans certains pays, l'utilisation de multiples sources d'information sur les caractéristiques des cancers survenant dans la population cible représente un élément essentiel des registres de population. L'intérêt est de pouvoir identifier le plus de cas possible parmi ceux qui sont diagnostiqués chez les habitants de la zone géographique couverte à des fins d'exhaustivité [4,12].

Dès lors, le registre s'est attaché à avoir accès au plus grand nombre de sources d'informations, tout en veillant à mettre en œuvre des procédures performantes pour rassembler toutes les données utiles concernant un même individu. Des réflexions ont été engagées pour privilégier une uniformisation des modalités d'export au sein de chaque source de données, garant d'une simplification des méthodes de traitement mises en œuvre ultérieurement. L'enjeu était de pouvoir maintenir des exports et un flux régulier de données en provenance des sources de données, de sorte de pouvoir proposer aux utilisateurs du registre le niveau d'information le plus élevée possible au moment de l'enregistrement (*i. e.* offrir un recul optimal par rapport à la date de diagnostic, et disposer d'une représentation complète du parcours de soins du patient) et contribuer ainsi à augmenter les performances et l'efficience de l'enregistrement. Le succès du registre dépendant étroitement du partenariat et de la coopération du monde médical et des établissements, cela impliquait également un minimum d'investissements et de contraintes pour les structures participantes.

Le registre a mis en place un réseau de sources d'information lui permettant d'accéder de par son régime d'autorisation spécifique[5] aux données de 110 structures partenaires, réparties sur le Poitou-Charentes, les départements limitrophes et certains sites d'attraction spécifique en région Parisienne :

1. Prélèvements et comptes rendus d'anatomie et de cytologie pathologique (ACP),
2. Résumés PMSI des établissements de santé partenaires (PMSI),
3. Séances de radiothérapie en secteur libéral,
4. Compte rendus des réunions de concertation pluridisciplinaire en oncologie des réseaux de cancérologie (RCP),
5. Analyses de biologie des cancers (BIO),
6. Affections de longue durée de l'Assurance Maladie (AMA),
7. Données des registres hospitaliers des Centres Régionaux de Lutte contre le Cancer (EPC).

[5] Le RGCPC est autorisé par la Commission Nationale de l'Informatique et des Libertés (CNIL) à recevoir toute information nominative d'ordre médicale pour l'enregistrement de routine des cancers (N°907303, 15/02/2008). Il dispose également d'un avis favorable du Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé (CCTIRS) (N° 07.374, 04/10/2007).

2.2.1. Anatomie et cytologie pathologiques

L'anatomie et cytologie pathologiques constitue la source de données essentielle pour l'enregistrement des tumeurs. Dans un grand nombre de cas, c'est le médecin pathologiste qui par son analyse confirme le diagnostic et établit la typologie de la tumeur. Ainsi, l'histologie constitue la base diagnostique la plus valide pour l'enregistrement des cas incidents de cancer. Notons que la biologie moléculaire est de plus en plus utilisée en lien avec l'anatomie et cytologie pathologiques compte tenu de son intérêt à visée diagnostique et théranostique (médecine de précision).

Fort de l'expérience malheureuse du registre du Val-de-Marne et de l'utilisation méthodique du codage ADICAP[6] des tumeurs par les pathologistes du Poitou-Charentes, le développement et l'installation de modules d'export spécifiques ont été déployés au sein des différents logiciels métiers des structures participantes. Les requêtes ont été conçues et testées de façon à extraire les enregistrements des cancers répondant aux critères d'éligibilité du registre, à partir d'une sélection établie selon la structuration du code ADICAP utilisé par les pathologistes. Aux données d'identité et aux données tumorales associés à ces enregistrements étaient systématiquement attachés le contenu textuel des comptes rendus d'anatomo-cytopathologie (CRAP) et les codes diagnostics ADICAP correspondants. Disposer du contenu textuel des CRAP présentait l'intérêt de pouvoir instaurer des requêtes de recherche textuelle et des analyses de contenu (mots, expressions, annotations, sémantiques). Concernant les départements limitrophes, les critères de sélection ont pu être élargis selon les laboratoires, pour tenir compte des cas où la structure n'utilisait pas le codage ADICAP des tumeurs ou ne renseignait pas systématiquement le lieu de résidence du patient (*e. g.* lorsque le laboratoire facture l'analyse du prélèvement directement auprès de l'établissement prescripteur).

Malgré la diversité initiale des logiciels métiers, le RGCPC réceptionne aujourd'hui des fichiers sous un format hautement structuré et normalisé, simplifiant leur importation dans le SI-RGCPC. Le registre reçoit en moyenne 45 000 lignes prélèvements et 37 000 comptes rendus annuels, répartis sur 31 structures. Le rythme d'export est mensuel (au minimum semestriel) en Poitou-Charentes et semestriel (au minimum annuel) sur les départements limitrophes.

- **Classification automatisée des comptes rendus d'anatomopathologie** [32]

Près de 9% des prélèvements importés tous les ans dans le SI-RGCPC ne sont pas codés en ADICAP. Ces prélèvements impliquent d'être sélectionnés individuellement via la relecture systématique des comptes rendus en amont de leur intégration dans le SI-RGCPC. Pour subvenir à cette éventualité d'absence de codage des prélèvements anatomopathologiques, le registre a publié en 2012 les

[6] Association pour le Développement Informatique et Statistique en Cytologie et en Anatomie Pathologiques. Voir pour plus d'informations le chapitre « Nomenclatures diagnostiques et serveur de terminologies »

résultats d'une étude visant à construire et évaluer des fonctions appelés "classifieurs", dont l'optique était de catégoriser de façon automatique les comptes rendus d'anatomo-pathologie (CRAP) à partir de leur contenu textuel. Le principe était de pouvoir déterminé la topographie et l'histologie du cancer en procédant à un apprentissage artificiel supervisé (technique de machine learning) sur un lot de comptes rendus préalablement codés selon la Classification Internationale des Maladies en Oncologie 3ème édition (CIM-O3, classification en vigueur au niveau international utilisés par les registres)[7]. En pratique, le classifieur apprend de façon autonome la sémantique (associations de certains termes spécifiques) classiquement utilisée par les pathologistes pour décrire un cas de cancer.

Deux niveaux de granularité ont été étudiés :

1. L'attribution d'une catégorie définie par les recommandations de 2004 relatives à l'enregistrement des sites primitifs multiples (groupes de Berg, *i. e.* le niveau recommandé par le CIRC (Centre Internationale de Recherche sur le Cancer) pour définir l'existence ou non d'un nouveau cas de cancer) : topographie générique et morphologie générique ;
2. L'attribution d'un code selon la CIM-O3 afin de fournir une typologie précise de la tumeur : topographie complète et morphologie complète.

Les méthodes développées ont permis de positionner correctement 96% des comptes rendus au niveau des groupes de Berg. En revanche, le codage complet de la tumeur était plus délicat, pour la topographie en particulier (respectivement 72% et 85% de topographie et de morphologie complètes), pointant le fait que le pathologiste ne possédait pas toujours les informations précises concernant la topographie exacte de la tumeur. Ceci corroborait les excellents résultats reportés pour la prostate et le sein (> 99%), et les moins bonnes performances relevées pour certaines localisations comme le côlon-rectum (82%), pour laquelle ces deux topographies contiguës du tube digestif étaient souvent difficiles à distinguer.

Ces résultats soulèvent ainsi un intérêt pratique à pouvoir être appliquer au niveau des structures ACP n'utilisant que tout ou partie d'un système de classification diagnostique. Des travaux complémentaires seraient nécessaires toutefois pour évaluer la généralisation de ces méthodes influencées inévitablement par la variabilité sémantique textuelle entre pathologistes.

Lire Article 1 : *Jouhet V, Defossez G, Burgun A, le Beux P, Levillain P, Ingrand P, Claveau V. Automated classification of free-text pathology reports for registration of incident cases of cancer.* ***Methods Inf Med****. 2012;51(3):242-51. doi: 10.3414/ME11-01-0005. Epub 2011 Jul 26. PMID: 21792466.*

Accéder à la suite « Résumés PMSI des hôpitaux et cliniques »

[7] Voir pour plus d'informations le chapitre « Nomenclatures diagnostiques et serveur de terminologies »

Automated Classification of Free-text Pathology Reports for Registration of Incident Cases of Cancer

V. Jouhet[1]; G. Defossez[1]; A. Burgun[2]; P. le Beux[2]; P. Levillain[3,4]; P. Ingrand[1,5]; V. Claveau[6]
[1]Unité d'épidémiologie, biostatistique et registre des cancers de Poitou-Charentes, Faculté de médecine, Centre Hospitalier Universitaire de Poitiers, Université de Poitiers, Poitiers, France;
[2]INSERM U936, Faculté de médecine, Université de Rennes 1, Rennes, France;
[3]Anatomie et cytologie pathologiques, Centre Hospitalier Universitaire de Poitiers, Poitiers, France;
[4]Centre de Regroupement Informatique et Statistique en Anatomo-Pathologie de Poitou-Charentes, Faculté de médecine, Université de Poitiers, Poitiers, France;
[5]INSERM, CIC 802, Poitiers, France;
[6]IRISA – CNRS UMR 6074, Rennes, France

Keywords
Medical Informatics, neoplasm, pathology, free text, automated classification

Summary

Objective: Our study aimed to construct and evaluate functions called "classifiers", produced by supervised machine learning techniques, in order to categorize automatically pathology reports using solely their content.

Methods: Patients from the Poitou-Charentes Cancer Registry having at least one pathology report and a single non-metastatic invasive neoplasm were included. A descriptor weighting function accounting for the distribution of terms among targeted classes was developed and compared to classic methods based on inverse document frequencies. The classification was performed with support vector machine (SVM) and Naive Bayes classifiers. Two levels of granularity were tested for both the topographical and the morphological axes of the ICD-O3 code. The ability to correctly attribute a precise ICD-O3 code and the ability to attribute the broad category defined by the International Agency for Research on Cancer (IARC) for the multiple primary cancer registration rules were evaluated using F1-measures.

Results: 5121 pathology reports produced by 35 pathologists were selected. The best performance was achieved by our class-weighted descriptor, associated with a SVM classifier. Using this method, the pathology reports were properly classified in the IARC categories with F1-measures of 0.967 for both topography and morphology. The ICD-O3 code attribution had lower performance with a 0.715 F1-measure for topography and 0.854 for morphology.

Conclusion: These results suggest that free-text pathology reports could be useful as a data source for automated systems in order to identify and notify new cases of cancer. Future work is needed to evaluate the improvement in performance obtained from the use of natural language processing, including the case of multiple tumor description and possible incorporation of other medical documents such as surgical reports.

Correspondence to:
Vianney Jouhet
Unité d'épidémiologie, biostatistique et registre des cancers de Poitou-Charentes
Faculté de médecine
Centre Hospitalier Universitaire de Poitiers
Université de Poitiers
6, rue de la milétrie BP 199
86034 Poitiers Cedex
France
E-mail : vianney.jouhet@gmail.com

Methods Inf Med 2011; 50: ■–■
doi: 10.3414/ME11-01-0005
received: January 14, 2011
accepted: May 30, 2011
prepublished: July 26, 2011

1. Introduction

Most clinical databases in use today are special-purpose and restricted, and were not designed for interoperability [1]. In a White Paper by Safran et al., the secondary use of health data is discussed [2]. Secondary use of health data applies personal health information for uses outside direct health care delivery, which include, among other activities, research, quality and safety measurement, and public health. Despite the development of clinical repositories (e.g. [3]), secondary use of medical data for research and public health remains a challenge [4]. Kohane [5] contrasts the ease of access and sharing of biological data, which is mainly structured and standardized data, with the difficulty of utilizing patient information generated from routine care procedures, which is often in an unstructured free-text format. Integrating structured and unstructured information remains a challenge [6].

Pathology reports form one of the essential sources of data in order to include cases in cancer registries. These documents provide histological evidence of cancer on which the inclusion is mainly based. It is in many cases the pathologist who, via the analyses performed, confirms diagnosis and establishes the type of tumour, and when available, information contained in pathology reports is the most reliable source of data for the registration of incident cases of cancer [7].

For this registration procedure, a large part of the task involves classification targeting anatomical topography and tumour morphology. In order to harmonise data

collection, the International Association for Cancer Registries (IACR), the International Agency for Research on Cancer (IARC) and the World Health Organisation (WHO) specify that registered cases should be coded according to International Classification of Disease in Oncology, 3rd edition (ICD-O3) [8, 9]. Further to this, recommendations have been issued in collaboration with IACR, IARC, WHO and European Network of Cancer Registries (ENCR) concerning the registration rules for multiple primary cancers [10]. These recommendations specify the topographies and tumour morphologies that are to be recorded separately for one and the same individual so as to enable data comparisons across different populations. The recommendations define when a record should be considered to contribute to a new case or when it contributes to an already registered case, and the level at which data are to be agregated for follow-up of incidence and survival data.

As in numerous fields, the mass of information available in cancer registries is constantly increasing, so that manual processing is both tedious and costly. Automated cancer registration seems attractive as it may help to reduce delays in data production and allow personnel to devote more time to analysis and research. One of the major challenges is the development of automated analysis procedures for the data available so as to submit tumours for registration to the manual validation phase with the most plausible typology, defined both by anatomical site and histology.

Free-text pathology reports, often available in digital form, can be used for this data processing, but are not often available with an underlying terminology for anatomical sites and histology.

The Cancer Text Information Extraction System (caTIES) [11] has been developed in the framework of the caBIG project with focus on information extraction from pathology reports. Specifically, caTIES extracts information from free text Surgical Pathology Reports (SPRs), using the NCI Metathesaurus to populate caBIG-compliant data structures. Evaluation of CaTIES performance shows great precision [11] but recall has only been evaluated on the ability to retrieve ovarian cancer from radiology reports [12]. The Medical Text Analysis System/Pathology (MedTAS/P) [13] instantiates the Cancer Disease Knowledge Representation Model (CDKRM). MedTAS/P extracts cancer disease characteristics such as anatomical sites and histology from pathology reports. This system was evaluated on reports of colon cancer and showed good performance in terms of both recall and precision. However, to our knowledge, there is no efficient system enabling automatic extraction of tumour type for a tumour described in a French pathology report, and no study has evaluated recall and precision for more than one localisation.

Free-text pathology reports were also used by way of the stage coding in lung cancer [14, 15]. The CaFE (Case-Finding Engine) identifies cases of cancer in sets of pathology reports in English. This instrument, using lists of predefined terms, demonstrates a recall of 1 but a precision of 0.85 [16]. In the biomedical field, a large body of research has attempted to "map" biomedical concepts in clinical documents [17–19]. These approaches, mainly based on term-matching, have two main purposes: indexation, and the development of meta-data for use in automatised systems.

Little research has sought to apply automatic text classification (or categorisation) techniques on pathology reports to extract revelant informations. Li et al. [20] evaluated it over 203 pathology reports of colorectal cancer. To our knowledge, no study evaluated automatic text classification methods over more than one localisation. The automatic classification of text consists in annotation of the text so as to attribute a category solely on the basis of its content [21]. Since the 1990s, this discipline has developed considerably, and performances have improved, in particular since the development of machine learning techniques. One of the key resources for using learning machines is the availability of annotated documents [21]. These documents can be used in a learning process to construct a function by way of inference that enables categorisation of a new and unknown document. In this case the learning machine is said to be "supervised".

One of the tasks performed by physicians in the cancer registry in Poitou-Charentes (west France) consists in the annotation of pathology reports that contribute to the process of coding tumours. This provided us with the required database for the implementation of a supervised machine learning process for the classification of pathology reports in cancer units.

The aim of this study was to use pre-annotated data to construct and evaluate functions known as "classifiers", enabling the automatic categorisation of pathology reports in cancer unit solely from their textual content. Different classifiers were constructed, using supervised learning machines, according to two granularities, thus enabling two operational objectives to be envisaged:

- the attribution of a category defined by the 2004 recommendations on registration of multiple primary cancers [10]: generic anatomical site (IARC topography) and generic histology (IARC morphology)
- the attribution of a ICD-O3 code so as to provide a precise typology for the tumour: complete topography and complete morphology.

2. Methods

2.1 Classification Targets

The target class is defined as the annotation decided on by the human annotator for a given level of granularity. The ICD-O3 classification has two axes (►Fig. 1), topography and morphology. For each, two levels of granularity were studied so as to comply with the two operational objectives of the annotation. Thus the granularities targeted were:

2.1.1 Topography

2.1.1.1 Coding of the complete topography according to ICD-O3

In ►Figure 1, complete topography corresponds to C50.2 (upper inner quadrant of the breast)

2.1.1.2 Coding according to the recommended level for multiple primary cancers (IARC topography)

This level comprises 54 target classes of the 330 classes in the complete topography. For certain tumour morphologies (Kaposi sarcoma and tumour of the haemato-poietic system), a single tumour is registered, independently from topography. In Figure 1, IARC topography corresponds to breast.

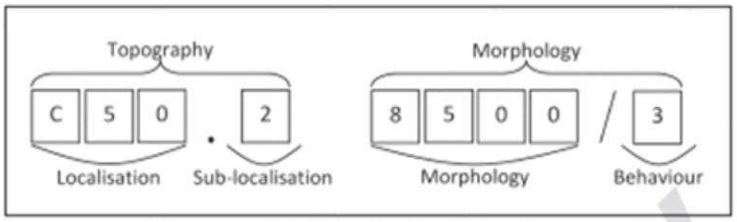


Fig. 1 Structure of the ICD-03 code (with example of an Infiltrating duct carcinoma on the upper inner quadrant of the breast)

2.1.2 Morphology

2.1.2.1 Coding of the complete morphology according to ICD-03

In ▶Figure 1, complete morphology corresponds to 8500/3 (Infiltrating duct carcinoma).

2.1.2.2 Coding according to the level recommended for multiple primary cancers (IARC morphology).

This level comprises 17 target classes of the 553 in the complete morphology. It corresponds to an adaptation of the morphology groups defined by Berg [22]. In ▶Figure 1, IARC morphology corresponds to adenocarcinoma.

2.2 Data Used

This study was conducted on data available in the Poitou-Charentes cancer registry for the year 2008. The collection and analysis of medical data by the cancer registry received the approval of the French regulatory authorities. The reports were written in French. Four main elements require prior definition:

- the distribution of topographies and tumour morphologies,
- the sources of the pathology reports,
- the manual production of annotations by physicians in the registry,
- and the relationship between the reports and ICD-O3 codes in the Poitou-Charentes cancer registry database.

The incidence of tumours according to localisation and tumour type is very uneven. For instance, in 2005, out of an estimated 320 000 cases in France, breast cancer, prostate cancer, colorectal cancer and lung cancer accounted for 37%, 34%, 12% and 10% of cases respectively, while pancreatic cancer, which ranks tenth out of 25, accounted for 2% of incident cases [23]. The data available for each targeted class relied on cases registered in the Poitou-Charentes cancer registry and the numbers of documents available was therefore varying a lot among the targeted class.

The pathology reports are in the form of free text in French, they are collected routinely from all the pathology laboratories in the Poitou-Charentes region, and annotated by the registry physicians. ▶Figure 2 gives an example of a report along with its translation. In the Poitou-Charentes registry database, one individual can have several tumours, and one tumour can be described in several reports. Likewise, it should be noted that one and the same report may contain the descriptions of several tumours. In addition, the reports describing secondary tumours (lymph glands or bowel) are directly linked to the primary tumour.

Finally a considerable proportion of the reports are made up of complementary notes to be attached to the earlier report (immuno-histo-chemical analyses and other expertise). These reports are linked to the tumours that they describe, but are not attached to the reports to which they contribute.

In this setting, and so as to ensure a single link between the report and its annotations, we included all pathology reports of individuals meeting the all following criteria:

1. a single identified tumour
2. an invasive tumour (behaviour coded /3 in ICD-O3)
3. a tumour having a manually validated, complete ICD-O3 coding
4. a tumour described by at least one pathology report
5. no identified organ or lymph node metastatic site

All the steps to prepare and represent the reports were performed using SAS v9.1. The construction and the evaluation of the different "classifiers" was performed on WEKA v3.6.2, a program developed in Java which provides classification tools [24].

2.3 Preparation of the Reports

Pre-processing of the reports had to take into account their heterogeneous forms, depending on the laboratory issuing them and even the pathologist who drafted the report. All the reports available for a given tumour were concatenated. Non-useful characters were replaced by spaces (special characters, parentheses and brackets, figures and operators). Non-informative words (e.g. certain articles and pronouns such as le, la, lui, elle) were removed using a "stop list". Lower case was used for all words in all reports, and accents, variably used, were removed. Finally a stem identification process removed plural and feminine forms [25].

2.4 Construction of the Training and Test Sets

To enable the evaluation of the automatic classification, the available data were split into two separate sets: a training set and a test set. Only the training set is used in all the construction phases of the classifier [21]. Separation into a training set and a test set was performed independently for

RENSEIGNEMENTS CLINIQUES :
Masse de plus de 2 cm de découverte récente, à l'union des QS du sein droit.
ACR5.
2 prélèvements 16 G.
MICROSCOPIE : MICROBIOPSIES MAMMAIRES DROITES (UNION DES QS) (2 carottes de 7 et 9 mm)
Les 2 carottes biopsiques examinées sont lésionnelles et d'aspect microscopique superposable.
Il existe en effet une prolifération cellulaire agencée en travées dissociant largement les tissus fibro-collagéniques présents.
Les cellules qui constituent cette prolifération montrent une anisocaryose, quelques mitoses.
Les immunomarquages pratiqués montrent :
- P63 négative en périphérie des structures épithéliales proliférantes, versus témoin interne positif.
- Ecadhérine légérement positive.
Ces aspects morphologiques correspondent à un CARCINOME CANALAIRE INFILTRANT dont le SBR est estimé à 2 dans la limite du matériel lésionnel examiné.
A noter la présence d'un embole vasculaire péri-tumoral.

CLINICAL INFORMATION
Recently discovered mass of over 2cm in diameter at junction between the upper quadrants of the right breast.
ACR5.
Two samples taken 16G.
MICROSCOPIC EXAMINATION: RIGHT BREAST MICRO-BIOPSIES (JUNCTION UPPER QUADRANTS) (2 core samples of 7 and 9 mm)
The two biopsy core samples examined are lesional and microscopic aspect is identical.
There is a cellular proliferation in bands cutting across the fibro-collagen tissues present.
The cells composing this proliferation demonstrate anysocaryosis, and some mitoses.
The immuno-marking performed showed:
Negative P63 on the periphery of the proliferating epithelial structures, versus positive internal control.
Slightly positive Ecadherine
These morphological aspects correspond to an INFILTRATING DUCT CARCINOMA, for which SBR is estimated to be 2 on the basis of the lesional material examined.
The presence of a peri-tumoral vascular embolus can also be noted.

Fig. 2 Example of a French pathology report and its English translation for the purpose of the article

each level of granularity. Depending on the prevalence of reports available for each target class in the granularity being processed, the reports were allocated to one or other of the sets of data. The reports were randomly allocated, 75% to the training set and 25% to the test set. We set the minimum at 25 reports for the training set and 5 for the test set, i.e. at least 30 reports were available for a given class. The reports belonging to target classes comprising smaller numbers than this were re-allocated to a class labelled "Others". With regard to IARC topography, morphologies for which the topography was non-informative (Kaposi sarcoma and tumours of the haematopoietic system) were grouped in a class labelled "systemic tumour".

2.5 Representation of Reports

In order to perform automatic classification tasks, a representation of the reports has to be produced, in a form that can be interpreted by the learning algorithms. The most widely used representation is the projection of the report descriptors (basic forms that will be used for the purpose of representation, for instance words) within a vector space [26, 27]. Here the documents are represented by a vector, for which the number of dimensions corresponds to the number of different descriptors identified from the set of documents.

The representation of the reports requires three successive procedures [21]:

- The choice of descriptors
- The reduction of the number of dimensions (optional)
- The choice of a quantifiable representation.

2.5.1 Choice of descriptors and dimensionality reduction

We chose a representation taking the form of a "bag of words". This is the most classic and simplest representation, whereby the text is divided up into the words that it comprises [21]. Some of these words will be used as dimensions for the vector representing the document. The number of dimensions (words) extracted from the training data is potentially very large (easily 10,000, even for texts of only moderate length). To reduce processing time, and also to avoid "overfitting" [21, 28], the number of dimensions was reduced by term selection using the multivariate Chi^2 method [29]. This method selects the terms that are the most strongly related to each of the target classes.

2.5.2 Quantifiable representation of descriptors

The aim here is to represent the report by a vector of numerical descriptors. We tested different methods of representation:

- Classic methods
  - *tf* for "Term Frequency", where it is considered that the more often a term occurs in a document, the more representative it is of text content, so that its weight should be of a high magnitude [21]
  - *tf.idf* for Term Frequency with Inverse Document Frequency, where it

is considered that a term that appears in numerous documents in the training corpus has little discriminating power for the task of classification, and should have a correspondingly small weight [21, 27, 29, 30]. The weight of a descriptor t_k in a document d_j, noted $w(t_k,d_j)$, is given by ▶Equation 1 [30] where |D| is the total number of documents, $\#(t_k,d_j)$ is the number of occurrences of term t_k in document d_j and $\#Tr(t_k)$ is the number of documents in which t_k appears (see ▶Eq. 1). Given the imbalance among classes in the training data available, we made the hypothesis that the use of *tf.idf* was liable to decrease the weight of the terms associated with classes in which number of training documents available were large. We therefore developed and evaluated a new method of representation.

- *tf.icf* for Term Frequency with Inverse Class Frequency. This method proposes to use the relationship between the terms and the classes in the corpus in order to generate a weighting index. This index, which we call "*icf*" depends on the distribution of a term across the different classes in the corpus. It is considered that a term with a homogenous distribution across classes is not discriminant for the task of classification, and should therefore have a correspondingly small weight. ▶Equation 2 shows an application of the *tf.icf* function, where $w(t_k,d_j)$ is the weight obtained by applying the *tf.icf* function, |C| is the total number of classes, $\#T(C_i)$ is the number of term in the class C_i, $\#(t_k,C_i)$ is the number of occurrences of the term tk in class C_i.

For the representation of the descriptors to range from 0 to 1, the weight of each descriptor is standardized by l^2-norm of the document vector (▶Eq. 3) [21] where

$$w(t_k,d_j) = \#(t_k,d_j).\log\frac{|D|}{\#T_r(t_k)}$$

Equation 1 *tf.idf*

$$\text{weight}(t_k,C_r) = \frac{\frac{\#(t_k,C_r)}{\sum_{j=1}^{\#T(C_r)}\#(t_j,C_r)}}{\max_{i=1}^{|C|}\left(\frac{\#(t_k,C_i)}{\sum_{j=1}^{\#T(C_i)}\#(t_j,C_i)}\right)}$$

$$w(t_k,d_j) = \#(t_k,d_j).\log\frac{|C|}{\sum_{i=1}^{|C|}\text{weight}(t_k,C_i)}$$

Equation 2 *tf.icf*

$w(t_k,d_j)$ is the weight of a descriptor k in a document j calculated using tf.idf or *tf.icf* and $|T|$ is the total number of terms in the document.

2.6 Classification

The classification was performed using machine learning techniques. The learning is said to be "supervised", since both the reports and their corresponding coding were available for all the data used in the development and validation stages. We compared two learning algorithms commonly used in classification tasks applied to text:

- the Naïve Bayes classifier [21, 31]
- the Support Vector Machine (SVM) classifier [21, 32, 33].

2.7 Evaluation

The classification was evaluated on the test data set. The documents in the test sets were represented and classified according to the parameters calculated and the classifier constructed from the training data. The results of this classification were compared with the manual annotations of the registry physicians. Three indicators were used:

- Percentage of correctly classified reports
- Mean F1-measure (harmonic mean of precision and recall [21])
- Kappa agreement coefficient [34]

These measures are classically used to assess the performance of classifiers, in the field of automatic classification (F1-measure), or for agreement between two methods, in particular in the biomedical field (Kappa coefficient).

All these steps were reiterated 10 times. Each time a training set and a test set were generated randomly from the overall report corpus.

A qualitative assessment of error was performed by way of systematic perusal of the wrongly classified reports. The main sources of error are discussed in Section 3.2.

3. Results

▶Tables 1 and 2 present the detailed results for each representation, classification method and granularity used.

In all, the complete corpus of data comprised 5121 documents. The reports used were derived from 16 laboratories and 35 pathologists.

$$r_{kj} = \frac{w(t_k,d_j)}{\sqrt{\sum_{s=1}^{|T|}(w(t_s,d_j))^2}}$$

Equation 3 Normalization (r_{kj}) of a descriptor k in a document j

Representation and model	Correct (%)	F1-measure	Kappa
IARC topographic class (16 categories)			
TF			
Naïve Bayes	85.5	0.802	0.799
SVM	96.4	0.963	0.952
TF.IDF			
Naïve Bayes	88.4	0.844	0.843
SVM	96.6	0.966	0.955
TF.ICF			
Naïve Bayes	88.8	0.848	0.847
SVM	**96.7**	**0.967**	**0.957**
Complete topography (26 categories)			
TF			
Naïve Bayes	60.9	0.507	0.486
SVM	72.4	0.713	0.655
TF.IDF			
Naïve Bayes	65.5	0.567	0.556
SVM	**72.6**	**0.716**	**0.659**
TF.ICF			
Naïve Bayes	65.7	0.569	0.559
SVM	72.5	0.715	0.657

Table 1 Classification among granularities of the tumoral topography

After stemming and "stop words" suppression, 10193 unique terms were identified in the overall corpus. This corresponds to the maximum dimensionality of representation's vectors before applying the dimensionality reduction (e.g. the maximum final dimensionality after reduction for the IARC morphology would be 8 targeted classes × 100 = 800 terms).

The comparison of performance of the different methods of representation and classification is provided in the annexes. In the experimentations performed overall, SVM was the classifier that demonstrated the best performances. Representations of the tf type provided the least satisfactory results. The two other methods of representation were comparable, with slightly better results for tf.icf (▶Table 3).

We chose to use the most efficient method across all the granularities targeted. The stages in this model were: a representation of the descriptors of the tf.icf type and a classifier of the SVM type. The following results present the quantitative and qualitative evaluation of this method for the different granularities targeted.

3.1 Quantitative Evaluation

Overall, good consistency was observed among the different evaluation measures implemented. However, regarding IARC morphology, there is a discrepancy between the F1-measure (demonstrating the best values for the granularities overall) and the slightly lower kappa values.

Irrespective of the axis considered (topography or morphology) the classification at the level recommended for multiple primary cancer coding showed good performances, with F1-measures at 0.967 and kappa coefficients of 0.957 to 0.892 for topography and morphology respectively. In contrast, the classification according to the ICD-O3 level of precision proved more delicate, in particular for the topography axis (▶Table 4).

▶Tables 5 and 6 give details of the quantitative indices for each target class in IARC topography and Berg's morphological groups.

Unsurprisingly, the "Others" class systematically showed a low F1-measure, while the classes with the largest prevalence available for the training process showed better performances. A tendency was noted towards a decrease in the F1-measure proportionately to the decrease in prevalence available for training. However, precision was systematically greater than 0.7 for all the target classes studied (with the exception of the "Others" for the two axies and "Unspecified type of haematopoietic and lymphoid tissue" classes in IARC morphology); the small prevalence in the training set tended above all to alter recall. The examination of confusion matrices showed that the most frequently encountered errors consisted in allocation of a report belonging to a class with small prevalence to a class comprising large prevalence.

Regarding IARC topography, "Colon" and "Rectum and rectal-sigmoid junction" showed poorer performances independently of the prevalence of training data available. They correspond to difficulties to distinctly identify these two contiguous topographies of the lower digestive tract, as pointed out during the qualitative examination of errors.

3.2 Qualitative Evaluation of Classification Errors

3.2.1 Reports Describing Several Examinations

Certain situations can occur in which a single report concerns the examination of several samples. Most often, this relates to series of samples taken within a single diagnostic phase, examined at the same time and included in a single pathologist's report. For instance, for the diagnosis of cancers of the digestive tract, a colonoscopy gives rise to several samples being taken and can often be accompanied by upper

gastrointestinal endoscopy which will also generate sampling procedures. In these reports, descriptions of healthy and cancerous tissue can coexist. For each of these descriptions there is a corresponding topography, and the human annotator in this case retains the topography relating to the tumour. The automatic classification of the topography may however retain a topography that does not correspond to the tumour.

3.2.2 Semantic Influence of a Corpus-based Training Data

As our method uses corpus-based training data, the classifier learned the strong association that some specific terms are classically used by majority of pathologists to describe cancer of certain anatomical location. As an influence in testing, an unspecific term that deviates from the classical usage may cause classification error, even if the unspecific term implies the associated location. This is the case for instance for the term "spinocellular epithelioma", which is classically used to describe epidermoid carcinoma of the skin. On account of this usage, there was a very strong association in the training data between this term and a cutaneous localisation. In the test set, the use by a pathologist of the term "epidermoid carcinoma" for the skin could thus lead to classification errors.

4. Discussion

The system of classification according to the level recommended by IARC for the registration of multiple primary cancers demonstrated very good performance. However the complete encoding of the tumour at a level of detail as refined as that of the ICD-O3 remains a delicate task. This is particularly true for complete topography, but we also need to take into account the fact that the pathologist does not always possess the information concerning the exact topography of the tumour, so that this data is not always present in the reports, and the practitioner who codes the tumour may, if needed, have used information derived from another document.

Table 2 Classification among granularities of the tumoral morphology

Representation and model	Correct (%)	F1-measure	Kappa
IARC morphological class (8 categories)			
TF			
Naïve Bayes	91.8	0.893	0.639
SVM	97.0	0.965	0.887
TF.IDF			
Naïve Bayes	94.0	0.924	0.759
SVM	97.0	0.966	0.887
TF.ICF			
Naïve Bayes	94.5	0.929	0.782
SVM	**97.1**	**0.967**	**0.892**
Complete morphology (18 categories)			
TF			
Naïve Bayes	75.8	0.683	0.646
SVM	86.3	0.853	0.812
TF.IDF			
Naïve Bayes	77.9	0.710	0.685
SVM	86.3	0.852	0.811
TF.ICF			
Naïve Bayes	78.1	0.714	0.691
SVM	**86.4**	**0.854**	**0.813**

Our method of representation (*tf.icf*) showed the best performance. It introduces the notion of relationships between terms and classes for quantitative representation, and reduces the effects of the imbalance between classes in the training process. Although differences observed were small, an improvement in performance was nonetheless noted with icf, and this method should therefore be evaluated in different settings and on different types of textual data. Lertnattee et al. investigated inverse class frequency in centroid-based text classification [35]. The functions applied were

Table 3 Comparison of performance by tf.idf and tf.icf according to granularity using a SVM-type classifier

Granularities	F1-measure mean (sd)*		Kappa	
	tf.idf	tf.icf	tf.idf	tf.icf
Topography				
IARC**	0.965 (0.003)	0.967 (0.003)	0.956	0.957
Complete topography	0.716 (0.008)	0.715 (0.008)	0.659	0.657
Morphology				
IARC**	0.966 (0.003)	0.967 (0.003)	0.887	0.892
Complete morphology	0.852 (0.007)	0.854 (0.006)	0.811	0.813

* Mean and standard deviation over the ten iterations
** Coding according to level recommended for registration of multiple primary cancers

Table 4 Performance of the classification according to granularity

Granularity	Number of classes	Correct (%)	F1-measure	Kappa
Topography				
IARC*	16	96.7	0.967	0.957
Complete topography	26	72.5	0.715	0.657
Morphology				
IARC*	8	97.1	0.967	0.892
Complete morphology	18	86.4	0.854	0.813

* Coding according to level recommended for registration of multiple primary cancers

based on the number of classes that contained the term to weight and neither used the number of occurrences of this term nor the total number of terms in each targeted class. By introducing these two factors, our representation method tried to capture the distribution of terms among classes taking into account the size of the training set of each class.

Our evaluation required the production of reliable indicators in view of our objective. The selected reports formed a corpus of simple cases for which no secondary localisation or multiple primary sites had been identified. This made it possible to ensure that the reports contained conclusions for only one topography and one invasive morphology (according to the 2004 recommendations). This process is obviously unrealistic from the point of view of the routine functioning of registries. We left out 16% of reports with multiple primary localizations or organ or lymph node metastatic site identified. In addition, for the topographic axis, we noted the potential impact of the coexistence of tumoral and non-tumoral pathologies on report content. It is therefore essential to identify reports containing several topographies. These elements strongly suggest the potential benefit of a multi-label classification, so that several target classes could be allocated to a single report. Both Naïve Bayes and SVM can generate ranked class predictions and can be adapted for multi-label classification.

The reports were derived from a large sample of pathologists working in different laboratories. The dataset therefore reflected certain variability in the drafting. This suggests that the introduction of written reports from other pathologists should not generate extra difficulties, so long as reports from these pathologists are available for the training process.

The use of different quantitative indices enabled apprehension of the overall performances of the different classifiers. The coherence of the results in relation to IARC topography, complete topography and complete morphology confirmed a certain level of performance. However the examination of disagreement concerning IARC morphology puts some perspective on the high F1-measure. The examination of the confusion matrices concerning this granularity showed that most of the mistakes concerned false positives of the "adenocarcinoma" class. These false positives in fact amounted to only a small prevalence in relation to the true positives in the adenocarcinoma class. Thus the effect on preci-

Table 5 Detailed results for topography at level recommended by IARC for registration of multiple primary cancers

IARC topographic class	Training	Test	Recall	Precision	F1-measure
Prostate	1529	509	0.999	0.998	0.999
Breast	972	324	1.000	0.995	0.998
Systemic tumour	264	88	0.990	0.976	0.983
Skin	144	48	0.969	0.973	0.970
Uterus	54	18	0.917	0.928	0.922
Colon	405	134	0.940	0.902	0.921
Liver and intrahepatic bile duct	44	14	0.950	0.888	0.917
Ovary	26	5	0.900	0.930	0.910
Trachea, bronchi, lung	39	12	0.883	0.932	0.905
Oesophagus	44	14	0.850	0.934	0.887
Gallbladder and other parts of biliary duct	30	9	0.789	0.928	0.844
Pancreas	25	5	0.880	0.812	0.831
Rectum and rectal-sigmoid junction	166	55	0.802	0.850	0.823
Stomach	31	10	0.730	0.919	0.806
Female genital organs	26	5	0.660	0.883	0.728
Other	54	18	0.639	0.692	0.661

Table 6 Details of results for morphology at level recommended by IARC for registration of multiple primary cancers

IARC morphological class	Training	Test	Recall	Precision	F1-measure
Adenocarcinomas	3240	1079	0.997	0.984	0.990
Squamous and transitional cell carcinomas	244	81	0.944	0.967	0.956
B-cell neoplasms	179	59	0.958	0.852	0.902
Hodgkin lymphomas	25	8	0.850	0.947	0.889
Other specific carcinomas	63	21	0.743	0.852	0.791
T-cell and NK-cell neoplasms	25	8	0.500	0.880	0.627
Other	42	14	0.357	0.693	0.461
Unspecified types of haematopoietic and lymphoid tissues	25	8	0.113	0.430	0.167

sion and on the F1-measure was very small. Adenocarcinomas amounted to nearly 87% of Berg's morphologies. As a result of weighting procedures, this class had a preponderant role in the calculation of the global F1-measure. The kappa coefficient, in contrast, may underestimate concordance in this situation. As Kappa is a chance-corrected measure [36], it is affected by a skewed prevalence. In this situation, the probability of chance agreement on high prevalence cases is very high [36], this is the "Kappa paradox" [37, 38] and a low kappa measure may not necessarily reflect a low concordance [39]. Its fairly high value here reflected adequate performances, although they were below what the F1-measure might lead one to expect.

Difficulties of a semantic nature were identified. They highlight the limitations of using words as descriptors, and point to the need for processing of the classification upstream. The use of UMLS (Unified Medical Language System) and "mapping" of reports by way of tools such as "MetaMap" [17] in order to extract and use medical concepts as descriptors of the reports should enable the impact of these problems to be reduced. This approach would also enable identification of concepts described by several successive words (e.g."infiltrating duct carcinoma"). These words are at present treated as independent entities. In addition, concept-based feature aggregation could reduce feature dimensionality and remove dependent features simultaneously. However as yet no equivalent tool has been developed in French.

Moreover, text information extraction systems based on the NCI Enterprise Vocabulary System or the UMLS make use of the sets of synonyms and the semantic categorisation provided by these resources. In caTIES, concepts are first identified using MetaMap. Then after detecting those that are explicitly negated, concepts are categorized on the basis of vocabulary semantic types, and classified as Diagnosis, Procedure or Organ type. Evaluations show a high level of precision (0.94 over 30 different queries ranging from low to high complexity). This suggests that, if equivalent resources were available for French, these methods could be adapted to extract and qualify concepts before applying our categorisation methods using a "bag of concepts" representation.

The performance metrics using CaTIES show a 0.98 precision in retrieving cases of prostatic adenocarcimoma identified on a prostatectomy for 60–80-year-old men [11]. Our results are comparable (0.998 for prostate and 0.984 for adenocarcinomas). However, the recall was not evaluated in the study by Crowley et al. [11] and this measure remains essential in a registry setting to ensure exhaustiveness. CaTIES was used to code free-text radiology reports in order to retrieve reports describing ovarian cancers with a 0.82 recall [12]. Our method provided a 0.90 recall on pathology reports for ovary topographic class (table V). Li et al. proposed a method to extract information from pathology reports using machine learning algorithms. They achieved the best performances with feature selection and Naive Bayes with a 0.511 F1-measure for tumour site and 0.964 for microscopic type [20].

The prevalence of reports available for the training process was very variable according to the target classes, on account of the natural variability of distribution of cancer localisations and tumour morphologies. The data available in our study did not enable all the classes for each granularity to be envisaged. It is however clear from our results that certain target classes are easier to annotate than others. Therefore this research work needs to be widened to datasets in which prevalence in the different classes could enable use of a larger number of target classes for each granularity. We noted the influence of this prevalence on the performance of classifiers. The threshold of 25 reports used for the training process was probably the reason for the decrease in performances, and it is probable that an increase in this threshold would enable an improvement.

Numerous other elements contribute to the coding of tumours. It is therefore clear that the sole use of pathology reports for automatic coding deprives users of part of the information that is available to a human annotator. Other text documents are often available, such as surgical reports, letters, or notes from pluri-disciplinary team meetings, and these could possibly be used in the same manner. For instance, surgical reports would provide a better source for the coding of complete topography.

The present results make it possible to envisage the incorporation of pathology reports as sources of data for the automated processing of information upstream of the manual validation phases in cancer registries. Adjustments to allow for semantic factors and for reports that describe several

samples should lead to the improvement of performance, and widen the field of application.

Acknowledgments

The authors would like to thank Pathologists of CRISAP Poitou-Charentes: Elisabeth Baltus , Dominique Battandier, François Baylac, Christine Blanchard, Françoise Bonneau-Hervé, Elisabeth Boudaud, Karine Boye, Philippe Debiais, Rony El Khoury, Catherine Emile, Catherine Fleury, Gaëlle Fromont Hankard, Jean-Michel Goujon, Harold Ip Kan Fong, Sylvain Labbé, Christian Lancret, Pierre Levillain, Didier Lhomme, Marie-Josée Loyer-Lecestre, Baudoin Mazet, Françoise Memeteau, Serge Milin, Olivier Nohra, Martine Paoli-Labbé, Dominique Petrot, Olivier Renaud, Mercédès Riols, Denis Roblet, Mokrane Yacoub who provided electronic full-text annotated pathology reports. They would also like to thank the medical records assistants, Nicolas Mériau and Sébastien Orazio of the *Registre des cancers de Poitou-Charentes* for their helpful contribution to the implementation of this research.

References

1. Maojo V, Kulikowski CA. Bioinformatics and medical informatics: collaborations on the road to genomic medicine? J Am Med Inform Assoc 2003; 10 (6): 515–522.
2. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc 2007; 14 (1): 1–9.
3. Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc 2007. pp 548–552.
4. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med 2009; 48 (1): 38–44.
5. Kohane IS. Bioinformatics and clinical informatics: the imperative to collaborate. J Am Med Inform Assoc 2000; 7 (5): 512–516.
6. Garcia-Remesal M, Maojo V, Billhardt H, Crespo J. Integration of relational and textual biomedical sources. A pilot experiment using a semi-automated method for logical schema acquisition. Methods Inf Med 2010; 49 (4): 337–348.
7. MacLennan R. Cancer registration: principles and methods. Items of patient information which may be collected by registries. IARC Sci Publ 1991; 1: 43–63.
8. Buemi A. Pathology of Tumours for Cancer Registry Personnel. IARC, Lyon; 2008.
9. Percy C, Fritz A, Jack A, Shanmugarathan S, Sobin L, Parkin D, et al. International Classification of Diseases for Oncology (ICD-O). 3rd ed. World Health Organization; 2000.
10. Curado M, Okamoto N, Ries L, Sriplung H, Young J, Carli M, et al. International rules for multiple primary cancers (ICD-O). 3rd ed. 2004.
11. Crowley RS, Castine M, Mitchell K, Chavan G, McSherry T, Feldman M. caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. J Am Med Inform Assoc 2010; 17 (3): 253–264.
12. Carrell D, Miglioretti D, Smith-Bindman R. Coding free text radiology reports using the Cancer Text Information Extraction System (caTIES). AMIA Annu Symp Proc 2007. p 889.
13. Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, et al. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. J Biomed Inform 2009; 42 (5): 937–949.
14. McCowan IA, Moore D, Fry MJ. Classification of cancer stage from free-text histology reports. Conf Proc IEEE Eng Med Biol Soc 2006; 1: 5153–5156.
15. McCowan IA, Moore DC, Nguyen AN, Bowman RV, Clarke BE, Duhig EE, et al. Collection of cancer stage data by classifying free-text medical reports. J Am Med Inform Assoc 2007; 14 (6): 736–745.
16. Hanauer D, Miela G, Chinnaiyan A, Chang A, Blayney D. The Registry Case Finding Engine: An Automated Tool to Identify Cancer Cases from Unstructured, Free-Text Pathology Reports and Clinical Notes. Journal of the American College of Surgeons. 2007; 205 (5): 690–697.
17. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001. pp 17–21.
18. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. Stud Health Technol Inform 2004; 107 (Pt 1): 268–272.
19. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004; 11 (5): 392–402.
20. Li Y, Martinez D. Information extraction of multiple categories from pathology reports. Australasian Language Technology Association Workshop (ALTA Workshop 2010): Australasian Language Technology Association (Melbourne); 2010. pp 41–48.
21. Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv 2002; 34 (2): 1–47.
22. Berg JW. Morphologic classification of human cancer. In: Shottenfeld D FJ, Jr., editor. Cancer epidemiology and prevention. 2nd ed. New York: Oxford University Press; 1996.
23. Belot A, Grosclaude P, Bossard N, Jougla E, Benhamou E, Delafosse P, et al. Cancer incidence and mortality in France over the period 1980–2005. Rev Epidemiol Santé Publique 2008; 56 (3): 159–175.
24. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor Newsl 2009; 11 (1): 10–18.
25. Savoy J. A stemming procedure and stopword list for general French corpora. J Am Soc Inf Sci 1999; 50 (10): 944–952.
26. Salton G, Wong A, Yang C. A vector space model for information retrieval. Communications of the ACM 1975; 18 (11): 613–620.
27. Laroum S, Béchet N, Hamza H, Roche M. Classification automatique de documents bruites a faible contenu textuel. RNTI: Revue des Nouvelles Technologies de l'Information 2009; 1: 25.
28. Yang Y, Jan P. A comparative study on feature selection in text categorization. In: Proceedings of ICML-97, 14th international conference on machine learning, Nashville, TN; 1997. pp 412–420.
29. Clech J, Rakotomalala R, Jalam R. Séléction multivariée de termes. XXXVèmes Journées de Statistiques. Lyon, France; 2003. pp 933–936.
30. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manage 1988; 24 (5): 513–523.
31. John GH, Langley P. Estimating Continuous Distributions in Bayesian Classifiers. Eleventh Conference on Uncertainty in Artificial Intelligence. San Mateo: Morgan Kaufmann; 1995. pp 338–345.
32. Boser BE, Guyon I, Vapnik NV. A training algorithm for optimal margin classifiers. 1992. pp 144–152.
33. Platt JC. Fast training of support vector machines using sequential minimal optimization. 1999. pp 185–208.
34. Fleiss JL. Statistical methods for rates and proportions. New York: Wiley; 1981.
35. Lertnattee V, Theeramunkong T. Analysis of inverse class frequency in centroid-based text classification. IEEE International Symposium on Communications and Information Technology 2004; 2: 1171–1176.
36. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005; 12 (3): 296–298.
37. Feinstein A, Cicchetti D. High agreement but low kappa: I The problems of two paradoxes. J Clin Epidemiol 1990; 43 (6): 543–549.
38. Cicchetti D, Feinstein A. High agreement but low kappa: II Resolving the paradoxes. J Clin Epidemiol. 1990; 43 (6): 551–558.
39. Viera A, Garret J. Understanding interobserver agreement: the kappa statistic. Fam Med 2005; 37 (6): 543–549.

2.2.2. Résumés PMSI des hôpitaux et cliniques

Le PMSI (ou Programme de Médicalisation des Systèmes d'Information) dispose des données relatives aux séjours hospitaliers de l'établissement et constitue une source de données classiquement utilisée pour tenter d'identifier les tumeurs dont la prise en charge nécessite au moins une hospitalisation.

Une procédure d'interrogation des Départements d'Information Médicale (DIM) a été développée en lien avec le Collège Régional de l'Information Médicale de Poitou-Charentes en vue de recueillir des fichiers standardisés de Résumés d'Unités Médicales (RUM) répondant aux critères d'éligibilité du registre. Cette procédure, simple et reproductible, vise à rendre nominatif le fichier de RUM groupés envoyé mensuellement aux tutelles et permet d'appliquer un format unique à l'ensemble des établissements à activité MCO (Médecine Chirurgie Obstétrique) partenaires du registre. Certaines données (données de résidence, communes de naissance) sont complétées selon l'organisation interne de l'établissement par le biais des services administratifs en vue d'un géocodage au niveau géographique le plus fin. Les patients sont sélectionnés par la présence d'un code de tumeur en diagnostic principal (DP), relié (DR) ou associé (DAS) selon la Classification Internationale des Maladies 10ème révision (CIM-10)[8].

De façon similaire à la logique appliquée pour l'ACP, les fichiers sont réceptionnés selon un format structuré et normalisé (applicable à toute la France), simplifiant dès lors nettement leur importation dans le SI-RGCPC. Le registre traite en moyenne 251 600 RUM tous les ans, répartis sur 67 établissements de santé. Le rythme d'export est trimestriel en Poitou-Charentes (au minimum annuel) et semestriel sur les départements limitrophes (au minimum annuel).

Cette démarche garantit l'accès aux données complètes d'hospitalisation et particulièrement aux actes médicaux codés selon la CCAM, qui confèrent un niveau élevé d'information et qui, nous le verrons, seront très utiles dans la notification des cas incidents de cancer et l'identification d'évènements traceurs de la trajectoire de soins.

2.2.3. Séances de radiothérapie en secteur libéral

La radiothérapie privée étant facturée en consultation (soins externes), elle ne fait pas parti champ du PMSI et doit faire l'objet d'un processus de collecte spécifique. Une requête permet d'extraire une liste nominative d'actes CCAM d'irradiation datés auxquels sont reliés la localisation cancéreuse irradiée (à partir d'un code CIM-10). Les données sont intégrées dans le SI-RGCPC selon la même trame que les données PMSI et contribuent à enrichir les parcours de soins des patients atteints de

[8] Voir pour plus d'informations le chapitre « Nomenclatures diagnostiques et serveur de terminologies »

cancer. Le registre traite en moyenne 45 000 séances de radiothérapie libérale tous les ans, répartis sur 3 établissements.

2.2.4. Réunions de concertation pluridisciplinaire (réseaux de cancérologie)

Le Dossier Communicant de Cancérologie (DCC) est l'outil métier destiné à faciliter le partage des données médicales des patients entre les professionnels de santé. Le DCC est souvent décrit du point de vue de l'organisation des réunions de concertation pluridisciplinaire (RCP). La RCP vise à choisir le traitement le plus adapté pour le patient en prenant en compte les dernières recommandations sur la pathologie, mais aussi les bénéfices, les risques et les conséquences sur la qualité de vie du patient. Elle représente ainsi une source privilégiée pour le registre pour appréhender les informations relatives aux stades d'extension au diagnostic et aux recommandations de traitement, contribuant à augmenter la mise à disposition passive d'informations pour la validation des dossiers. Un volume moyen de 32 000 fiches RCP sont exportées tous les ans à partir de 4 des 5 réseaux de cancérologie participants.

2.2.5. Laboratoires d'hématologie biologique

Les tumeurs hématologiques représentent une part importante des pathologies cancéreuses. Or la prise en charge d'un patient atteint d'hémopathie maligne, leucémie ou lymphome, nécessite une certitude diagnostique qui relève aussi de l'interprétation de données issues de l'hématologie, la cytogénétique et la biologie moléculaire.

Le registre collecte à cet effet les myélogrammes et immunophénotypages de cellules tumorales, et traite en moyenne 2 500 examens tous les ans, répartis sur 12 sites. Une des principales difficultés pour cette source réside dans l'absence de codage des prélèvements nécessaire à un export ciblé de la seule pathologie tumorale, et implique, comme nous l'avons vu pour l'ACP, de procéder à une relecture systématique des examens diagnostiques en amont de l'import dans le SI-RGCPC. Notons que les données de cytogénétique et de biologie moléculaire, qui prennent aujourd'hui une part importante dans la classification des tumeurs à visée pronostique et théranostique, ne font pas l'objet d'une collecte systématique pour l'enregistrement de routine.

2.2.6. Affections de longue durée (assurance maladie)

Le cancer est une maladie qui nécessite un suivi et des soins coûteux prolongés, et est considéré comme une affection de longue durée (ALD). Cette ALD ouvre droit à la prise en charge à 100 % des soins liés à la pathologie, dès lors que le médecin traitant en rédige la demande.

Les services médicaux des 3 principaux régimes d'Assurance Maladie (Régime Général, Mutualité Sociale Agricole et Régime Social des Indépendants) transmettent régulièrement au RGCPC la liste des ALD admises parmi les patients domiciliés en Poitou-Charentes pour les cancers étudiés. Le registre traite en moyenne 10 500 ALD tous les ans.

2.2.7. Enquête permanente cancer (registres hospitaliers)

Les registres hospitaliers compilent les données sur les cas de cancer diagnostiqués et/ou traités dans un établissement donné, selon des règles de classification analogues à celles des registres de population. Toutefois, les données collectées sont dépendantes de la fréquentation de l'établissement concerné, et ne permettent pas de fournir un profil non biaisé du poids du cancer dans la population et de son évolution au cours du temps. Les Centres Régionaux de Lutte Contre le Cancer (CRLCC) qui participent à l'Enquête Permanente Cancer (EPC) communiquent leurs données au RGCPC annuellement (volume moyen de 850 tumeurs traitées tous les ans à partir des 3 CRLCC limitrophes : Institut Bergonié de Bordeaux, Institut de Cancérologie de l'Ouest de Nantes et Angers). Il n'y a pas de CRLCC en Poitou-Charentes. Cette source tend à disparaître progressivement.

2.2.8. Au global

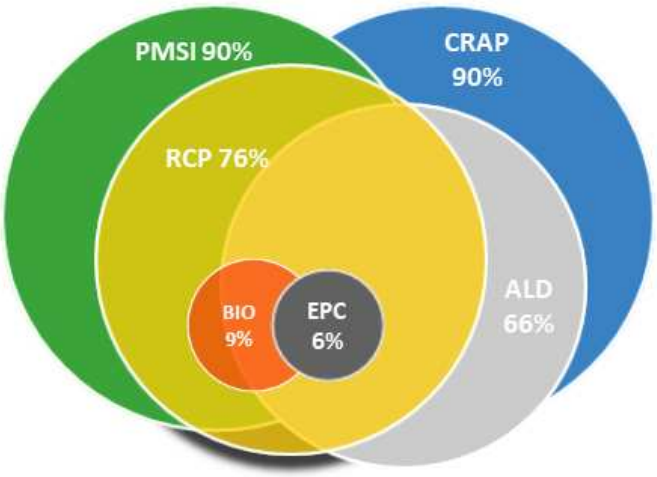
La distribution des 110 partenaires au sein de chaque source est résumée Tableau 1.
Le taux de participation globale des structures est supérieur à 92% si l'on se réfère à l'ensemble des sources éligibles par le RGCPC. La participation est complète pour le Poitou-Charentes (47 sources sur 47) et de 87% hors Poitou-Charentes (60 sources sur 69). Notons que le volume d'activité des centres hors Poitou-Charentes non participants est faible.
Chaque source de données contribue à enrichir l'information utile à l'enregistrement des cas et il est rare de disposer d'une seule source de données pour un cas de cancer (Figure 2). Certaines sources sont plus souvent disponibles que d'autres. Sur les 117 298 cas incidents de tumeurs malignes invasives enregistrés entre 2008 et 2017, 90% disposent d'un prélèvement anatomopathologique (95% si l'on se limite aux tumeurs solides et 53% pour les hémopathies malignes) ou d'un résumé PMSI (inclut les séances de radiothérapie en secteur libéral), 76% d'une RCP et 66% d'une ALD. Plus de la moitié des cas (54%) sont notifiés à partir d'au moins 4 sources de données différentes (82% à partir d'au moins 3 sources).
Si l'on se réfère aux structures cette fois parmi les sources qui ont contribué à l'enregistrement des cas (*e. g.* une tumeur peut avoir été notifiée à partir de 2 laboratoires d'anatomopathologies différents), ce nombre est en moyenne de 4 structures par cas et peut varier jusqu'à 14 structures différentes pour une seule tumeur (Tableau 2), illustrant la richesse des données collectées et la pluridisciplinarité sous-jacente des parcours de soins empruntés par les patients.

Tableau 1 : Sources de signalement des cas au sein du SI-RGCPC

Sources de données	Nombre de structures	Rythme de consultation
EN POITOU-CHARENTES	**50**	
ACP : Laboratoires d'anatomie et de cytologie pathologiques	8	Entre 1 et 6 fois / an
PMSI: Services d'Information Médicale	28	Entre 1 et 4 fois / an
AMA : Services médicaux de l'Assurance Maladie	3	1 fois / an
RCP: Réseaux de cancérologie	1	2 fois / an
PMSI: Radiothérapie privée*	2	1 fois / an
BIO: Laboratoires d'hématologie biologique	8	1 fois / an
HORS POITOU-CHARENTES	**60**	
ACP : Laboratoires d'anatomie et de cytologie pathologiques	21	Entre 1 et 6 fois / an
PMSI: Services d'Information Médicale	30	Entre 1 et 2 fois / an
EPC : Centres régionaux de lutte contre le cancer	1	1 fois / an
RCP: Réseaux de cancérologie	3	1 fois / an
PMSI: Radiothérapie privée*	1	1 fois / an
BIO: Laboratoires d'hématologie biologique	4	1 fois / an
TOTAL	**110**	

** Les séances de radiothérapie en secteur libéral sont reliées au sein du SI-RGCPC à la source de données PMSI*

Figure 2 : Proportion des sources de données disponibles reliées aux cas incidents (tumeurs malignes) recensés sur la période 2008-2017 (n=117 298 cas) au sein du SI-RGCPC



Nombre de sources	Nombre de cas	Proportion
0	504*	0,4%
1	6 376	5%
2	13 928	12%
3	32 918	28%
4	57 174	49%
5	6 148	5%
6	250	0,2%
TOTAL	117 298	-

** 504 tumeurs (0,4%) ne sont reliées à aucune donnée source car elles ont été notifiées par les opérateurs du registre à travers la validation des autres tumeurs et autres sources de notification.*

Tableau 2 : Nombre moyen de structures disponibles (parmi chaque source) reliées aux cas incidents (tumeurs malignes) recensés sur la période 2008-2017 (n=117 298 cas) au sein du SI-RGCPC

	Nombre moyen de structures par cas +/- écart-type (min-max)
ACP : Laboratoires d'anatomie et de cytologie pathologiques	1,09 +/- 0,57 (0-6)
PMSI: Services d'Information Médicale/	1,36 +/- 0,81 (0-7)
AMA : Services médicaux de l'Assurance Maladie	0,67 +/- 0,49 (0-3)
EPC : Centres régionaux de lutte contre le cancer	0,06 +/- 0,24 (0-3)
RCP: Réseaux de cancérologie	0,81 +/- 0,52 (0-4)
BIO: Laboratoires d'hématologie biologique	0,10 +/- 0,30 (0-4)
TOTAL	**4,08 +/- 1,64 (0-14)**

** Les séances de radiothérapie en secteur libéral sont reliées au sein du SI-RGCPC à la source de données PMSI.*

2.3. Traitement de l'information

Une fois les données collectées, le traitement de l'information représente une étape essentielle en amont de l'enregistrement des tumeurs. Le processus intègre quatre étapes successives : 1) import et alignement des données, 2) interconnexion des enregistrements, 3) identito-vigilance et 4) préparation de l'information (notification des cas et représentation du parcours de soins). La Figure 3 présente le circuit de l'information au sein du SI-RGCPC.

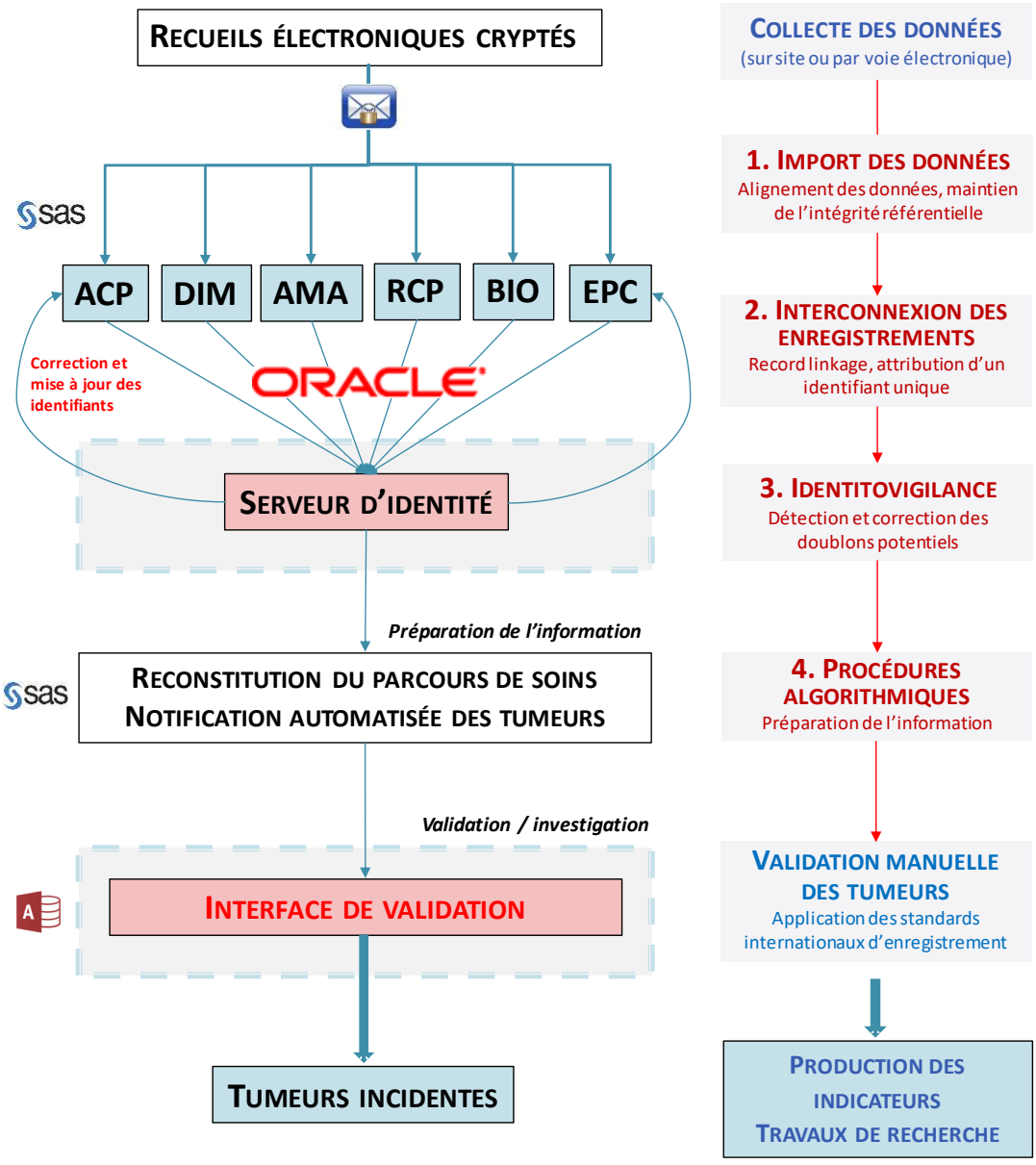


Figure 3 : Processus de circulation de l'information au sein du système d'information du registre général des cancers de Poitou-Charentes (SI-RGCPC)

ACP : Anatomie et Cytologie Pathologiques
DIM : Départements d'Informations Médicale (PMSI)
AMA : Assurance Maladie
RCP : Réunions de Concertation Pluridisciplinaire en oncologie
BIO : Laboratoires d'hématologie biologiques
EPC : Enquête Permanente Cancers (CRLCC)

2.3.1. Import des données dans le SI-RGCPC

Les données émanant des différentes sources interrogées sont structurées après réception au sein d'une base de données Oracle. La procédure d'import a pour objet de garantir l'alignement des noms, formats de variables, et la granularité (unité statistique) des données entre chaque fichier source et le schéma structurel de la base de données. Cette étape implique la production des variables structurelles nécessaires au maintien de l'intégrité référentielle de la base de données.

2.3.2. Interconnexions entre les enregistrements (« record linkage »)

Le registre recevant des informations relatives à un patient de plusieurs sources, ces dossiers (enregistrements) doivent tous être liés au même patient afin que les détails de chaque patient soient complets et qu'il n'y ait pas de doublons dans les enregistrements pour la même tumeur. Cette mise en relation des données est un préalable nécessaire à la phase de notification des cas. L'intégration des données sources au sein du SI-RGCPC se fait autour de l'individu, et fait appel aux techniques de « computerized record linkage » (interconnexion automatique des enregistrements) au sein du serveur d'identité. Un individu y est identifié par son nom (nom de jeune fille et/ou nom marital pour les femmes mariées), prénom, sexe et date de naissance. L'objectif est de rattacher chaque enregistrement source à l'individu auquel il appartient. Cela se produit selon 2 cas de figures (couplage déterministe) : 1) Si un nouvel enregistrement correspond à un individu déjà connu du registre, il est automatiquement rattaché à cet individu ; 2) Sinon, une nouvelle identité est créée avec son identifiant unique, puis est reliée aux enregistrements sources correspondants.

2.3.3. Identito-vigilance

Toutefois, des doublons d'enregistrement peuvent se produire à la suite d'une défaillance dans le processus de mise en relation des sources, en raison d'erreurs administratives comme la mauvaise orthographe des noms. L'identito-vigilance vise ainsi à limiter les risques de doublons en mettant en œuvre une phase de détection visant à identifier les doublons potentiels (couplage probabiliste), et les soumettre à une phase de regroupement manuel auprès d'un opérateur du registre. Celui-ci procède alors au recodage (ou non) des identités approchantes, selon les traits de similarité identifiés à partir des informations administratives et médicales du dossier. Il n'y a pas de recodage automatique pour éviter le risque de collision. Une fois le dossier traité, un numéro patient unique est attribué aux identités jumelées, puis est mis à jour en cascade à chacun des enregistrements source reliés à cet unique individu. Le serveur d'identité conserve l'historique de toutes les identités initiales, de sorte de pouvoir augmenter les probabilités d'attribution automatique (déterministe) de nouveaux enregistrements à cet individu.

2.3.4. Préparation de l'information (approche algorithmique)

2.3.4.1. Algorithme 1 : Notification des cas [33]

Alors que l'objectif de l'interconnexion des enregistrements est de relier les données sources des patients entre eux, la notification des cas vise à individualiser les cancers primitifs parmi toutes les informations disponibles pour un individu. Tandis que la notification repose le plus souvent sur un processus manuel actif engagé par le personnel des registres, l'approche privilégie ici des développements algorithmiques pour produire automatiquement le « meilleur » diagnostic à partir des données disponibles. Cette résolution des données est une opération cruciale soulevant les concepts de base de notifications multiples et de tumeurs multiples.

Concepts de notifications multiples et de tumeurs multiples

- Notifications multiples

L'objectif est de déterminer si un enregistrement pris individuellement concerne une tumeur (cas) déjà connue du registre ou une nouvelle tumeur primitive. Si un patient est diagnostiqué comme ayant un cancer dans un hôpital et qu'il est adressé à un autre pour y être traité, il se peut que les deux hôpitaux signalent le cas au registre. Le registre doit reconnaître en amont ces notifications comme des notifications multiples, et garantir en aval la notification d'un cas unique.

- Tumeurs multiples

Il arrive qu'un patient atteint de cancer développe plus d'une tumeur primitive et c'est l'habitude de faire un enregistrement indépendant pour chacun, puisque les registres du cancer comptent en fait le nombre de cancers primitifs plutôt que le nombre de patients atteints de cancer. Il est donc important pour un registre d'avoir une définition claire de ce qui constitue les tumeurs multiples, pour éviter à la fois le sur- et le sous-enregistrement et garantir ainsi la comparabilité de ses données et leur uniformité au cours du temps. Ceci impose un suivi régulier et une gestion stricte des nomenclatures au regard des évolutions de classification (voir le chapitre « Nomenclatures diagnostiques et serveur de terminologies »).

L'étude des tumeurs multiples se réfère ainsi à une série de règles internationales mises à disposition par le CIRC [34], afin que les taux d'incidence soient comparables entre les différentes populations. En résumé, ces règles sont les suivantes :

- L'identification de 2 tumeurs primitives (ou plus) ne dépend pas du temps.
- Un cancer primitif est une tumeur qui prend naissance dans un organe ou un tissu ; il ne s'agit donc pas d'une extension, d'une récidive ni d'une métastase.

- Une seule tumeur sera prise en compte dans un organe ou une paire d'organes ou un tissu, en-dehors des deux circonstances suivantes :
    - ✓ La survenue d'une tumeur de type histologique différent dans un organe précédemment atteint est comptée comme un nouveau cas.
    - ✓ Un seul cas sera enregistré par individu pour les tumeurs systémiques ou cancers multicentriques atteignant plusieurs organes. Trois groupes sont concernés : les lymphomes, les leucémies et les sarcomes.

Le CIRC définit ainsi des groupes de codes topographiques et types histologiques (appelés groupes de Berg) considérées comme étant des entités similaires/différentes dans le cadre de la définition des tumeurs multiples (voir encadré ci-dessous).

Le système de classification des tumeurs utilisé internationalement par les registres du cancer est la Classification Internationales des Maladies en Oncologie, 3ème édition (voir aussi le chapitre « Nomenclatures diagnostiques et serveur de terminologies »)

Chaque cas est classé selon 3 axes :

Le *code topographique* est utilisé pour identifier le site primitif. Il est constitué de 4 caractères (de C00.0 à C80.9). Un point (.) sépare les sous-divisions des catégories à trois caractères. Une tumeur du lobe supérieur du poumon se code par exemple C34.1 (C34 : Poumon ; .1 : Lobe supérieur). Lorsque la présence de métastases est évidente mais que le véritable point d'origine de la tumeur ne peut être déterminé, ce dossier est codée en localisation primitive inconnue (C80.9).

Le *code histologique* traduit le type cellulaire de la tumeur et est constitué de 4 chiffres (de 8000 à 9989). Le diagnostic histologique doit être le plus détaillé possible et être déduit du résultat définitif de la biopsie ou de la chirurgie.
Le *comportement tumoral* traduit l'activité biologique de la tumeur et est précisé par le 5ème chiffre du code de morphologie. Il est utilisé pour distinguer les tumeurs bénignes (/0) des tumeurs malignes (/3) et les stades intermédiaires : les tumeurs in situ (/2) et les tumeurs de bénignité ou de malignité non assurées (/1).
Un adénocarcinome se code par exemple 8140/3 (8140 : adéno ; /3 : carcinome).

Nomenclatures diagnostiques et serveur de terminologies

Plusieurs terminologies sont utilisées pour coder des diagnostics de cancer (CIM-10 pour le PMSI, ADICAP pour l'anatomo-pathologie, CIM-O3 pour les registres). Chacune de ces terminologies implique une gestion stricte qui suppose de pointer et de reporter rigoureusement les évolutions de classification au cours du temps (nouveaux codes, nouveaux libellés, changements de comportement, suppression de codes), de veiller à adapter les règles décisionnelles d'enregistrement (guides de recommandations Francim), d'assurer la traçabilité des modifications (antériorité) et d'en tenir compte au moment de l'analyse (règles de comptage des cas incidents selon qu'ils sont diagnostiqués avant ou après la mise en œuvre de la nouvelle classification).

- CIM-10

La CIM-10 est la Classification Internationale des Maladies (10ème révision) publiée par l'OMS pour l'enregistrement des causes de morbidité et de mortalité touchant le domaine de la médecine. Elle permet de classer les maladies, mais également les signes, symptômes, lésions traumatiques, empoisonnements, circonstances sociales et causes externes de blessures ou de maladies. Elle est principalement utilisée par les établissements hospitaliers français (PMSI) pour le codage des diagnostics et des motifs de recours aux services de santé, par l'Assurance Maladie pour le codage des Affections Longue Durée (ALD) et par le CépiDc (Inserm) pour le codage des causes médicales de décès.

La CIM-10 est une terminologie qui représente le diagnostic sous la forme d'un seul code alphanumérique de trois à cinq caractères. Le chapitre 2 de la CIM-10 est entièrement dédié aux tumeurs (C00-C97 : Tumeurs malignes, D00-D09 : Tumeurs in situ, D10-D36 : Tumeurs bénignes, D37-D48 : Tumeurs à évolution imprévisible ou inconnue). La CIM-10 ne propose pas de détails de la morphologie tumorale, en-dehors de certains types histologiques particuliers : le sarcome, le mélanome cutané, le mésothéliome, les hémopathies malignes et certaines catégories morphologiques de cancer du foie.

La CIM-10 est soumise à des révisions périodiques sous l'égide de l'OMS (à noter que la CIM-11 doit rentrer en application à compter du 1er janvier 2022).

Exemple : Une tumeur maligne du lobe supérieur du poumon se code « C34.1 » (C : Tumeur maligne ; 34 : Poumon ; .1 : Lobe supérieur).

- ADICAP

L'ADICAP est une terminologie française élaborée par l'Association pour le Développement de l'Informatique en Cytologie et en Anatomie Pathologique (ADICAP) qui recouvre tous les domaines

de l'anatomopathologie dont la cancérologie. L'ADICAP est fréquemment intégré aux logiciels métiers d'ACP pour le suivi de la facturation et de l'activité.

L'ADICAP représente le diagnostic d'une tumeur à partir d'un code sur 15 caractères alphanumériques réparties sur 2 zones :

- Une première zone obligatoire sur les 8 premiers caractères au sein de laquelle sont renseignés le mode de prélèvement (1 caractère), le type de technique (1 caractère), l'organe (2 caractères) et la lésion (4 caractères, le 1er caractère identifiant la famille histogénétique de la tumeur, le 2ème caractère correspondant au type de comportement, et les 3èmes et 4èmes caractères désignant les groupes et variétés de tumeurs dans ces groupes) ;
- Une deuxième zone facultative sur les 7 caractères suivants au sein de laquelle peuvent être renseignées la topographie et la latéralité de la tumeur.

L'ADICAP fait l'objet de révisions périodiques en pratique par l'association, mais aucune mise à jour n'est disponible depuis l'année 2009, en-dehors d'une révision établie en 2018 par le groupe francophone d'hématologie cellulaire ciblée sur les hémopathies.

Exemple : Un adénocarcinome invasif moyennement différencié du lobe supérieur du poumon droit diagnostiqué sur une histologie de pièce opératoire se code « OHRPA7A2 LSD » (O : Pièce opératoire avec exérèse complète de l'organe ; H : Histologie et cytologie par inclusion ; RP : Appareil respiratoire / Poumon; A7A2 : Adénocarcinome invasif moyennement différencié ; LS : Lobe supérieur ; D : Droit).

- CIM-O3

La CIM-O3 ou Classification Internationales des Maladies en Oncologie (3ème édition) est le système de classification des tumeurs publié conjointement par l'OMS et le CIRC et utilisé internationalement par les registres du cancer. Elle constitue une extension historique du chapitre « *Tumeurs* » de la CIM-10 visant à prendre en compte la morphologie (histologie) de la tumeur.

La CIM-O3 est une terminologie qui représente le diagnostic d'une tumeur selon 2 axes principaux : le code topographique (site primitif / 4 caractères) et le code histologique (qui associe le type cellulaire de la tumeur sur 4 chiffres et son comportement tumoral sur 1 chiffre en 5ème position).

La CIM-O3 fait l'objet de mises à jour régulières sous la responsabilité du CIRC et connues sous le nom de « Blue books »[9] qui couvrent tous les sites d'organes en 12 volumes.

Exemple : Un adénocarcinome invasif du lobe supérieur du poumon droit se code « C34.1 » « 81403 » (Topo C34 : Poumon ; .1 : Lobe supérieur ; Morpho 8140 : Adéno ; 3 : Carcinome).

[9] WHO Classification of Tumours: https://whobluebooks.iarc.fr/

- Serveur de terminologies

L'utilisation de plusieurs terminologies implique de pouvoir les mettre en correspondance au sein d'un système cohérent et hiérarchisé, *i. e.* le serveur de terminologie. Son objectif est de fournir une solution générique capable de faire correspondre les diagnostics issus des sources de données utilisant des systèmes de codage différents. Le serveur de terminologies du RGCPC fait référence à 4 tables de transcodage unidirectionnel, permettant de passer de la terminologie CIM-10 ou ADICAP aux 2 axes de la terminologie CIM-03 (*i. e.* cible des registres du cancer) :

1. Transcodage de l'ADICAP vers l'axe topographique de la CIM-O3 (un ou plusieurs codes « organe » de l'ADICAP +/- associé à un code « topographie » (si renseigné en zone facultative) pointe vers un code « topographie » de la CIM-O3) ;
2. Transcodage de l'ADICAP vers l'axe morphologique de la CIM-O3 (un ou plusieurs codes « lésion » de l'ADICAP pointe vers un code « morphologie » de la CIM-O3) ;
3. Transcodage de la CIM-10 vers l'axe topographique de la CIM-O3 (un ou plusieurs codes CIM-10 pointe vers un code « topographie » de la CIM-O3) ;
4. Transcodage de la CIM-10 vers l'axe morphologique de la CIM-O3 (un ou plusieurs codes CIM-10 pointe vers un code « morphologie » de la CIM-O3).

Rappelons que la CIM-10 ne propose pas de détails de la morphologie tumorale, en-dehors de certains types histologiques particuliers. Dans ce contexte, un code morphologique générique est affecté selon le comportement de la tumeur (exemple : « C341 » – Tumeur maligne du lobe supérieur du poumon renvoie à un code morphologique « 80003 » – Tumeur maligne SAI (sans autre indication) associé au code topographique « C341 » – Lobe supérieur du poumon). A noter également la prise en compte des tumeurs malignes secondaires (C77-C79) auxquelles sont affectées le code morphologique générique « 80006 » – Tumeur métastatique.

En revanche pour les types histologiques particuliers (sarcome, mélanome cutané, mésothéliome, hémopathies malignes et certaines catégories morphologiques de cancer du foie), un code morphologique spécifique est affecté selon l'entité, associé à un code topographique générique (« C809 » – Siège primaire non précisé) pour les hémopathies malignes ou un code topographique spécifique pour les autres cas (exemple : « C434 » – Mélanome malin du cuir chevelu et du cou renvoie à un code morphologique « 87203 » – mélanome malin SAI (sans autre indication) associé au code topographique « C444 » – Peau du crâne et du cou).

Algorithme de notification

L'algorithme de notification constitue le cœur du système d'information [33]. Il permet de créer de façon autonome une tumeur et ses diverses composantes à partir des données sources.

- Principe général de fonctionnement

L'algorithme de notification a pour but de définir en premier lieu la topographie et la morphologie de la tumeur pour chaque ligne source reliée à un individu selon ses métadonnées codées. Cette phase implique au préalable le transcodage et l'alignement des données sources vers une nomenclature unique, la CIM-O3, via le serveur de terminologies (ADICAP → CIM-O3 pour la source ACP, CIM-10 → CIM-O3 pour les sources DIM, AMA et RCP).

Dans un second temps, l'algorithme détermine pour chaque individu, en application des critères du CIRC relatifs à l'enregistrement des sites primitifs multiples (groupes de Berg), le besoin de notifier une ou plusieurs tumeurs primitives selon la granularité déduite de la phase d'agrégation des données, en tenant compte de l'existence d'une tumeur déjà connue ou non du registre. Ceci implique dès lors qu'il n'y ait pas de nouvelles notifications en cas de récidive ou d'entités similaires entrant dans le cadre de la définition des tumeurs multiples.

- Vers une sélection hiérarchisée de l'information

En raison du codage divergent parfois utilisé par les sources, l'algorithme de notification peut créer par erreur plusieurs cas pour une seule et même personne. En vue de limiter le bruit induit par les imprécisions ou les erreurs de codage, la démarche a été d'instaurer une notification autonome pour la topographie et pour la morphologie tumorale, en sélectionnant l'information pertinente sur la base d'une typologie préalable des enregistrements (hiérarchisation de la source de données, identification d'évènements traceurs). Les hypothèses sous-jacentes étaient de considérer : i) la véracité du codage du clinicien pour notifier la topographie tumorale, celui-ci étant le plus à même (le chirurgien le plus souvent) à préciser l'organe prélevé ; ii) la véracité du codage du médecin pathologiste pour notifier la morphologie tumorale, celui-ci étant le plus à même à caractériser l'histologie de la tumeur. Notons que la connaissance de l'organe prélevé pour le pathologiste est basée le plus souvent sur les renseignements cliniques fournis par le clinicien.

Par exemple, un patient opéré d'un cancer du côlon sigmoïde (PMSI : tumeur maligne du côlon sigmoïde « C18.7 » associée à un acte CCAM de colectomie) et dont le compte-rendu anatomopathologique précisera qu'il s'agit d'un adénocarcinome de localisation recto-sigmoïdienne (ADICAP : OHDRA7A0) sera notifié « C187 » (côlon sigmoïde) et « 81403 » (adénocarcinome), l'hypothèse sous-jacente étant qu'il s'agit de la même tumeur et que la topographie de l'organe est préférentiellement celle déduite du clinicien.

Cette approche par sélection hiérarchisée permet ainsi de traiter chaque individu de façon différenciée en adaptant la sélection à la complétude de l'information disponible, et a pour effet de réduire la notification de faux sites primitifs multiples.

- Evaluation des performances de l'algorithme

Une évaluation de cet algorithme a été produite sur les 12 346 tumeurs malignes enregistrées en 2008 (une année entière incluant les tumeurs multiples). Les résultats ont démontré une nette amélioration de la performance de l'algorithme via la mise en œuvre de l'approche sélective décrite ci-dessus. La F-mesure, moyenne harmonique du rappel (ou sensibilité) et de la précision (ou valeur prédictive positive), était de 0,926 avec sélection versus 0,857 sans sélection pour la topographie tumorale, et de 0,805 avec sélection versus 0,750 sans sélection pour la morphologie tumorale (0,725 versus 0,542 pour la tumeur complète).

La stratégie visant à notifier la topographie d'une tumeur en privilégiant l'identification d'un acte traceur améliorait nettement la précision de l'algorithme (réduction du nombre de faux positifs) sans altérer le rappel (pas plus de faux négatifs). Ce qui sous-tend que les topographies retenues à partir des typologies d'enregistrement appliquées sont pertinentes, et que la perte d'information résultante de la sélection est faible. Quant à la stratégie de notifier la morphologie d'une tumeur à partir de l'anatomopathologie, elle est évidente. Toutefois, l'utilisation de sources de données complémentaires permet de minimiser les dommages en termes de précision lorsque le dossier est incomplet.

Notons qu'il est logique de s'attendre en l'absence d'approche sélective à un meilleur rappel, la probabilité étant plus forte de notifier correctement un cas au vu des multiples notifications. En revanche, la proportion des tumeurs notifiées à tort (faux positifs) est plus importante et explique la moins bonne précision de l'algorithme.

Lire Article 2 : *Jouhet V, Defossez G; CRISAP; CoRIM, Ingrand P. Automated selection of relevant information for notification of incident cancer cases within a multisource cancer registry.* ***Methods Inf Med****. 2013;52(5):411-21. doi: 10.3414/ME12-01-0101. Epub 2013 Apr 24. PMID: 23615926.*

Accéder à la suite « Tous les cas créés par l'algorithme de notification … »

Automated Selection of Relevant Information for Notification of Incident Cancer Cases within a Multisource Cancer Registry

V. Jouhet[1,2]; G. Defossez[1]; CRISAP[3]; CoRIM[3]; P. Ingrand[1,5]

[1]Registre général des cancers de Poitou-Charentes, Faculté de médecine, Centre Hospitalier Universitaire de Poitiers, Université de Poitiers, Poitiers, France;
[2]CHU de Bordeaux, Pole de santé publique, Service d'information medicale, Bordeaux, France;
[3]Centre de Regroupement Informatique et Statistique en Anatomie Pathologique, Faculté de Médecine, Université de Poitiers, Poitiers, France;
[4]Collège Régional de l'Information Médicale de Poitou-Charentes, Centre Hospitalier Général Georges Renon, Niort, France;
5INSERM, CIC 802, Poitiers, France

Keywords
Multisource information system, cancer registry, tumour notification, selection algorithm, relevant information

Summary

Objective: The aim of this study was to develop and evaluate a selection algorithm of relevant records for the notification of incident cases of cancer on the basis of the individual data available in a multi-source information system.

Methods: This work was conducted on data for the year 2008 in the general cancer registry of Poitou-Charentes region (France). The selection algorithm hierarchizes information according to its level of relevance for tumoral topography and tumoral morphology independently. The selected data are combined to form composite records. These records are then grouped in respect with the notification rules of the International Agency for Research on Cancer for multiple primary cancers. The evaluation, based on recall, precision and F-measure confronted cases validated manually by the registry's physicians with tumours notified with and without records selection.

Results: The analysis involved 12,346 tumours validated among 11,971 individuals. The data used were hospital discharge data (104,474 records), pathology data (21,851 records), healthcare insurance data (7508 records) and cancer care centre's data (686 records). The selection algorithm permitted performances improvement for notification of tumour topography (F-measure 0.926 with vs. 0.857 without selection) and tumour morphology (F-measure 0.805 with vs. 0.750 without selection).

Conclusion: These results show that selection of information according to its origin is efficient in reducing noise generated by imprecise coding. Further research is needed for solving the semantic problems relating to the integration of heterogeneous data and the use of non-structured information.

Correspondence to:
Vianney Jouhet
Unité d'épidémiologie, biostatistique et registre des cancers de Poitou-Charentes
Faculté de médecine
Centre Hospitalier Universitaire de Poitiers
Université de Poitiers
6, rue de la milétrie – BP 199
86034 POITIERS Cedex, France
E-mail: vianney.jouhet@gmail.com

Methods Inf Med 2013; 52: ■-■
doi: 10.3414/ME12-01-0101
received: October 25, 2012
accepted: March 27, 2013
prepublished: ■■■

1. Introduction

Cancer registries have the task of exhaustively recording incident cases of cancer in a given territory. In order to harmonise data collection, the International Association for Cancer Registries (IACR), the International Agency for Research on Cancer (IARC) and the World Health Organization (WHO) specify that registered cases should be coded according to the International Classification of Diseases in Oncology, 3rd edition (ICD-O-3) [1, 2]. Further to this, recommendations have been issued in collaboration with IACR, IARC, WHO and the European Network of Cancer Registries (ENCR) concerning registration rules for multiple primary cancers [3]. These recommendations define when a record should be considered to contribute to a new case and when it should be considered to contribute to an already registered case, and the level at which data are to be aggregated for follow-up of incidence and survival data.

The job of cancer registries extends well beyond the mere recording of incident cases. In France, the National Committee of Registers (NCR) evaluation grids to be applied for the accreditation of registries include not only the methods used and the quality of the records, but also the use made of the data, the interest and the originality of the research work conducted [4].

To enable the registries to make full use of their expertise and research function in

the area of cancer epidemiology, the optimization of registration procedures for incident cases of cancer is crucial, and has been recalled in the two French national cancer plans. This optimization is in particular necessary for registries covering large territories and populations. The implementation of automated and semi-automated procedures to assist in detecting and documenting incident cases of cancer is therefore an attractive approach in this setting.

In order to be able to register cases, cancer registries staff, must be informed that possible new cases should be registered. This process, called notification, was historically based on voluntary practitioners that declared all new cases they encountered. As early as 1998, an IARC technical report was drawn up describing the methods used by different registries for the establishment of automated notification procedures [5]. To ensure adequate cover of a population of several million people, at a very early stage the Ontario registry was obliged to develop methods for data acquisition [6]. Both notification and record tasks in this registry are automated with no clerical interventions in routine processing [5, 7]. The integration of multi-source data into information systems that are structured around the patient makes it possible to optimize automatic processing for the notification and registration of incident cases, and the recording of complementary data [8]. The use of this accumulated information for case notification is a logical strategy in a perspective of exhaustiveness. This view has led registries to increasingly diversify notification sources. The mean number of notification sources has thus become a criterion for exhaustiveness, and a percentage of histological confirmation of cases that is too high should lead to suspicion of non-exhaustiveness [9]. The price to pay for this approach is excess notification of false incident cases following coding inaccuracies by the different data sources. These false cases require manual processing to be removed. Reducing their numbers without affecting exhaustiveness would enable time to be saved (and hence cost) for each case finally registered and would improve the quality of the automated data produced.

Olive et al. have presented a critical analysis of French hospital discharge data for the epidemiology of cancers. In particular, they noted difficulties relating to the use of data in isolation to detect incident cases [10]. This finding underlines the importance of using diverse sources for the notification of new incident cases. Deterministic [8, 11, 12] or probabilistic algorithms using artificial learning techniques [13] have been developed and evaluated. These works were mainly based on interrelating and aggregating data from different sources, but without any selection of relevant information according to its source. Couris et al. used coded information on surgical procedures performed at the time of hospitalisation as the criterion for the selection of relevant records in hospital discharge data, enabling a 30% reduction in false positive rates for the notification of breast cancer. Nevertheless, sensitivity, already poor, dropped from 69% to 64% [14] so that it was impossible to envisage exhaustive notification solely on the basis of hospital discharge data.

To our knowledge, no research has been conducted on the implementation of a selection among the variety of information available for a given individual from different data sources. Our hypothesis is that adding a selection step could reduce the noise resulting from incorrect or imprecise coding, and would enable the extraction of relevant information upstream from case notification. The aim of this study was therefore to develop and evaluate an algorithm selecting relevant records for the notification of incident cases of cancer according to the information available for each individual.

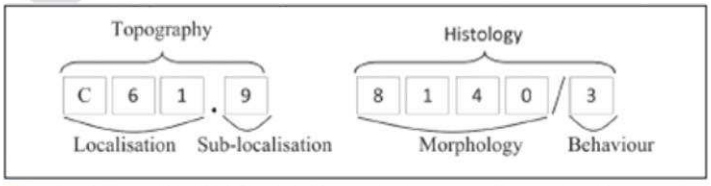


Figure 1 Structure of the ICD-O-3 code (with example of a prostate adenocarcinoma)

2. Background

2.1 Registration Rules for Multiple Primary Cancers

►Figure 1 presents the structure of the ICD-O-3 classification. The ICD-O-3 is a multi-axial classification used in cancer registries in order to record the anatomical site (topography) and the histology of a neoplasm. The histology axis is coded on five digit. The first four digits correspond to the morphology (histological description) and the fifth digit indicates whether a tumour is malignant, benign, in situ, or uncertain (whether benign or malignant).

For the purpose of defining multiple tumours, groups of topography codes considered as a single site and groups of morphology codes considered as a single neoplasm are defined [3]:

- **IARC topography (Topographic axis).** This level aggregates the topography codes in 54 target classes.
- **IARC morphology (Morphological axis).** This level aggregates the morphology codes (regardless of behaviour) in 17 target classes. It corresponds to an adaptation of the morphology groups defined by Berg [15].

If the morphological diagnoses fall into one category of the 17 target classes, and arise in the same primary site, they are considered to be the same tumour for the purpose of counting multiple primaries. For paired organs such as breast, the same rules are applied so that bilateral breast cancers are counted as one tumour even if asynchronous. If the morphological diagnoses fall into two or more of the 17 target classes, even if they concern the same site, the morphology is considered to be different, and two or more cases should be counted. For certain tumour morphologies (Kaposi sarcoma and tumour of the haematopoietic system), a single tumour is registered, independently from topography.

We used this granularity given as the recommended level for multiple primary cancers as the target classes to be identified by the notification algorithm.

2.2 Validation Process in the General Cancer Registry of Poitou-Charentes

The collection and analysis of medical data by this cancer registry received the approval of the French regulatory authorities. In compliance with national and international recommendations, since January 1, 2008 the general cancer registry of the Poitou-Charentes region (western France) has included any incident case of malignant invasive tumour (haematological malignancy and solid tumours not including baso-cellular skin carcinomas), in situ tumour, borderline ovarian tumour, and tumour of the brain or the bladder that are benign or where evolution is unpredictable, involving a subject regularly residing in the Poitou-Charentes region at the time of diagnosis. The general cancer registry of Poitou-Charentes is qualified by the French NCR since January 2013 based on the 2008 and 2009 registered data.

▶ Figure 2 presents the validation process in the Poitou-Charentes cancer registry that was used in 2008 for notification, registration and validation of incident

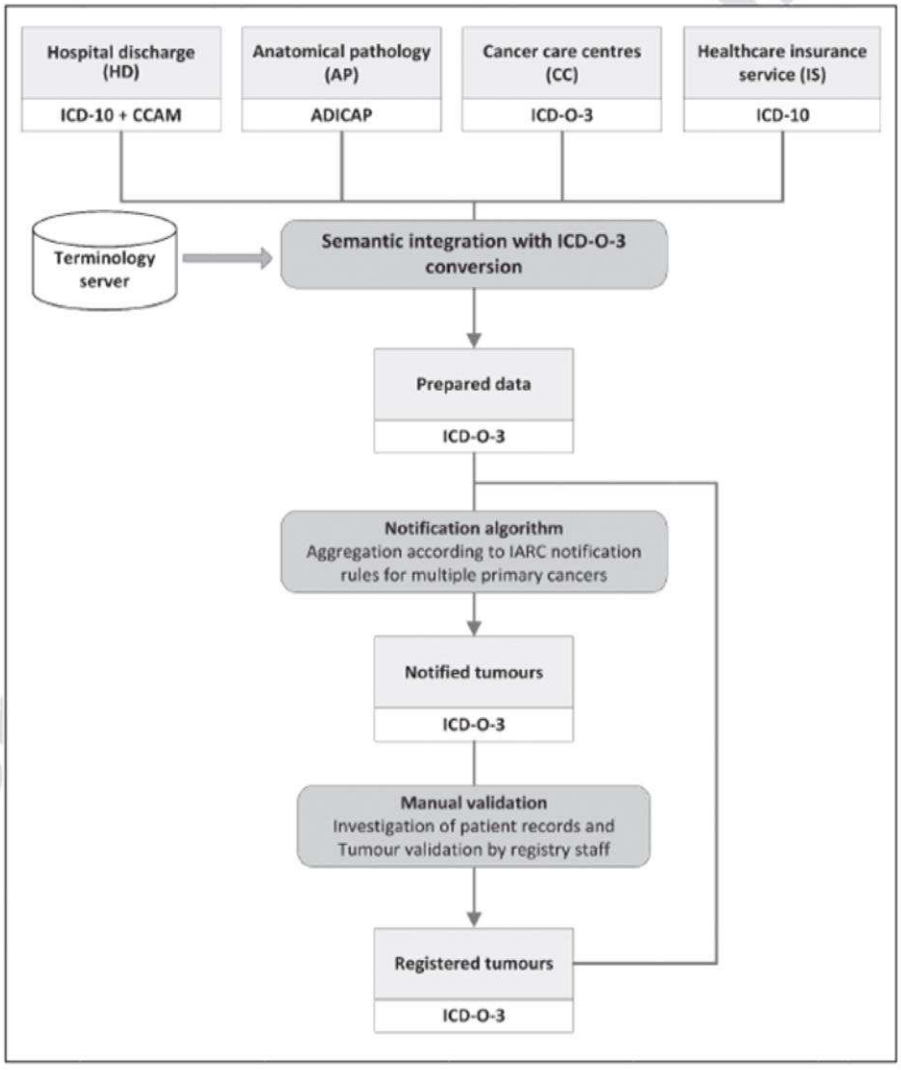


Figure 2
Validation process in the Poitou-Charentes cancer registry

cases of cancer. The registry routinely collects four types of data sources using various terminologies to describe diagnosis:

- Anatomical pathology (AP) data that includes free-text reports related to one or several ADICAP diagnostic codes (Association pour le Développement de l'Informatique en Cytologie et en Anatomie Pathologique – French classification of lesions with topographical and histological axis).
- Hospital discharge (HD) data recorded in the French medical information program that includes ICD-10 diagnostic codes and CCAM medical procedure coded fields (Classification Commune des Actes Médicaux – the health insurance classification).
- Reimbursement of a dispensation for a cancer granted by the French healthcare insurance service (IS) that includes ICD-10 diagnostic codes.
- Data from the continuous survey in cancer care centres ("les Centres de Lutte contre le Cancer") that includes ICD-O-3 tumour codes. The continuous survey is a system of data collection in oncology promoted in France in 1975 by the National Federation of Cancer Care Centres. The continuous survey allows the identification of topographic and histological diagnosis of tumours, their initial extension, as well as thera-

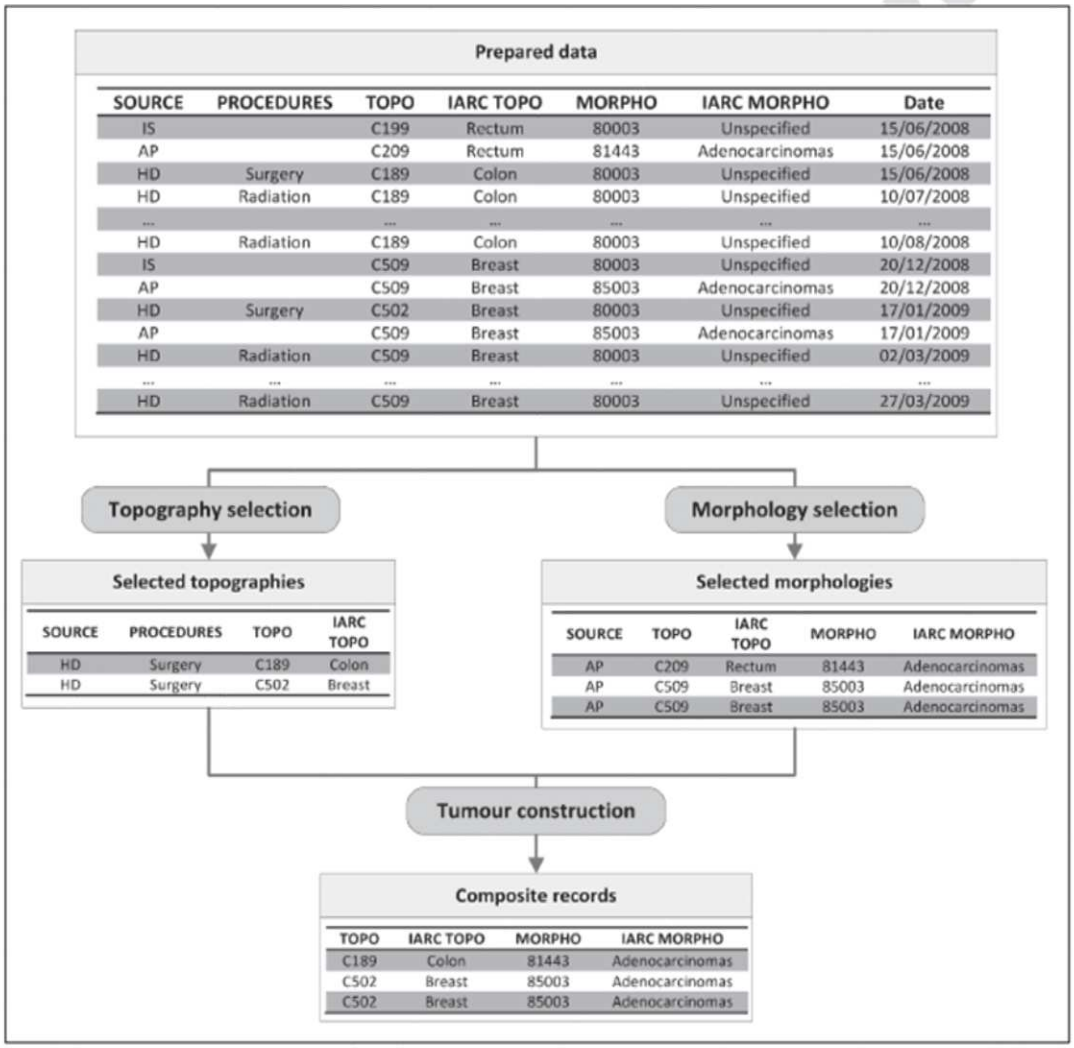


Figure 3 Example of execution of the selection algorithm for a given individual. In this example, all data are related to a unique individual.

peutic and evolutionary data for all cases in a given cancer care centre.

Data extractions include patient identity (name, surname, birthdate …). When loading data in the registry's information system, patient's identities are integrated in an identity server which identifies with a semi-automated process data that are related to a single patient. The patient's identification process, based on computerized record linkage [5], allows automated linkage (based on deterministic rules), ambiguity detection, duplicate search and manual patient grouping or separation. This process allows all eligible records from data sources integrated in the information system to be related to single patients.

In order to permit a semantic integration, all diagnostic codes from data sources are transformed into ICD-O-3 using a terminology server. Then, the notification algorithm processes every record prepared by the semantic integration process and uses ICD-O-3 diagnostic codes to determine, for each patient, tumours that should be notified to cancer registry staff regarding registration rules for multiple primary cancers [3] and already validated tumours.

All tumour records automatically created by the notification algorithm contain related information including date of diagnosis, ICD-O-3 topography, ICD-O-3 morphology, basis of diagnosis and a zero level of validity. Then each case is manually checked by registry staff by visual inspection of information sources, assessing the need of a patient's medical record investigation for recording the tumour.

Only tumours that have never been manually modified can be updated through the notification process. The notification algorithm is executed each time new records are integrated from a data source and uses the entire information available for individuals that have at least one new record.

3. Materials and Methods

3.1 Selection Algorithm for Relevant Records

Our approach was to add a selection step of relevant records between the semantic integration step and the notification algorithm (►Figure 2). This approach was guided by the knowledge of data sources that the registry staff has acquired during manual data validation.

Throughout patient care several practitioners from different specialities and different structures can have occasion to implement coding procedures resulting in multiple tumour occurrence in the data extracted by the registry for secondary use. The accuracy of this coding, whether in terms of topography or morphology of the tumour, may be highly dependent on the way in which the practitioner approaches the case. We therefore set out some simple hypotheses concerning the choice of records liable to be used for notifying tumours:

- Cancer care centres' data, where the registration rules are close to those used by the registries themselves, provide a satisfactory synthesis of the patient file, thus enabling notification of topography and morphology.
- It is the clinician who is the best suited to notifying tumoral topography, in particular the clinician who performed the biopsy or the surgery. Indeed, the surgeon knows exactly the location of the tumour and is used to describe it.
- It is the pathologist who is best suited to notifying tumoral morphology and behaviour. Indeed the pathologist knows exactly the histological type of the tumour and is used to describe it.
- The existence of multiple primary sites in an individual is not the norm, and notification requires an adequate level of proof.
- The remaining information should be taken into account when the data for a given individual is incomplete. This rule is essential in order to prevent from non-notification when some pertinent data are available for an individual.

Based on these assumptions the algorithm selects the best information at the patient level according to the data source providing the diagnostic code for tumoral topography and tumoral morphology independently.

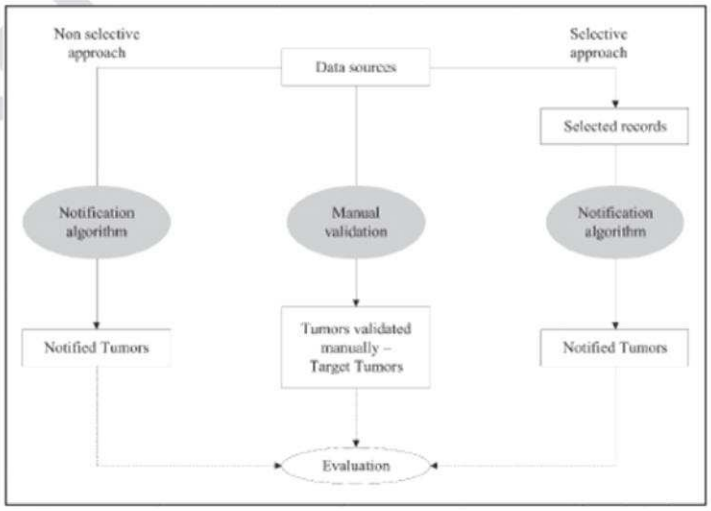


Figure 4 Strategy of evaluation of the selection algorithm. Target tumours: tumours validated by the physicians in the registry and included in the evaluation. Notified tumours: tumours that are automatically produced by an algorithm from source data

3.1.1 Selection of Qualifying Topographies

►Figure 3 presents an example of execution of the selection algorithm for a given individual.

The selection algorithm is based upon the following hierarchy:

1. Cancer care centres' data or Hospital discharge data → Diagnosis associated with surgical procedure or biopsy
2. Hospital discharge data → Diagnosis associated with lymph node dissection
3. Hospital discharge data → Diagnosis with no associated tracer procedure
4. Anatomical pathology data
5. Other diagnostic codes

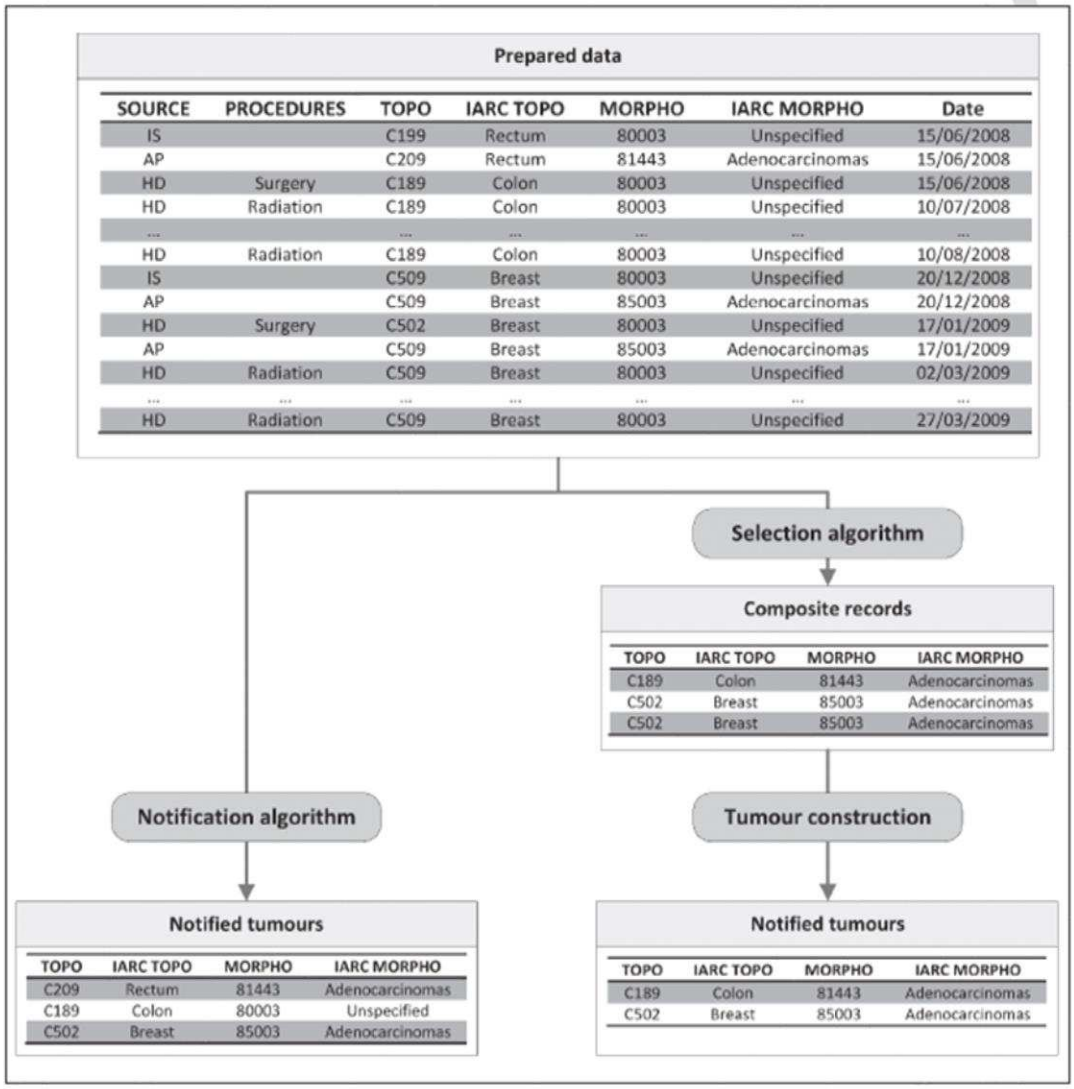


Figure 5 Example of execution of the notification algorithm with and without the selection step. Composite records derived from the selection algorithm are grouped to eliminate duplicate records. Notification algorithm without selection produced three tumours in application of registration rules for multiple primary cancers. The addition of the selection step produced only two tumours by excluding "rectum topography" reducing the noise generated by imprecise coding.

The aim of this hierarchy is to use, if available, the diagnostic code that was produced by the surgeon (or his staff) and the cancer care data. Given the data available for an individual, the algorithm retains only topographies derived from the source records corresponding to the highest hierarchical level available for the individual (level 1 corresponds to the highest possible level). Thus the different topographies derived from records at a lower level are ignored. For instance, in ▶ Figure 3 only two records associated with surgery are selected for topography notification. If the patient did not have surgery records then hospital discharge data with radiation procedure would have been selected for topography notification.

3.1.2 Selection of Qualifying Morphologies

The selection of morphologies is performed according to the same principle as for the topography, using the following hierarchy:

1. Cancer care centres' data or Anatomical pathology data
2. Hospital discharge data → Diagnosis associated with surgical procedure or biopsy
3. Hospital discharge data → Diagnosis associated with lymph node dissection procedure
4. Hospital discharge data → Diagnosis with no associated tracer procedure
5. Other diagnostic codes

The aim of this hierarchy is to use, if available, the diagnostic code that was produced by the pathologist (or his staff) and the cancer care data. For instance, in ▶ Figure 3 only three records from anatomical pathology data are selected for morphology notification.

3.1.3 Construction of Tumours

From the topographies and morphologies selected independently, the records are reconstructed for notification. Morphologies and topographies are linked one to the other, using following rules:

- **If, for a given individual, there is strictly only one topography and only one morphology**
  - The notified tumour is constructed by associating topography and morphology
- **Otherwise**
  - **If a selected topography is available in selected records for morphologies**
    - A notified tumour is constructed by associating this topography and this morphology
    - This topography and this morphology are excluded for further processing
    - The algorithm restarts from the beginning with remaining records
- **Among the remaining records**
  - All possible tumours are constructed by connecting the available topographies and morphologies.

For instance, in ▶ Figure 3 "breast topography" is available in the selected morphology records describing "adenocarcinoma". As a result the selected topography record (C502) is combined to the selected morphologies records (85003) to build two composite records. There is no more ambiguity among the remaining topography (C189) and morphology (81443). As a result, they are then combined in a unique composite record. In that example, the tumour construction, leads to three records available for the notification algorithm.

Once this processing is complete, a record is a composite of topography and morphology derived from different data sources. The records derived from the selection algorithm are then formatted to be fed directly into the registry notification algorithm.

3.2 Evaluation

The analysis was performed on the manually validated data derived from a full year registration (2008). Validated data comprise all the records present in the source data and attached to a tumour for which at least the topography, the morphology, the incidence date and the diagnostic basis have been finally validated by one of the physicians in the registry. In 2008, the registration process was as described in ▶ Figure 2. Every validated tumour had been notified by the automated notification process (without selection) and manually validated by the registry staff.

3.2.1 Evaluation Strategy for the Performance of the Algorithms

Target tumours are defined as the set of tumours validated by the physicians in the registry and included in the evaluation. Notified tumours are defined as the tumours that are automatically produced by an algorithm from source data. Concordant tumours are the notified tumours for which IARC topography, IARC morphology and tumoral behaviour correspond to the target tumour validated for an individual.

The evaluation consisted in comparing performances of an approach that does not include selection of relevant records (non-selective approach) with an approach including selection (selective approach). In either case tumours were notified from the data available, using the notification algorithm that was in routine use in the Poitou-Charentes cancer registry in 2008.

For each approach (selective and non-selective) the evaluation was performed by comparing the "target tumours" with the"notified tumours" for each patient (▶ Figure 4). ▶ Figure 5 presents the execution of the two methods for the example presented in ▶ Figure 3. Composite records derived from the selection algorithm are grouped to eliminate duplicate records. Notification algorithm without selection produced three tumours in application of registration rules for multiple primary cancers. Elsewhere notification algorithm with a step selection produced only two tumours reducing the noise generated by imprecise coding. This evaluation involves the production of the objective measures that are presented in the next paragraph.

3.2.2 Evaluation Measures: Recall, Precision, F-measure

For each individual a sparse binary matrix was constructed representing presence or absence of each type of tumour. From this

Table 1 Type of data available per individual

Type of data	N	%
Anatomical pathology data	10,757	89.9
Diagnosis with tracer procedure	8254	68.9
Healthcare insurance service	7241	60.5
Diagnosis with no tracer procedure	1445	12.1
Cancer care centres' data	664	5.5

matrix, the contingency table was derived for each type of tumour so as to calculate locally recall (sensitivity) and precision (positive predictive value). The F-measure, the harmonic mean of recall and precision, was deduced from these values [16].

3.2.3 Analyses Performed

1. We compared performances obtained by the use of the selection step along with the notification algorithm with performances obtained using the notification algorithm on its own, collating the results for "complete" tumours and the three axes defining a tumour (IARC topography, IARC morphology, behaviour). In order to produce synthetic global measures from measures calculated locally for each type of tumour, we used the micro-averaging method [16]: recall and precision were weighted by the corresponding number of individuals.
2. We compared the performances of the two approaches according to the types of data available for each individual. For this comparison, a single type of data was allocated hierarchically to each individual. Different hierarchies were used according to the axis studied :
   - Complete tumour
     1. Cancer care centres' data
     2. Diagnosis with tracer procedure and anatomical pathology data
     3. No tracer procedure and anatomical pathology data
     4. Other
   - Topography
     1. Cancer care centres' data
     2. Diagnosis with tracer procedure
     3. Diagnosis without tracer procedure
     4. Other
   - Morphology
     1. Cancer care centres' data
     2. Anatomical pathology data
     3. Other

For instance, for the analysis of the notification of complete tumours, an individual with the source type "Diagnosis with tracer procedure and anatomical pathology data" has no record of the type "Cancer care centres' data".

4. Results

4.1 Data Description

For the evaluation, 11,971 individuals with 12,346 manually validated tumours were included. Among these individuals, 11,604 had a single tumour (97%) and 367 had multiple primary sites (3%). The mean number of sources per individual (among hospital discharge data, cancer care centres' data, healthcare insurance data and anatomical pathology data) was 2.4. The mean number of records per individual was 11.2 (minimum 1 record, maximum 129 records). To execute the notification algorithms the following were used:

- 104,464 records for hospital discharge data
- 21,851 records for anatomical pathology data related to 17,205 pathology reports
- 7508 records for healthcare insurance data
- 686 records for cancer care centres' data.

▶ Table 1 presents the types of data available per individual. Almost 90% of the individuals had an anatomical pathology record available, and almost 69% an eligible diagnosis associated with a tracer procedure.

The validated tumours corresponded to 48 different topographies and 17 different morphologies, or 220 different types of tumour.

4.2 Algorithm Performances

4.2.1 Global Performances

The selective approach notified 12,493 tumours as compared to 16,014 from the notification algorithm on its own. Among the 12,493 tumours notified by the selective approach, 8996 (72%) were concordant with the target tumours. Among the errors identified, 1720 (49%) were due to the production of tumours for which the topography was correct but the morphology imprecise (rather than incorrect). An anatomical pathology record was available for 1283 of these tumours encoded with an imprecise morphology, or 37% of the discordant instances identified overall. Regarding the results for the notification algorithm on its own, among the 16 014 tumours notified, 7684 (48%) were concordant.

▶ Table 2 presents the performances of the two approaches for the notification of the different coding axes (IARC topography, IARC morphology, behaviour), and

Table 2 Recall, precision and F-Measure for different granularities. Bold font indicates the best performing method.

Granularity	Recall		Precision		F-measure	
	Select[a]	N-select[b]	Select[a]	N-select[b]	Select[a]	N-select[b]
Tumour topography	0.927	0.965	0.925	0.753	0.926	0.857
Tumour morphology	0.805	0.809	0.804	0.680	0.805	0.750
Tumour behaviour	0.894	0.826	0.909	0.772	0.901	0.806
Complete tumour	0.729	0.622	0.721	0.480	0.725	0.542

[a] With selection algorithm, [b] without selection algorithm

Table 3 Performances of the algorithms according to the types of data available per individual. Bold font indicates the best performing method.

Type of data	N [a]	Recall		Precision		F-Measure	
		Select[b]	N-select[c]	Select[b]	N-select[c]	Select[b]	N-select[c]
Notification of complete tumour							
Cancer care centre's data	664	**0.904**	0.836	**0.785**	0.570	**0.840**	0.677
Diagnosis with tracer procedure and anatomical pathology data	7380	**0.752**	0.632	**0.742**	0.483	**0.747**	0.548
Diagnosis with no tracer procedure and anatomical pathology data	2798	**0.713**	0.593	**0.735**	0.471	**0.724**	0.525
Other	1129	0.504	**0.504**	**0.501**	0.414	**0.502**	0.454
Notification of tumour topography							
Cancer care centre's data	664	0.981	**0.987**	**0.893**	0.703	**0.935**	0.821
Diagnosis with tracer procedure	7725	0.948	**0.985**	**0.940**	0.764	**0.944**	0.861
Diagnosis with no tracer procedure	1549	0.878	**0.941**	**0.873**	0.643	**0.876**	0.764
Other	2033	0.868	**0.898**	**0.918**	0.835	**0.892**	0.865
Notification of tumour morphology							
Cancer care centre's data	664	0.939	**0.942**	**0.862**	0.717	**0.899**	0.814
Anatomical pathology data	10178	0.825	**0.830**	**0.826**	0.697	**0.825**	0.758
Other	1129	**0.548**	0.543	**0.557**	0.487	**0.553**	0.514

globally for complete tumours. The selective approach demonstrates better performances for the notification of complete tumours (0.72 and 0.54 f-measures for topography and morphology respectively). This improvement holds for all the tumour components (IARC topography, IARC morphology, behaviour).

Regarding topography, the gain in performance is in terms of precision at the expense of recall, but with an improvement in the F-measure. For IARC morphology, recall is altered only slightly, and there is a clear gain in precision. For behaviour, both recall and precision are improved by the new algorithm. The morphological axis is the one that presents the least satisfactory performances.

4.2.2 Performances According to the Type of Data Available

▶ Table 3 presents the performances of the algorithms for the different coding axes and for complete tumours according to sources available for the different individuals. The selective approach shows better performances for the notification of complete tumours, with an improvement in recall and precision whatever the sources available.

Regarding IARC topography, a clear improvement is observed in the F-measure in presence of a diagnosis with tracer procedure, and a diagnosis without any procedure. This improvement results from a considerable increase in precision, alongside a slight decrease in recall. The selective approach exhibits better performances for the notification of IARC morphology when anatomical pathology records and cancer care centres' records are available. This improvement is largely due to a gain in precision, without markedly affecting recall.

5. Discussion

Our results demonstrate a clear improvement in performance with the implementation of the selective approach. We have shown the advantage of not using all the records for the notification phase. An approach involving a hierarchized selection makes it possible to process each individual case in differentiated manner, adapting the selection to the information available. In a large number of cases this enables the reduction of noise resulting from imprecision or inaccuracies in coding procedures. The notification of false multiple primary sites was reduced, and as a result the number of tumours produced after application of the hierarchical selection algorithm was far more realistic than when there was no selection procedure.

The initial hypotheses concerning the quality of coding according to the practitioner appear justified. Cancer care centres' data constitutes a very reliable source for notification, as the synthesis that is routinely produced by these centres is very close to the registry targets. However the absence of anti-cancer centres on the Poitou-Charentes region resulted in a small proportion of cases (5.5%) being detected via cancer care centres' data. For the notification of tumoral topography, the strategy, where a tracer procedure exists, of removing all the other records without a tracer procedure enables a marked improvement in precision with no major change in recall. This shows that the topographies associated with these records are accurate, and that the loss of information as a result of this selection is small. For the notification of morphology, anatomical pathology data is an essential source when there is no cancer care centre's data coding. Nevertheless, the use of complementary sources when the case file is incomplete enables a degree of exhaustiveness to be preserved, minimising damage in terms of precision.

There is discordance between the decrease in recall for the two axes, topography and morphology, and increase in recall for the complete tumour. This observation results from the construction by the selection algorithm of a composite record describing the tumour according to these two axes. In this composite record, morphology and topography can derive from different sources. If there is no selection algorithm, the notification algorithm creates two tumours. In ▶Figure 5, without selection, two tumours are generated (colon-unspecified and rectum-adenocarcinomas). If the target tumour is (colon-adenocarcinomas), no tumour matches this target whereas the record generated with the selection algorithm lead to create the target tumour thus improving both recall and precision. The recall discordance observed clearly shows that although the selection procedure generates a loss of information for each axis, the synthesis is more relevant for the notification of a complete tumour.

In anatomical pathology data, there are non-coded records (the practitioner has not coded the tumour) or else the coding is not useable (coding irregularity). In this situation, our algorithm notifies an imprecise morphology coding. In the data analysed here, this type of error accounted for 37% of the discordant tumours observed in the selective approach. An earlier study implemented a text categorisation method which demonstrated very good performances in the case of single non-metastatic tumours, with an F-measure of 0.97 for IARC morphological coding [17]. Its integration upstream from the selection algorithm could automatically generate a morphological group for these records, and would probably improve the global notification performances.

In Italy, the Varese Cancer Registry estimated the proportion of wrongly-coded topographies from their data sources to be 1.6% on a full year registration [8, 11]. Another evaluation performed on 1539 files in the Venetian Tumour Registry showed 85.7% concordance on IARC topography and IARC morphology and only 39 false positives (2.5%) [12]. These evaluations concerned cases that had several concordant sources leading to describe only one single tumour for a given patient and were then eligible to be registered automatically in these registries. These tumours were representing about 60% of the total number of registered tumours in 1997. These results are difficult to compare with ours because we used a complete year of registration. Our aim was to deal with coding inconsistencies that led to non-concordant codding for a single individual. These complex cases were excluded from evaluations performed in Italy. The Ontario Cancer Registry estimated the proportion of error to be 6.7% for notification of IARC topographies but the analysis was restricted to cases in which only a single primary had been registered for the patient [5, 18] corresponding to 91% of all registered tumours in this registry. To our knowledge, no evaluation is available about patient with automated multiple primary cancers. Our algorithm produces 7.5% errors for IARC topographic groups and 72% of complete tumours notified were concordant. Our evaluations were performed on a full year registration including multiple primary cancers.

The method developed in Italy in order to automatically register cases is based on diagnostic codes : *"To choose the best code a hierarchy was defined whereby a specific site (ICD-9) code takes precedence over a generic site code, which in turn takes precedence over a metastatic or ill-defined site code, or a code indicating secondary unspecified lymph node involvement"* [8]. The registration algorithm developed in Italy aims at determining the concordance between diagnostic codes available through multiple data sources. The selection algorithm we developed is an additional step that uses information associated to diagnostic codes (such as source or associated procedures) rather than diagnostic codes themselves. By selecting, for a given individual, relevant records (and excluding non-relevant record) before evaluating diagnostic codes concordance, the selection algorithm reduces the effect of coding inconsistencies. These inconsistencies might represent a significant part of the cases that cannot be registered automatically due to non-concordant diagnostic codes. Tognazzo et al. [13] proposed a probabilistic approach in order to find new cancer cases. Their evaluations showed that the source of a discordant diagnostic code was significantly influencing both random forest and multilogit model. The deterministic selection process that we developed shows that information associated with diagnostic codes (such as data source and procedures) plays a significant role in new cases definition.

Notification of tumours as accurately as possible is a genuine asset for the manual validation phase. By decreasing the number of alterations required during validation as far as possible, while at the same time allowing the operator access to all the non-selected information, the approach necessarily improves the efficiency and the quality of the registration procedures.

The year of incidence was not included in the definition of the tumour. One of the main tasks in future work will be, from the data available, to detect situations in which a case is likely to be previous to the date of the first notification by one of the sources (delayed relapse, late instatement of care, therapeutic abstention). As the rules we applied were based on the registry staff knowledge of data sources, we did not evaluate other hierarchical selection. The result of this study suggests that the context of a diagnostic code plays a significant role. The Poitou-Charentes cancer registry information system provides a manually validated relation between tumours and data sources. This data set will be used to build learning and test sets in order to develop and evaluate probabilistic approaches with development of context encoding.

The ongoing work on improvement of the quality of data in the information system of the Poitou-Charentes cancer registry has led to the development of a selection algorithm for relevant records. The present work has made it possible to demonstrate the value of a selective approach compared to an exhaustive approach in the definition of tumours to be notified. The use of only part of the information available for notification, via the hierarchical selection of qualifying records, enables the removal of a large part of the noise generated by imprecise or incorrect records. In a multi-source system, it is coherent to use a typology of information to determine its qualifying relevance. It should however be noted that this approach is envisaged solely in the context of a manual validation of

case files by an operator who has access to all the information (not merely the hierarchized selection). Although the approach consisting in selecting information for notification might be viewed as risky, it can nevertheless be considered coherent and amply justified when implemented in association with the development of an exhaustive summary of care trajectory, enabling the operator to have access to all the available information for case validation.

Acknowledgements

The authors would like to thank the medical records assistants, Nicolas Mériau, Sébastien Orazio and Soizic Lellouch of the *Registre des cancers de Poitou-Charentes* for their helpful contribution to the implementation of this research. We also wish to thank Angela Swaine Verdier for the translation of the manuscript.

References

1. Buemi A. Pathology of Tumours for Cancer Registry Personnel. Lyon: IARC; 2008.
2. Percy C, Fritz A, Jack A, Shanmugarathan S, Sobin L, Parkin D, et al. International Classification of Diseases for Oncology (ICD-O). Third ed: World Health Organization; 2000.
3. Curado M, Okamoto N, Ries L, Sriplung H, Young J, Carli M, et al. International rules for multiple primary cancers. ICD-O Third Edition, 2004.
4. Comité national des registres (CNR). Appel à qualification et procédure d'évaluation. Institut de veille sanitaire; 2010 (updated 10/06/2010; cited 26/05/2011). Available from: http://www.invs.sante.fr/surveillance/comite_national_des_registres/fonctionnement.htm.
5. Black RJ, Simonato L, Storm HH, Démaret E. Automated Data Collection in Cancer Registration. Lyon: IARC; 1998.
6. MacKay EN, Sellers AH. The Ontario cancer incidence survey, 1964–1966: a new approach to cancer data acquisition. Can Med Assoc J 1973; 109 (6): 489 passim.
7. Clarke EA, Marrett LD, Kreiger N. Cancer registration in Ontario: a computer approach. In: Jensen OM, Parkin DM, MacLennan R, Muir CS, Skeet RG, editors. Cancer Registration: Principles and Methods. IARC Sci Publ 1991; 95: 246–257.
8. Contiero P, Tittarelli A, Maghini A, Fabiano S, Frassoldi E, Costa E, et al. Comparison with manual registration reveals satisfactory completeness and efficiency of a computerized cancer registration system. J Biomed Inform 2008; 41 (1): 24–32.
9. Cancer incidence in five continents. Volume VIII. IARC Sci Publ 2002; 155: 1–781.
10. Olive F, Gomez F, 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 Santé Publique 2011; 59 (1): 53–58.
11. Tagliabue G, Maghini A, Fabiano S, Tittarelli A, Frassoldi E, Costa E, et al. Consistency and accuracy of diagnostic cancer codes generated by automated registration: comparison with manual registration. Popul Health Metr 2006; 4: 10.
12. Tognazzo S, Andolfo A, Bovo E, Fiore AR, Greco A, Guzzinati S, et al. Quality control of automatically defined cancer cases by the automated registration system of the Venetian Tumour Registry. Cent Eur J Public Health 2005; 15 (6): 657–664.
13. Tognazzo S, Emanuela B, Rita FA, Stefano G, Daniele M, Fiorella SC, et al. Probabilistic classifiers and automated cancer registration: an exploratory application. J Biomed Inform 2009; 42 (1): 1–10.
14. 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): 660–666.
15. Berg JW. Morphologic classification of human cancer. In: Shottenfeld D FJ, Jr., editor. Cancer epidemiology and prevention. 2nd ed. New York: Oxford University Press; 1996.
16. Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv 2002; 34 (2): 1–47.
17. Jouhet V, Defossez G, Burgun A, Le Beux P, Levillain P, Ingrand P, et al. Automated Classification of Free-text Pathology Reports for Registration of Incident Cases of Cancer. Methods Inf Med 2012; 51 (3): 242–251.
18. Holowaty EJ, Lee G, Moravan V, Chong N, Dale DJ. A Reabstraction Study to Estimate the Completeness and Accuracy of Data Elements in the Ontario Cancer Registry. Report submitted to Health Canada. Cancer Care Ontario. Toronto, Canada, 1996.

Tous les cas créés par l'algorithme de notification se voient affecter une date de diagnostic, une topographie et une morphologie CIM-O3, une base de diagnostic, un code commune Insee de résidence et un niveau de validité nulle. Ce niveau stipule que la tumeur et les données sources qui ont permis sa création n'ont fait l'objet d'aucune validation manuelle. Les informations associées peuvent ainsi faire l'objet d'une mise à jour lors de l'ajout de nouvelles données sources, cet ajout déclenchant l'actualisation de la procédure de notification.

Tout cas jugé hors d'intérêt du RGCPC est signalé de manière appropriée (patient résidant en-dehors du Poitou-Charentes au diagnostic, cas diagnostiqués avant le 1er janvier 2008, *i. e.* considérés comme « prévalent »). Tous les autres cas sont considérés comme incidents et entrent dans une phase de revue manuelle par les opérateurs du registre.

Contrôle et notification manuelle des cas suspects

Si la très large majorité des tumeurs sont créées algorithmiquement, certaines situations :

1. Peuvent être à l'origine d'erreurs de notification (ex : erreurs ou imprécisions de codage, disponibilité partielle des données sources au moment de la notification),
2. Sont des motifs connus de non-notification au sein du RGCPC :
    a. Cas n°1 : Données de biologie = pas de notification automatique compte tenu de l'absence de métadonnées codées;
    b. Cas n°2 : Données d'ALD de l'Assurance Maladie = pas de notification automatique engagée compte tenu du faible rappel et précision de ces données lorsqu'elles sont isolées.

Dès lors, des programmes de contrôle sont exécutés précocement sur la base de données (au début de chaque nouvelle année d'enregistrement) pour individualiser ces dossiers et les repositionner au terme de leur relecture dans le schéma de validation classique :

i) comme des cas incidents de cancer suspectés (notification manuelle),
ii) comme une récidive
iii) comme un cas faux positif (cas prévalent, hors territoire, tumeur bénigne, sans objet …).

Sont revues à cette étape toutes les métadonnées non codées (ACP, RCP, BIO) et les informations émanant des sources de données qui n'ont fait l'objet d'aucune pré-affiliation par l'algorithme à une tumeur à valider. A ces dossiers s'ajoutent les cas notifiés l'année X+1 et pour lesquels certaines données sources sont datées sur l'année X (contrôle de la date de diagnostic). Une fois ces listes de contrôles épuisées, les cas incidents de cancer suspectés (*i. e.* ceux notifiés manuellement) s'ajoutent aux tumeurs notifiées algorithmiquement.

2.3.4.2. Algorithme 2 : Représentation du parcours de soins [35]

La seconde phase algorithmique du processus de préparation de l'information consiste à modéliser le parcours de soins de chaque individu, et de mettre à la disposition des opérateurs du registre les outils de visualisation et de saisie adaptés pour un processus de validation manuelle optimisé.

- Principe général de fonctionnement

La méthode consiste à modéliser le parcours de soins comme une succession ordonnée (séquence) d'évènements horodatés, qui sont ensuite agrégés pour rendre compte de chaque épisode de soins significatif survenu durant la prise en charge des patients. Pour ce faire, une typologie des évènements traceurs est dressée à partir des codes diagnostics et des actes CCAM (médicaux et chirurgicaux) identifiables dans les données PMSI (voir encadré sur la CCAM page suivante). A ces données sont ensuite associées, une fois agrégées, les informations issues des autres sources (ACP, AMA, RCP, EPC, BIO).

Cette synthèse chronologique est rendue consultable pour chaque patient au sein d'une interface sécurisée, offrant une lecture synthétique et un accès ergonomique à l'ensemble des informations disponibles pour un individu dans les données sources (voir illustration Figure 4). Cette représentation du parcours fournit aux opérateurs des informations fondamentales pour la validation des cas :

- Analyse et repérage du début de la séquence de soins (élimination des cas prévalent, identification des récidives, des seconds cancers),
- Confrontation et analyse de la concordance des informations des différentes sources (appréciation des erreurs ou imprécisions de codage, des faux négatifs et faux positifs),
- Analyse de la cohérence des parcours et des séquences thérapeutiques empruntées par les patients (recommandations de bonne pratique, stades d'extension),
- Identification des sites et établissements empruntés par les patients (optimisation de l'investigation, sélection du site pour le retour au dossier médical).

Cette phase de confrontation et d'analyse des informations au moment de la validation permet aux opérateurs du registre d'opérer un alignement de l'information, et de transformer une base de signalements - de niveau de validité initiale faible (données brutes) - en une base de données exhaustive en population - avec un niveau de validité finale élevée (données qualifiées).

La Classification Communes des Actes Médicaux (CCAM)

La CCAM est une nomenclature française de la Sécurité Sociale répertoriant et identifiant les gestes pratiqués par les médecins, les chirurgiens-dentistes et les sages-femmes. Elle sert à établir la tarification des séjours hospitaliers dans les hôpitaux publics et privés dans le cadre de la T2A (tarification à l'activité), les honoraires pour les interventions réalisées dans les cliniques privées et les honoraires des actes techniques réalisés lors des consultations en médecine libérale et en milieu hospitalier.

La codification des actes médicaux est établie sur la base d'un code alphanumérique composé de 4 lettres et de 3 chiffres (illustration ci-dessous) :

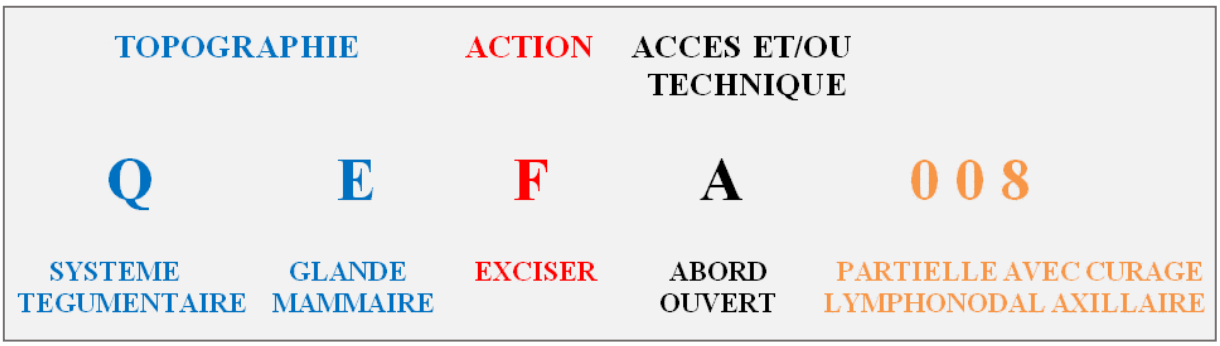
la 1ère lettre désigne un grand appareil anatomique ;

la 2nde précise l'organe (ou la fonction) dans l'appareil correspondant à la première lettre ;

la 3ème lettre désigne l'action effectuée ;

la 4ème lettre identifie la voie d'abord ou la technique utilisée ;

les 3 chiffres suivants servent à différencier les actes possédant 4 lettres clefs identiques.



A chaque acte CCAM est affecté un « type d'évènement » parmi 15 catégories définies par le RGCPC. Huit des 15 catégories préfigurent d'évènements hiérarchiques dits « traceurs » dans le parcours de soins : 1/ « Exérèse et curage », 2/ « Exérèse », 3/ « Curage », 4/ « Biopsie ostéo-médullaire », 5/ « Myélogramme », 6/ « Prélèvement anatomopathologique », 7/ « Radiothérapie », 8/ « Réparation, reconstruction ».

Chaque RUM (PMSI) se voit attribué par l'algorithme de représentation du parcours de soins un évènement traceur unique (fait marquant se produisant à un instant t). En premier lieu, l'évènement traceur est déduit du motif d'hospitalisation s'il fait référence à un diagnostic principal (DP) de « Radiothérapie » (Z510), de « Chimiothérapie » (Z511) ou de « Soins palliatifs » (Z515). Sinon, l'évènement traceur est déduit de la catégorie d'appartenance de l'acte CCAM. La sélection est hiérarchique en présence de plusieurs actes CCAM sur un même RUM.

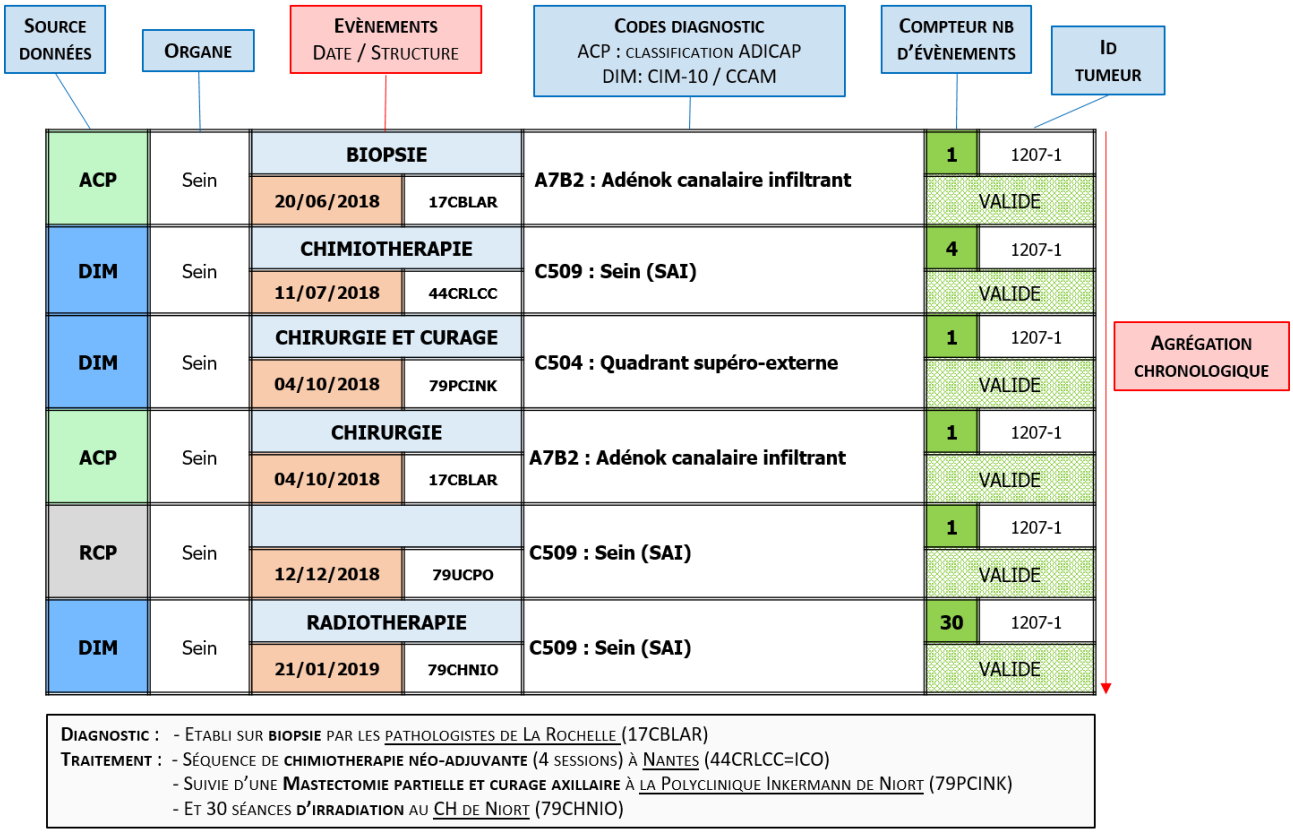


Figure 4 : Représentation du parcours de soins d'une patiente diagnostiquée en 2018 pour un cancer du sein

- Evaluation des performances de l'algorithme : application au cancer du sein

Au-delà de l'aide que représente cette modélisation du parcours de soins pour l'enregistrement des cas incidents de cancer, l'objectif sous-jacent était de pouvoir déduire de l'analyse des parcours de soins des indicateurs d'évaluation des pratiques (ex. délais et adéquation de prise en charge), en sus des objectifs classiques de surveillance épidémiologique des cancers. L'enjeu était de réutiliser les données médicales collectées et le travail d'alignement de l'information produit en routine par les opérateurs du registre au moment de la validation.

Une première application de l'algorithme a été mise en œuvre dans le cancer du sein. La performance de l'algorithme à reproduire les données réelles de trajectoires des patientes a été évaluée à l'aide de mesures de dissimilarité (mesure d'édition de Levenshtein généralisée), en confrontant les séquences générées par l'algorithme aux séquences réelles reconstruites à partir des éléments recueillis dans les dossiers médicaux.

Les séquences étaient disponibles sous deux formes :

i) Une forme « simple », représentative de la présence et de l'ordre des évènements,

ii) Une forme « étendue », faisant intervenir la durée des états (avec des périodes sans évènements).

L'organisation structurée en base de données relationnelles datées (« time-stamped database ») du RGCPC était propre à faciliter la mise en application de telles techniques. La forme étendue était générée sous la forme d'une chaîne de caractères, où chaque caractère était répété en fonction de la durée de l'état. Ainsi, la durée de chaque évènement et le calcul des délais pouvait être mis en œuvre grâce à l'utilisation d'expressions régulières sur les chaînes de caractères (Perl regular expressions, SAS).

L'évaluation a porté sur un échantillon de 159 patientes atteintes de cancer du sein non métastatique. Au global, 98% des parcours étaient correctement reconstruits sur l'ordonnancement des états (forme simple), et 94% des parcours étaient fidèles à la réalité à 3 jours près et 88% à 1 jour près en prenant en compte la durée des états (forme étendue). Les dissimilarités entre séquences étaient liées dans la majorité des cas à des erreurs de dates ou des séances de chimiothérapie ou de radiothérapie manquantes en début ou fin d'évènement. Dans ce dernier cas de figure, la variabilité des dissimilitudes observées était mécaniquement liée aux schémas de traitement, les dissimilarités étant plus faible pour la radiothérapie que la chimiothérapie (le traitement de référence s'étalant sur 33 jours en moyenne pour la radiothérapie contre 9 à 15 semaines pour la chimiothérapie).

Dans la continuité de ces travaux, l'algorithme a été appliqué à de larges populations de patients, afin d'en évaluer l'applicabilité et d'éprouver la prédisposition du registre à produire les indicateurs clés

d'évaluation des pratiques de soins. Un projet nommé TRAJAN[10] a été déposé et financé par l'INCa en 2013 dans le cadre de l'appel à projets libres SHS-E-SP (Sciences Humaines et Sociales, Epidémiologie et Santé Publique) pour l'analyse à grande échelle des trajectoires de soins des patients atteintes de cancer du sein (voire section 3.3.1 « Surveillance des délais d'accès aux traitements »). La logique de modélisation a été appliquée ensuite progressivement à d'autres localisations cancéreuses (cancer colorectal, cancer broncho-pulmonaire), en prenant soin d'adapter la granularité des évènements aux référentiels HAS-INCa publiés.

De l'utilisation de cet algorithme a finalement découlé l'intérêt d'ouvrir le champ des applications, aussi bien dans le domaine de l'évaluation des pratiques de soins que de l'évaluation des services de santé (health services research). L'enjeu était de réutiliser des données électroniques externes pour enrichir la connaissance et la contextualisation des parcours des patients en réponse aux objectifs d'évaluation et de recherche fixés.

Lire Article 3 : *Defossez G, Rollet A, Dameron O, Ingrand P. Temporal representation of care trajectories of cancer patients using data from a regional information system: an application in breast cancer.* ***BMC Med Inform Decis Mak****. 2014 Apr 2;14:24. doi: 10.1186/1472-6947-14-24. PMID: 24690482; PMCID: PMC3983896.*

Accéder à la suite « Vers une démarche intégrative de données … »

[10] TRAJAN. Evaluation du parcours de soins de patientes atteintes de cancer du sein. Application d'un algorithme de représentation temporelle synthétique des trajectoires de soins à partir des données du système d'information régional du registre des cancers de Poitou-Charentes (coordonnateur du projet : Gautier DEFOSSEZ)

Defossez et al. BMC Medical Informatics and Decision Making 2014, 14:24
http://www.biomedcentral.com/1472-6947/14/24

BMC Medical Informatics & Decision Making

RESEARCH ARTICLE Open Access

Temporal representation of care trajectories of cancer patients using data from a regional information system: an application in breast cancer

Gautier Defossez[1*], Alexandre Rollet[1], Olivier Dameron[2] and Pierre Ingrand[1,3]

Abstract

Background: Ensuring that all cancer patients have access to the appropriate treatment within an appropriate time is a strategic priority in many countries. There is in particular a need to describe and analyse cancer care trajectories and to produce waiting time indicators. We developed an algorithm for extracting temporally represented care trajectories from coded information collected routinely by the general cancer Registry in Poitou-Charentes region, France. The present work aimed to assess the performance of this algorithm on real-life patient data in the setting of non-metastatic breast cancer, using measures of similarity.

Methods: Care trajectories were modeled as ordered dated events aggregated into states, the granularity of which was defined from standard care guidelines. The algorithm generates each state from the aggregation over a period of tracer events characterised on the basis of diagnoses and medical procedures. The sequences are presented in simple form showing presence and order of the states, and in an extended form that integrates the duration of the states. The similarity of the sequences, which are represented in the form of chains of characters, was calculated using a generalised Levenshtein distance.

Results: The evaluation was performed on a sample of 159 female patients whose itineraries were also calculated manually from medical records using the same aggregation rules and dating system as the algorithm. Ninety-eight per cent of the trajectories were correctly reconstructed with respect to the ordering of states. When the duration of states was taken into account, 94% of the trajectories matched reality within three days. Dissimilarities between sequences were mainly due to the absence of certain pathology reports and to coding anomalies in hospitalisation data.

Conclusions: These results show the ability of an integrated regional information system to formalise care trajectories and automatically produce indicators for time-lapse to care instatement, of interest in the planning of care in cancer. The next step will consist in evaluating this approach and extending it to more complex trajectories (metastasis, relapse) and to other cancer localisations.

Keywords: Epidemiology, Evaluation, Care trajectory, Temporal reasoning, Data integration, Cancer

* Correspondence: gautier.defossez@univ-poitiers.fr
[1]Unité d'épidémiologie, biostatistique et registre général des cancers de Poitou-Charentes, Faculté de médecine, Centre Hospitalier Universitaire de Poitiers, Université de Poitiers, 6, rue de la milétrie, Poitiers, Cedex BP 199 86034, France
Full list of author information is available at the end of the article

Background

Care trajectories and guidelines

The care provided for patients with cancer requires a multi-disciplinary approach. Throughout the course of their illness, patients undergo a series of diagnostic investigations and surgical and medical treatments which are all occasions for contact with the care system. They follow care trajectories that vary according to geographical, temporal, institutional, medical, economic or social factors. The trajectory can become more complex in case of relapse or intercurrent illnesses, or it can be simpler if the patient is cured or stabilises. In addition, an optimal trajectory will depend on patient compliance, avoidance of redundancy in investigations, prevention of complications and delivery of appropriate therapies. Clinical practice guidelines have been developed for this purpose by groups of experts using updated information from medical research along the lines of Evidence-Based Medicine. The purpose of these guidelines is to reduce variability in practices, control costs, and above all improve the quality of care. Nevertheless, in actual operation, these recommendations are not always easy to implement in settings where multidisciplinary care involves several different professionals, or even several institutions. Indeed, the fact that recommendations exist does not always mean that they are put into practice, and there is the question of awareness of updates, and of the need to comply with them [1].

Interest of modeling care trajectories

There is a need to describe, analyse and understand care trajectories by modeling the itineraries followed by patients. This process could also enable the evaluation of the appropriateness of care provision in relation to reference standards, and thus contribute to improving the organization of healthcare and to determining strategic choices [2]. The orientation of a patient through an optimized trajectory does indeed depend on satisfactory coordination among the different protagonists, and adequate planning of the care itself. In France, cancer care is evolving towards the formation of regional networks, so as to coordinate expertise, services and resource allocation. The regional health agencies (in France ARS –*Agences Régionales de Santé*) define and implement regional plans for hospital care aiming to meet requirements of accessibility and quality. Time is often a central element in care provision, which is why the reduction of the time-lapse to instatement of care is a major strategic orientation [3]. Modelling care trajectories offers an explicit process-oriented view of healthcare and will enable routine evaluation of the compliance of observed care trajectories with those set out in guidelines. The production of care trajectories and waiting-time indicators at the regional population level should contribute to improve care planning, ultimately ensuring that all patients have access to the appropriate treatment within an appropriate time-lapse.

International context, computerization of medical data and interoperability

The observation of the actual functioning of the care system requires the existence of information systems suited to following up patient care trajectories in a given environment. Although numerous health information systems have been set up in developed countries, they were not specifically designed for this particular purpose. Every year the care system generates enormous amounts of data. The information required is fragmented and spread across a number of sources. Yet there are few tools able to mobilise and integrate this data for the purpose of describing and modeling a set of care trajectories that are characteristic of real-life patient trajectories.

The first work conducted in this area of health and on the scale of a population was certainly that concerning the classifications of patients, one of the first of which was the DRG (Diagnosis Related Groups). These were developed from the 1960s by Fetter [4], and the aim was to define comparable care provision groups in which individuals were expected to use the same level of hospital resources. These DRG led to numerous adaptations across the world, for instance the PMSI in France (*Programme de Médicalisation des Systèmes d'Information*), and to a whole body of related research [5-7].

The representation of care trajectories using data mining methods is a dynamic research area. These methods are useful to seek sequential patterns corresponding to the most frequent patient trajectories and to conduct formal analysis of concepts enabling the description of patient flows generating easily understandable graphical representations [6-15]. While methods of data mining aim to discover details in clinical trajectories or clinical pathways, the number of patterns discovered need to be restricted to the main stages defined by guidelines when the objective is to provide routine evaluation indicators of the compliance of observed care trajectories to guidelines. Clinical trajectories usually yield models restricted to a single piece of hospital information system. To our knowledge, there is no approach to date that has integrated multiple-source data from all hospitals and health structures involved in cancer care.

Cancer registry data

Internationally, cancer registries have achieved a high degree of standardization of definitions, classification systems, methods of analysis of data. This has been important in ensuring the comparability of incidence data from cancer registries, and has enabled their increasing use in epidemiological, clinical and health services management studies [16]. Most modern cancer registries use

information sources on computer media at some point in the data collection process. Work on the development of a multi-source information system centred on the patient has been conducted to collect relevant information for cancer case registration according to international rules [17]. The General Cancer Registry of Poitou-Charentes covers an administrative region of 1.8 million people in south-western France. The increasing availability of computerized information on cancer patients from pathology databases, hospital administration systems and other computerized data sources has led this registry to extend its computer systems to exploit these new opportunities.

Objectives

The main objective of this study was to develop a representation of care trajectories over time for new cancer patients, using the data from the Poitou-Charentes Regional Cancer Registry information system, and to assess the reliability and validity of this representation by confronting trajectories derived from cancer registry data used by the algorithm with observed care trajectories documented from medical records. Since breast cancer is the most frequent cancer in women in France [18] and since it has been subject for many years to recommendations based on updated survey data [19,20], we chose to make a first illustration of an application of the algorithm in the setting of non-metastatic breast cancer.

International registration rules

In compliance with national and international recommendations [17], since January 1st 2008 the General Cancer Registry of the Poitou-Charentes region (south-western France) has included all the incident cases of malignant tumour (haematological malignancies and solid tumours not including baso-cellular skin carcinomas), involving subjects regularly residing in the Poitou-Charentes region at the time of diagnosis.

Information system

The information system uses an Oracle 11 g database. The data management procedures were developed in the SAS environment (version 9.3).

The database of the cancer registry is a time-stamped relational database storing medical data with the date of occurrence of every event (such as date of admission, date when a biopsy was performed). Temporal intervals (such as the start, end and duration of a treatment) are not directly available but can be computed from the data.

The cancer registration database results from the investigation of five types of electronic data sources in the Poitou-Charentes area and in neighboring départements (Data sources), where each data source comprises all facilities, centers or establishments involved in cancer care to ensure the completeness and the coverage of all areas of cancer care for patients residing in Poitou-Charentes area.

A single program manages data capture in the system, with distinction according to the formats of each source. The integration process of data from different sources for a given individual enables each source record to be linked to the individual to which it belongs, performed by an identity server (Computerised record linkage).

A single dataset has been created for the study of temporal representation of care trajectories by extracting from two of the five data sources (anatomical pathology data and hospital discharge data), which concern 93 facilities or hospitals.

Data sources

The Registry has established collaboration with more than 100 partners in the Region and in surrounding departments. Five types of data sources, using various terminologies to describe diagnosis, are routinely collected:

- Anatomical-pathology (AP) data (pathology data) which includes free-text reports related to one or several ADICAP diagnostic codes (*Association pour le Développement de l'Informatique en Cytologie et en Anatomie Pathologique* – the French classification of lesions with topographical and histological axis [21]) (n = 28 facilities).
- Hospital discharge (HD) data recorded in the French medical information program (PMSI) [22] which includes ICD-10 diagnostic codes and CCAM medical procedure coded fields (*Classification Commune des Actes Médicaux* – the health insurance classification [23]) (n = 65 hospitals).
- Full reimbursement for cancer care granted by the French healthcare insurance service (IS) which includes ICD-10 diagnostic codes (n = 3 facilities).
- Data from cancer surveillance (CS) in cancer care centres ("*Centres de Lutte contre le Cancer*") which includes ICD-O3 tumour codes. Cancer surveillance is a system of data collection in oncology promoted in France in 1975 by the National Federation of Cancer Care Centers. It enables the identification of topographic and histological diagnosis of tumours, their initial extension, as well as therapeutic and outcome data for all cases in a given cancer care center (n = 3 centres).
- Data from Multidisciplinary Consulting Meetings (MCM) which may or may not include ICD-10 diagnostic codes. Multidisciplinary Consulting Meetings are held and serve for exchange among specialists from different disciplines concerning the diagnostic and therapeutic strategies to adopt for cancer patients. They are an essential part of the organisation

of cancer care (n = 2 regional networks). These meetings are known as Institutional Tumor Boards in the United States.

Each of these bodies regularly transfers the information required by the Registry in standardized structured encrypted files. A single program manages data capture in the system, distinguishing sources. The source data is the most detailed level of information available in the system.

Computerised record linkage

A primary function in the operation of a cancer registry is to bring together information describing the same individual from a variety of data sources. Because multiple notifications of the same tumour are expected if several sources of information are used, effective procedures for linking data on the same individual are very important, minimizing duplicate registrations of the same tumour and/or individual. So data extractions include patient identity (name, surname, birthdate ...). When loading data in the registry information system, patient identities are integrated into an identity server which by way of a semi-automated process identifies data related to one and the same patient. The patient identification process, based on computerised record linkage [16], enables automated linkage (using deterministic rules), detection of ambiguities, duplicate searches and manual patient grouping or separation. This process enables all eligible records from data sources integrated into the information system to be related to single patients.

Cancer registration

A notification algorithm [24] determines, for each patient, the tumours that should be notified to cancer registry staff according to registration rules for multiple primary cancers [25] and already validated tumours. Then each case is manually checked by registry staff by visual inspection of information sources, assessing the need to refer to a patient's medical record to register the tumour. Finally, all registered tumours contain relevant information, including date of diagnosis, ICD-O3 topography, ICD-O3 morphology and basis of diagnosis, and they are systematically related to the data sources. Following this step, the source data related to non-metastatic breast cancers can be selected on the basis of this relationship to enable the representation of the care trajectories.

Methods

Definitions

Care trajectory

The care trajectory is the term used to refer to the itinerary of a patient through the healthcare system and among the different actors over a continuous period from the onset of the illness to its resolution [26]. In cancer care, the care trajectory can comprise several successive episodes of care, and this entails a distinction between the phases of diagnosis, treatment, consolidation, cure or local or distant relapse.

Event and state

The representation algorithm developed is able to take account of each stage in the initial care provision for the tumour. The method chosen consists in modelling the care trajectory as an ordered succession of dated events aggregated into states, the granularity of which is defined from care provision guidelines.

An event is a phenomenon considered to be local and instantaneous, occurring at a particular time-point. An event is a time-stamped attribute (like a chemotherapy administration, regardless of events occurring before of after).

When the event lasts over time, we refer to a state, which is as a temporal interval with a start and an end. A state corresponds to the aggregation of repeated occurrence of several events over time according to their type and their chronology. A state corresponds to a phase or a period in treatment, defined as a stage in the initial care provision previously derived from guidelines.

For instance, the repeated occurrence of several events in chemotherapy before surgery treatment is equivalent to the definition of the state "neoadjuvant chemotherapy".

Time scale

The time scale is determined by the information systems searched, which here, as in most areas of health, date the events occurring in the course of patient care according to calendar days. Consequently, the smallest measurable interval in the present study is the calendar day, and a point event is represented as an interval of one day.

Identification of standard sequences of initial care provision for non-metastatic breast cancer

The main stages in the initial care provision for non-metastatic breast cancer identified in the national guidelines [20] are listed below:

1. A diagnostic period involving pathology investigations to confirm any suspected malignancy following clinical and/or radiological examination.
2. A period of pre-surgery treatment for infiltrating, voluminous and/or inflammatory cancers, indicated to obtain initial reduction of tumour volume. It may be envisaged for cancers that are non-operable at the outset, or according to the size of the tumour, so as to enable partial surgery . The reference treatments are hormone therapy and chemotherapy.

3. A surgery period, since as in the case of most cancers the treatment of breast cancer is ideally based on surgical removal of the tumour. This removal should be accompanied by homolateral axillary lymph node dissection. In case of an infiltrating tumour of small size and in the absence of palpable axillary adenopathy or suspect ultrasound scan image, the sentinel lymph node technique can be used. The surgical period also involves pathology examination of the surgical piece, conducted extemporaneously or following surgery, and this enables confirmation of malignancy.
4. A post-surgery medical period covering different therapies:
   - Chemotherapy, mainly anthracyclines and taxanes, is initiated 3 to 6 weeks after surgery, generally in 4 to 6 administrations 21 days apart, although patterns can vary according to treatment protocols. Adjuvant chemotherapy therefore lasts 9 to 15 weeks.
   - HER2-targeted therapies, such as trastuzamab, are indicated when there is significant HER2 over-expression by the tumour. Depending on protocols, the administration cycle can vary, and can be either sequential, i.e. initiated after the chemotherapy, or concomitant with the administration of taxanes. The duration of administration is generally 1 year.
   - Hormone therapy is only indicated in hormone-sensitive tumours. The duration of treatment is generally 5 years. Hormone therapy is initiated after any chemotherapy and radiotherapy.
5. A radiotherapy period, involving irradiation of the tumour site and of axillary lymph nodes for non-metastatic breast cancer. The reference for all volumes treated is 50 gy in 25 fractions over a period of 33 days. The duration of radiotherapy can be increased by one or two weeks in cases where the patient presents a risk of relapse. It is therefore recommended that radiotherapy should be initiated not more than 12 weeks after surgery if no chemotherapy is planned. If chemotherapy is indicated, radiotherapy should be started not more than 5 weeks after the chemotherapy, and not more than 6 months after surgery. Neither immediate breast reconstruction nor prescription of targeted therapy should alter these time lapses.

If these recommendations are synthesised, omitting targeted therapies and hormone therapy which are administered over long periods, it is possible to distinguish three standard care sequences for non-metastatic breast cancer, taking account of presence or absence of neo-adjuvant or adjuvant medical treatment (Figure 1). These standard sequences were constructed manually to define the different states in non-metastatic breast cancer management that are expected to occur in cancer care trajectories, and which are to be identified by the algorithm.

Representation of care trajectories

The procedures for the temporal representation of care trajectories involve the following three stages:

1. Identification of tracer events in the care trajectory for each individual within the source datasets.
2. Chronological aggregation of events into states according to the level of granularity defined in step 1.
3. Representation of care sequences in a simple form (showing presence and order of states) and an extended form (including the duration of states) from which time lapses are calculated.

Identification of tracer events

Events in the course of chemotherapy or radiotherapy are identified in HD data using ICD-10 diagnostic codes for chemotherapy (Z511) and radiotherapy (Z510) linked to the ICD-10 diagnostic code for breast cancer. For instance (Z511-C509) in source data refers to a "chemotherapy" event.

Surgical events – tumour removal, tumour and lymph node removal, and lymph node removal alone are identified in HD data from the 3rd alphanumerical character in CCAM codes (defining the action performed) and linked to ICD-10 code for breast cancer. For instance (C502-QEFA001) in the source data refers to the event "tumour and lymph node removal".

Biopsy and surgery events are identified in AP data from the first character in ADICAP codes (identifying the sampling mode) for breast cancer. For example, (ADICAP-OHGSA7B2) in the source data refers to the event "surgery".

In summary, HD data provide information on the nature of the surgical act performed, and AP data enables precise description of the act.

Chronological aggregation of events into states

So as to be able to model sequences and position them on a temporal axis, we used a set of predefined relationships according to a linear approach which includes the notion of the time lapse [27,28]. The choice then was to retain only one characterised event per point in time before aggregating, rather than generating more numerous different states. This choice obeys the following hierarchy: tumour removal and/or lymph node removal > chemotherapy and/or radiotherapy > pathology sampling.

The aggregation takes account of the chronology of events and complies with the following two rules:

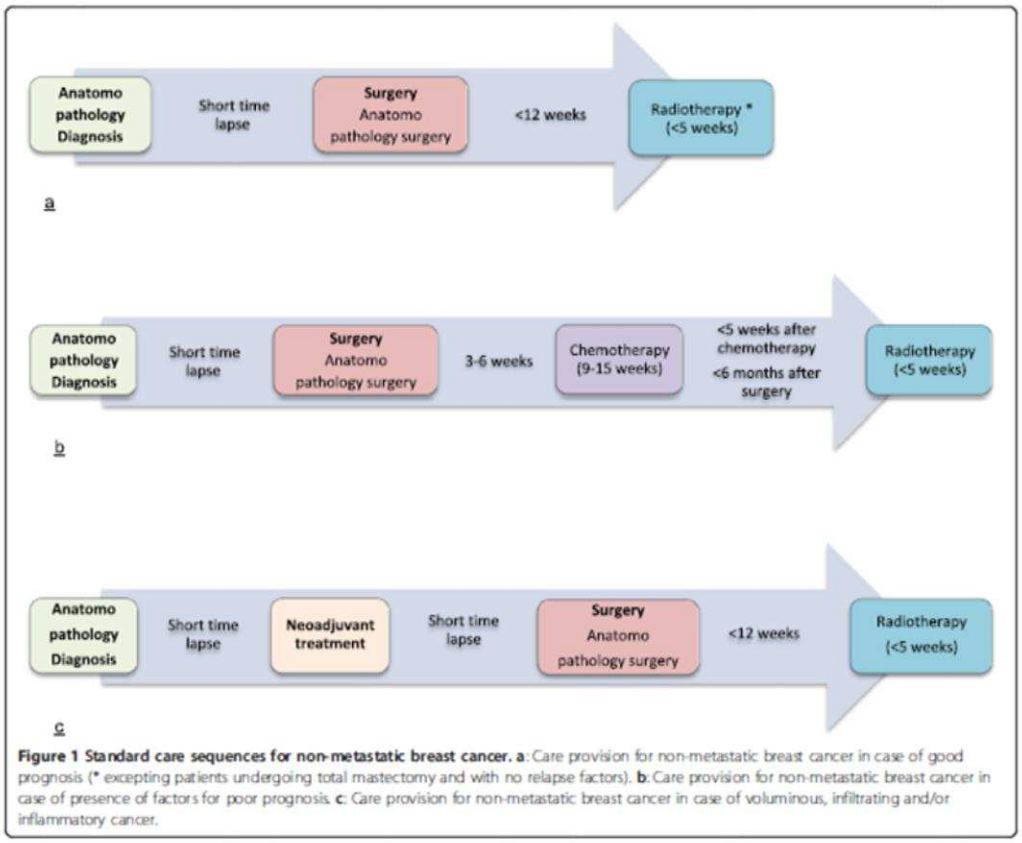


Figure 1 Standard care sequences for non-metastatic breast cancer. a: Care provision for non-metastatic breast cancer in case of good prognosis (* excepting patients undergoing total mastectomy and with no relapse factors). **b**: Care provision for non-metastatic breast cancer in case of presence of factors for poor prognosis. **c**: Care provision for non-metastatic breast cancer in case of voluminous, infiltrating and/or inflammatory cancer.

- If an event occurs between two records that would normally have been aggregated, the aggregation is not performed. For instance, if the patient undergoes surgery between two chemotherapy sessions, the aggregation of these two sessions into a "chemotherapy" state does not occur.
- If the time-lapse between two events of the same type is too long, the aggregation is not performed, and this enables the differentiation of two distinct care episodes (six months between two chemotherapy administrations, one month between two radiotherapy sessions, three months between two surgical acts, three months between two pathology samples). For instance if the patient has 3 months' chemotherapy following surgery, and then 1 year later another 3 months chemotherapy for a relapse, the algorithm will differentiate the two successive chemotherapy periods, and they will not be aggregated.

The rules for aggregating events into states are presented in Table 1. Because the main stages in the initial care provision for the tumour are defined from care provision guidelines, the relevance and completeness of the states illustrated in Table 1 are ensured by expert consensus.

It should be noted that a distinction is made for the concomitant administration of radiotherapy and chemotherapy in one and the same report. This situation can arise when HER2-targeted therapy is administered on the occasion of a visit by the patient for radiotherapy. The algorithm produces the specific state "chemotherapy-radiotherapy" (CT-RT).

Representation of sequences

The sequences of states produced by the algorithm are stored vertically in SPELL format [29], where a line represents a state and each state is characterised by a start and an end. The data for each patient is then transposed

Table 1 Main rules for aggregation of events into states

State code*	State	Type of event to aggregate (time-lapse between two events for aggregation)	Dating
A	AP	≥ 1 pathology investigations (less that 3 months apart)	Sampling Date If several: Start: date of first investigation End: date of last investigation
C	SURG	≥ 1 surgical acts (less than 3 months apart)	Date of the act If several: Start: date of first surgery End: date of last surgery
D	SURG_AP	≥ 1 surgical acts (less than 3 months) apart AND ≥ 1 pathology investigations	Date of the act, or else date of the sample. If several: Start: date of first surgery End: date of last surgery
N	CT_NEO	≥ 1 administrations of chemotherapy occurring before the first surgery	Start: date of the first administration End: date of the last administration
K	CT	≥ 1 administrations of chemotherapy (less than 6 months apart)	Start: date of the first administration End: date of the last administration
R	RT	≥ 1 radiotherapy sessions (less than 1 month apart)	Start: date of the first session End: date of the last session
O	CT_RT	Period of concomitant radiotherapy and chemotherapy	Start: date of the first session of intercurrent RT or CT End: date of the last session of intercurrent RT or CT

*Each state is recoded by a character so as to represent the overall sequence in the form of an ordered chain of characters.

so as to present the whole sequence in the form of an ordered chain of characters. An example of a representation of a care trajectory for a patient with infiltrating ductal adenocarcinoma of the breast is shown in Figure 2.

The sequences produced are available in two forms:

- A simple form in which the presence and order of the states are shown.
- An extended form which integrates the duration of states. This form includes periods without any event. The chains of characters are extended by repeating each character according to the duration of the state.

The choice to represent care trajectories as chains of characters where the length is directly proportional to the duration of care provision enables simple calculation of time-lapses via the use of PERL regular expressions [30,31].

Evaluation of the representation of care trajectories

The quality of the representations of care trajectories produced by the algorithm was assessed on a sample of patients for whom the care trajectories were constructed manually from data collected in the medical record (validation dataset). This dataset served as the reference standard for the evaluation of the representation of care trajectories. Each sequence was generated twice and independently from regional cancer registry data (automatically as "algorithm sequences") and from medical record (manually as "observed sequences"). The performance of the algorithm was assessed by confrontation of the sequences generated by the algorithm with the observed sequences for the patients (Figure 3).

Study population

The sample was formed by random selection stratified on TNM extension stage at diagnosis. A minimum of 50 patients per stratum (TNM stages I, II and III) was requested to conduct the evaluation of the algorithm in order to cover a variety of care trajectories of patients with breast cancer. The sample comprised 159 subjects with unilateral non-metastatic breast cancer (TNM stages I, II and III) diagnosed in 2008 among patients residing in the Poitou-Charentes region at the time of diagnosis, and who received all their care in three of the five health territories in the region covered by the

Figure 2 Algorithm representing care trajectories. Figure 2 presents an illustration of the representation of the care trajectory for a patient with infiltrating ductal adenocarcinoma of the breast. The diagnosis was established on examination of a biopsy by pathologists in a laboratory in La Rochelle, Charente-Maritime (17CBLAR). Treatment was a sequence of neo-adjuvant chemotherapy (4 sessions) in Nantes cancer center, Loire Atlantique (44CLRCC), followed by partial mastectomy and lymph node removal performed in Niort private hospital, Deux-Sèvres (79PCINK) and then 30 sessions of irradiation in Niort general hospital, Deux Sèvres (79CHNIO). * LNR: Lymph Node Removal, (I) The extended form integrates periods without any event noted I, where the start corresponds to the day following the previous state and the end to the day preceding the following state.

public radiotherapy unit (Vienne, Deux-Sèvres, Charente-Maritime Sud).

Reconstruction of observed sequences

Whereas the algorithm-based sequences were produced using solely the coded data from two electronic sources, the 159 observed sequences were manually reconstructed from the medical records, which collect all patient information produced and documented by staff involved in patient care. For this work the cancer registry staff systematically collected a copy of original pathology reports, surgery reports, hospitalisation reports or other relevant reports needed for the documentation of the main events in patient care trajectories. Each event occurring between diagnosis and the end of initial care provision for cancer was precisely documented (date and result of pathological investigation, date and nature of the surgical act, dates of start and end of chemotherapy, radiotherapy, HER2-targeted therapy, hormone therapy, number of sessions, alterations in treatment, occurrence of complications). This documentation may be derived from several different institutions depending on the number of establishments frequented by the patient.

The states making up the sequence were then manually captured in a single data table using the same rules implemented by the algorithm for the production of

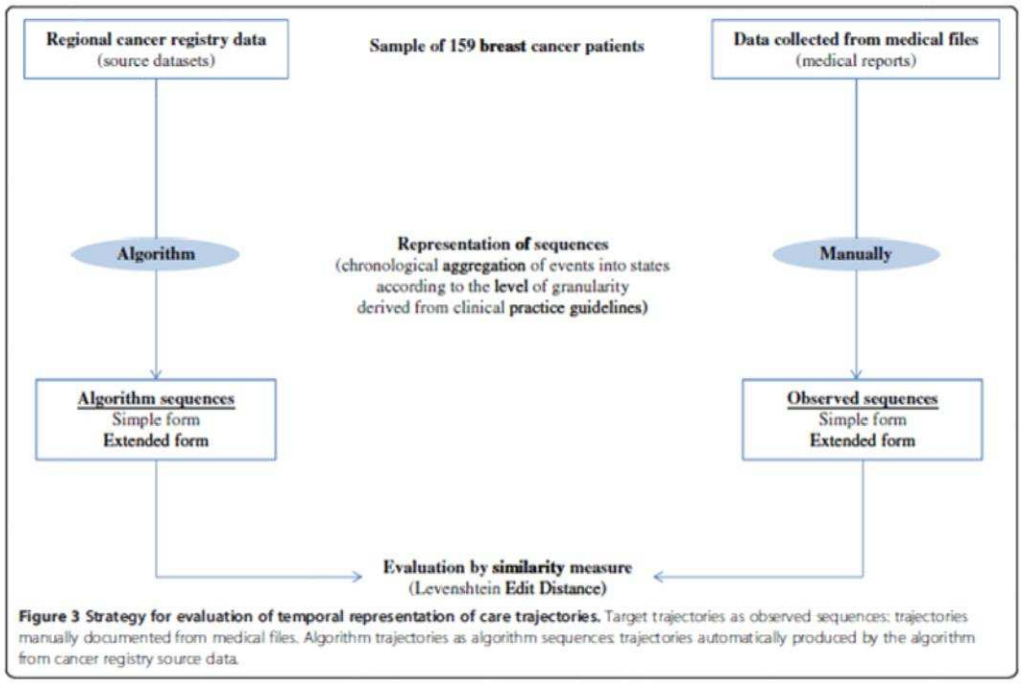


Figure 3 Strategy for evaluation of temporal representation of care trajectories. Target trajectories as observed sequences: trajectories manually documented from medical files. Algorithm trajectories as algorithm sequences: trajectories automatically produced by the algorithm from cancer registry source data.

states. The observed trajectories were structured according to the simple form and the extended form, to enable confrontation with the trajectories produced by the algorithm.

An endpoint for the end of treatment for each patient was determined manually from the last state in the initial care provision for the tumour. The endpoint of a trajectory can correspond to the death of the patient if it occurs before the end of the initial course of treatment of the cancer.

Similarity measures

A similarity measure was performed on all pairs of trajectories using the Levenshtein Edit Distance (LED) [32], which enables comparison of chains of characters of different lengths using identical cost operations. The LED is equal to the minimum number of characters that need to be added, removed or replaced to switch from one chain of characters to the other. Each of these elementary operations is associated with a cost equal to 1. The LED function is available on SAS software using the function COMPLEV.

Authorization and accreditation of the general cancer registry of Poitou-Charentes

In conformity with French law, the collection and analysis of medical data by the General Cancer Registry of Poitou-Charentes has received the approval of the French regulatory authorities: the *Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le Domaine de la Santé* and the *Commission Nationale Informatique et Libertés* (authorization number 907303).

In France, the *Comité National des Registres* (CNR) evaluation grids to be applied for the accreditation of registries include not only the methods used and the quality of the records, but also the use made of the data, and the interest and the originality of the research work conducted [33]. The General Cancer Registry of Poitou-Charentes has been approved by the French CNR since January 2013 based on the 2008 and 2009 registered data.

Results

Study population

The sample comprised 159 subjects with non-metastatic breast cancer. The mean age at diagnosis was 64.0 ± 15.3 (range 30–94). Fifty-two per cent were between the ages of 50 and 75, corresponding to the age group eligible for breast cancer screening from 2008. Fifty-five patients were TNM stage I, 52 were TNM stage II and 52 were TNM stage III.

Description of the sequences observed in the sample

The sequences observed in the sample are shown in Table 2. An individual care sequence belongs to one of the three categories illustrated in Figure 1 when the main tracer events are present in the same order as in the standard care sequence. Variants can be expected according to the presence or the absence of a pathology investigation, such as biopsy ("A") or a HER-targeted therapy ("O"). Otherwise, sequences are classified as non-standard sequences, and result in the particular cases being described as the fourth category "Other (non-standard sequences)".

Ninety per cent of these sequences were identifiable in the three standard sequences derived from recommendations (Figure 1). All the sequences involving a period of concomitant chemotherapy and radiotherapy (state "CT-RT" coded O) correspond to patients treated with HER2-targeted therapy.

More than half the observed (actual) sequences (51% or 81 patients) corresponded to the standard "ADR" sequence which combines pathology examination for diagnostic purposes, a surgical act for removal of tumour and/or lymph nodes, together with pathology examination of the surgical pieces, and radiotherapy treatment. Among these 81 sequences, 21 patients had total mastectomy, with biopsy ("AD") or without ("D"), and no relapse factor and hence absence of radiotherapy, while 13 patients had partial or total surgery on the basis of clinical parameters, and relapse risk factors providing indication for radiotherapy ("DR").

One third of the observed sequences (34% or 54 patients) corresponded to the standard "ADKR" sequence, combining diagnostic pathology investigation, a surgical act for removal of tumour and/or lymph nodes alongside pathology examination of the surgical pieces, chemotherapy treatment and radiotherapy treatment. Two patients underwent surgery at the outset ("DKR") and 10 were treated with HER2-targeted therapy, with biopsy ("ADKOK") or without ("DKOK").

Eight patients followed the "ANDR" standard sequence (5%) which associates diagnostic pathology examination, neo-adjuvant chemotherapy, a surgical act for removal of tumour and/or lymph nodes, with pathology investigation of the surgical pieces, chemotherapy and then

Table 2 Description of observed (actual) sequences in the sample

Care sequences	Numbers	Percentage
Care provision for non-metastatic breast cancer in case of good prognosis (Figure 1a)	*81*	*51%*
ADR	47	30%
AD	16	10%
DR	13	8%
D	5	3%
Care provision for non-metastatic breast cancer in case of presence of factors for poor prognosis (Figure 1b)	*54*	*34%*
ADKR	42	26%
ADKOK	8	6%
DKR	2	1%
DKOK	2	1%
Care provision for non-metastatic breast cancer in case of voluminous, infiltrating and/or inflammatory cancer (Figure 1c)	*8*	*5%*
ANDR	8	5.0
Other (non standard sequences)	*16*	*10%*
A	6	5%
ADK	2	0-1%
AK	2	0-1%
AR	2	0-1%
ADKDR	1	0-1%
ANDRK	1	0-1%
DK	1	0-1%
DKDKOK	1	0-1%

A Pathology investigation.
D Surgery and pathology examination of the surgical pieces.
N Neo-adjuvant chemotherapy.
K Chemotherapy.
R Radiotherapy.
O Concomitant chemotherapy and radiotherapy.

radiotherapy. Among these 8, 7 had a large and/or inflammatory carcinoma, and one patient expressed the wish to preserve the breast.

Ten per cent of the remaining sequences (16 patients) were particular cases:

- Ten patients who refused surgery, or whose advanced age constituted a counter-indication for surgery ("A", "AK", "AR").
- For three patients, with and without secondary total mastectomy (small breasts and areas surrounding the exeresis not clear) the decision, justified in MCM, was to administer chemotherapy but not radiotherapy ("ADK", "DK").
- Two patients having undergone secondary surgery at a late stage ("DKDKOK", "ADKDR").
- One patient who received HER2-targeted chemotherapy at a late stage "ANDRK").

Performance of the algorithm

Pair similarity in simple sequences (simple form)

Ninety-eight per cent of the sequences generated by the algorithm were similar to those observed when generated manually in terms of presence and ordering of states, i.e. there were only three dissimilar sequences among the 159 (LED = 1). These dissimilarities were related to absence of biopsy pathology report for one sequence (algorithm sequence "K" versus observed sequence "AK"), and absence of pathology report for the surgical pieces for two sequences (algorithm sequence "AC" versus observed sequence "AD").

Pair similarity in extended sequences (extended form)

Eighty-eight per cent of the identical sequences with respect to presence and ordering of states were also similar to one day for the duration of the states, giving 18 dissimilar sequences out of 159 (LED median = 7.5, range 1–51). The dissimilarities between sequences were in 89% (16 of 18 patients) date errors, or absence of coding for chemotherapy sessions (LED 21–51) or radiotherapy sessions (LED 1–3) at the start or the end of a state. One patient had no HD data on the lymph node removal she underwent secondarily (LED = 8). Finally, one patient presented a date error of one day in the date of surgery (LED = 1).

The percentage of match of pairs of extended sequences was 94% (10 dissimilar sequences out of 156) to the nearest 3 days (LED = 3).

Discussion

This study presents a method for the temporal representation of care trajectories for patients with non-metastatic breast cancer using data from a regional multi-source information system. The method proposed enables identification of the main tracer events in a care trajectory, and also integrates the duration of each event, and the time-lapse between events making up the trajectory. The results of the experiment show that the algorithm is able to reconstruct the care trajectories automatically from the registry data without implementing a collection of relevant information in medical records, which would be very resource-consuming in this setting.

Performance of the algorithm in the ordering of states

The ordering of treatment periods within the care sequences was correctly represented in 98% of cases. The three discrepancies observed were linked to absence of pathology evidence in the Registry data source, corresponding to non-coded sampling procedures in the pathology data (the practitioner did not code the tumour), or to coding errors or inconsistencies. Two of the dissimilar sequences nevertheless comprised all the treatment periods, because the absence of a pathology report on the surgical piece led to the creation of an intermediate surgery state in the sequence ("C" - surgery alone - rather than "D" - surgery and pathology evidence). An earlier study [34] implemented a text categorisation method using a machine-learning technique for the purpose of automatically categorising pathology reports solely on their content, which has demonstrated very good performances. It is therefore likely that the performance of the algorithm could be improved further by adding a supplementary check of the coding of pathology reports.

Performance of the algorithm on the duration of states

The performance of the algorithm relating to the duration of states was 88% to the nearest day, and 94% to three days, suggesting that the time-lapse indicators are excellent for the main tracer events in a care trajectory. The choice of representing care trajectories as chains of characters where the length is directly proportional to the duration of care provision makes it possible to extract and accurately calculate time-lapses in care trajectories as required, on the basis of regular expressions. Each regular expression needs to be drafted and adapted to the type of time-lapse or treatment duration to be calculated, thus enabling the main time-lapses to be produced for use in planning care in cancer units.

The dissimilarities between sequences were in 90% of cases related to dating errors or absence of coding for chemotherapy or radiotherapy sequences occurring at the start or at the end of a state. The variability of the dissimilarities observed was mechanically related to treatment patterns. The dissimilarities ranged from 21 to 51 days according to the sequence when the coding errors were linked to chemotherapy sessions, since the

reference administration pattern is 9 to 15 weeks. Dissimilarities were thus much smaller when the coding errors concerned radiotherapy sessions (1 to 3 days depending on sequences) since the reference pattern for the volumes treated is 50 Gy in 25 fractions over 33 days. Conversely, coding errors that persisted in the middle of a state before aggregation had no impact on its duration.

Limitations

First, one limitation of this study is its restriction to only three of the 5 health "territories" in the Poitou-Charentes region, because of absence of data available relating to radiology in private practice. Indeed, private radiology units bill their activities on the basis of CCAM-coded acts performed, and they are not covered by PMSI. The integration of this activity into a database is essential because the share of radiotherapy sessions performed on a profit-making basis by private facilities was estimated to be 55% over the territory as a whole in 2002 [5]. There are no consequences on the performances of the algorithm because the evaluation was conducted on a sample of patients who received all care in three of the five health territories in the region covered by the public radiotherapy unit. However, absence of radiotherapy data from private establishments in the cancer registry does not enable an evaluation of initial care trajectories extended to all breast cancer patients living in Poitou-Charentes area to be produced. We can however hope that this data will be integrated into the Registry database in the near future, since a reorganisation of the funding system for private-practice radiology is underway, and the units that at present code their activities according to CCAM will soon be required to produce standardised discharge reports as under PMSI [35].

Second, representing care trajectories by way of a linear approach generates difficulties when therapies are concomitant and when they are administered over long periods. These problems can be managed by creating composite states, with the risk of increasing the number of different sequences. In the ICD-10 there is no distinction between classic chemotherapy and HER2-targeted therapy, since both are coded Z511 for anti-tumour chemotherapy. A supplementary state "CT-RT" was therefore created to enable the representation, in this particular instance, of a treatment period relating the HER2-targeted therapy and concomitant radiology. This choice proved to be opportune, since all the sequences associating HER2-targeted therapy and radiotherapy were correctly identified and restored by the algorithm. Obviously, the creation of composite states will be discussed for each application of the algorithm to new cancer localization.

Third, not all the relevant sources of cancer data were included in modelling the trajectories. As can be seen from the 10% non-identifiable sequences among standard sequences from the different guidelines, the MCM are useful for the identification of specific instances relating to hospital care protocols, decisions linked to comorbidity, or patient choices. The integration of the MCM into the representation of sequences is also useful for the production of time-lapses to post-surgery therapeutic proposals. This source of data, although available at the time of study, was not assessed for the exhaustiveness of its integration into the Registry information system and therefore has not been included in the analysis.

Fourth, today the Registry does not routinely collect data relating to mammography, but it is working along these lines with screening centres in the Poitou-Charentes region. Bilateral mammography, the reference investigation in case of clinical warning signs or in screening campaigns, would be a very valuable element to include at the start of a trajectory, so as to analyse the time-lapse to diagnosis (time between mammography and biopsy) or the time-lapses more generally over the whole duration of care provision (for instance the time-lapse between mammography and the end of radiotherapy).

Comparison with the literature

Our modelling approach drew from work conducted in the socio-economic sphere using sequence analysis methods known as optimal matching, also known as sequence alignment, originally developed for rapid analysis of proteins and DNA sequences [36]. The first optimal matching algorithms appeared at the start of the 1970s, and their first application in the social sciences dates back to the article by Abbott and Forrest and their application to historical data [37,38]. Many techniques and tools, such as data mining, workflow mining or process mining [6-15] give a conceptual framework for clinical trajectory analysis. The aim of clinical trajectories is to offer a flowchart format for the decisions to be made and the care to be provided for a given patient or patient group. All published methods sought to enumerate regular medical behaviors that are expected to occur in patient-care trajectories from clinical workflow logs recorded by hospital information systems. But none to our knowledge has performed an application on patient care data in order to give an assessment of which regular medical behaviors occur in patient-care trajectories in real-life, rather than those expected. Moreover, they frequently implement models that provide too much detail to give a concise and comprehensive summary of the trajectory and are usually restricted to a single hospital information system. Our work focused on the representation of care trajectories through the main stages in the initial care provision for non-metastatic breast cancer identified in current updated guidelines, based on relevant and accurate information from medical source records from all hospitals and health facilities involved in

cancer care on a large geographical scale. To our knowledge this work is the first to assess the reliability and the validity of an algorithm for the representation care trajectories from cancer registry data. The method proposed in this work responds to needs expressed by institutions and health professionals of routine indicators aimed to improve the quality of patient care.

The structured organisation in the form of a relational time-stamped database in the Poitou-Charentes cancer Registry was well-suited to the application of these techniques. The value of our study is that it models the care trajectory from on-going routine collection of data for exhaustive recording of incident cases of cancer. This approach is useful for routine manual registration of incident tumours, because it makes relevant information available to registry staff, information that is dated and presented in chronological order, facilitating visual inspection of the case and the identification of structures possessing complementary information, thus enabling cases to be recorded according to international standards.

The routine production of care trajectories in cancer and waiting-time indicators for the purpose of evaluation, which are by-products of tumour notification to the Registry, opens up new perspectives in terms of cover, scope for comparison in time and space, and costs of producing information. A recent French study presented an overview on regional level of time-lapses between the different acts and key steps in the trajectories of patients with breast and lung cancer [39]. This study underlined the time required to collect data, imposed by the scatter of information for a given care trajectory, and the difficulties linked to the heterogeneity of data collection methods and practices, so that routine short-term follow-up of observed waiting times is not possible. Similar approaches have been implemented in other countries, for example in Ontario province, Canada, in the UK, or in New Zealand [40-43]. These countries have undertaken to study waiting times and estimate reductions in time-lapses to treatment, which has become one of the aims in their cancer Plans.

Certain specialised registries systematically record data concerning the initial treatment. This is the case, for instance, for breast cancer in the breast and gynaecological cancer Registry of the Côte d'Or area in France. This Registry used its database to supply information for a study conducted within the Francim network of cancer registries entitled "From diagnosis to first treatment: time-lapse to instatement of care for cancers recorded by the specialised registries in the Francim network 1999-2008", recently published by Inca [44]. It showed that there were variations in time-lapse to treatment over this period.

However, the majority of the general registries do not systematically collect information on care provision, and do not therefore have any scope for generalising these indicators unless a specific survey is conducted [45].

Perspectives

The routine production of these indicators will enable regular assessment of the match between care provided and official guidelines, the evaluation of time-lapse to treatment, and comparison of results with those reported in the international literature. This method can be applied by other cancer registries, conditionally on the availability of coded data sources using international classification of disease (ICD-O-3, ICD-10).

The results presented in this study are in favour of a continuation of this work on other cancer localizations. Before implementing the algorithm on the population-based cancer registry database for a new cancer localization, a further analysis of national guidelines and an evaluation on a new sample of patients are required. There are localisations where reducing time-lapse to treatment instatement is a major strategic orientation, and where providing care complying with guidelines requires the histological type to be identified, as in lung cancer. There are also rarer pathologies such as multiple myeloma, where the application of the algorithm would provide an overall picture of the organisation of the care trajectory, the diversity of trajectories according to patient characteristics and the different players in the care process. In our study the analysis was restricted to non-metastatic breast cancer and to the initial care provision for this cancer. Although rules were applied to avoid aggregating two events of the same type belonging to two different care episodes, the analysis needs to be extended to care episodes following the initial care trajectory, such as care relating to the occurrence of local or distant relapse, thus approaching the notions of cure, relapse, remission and consolidation.

Finally, one present limitation inherent in the representation and analysis of care trajectories in cancer is the absence of data on radiotherapy for private facilities, and the absence of data on the administration of treatment outside the hospital environment (oral chemotherapy and other). Subsequent work should therefore study the feasibility of obtaining and integrating SNIIR-AM data (French national insurance information system) to improve the representation of care trajectories among cancer patients.

Conclusions

This study presents a temporal representation of care trajectories in the form of a sequence of states ordered in time. The system enables accurate, routine identification of key dates and events in care trajectories, despite the fact that the data is initially fragmented across numerous sources in different territories. This information

is collated in a single base from which it can be readily extracted and exploited. The crossing of the trajectories produced with certain clinical data will enable routine evaluation of the compliance of observed care provision trajectories with those set out in guidelines. The production of indicators of this sort will contribute to significantly improving care planning at regional level, ultimately ensuring that all patients have access to the appropriate treatment within an appropriate time-lapse.

Consent

Patients were individually informed before the start of data collection of the nature of the information provided, the purpose of data processing, and their right of access, rectification or objection in conformity with the French law. The collection and analysis of medical data by the General Cancer Registry of Poitou-Charentes has received the approval of the French regulatory authorities: the Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le Domaine de la Santé and the Commission Nationale Informatique et Libertés (authorization number 907303).

Abbreviations

ADICAP: Association pour le Développement Informatique en Cytologie et Anatomie Pathologique; CCAM: Classification Commune des Actes Médicaux; CNR: Comité National des Registres; DRG: Diagnosis related groups; ICD-10: International classification of disease, 10th edition; ICD-O3: International classification of disease, oncology 3rd edition; MCM: Multidisciplinary consulting meeting; PMSI: Programme de Médicalisation des Systèmes d'information.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

GD and AR performed the data collection, data analysis and data interpretation. GD wrote the manuscript. AR, OD and PI contributed to data interpretation and drafting of the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors would like to thanks all the all the pathologists, departments of medical data processing in public and private hospitals, oncologists, and medical practitioners for their valuable contributions to the study. We also wish to thank Solzic Lellouch and Nicolas Mériau (data collection and monitoring) and A. Swaine Verdier (translation).

Author details

[1]Unité d'épidémiologie, biostatistique et registre général des cancers de Poitou-Charentes, Faculté de médecine, Centre Hospitalier Universitaire de Poitiers, Université de Poitiers, 6, rue de la milétrie, Poitiers, Cedex BP 199 86034, France. [2]Université de Rennes 1, IRISA UMR6074, Rennes, France. [3]INSERM, Poitiers CIC 802, France.

Received: 22 May 2013 Accepted: 27 March 2014
Published: 2 April 2014

References

1. Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PAC, Rubin HR: **Why don't physicians follow clinical practice guidelines? A Framework for improvement.** *JAMA* 1999, **282**(15):1458–1465.
2. Riou F, Jarno P: **Représentation et modélisation des trajectoires de soins.** *ITBM-RBM* 2000, **21**(5):313–317.
3. INCa: **Etude sur les délais de prise en charge des cancers du sein et du poumon.** In *Collection Etudes et Perspectives*. Boulogne-Billancourt: Institut National du Cancer; 2012.
4. Fetter R, Shin Y, Freeman J, Averill R, Thompson JD: **Case mix definition by diagnosis-related groups.** *Med Care* 1980, **18**(2):1–53.
5. Boinot L, Gautreau G, Defossez G, Daban A, Bourgeois H, Migeot V, Ingrand P: **Trajectoires hospitalières des patientes atteintes de cancer du sein en Poitou-Charentes.** *Rev Epidemiol Sante Publique* 2007, **55**(2):142–148.
6. Dart T, Cui Y, Chatellier G, Pattice D: **Analysis of hospitalised patient flows using data-mining.** *Stud Health Technol Inform* 2003, **95**:263–268.
7. Jay N, Napoli A, Kohler F: **Cancer patient flows discovery in DRG databases.** *Stud Health Technol Inform.* 2006, **124**:725–730.
8. Hu Z, Li JS, Zhou TS, Yu HY, Suzuki M, Araki K: **Ontology-based clinical pathways with semantic rules.** *J Med Syst* 2012, **36**(4):2203–2212.
9. Huang Z, Lu X, Duan H, Fan W: **Summarizing clinical pathways from event logs.** *J Biomed Inform* 2013, **46**(1):111–127.
10. Huang Z, Dong W, Ji L, Gan C, Lu X, Duan H: **Discovery of clinical pathway patterns from event logs using probabilistic topic models.** *J Biomed Inform* 2014, **47**:39–57.
11. Huang Z, Lu X, Duan H: **On mining clinical pathway patterns from medical behaviors.** *Artif Intell Med* 2012, **56**(1):35–50.
12. Rebuge A, Ferreira DR: **Business process analysis in healthcare environments: a methodology based on process mining.** *Inf Syst* 2012, **37**:99–116.
13. Bouarfa L, Dankelman J: **Workflow mining and outlier detection from clinical activity logs.** *J Biomed Inform* 2012, **45**(6):1185–1190.
14. Wang HQ, Li JS, Zhang YF, Suzuki M, Araki K: **Creating personalised clinical pathways by semantic interoperability with electronic health records.** *Artif Intell Med* 2013, **58**(2):81–89.
15. Combi C, Gozzi M, Oliboni B, Juarez JM, Marin R: **Temporal similarity measures for querying clinical workflows.** *Artif Intell Med* 2009, **46**(1):37–54.
16. Black RJ, Simonato L, Storm HH, Démaret E: *Automated Data Collection in Cancer Registration*. Lyon: IARC; 1998.
17. Jensen OM, Parkin DM, MacLennan R, Muir CS, Skeet RG: *Cancer Registration: Principles and Methods*. Lyon: IARC; 1991.
18. Belot A, Grosclaude P, Bossard N, Jougla E, Benhamou E, Delafosse P, Guizard AV, Molinié F, Danzon A, Bara S, Bouvier AM, Trétarre B, Binder-Foucard F, Colonna M, Daubisse L, Hédelin G, Launoy G, Le Stang N, Maynadié M, Monnereau A, Troussard X, Faivre J, Collignon A, Janoray I, Arveux P, Buemi A, Raverdy N, Schvartz C, Bovet M, Chérié-Challine L, *et al*: **Cancer incidence and mortality in France over the period 1980–2005.** *Rev Epidemiol Sante Publique* 2008, **56**(3):159–175.
19. Mauriac L: **Standards, options and recommendations for non metastatic breast cancer patients.** *Bull Cancer* 2002, **89**:207–224.
20. Haute Autorité de Santé (HAS) and Institut National du Cancer (INCA): *ALD 30 - Guide médecin sur le cancer du sein*. Saint-Denis: HAS-INCA; 2010.
21. Association pour le Développement de l'Informatique en Cytologie et Anatomo-Pathologie: *Thésaurus de la codification ADICAP*. 5th edition. Paris: ADICAP; 2009.
22. Agence Technique de l'Information sur l'Hospitalisation: *Guide méthodologique de production des résumés de séjour du PMSI en médecine, chirurgie et obstétrique*. Paris: Ministère du travail, de l'emploi et de la santé; 2011.
23. Ministère de la santé et des sports *Classification commune des actes médicaux - Guide de lecture et de codage*. Paris: Ministère de la santé, de la jeunesse et des sports; 2007.
24. Jouhet V, Defossez G, CRISAP CRM, Ingrand P: **Automated selection of relevant information for notification of incident cancer cases within a multisource cancer registry.** *Methods Inf Med* 2013, **52**(5):411–421.
25. Working Group Resort: **International rules for multiple primary cancers (ICD-O Third Edition).** *Eur J Cancer Prev* 2005, **14**(4):307–308.
26. Burgun A, Le Beux P: **Aspects sémantiques de la description des trajectoires de patients.** *ITBM-RBM* 2000, **21**(5):318–322.
27. Combi C, Shahar Y: **Temporal reasoning and temporal data maintenance in medicine: issues and challenges.** *Comput Biol Med* 1997, **27**(5):353–368.
28. Allen JF: **Towards a general theory of action and time.** *Artif Intell* 1984, **23**(2):123–154.
29. Gabadinho A, Ritschard G, Müller NS, Studer M: **Analyzing and visualizing state sequences in R with TraMineR.** *J Stat Softw* 2011, **40**(4):1–37.
30. Stephen GA: *String Searching Algorithms*. Singapour: World Scientific Press; 1994.
31. Cody R: **An introduction to Perl regular expressions in SAS 9.** Available at: http://www2.sas.com/proceedings/sugi29/265-29.pdf.

32. Levenshtein VI: **Binary codes capable of correcting deletions, insertions, and reversals.** *Soviet Physics Doklady* 1966, 10(8):707–710.
33. Institut de veille sanitaire: **Comité national des registres, Procédure de qualification par le CNR.** Available at: http://www.invs.sante.fr/Espace-professionnels/Comite-national-des-registres.
34. Jouhet V, Defossez G, Burgun A, Le Beux P, Levillain P, Ingrand P, Claveau V: **Automated classification of free-text pathology reports for registration of incident cases of cancer.** *Methods Inf Med* 2012, 51(3):242–251.
35. Agence Technique de l'Information sur l'Hospitalisation: *Notice technique n° CIM-MR/ME - 1116-3-2010.* Lyon: ATIH; 2010.
36. Needlman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *J Mol Biol* 1970, 48(443–453.
37. Abbott A, Forrest J: **Optimal matching methods for historical sequences.** *J Interdiscip Hist* 1986, 26:471–494.
38. Abbott A, Tsay A: **Sequence analysis and optimal matching methods in socio- logy.** *Sociol Methods Res* 2000, 29(1):3–33.
39. Institut National du Cancer: *Etude sur les délais de prise en charge des cancers du sein et du poumon.* Boulogne-Billancourt: INCa; 2012.
40. **Ontario Cancer Plan 2011–2015, Timely Access.** Available at: http://ocp.cancercare.on.ca/strategic_priorities/timely_access/.
41. Ontario Cancer Registry: **Wait Times.** Available at: https://www.cancercare.on.ca/ocs/wait-times/.
42. Department of Health: **Cancer reform strategy.** Available at: www.dh.gov.uk/.
43. Ministry of Health: *National Cancer Programme, Work Plan 2012/13.* Wellington; 2013. Available at www.health.govt.nz/.
44. Institut National du Cancer: **Du diagnostic au premier traitement : délais de prise en charge des cancers enregistrés par les registres spécialisés du réseau Francim 1999–2008.** In *Collection État des lieux & des connaissances*. Boulogne-Billancourt: Institut National du Cancer; 2012.
45. Molinié F, Leux C, Delafosse P, Ayrault-Piault S, Arveux P, Woronoff AS, Guizard AV, Velten M, Ganry O, Bara S, Daubisse-Marliac L, Tretarre B: **Waiting time disparities in breast cancer diagnosis and treatment: a population-based study in France.** *Breast* 2013, 22(5):810–816.

doi:10.1186/1472-6947-14-24
Cite this article as: Defossez et al.: Temporal representation of care trajectories of cancer patients using data from a regional information system: an application in breast cancer. *BMC Medical Informatics and Decision Making* 2014 14:24.

3. Vers une démarche intégrative de données à des fins d'évaluation et de recherche

Les efforts prodigués pour parvenir à rassembler un niveau élevé d'informations en réponse à l'objectif initial de surveillance épidémiologique nous ont incité à prolonger la réflexion sur la modélisation des trajectoires de soins dans une perspective d'évaluation des parcours en « vie réelle » et de recherche sur les services de santé (« Health services research »). De l'accès aux approches thérapeutiques les plus innovantes découlent pour le patient la qualité de sa prise en charge, les impacts sur sa qualité de vie et sur le pronostic de sa maladie, et se pose la question de la globalité du parcours de soins et des inégalités d'accès aux soins en cancérologie. Alors que la quantité d'informations disponibles a explosé, en santé comme dans bien d'autres domaines, l'enjeu était de pouvoir mobiliser et intégrer ces données, dans le cadre d'autorisations réglementaires, pour parvenir à enrichir la connaissance et la contextualisation des parcours de soins à des fins d'évaluation et de recherche.

Cette section décrit une série de travaux réalisés par le RGCPC, dont les résultats et indicateurs d'évaluation proviennent de la mise en œuvre d'un appariement à une base de données spécifique (data linkage) en réponse à l'objectif d'évaluation ou de recherche, ou sont directement issus de l'exploitation des données recueillies en routine par le registre.

Les domaines d'applications couvrent aussi bien le champ de la prévention, que celui du diagnostic, de la prise en charge et du suivi des patients atteints de cancer. Chaque domaine d'application est abordé de façon synthétique dans la suite du document et fait référence à un article original publié, qui est soit intégré dans le corps de texte soit référencé selon les choix motivés dans la section 1.6. Articles scientifiques.

3.1. Réutilisation des données de dépistage organisé des cancers [36]

Un premier domaine d'application a concerné l'évaluation des programmes de dépistage du cancer. L'intérêt était de privilégier la mise en relation des données du RGCPC à celles des structures de gestion du dépistage organisé (DO) du cancer pour répondre de façon efficiente à l'objectif d'évaluation. Le but était d'évaluer l'impact de la participation des femmes au DO du cancer du sein sur la précocité du diagnostic, le traitement et la survie à 5 ans, dans un contexte où le bénéfice du dépistage était régulièrement remis en cause sur les arguments de l'avance au diagnostic, du surdiagnostic et du surtraitement, engendrant des controverses et un feed-back négatif [37].

En pratique, la population des femmes âgées de 50 à 74 ans qui ont présenté un cancer du sein incident en 2008 et 2009 a été identifiée dans la base du RGCPC (1 613 patientes). Le rapprochement des données des structures de gestion (mammographies de dépistage) pour les 4 départements du Poitou-Charentes aux parcours de ces femmes a permis de les catégoriser en 3 groupes, selon qu'elles avaient bénéficié ou non d'une mammographie dans les 24 mois qui précédaient le diagnostic histologique du cas :

1. Les cancers détectés dans le cadre du dépistage, en présence d'anomalies révélées sur la mammographie (n=878),
2. Les cancers hors dépistage, en l'absence de mammographie (n=554),
3. Les cancers de l'intervalle, en cas de diagnostic histologique de cancer survenu dans les 24 mois au décours d'une mammographie normale (n=181).

L'analyse a permis de montrer la prédominance des formes localisées chez les participantes au programme (64% de stades I) et le caractère conservateur des traitements reçus (traitements moins mutilants avec près de 80% de chirurgie conservatrice et une survie à 5 ans supérieur à 96%). Le résultat majeur était le mode de présentation des cancers de l'intervalle, qui apparaissaient comme des tumeurs beaucoup plus agressives au diagnostic (48% de stades II / 12% de stades III, moins hormonosensibles) et dont les 2/3 étaient traitées par chimiothérapie. Leur survie spécifique à 5 ans était de 92%, mais était meilleure que les femmes hors DO (85%).

Ces résultats soulevaient l'intérêt de maintenir les efforts prodigués pour augmenter le taux de participation des femmes au programme de dépistage, compte tenu du bénéfice de survie attendu lié à la précocité du diagnostic. Le stade d'extension TNM était le déterminant majeur qui expliquait un bénéfice sur la survie en analyse multivariée.

Lire Article 4 : *Defossez G, Quillet A, Ingrand P. Aggressive primary treatments with favourable 5-year survival for screen-interval breast cancers.* ***BMC Cancer****. 2018 Apr 6;18(1):393. doi: 10.1186/s12885-018-4319-4. PMID: 29625602; PMCID: PMC5889614.*

Accéder à la suite « Réutilisation des données de génétique moléculaire des cancers »

Defossez et al. BMC Cancer (2018) 18:393
https://doi.org/10.1186/s12885-018-4319-4

BMC Cancer

RESEARCH ARTICLE Open Access

Aggressive primary treatments with favourable 5-year survival for screen-interval breast cancers

Gautier Defossez[1,2*], Alexandre Quillet[1] and Pierre Ingrand[1,2]

Abstract

Background: To assess the impact of the participation in screening programme according to the mode of detection on the early diagnosis, treatment, and specific survival outcomes in women with breast cancer.

Methods: Women diagnosed with invasive breast cancer in Poitou-Charentes region (France) between 2008 and 2009 were classified into three groups, using data linkage of cancer registry, vital statistics and French organized screening programme: the screening programme (SP), interval cancer (IC), and non-screening programme detected cancer (NSP) groups. Specific survival rates were analysed using the Kaplan–Meier method and Cox proportional hazard models.

Results: Among 1613 patients, 65.7% ($n = 1059$) participated in a screening programme. The interval cancer rate was 17.1% ($n = 181$). Tumours in the IC group were diagnosed at a more advanced stage, i.e. with further regional lymph node metastasis or local spread, than those in the SP group ($p < 0.001$), but with significantly fewer metastases at diagnosis than in the NSP group ($p < 0.001$). ICs underwent more aggressive primary treatments than the two other groups, with 28% of radical mastectomy and 67% undergoing chemotherapy. The five-year survival rate for IC group were 92.0% (95% CI, 89.9–94.0%).

Conclusions: Interval cancers had more aggressive features than screen-detected cancers but were diagnosed at a less advanced stage compared to non-screen detected cancers. Despite having cancers missed by the screening programme, women who participate in the screening process seem to benefit from early treatment. These results must be confirmed with long-term follow-up.

Keywords: Breast neoplasms, Mass screening, Treatment, Survival, Cancer registry, Interval Cancer, Data linkage

Background

Breast cancer is the leading cause of cancer death in women worldwide [1]. Prognosis is mainly determined by the tumour stage at diagnosis [2]. To facilitate early detection and access to effective treatment of breast cancers, screening programmes have been implemented gradually in European countries [3] and have mostly demonstrated the effectiveness of screening in reducing breast cancer mortality [4–9].

However, the benefit of mammogram screening is still regularly challenged by controversies, regarding overdiagnosis, false-positive results, possibly radiation-induced cancer or interval cancers [10–12]. Moreover, no reduction in breast cancer mortality was observed in a Canadian randomized mammography screening trial [13]. These controversies add complexity to informed decision making for clinicians and patients, and create negative feedback for screening programmes.

In France, a screening programme was implemented nationwide since 2004, and offers a physical examination and a bilateral mammogram biennially to women aged 50–74 years. Ten years after its establishment, only slightly more than half of women (52.1% in 2014) participate, and 10% of women choose an individual (opportunistic) screening with a mammography performed under medical prescription, outside the official programme [14, 15].

* Correspondence: gautier.defossez@univ-poitiers.fr
[1]Poitou-Charentes General Cancer Registry, Poitiers University Hospital, University of Poitiers, Poitiers, France
[2]INSERM, CIC 1402, Poitiers, France

Further investigations are needed to deliver an objective and comprehensible message to women and policy makers. The special case of interval cancers that have a potential influence on the effectiveness of screening should be considered. Interval cancers, although they are considered as false-negatives of the screening programme, usually become clinically evident shortly after the last normal screening result. Interval cancers often reflect aggressive tumours encountered in women already involved in the screening programme [16–18]. This fast-growing lump in the breast may be readily detected by self-examination and would also give rise to more anxiety than a slow-growing one, leading to clinical examination [19]. Whereas their clinical and biological characteristics are now better documented, inconsistent findings exist on their prognosis in the literature and no study to our knowledge has taken into account treatments as an indicator of screening programme effectiveness [20–25].

This study aimed to assess the impact of the participation in screening programme according to the mode of detection on the early diagnosis, treatment, and specific survival outcomes in women with breast cancer, using data linkage of cancer registry, vital statistics and French screening programme.

Methods

Patients

Women aged 50–74 years, residing in the Poitou-Charentes region (1.8 million inhabitants, South-West France) with the first diagnosis of invasive breast carcinoma between 1 January 2008, and 31 December 2009, were included in this study.

The French screening programme

In France, the breast cancer screening programme is offered biennially to women aged 50–74 years. It includes a physical examination and a bilateral mammogram, the results of which are based on the Breast Imaging-Reporting And Data System (BI-RADS) classification of the American College of Radiology. The BI-RADS 1 (negative) and 2 (benign finding) mammograms are systematically subjected to a second reading aimed to reduce the false-negative rate. For other patients as BI-RADS 0, 3, 4 or 5, follow-up or complementary diagnostic procedure (biopsy, ultrasonography, magnetic resonance imaging) with or without a specific mammographic follow-up are provided. The screening programme and data registration are conducted by the screening facilities located in each of the four French administrative departments (counties) of the Poitou-Charentes region. Screening mammograms performed outside the invitation of the screening programme (individual screening) are not included.

Data

Primary invasive breast carcinomas were identified from the Poitou-Charentes cancer registry between 1 January 2008, and 31 December 2009. For each case, patient, tumour, and healthcare data were routinely reported according to international rules [26]. In the present study, the prognostic variables included age, tumour stage classified according to the TNM classification of malignant tumours, histological Scarff-Bloom-Richardson grade, oestrogen and progesterone receptor status, and human epidermal growth factor receptor 2 (HER-2) expression. Cancer treatments (neoadjuvant treatment, surgery, adjuvant chemotherapy, and radiotherapy) were also recorded [27, 28]. Hormone therapy was not reported outside neoadjuvant hormone therapy.

The dates and results of screening programme mammograms were obtained from the four screening facility databases of Poitou-Charentes. Patients were classified according to the mode of detection as: Screen-detected cancers from the Screening Programme (SP group) were defined as women having a positive mammography (BI-RADS 3, 4, or 5) followed by complementary diagnostic procedures including histological confirmation of cancer. The interval cancer group (IC group) were defined as women having a negative mammography (BI-RADS 1 or 2) followed by a histological diagnosis of cancer occurring within the 24 months of the prior mammogram. The cut-off of 24 months corresponded to the waiting time, recommended in the French screening programme, between two screening mammograms. The non screen-detected cancer group (NSP group) were defined as women having a histological diagnosis of cancer without having participated in screening programme and could include opportunistic screening or breast cancers detected based on clinical signs or symptoms. A BI-RADS 0 (incomplete assessment) mammogram was a temporary classification that required complementary diagnostic action (extension, ultrasonography, biopsy). In the absence of reclassification for BI-RADS 1 or 2, these mammograms were considered positive and included in the SP group.

Patients' vital information until 31 December 2014 was obtained from data of the French national civil registration file RNIPP, maintained by the National Institute of Statistics and Economics Studies. French native patients who had not been reported dead were censored at December 31, 2014. Foreign patients were censored at the date of the last follow-up. A systematic review of the medical records was performed to identify the cause of death. Death was related to breast cancer in the presence of disease progression.

This study was approved by the French regulatory authorities (the "Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le Domaine

de la Santé" and the "Commission Nationale Informatique et Libertés", authorisation number 907303). According to French law, patients were informed of their data registration and given the right to deny access or to rectify their personal data.

Statistical analyses

Patient, tumour, and healthcare characteristics of IC were compared to the two other modes of detection using the Chi-squared or Fisher's exact test. Survival rates were estimated using the Kaplan–Meier method. Hazard ratios and their 95% confidence intervals were estimated from univariate and multivariate Cox proportional hazard models. For specific survival analysis, women who died of non-breast cancer-related causes were censored at the date of death. Searching for interactions and collinearity between included variables was performed in the multivariate analysis. Two multivariate analysis were performed, with and without adjustment on main prognostic factors, in order to highlight the important role of TNM stage at diagnosis in the interpretation of the results. The threshold p-value for including variables in the final multivariate Cox model was set at 5% except for the detection mode variable, which was forced into this model. Data management and statistical analyses were performed using SAS software version 9.4 (SAS Institute Inc., Cary, NC, USA).

Results

This registry-based study included 1613 patients aged 50–74 years at breast cancer diagnosis. Among these patients, 1059 (65.7%) underwent a screening programme mammogram within 24 months before diagnosis, which revealed tumours in 878 (82.9%) patients (Fig. 1). Therefore, the interval cancer rate was 17.1% ($n = 181$).

Comparison of prognostic and treatment characteristics

The distribution of characteristic and prognostic factors according to the mode of detection is shown in Table 1. Tumours in the IC group were diagnosed at a more advanced stage ($p < 0.001$), with higher-grade tumours ($p < 0.001$), and with a greater proportion of hormone receptor negative tumours ($p < 0.001$) compared with those in the SP group. Conversely, tumours in the IC group were diagnosed at a significantly less advanced stage ($p < 0.001$) compared with those in the NSP group. Distant metastasis were reported in 11.7% of the NSP group vs 1.0% in the SP group and 3.9% in the IC group.

The proportion of regional lymph node metastasis were significantly higher for non-metastatic cancers in the IC group than in the SP group (38.1% vs. 24.5%, $p < 0.001$). For tumours without lymph node involvement, 36.3% of tumours in the IC group were larger than 2 cm (T2-T3-T4) vs. 13.4% in the SP group.

Concerning treatment, breast conserving surgery was less frequently performed in the IC group compared with the SP group, with a higher proportion of mastectomy (28% of mastectomy in the IC group vs. 20% in the SP group, $p = 0.045$). Chemotherapy (adjuvant or neoadjuvant) was more frequently performed in the IC group compared with the two other groups (38% in SP group, 64% in IC group and 46% in NSP group, $p < 0.001$). Patients in the IC group underwent significantly less palliative care ($p < 0.001$) compared with the NSP group.

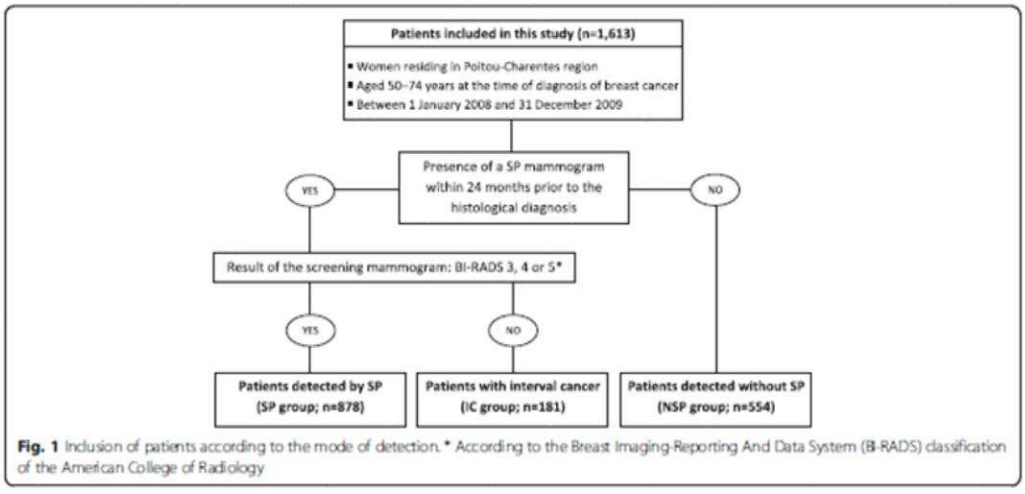


Fig. 1 Inclusion of patients according to the mode of detection. * According to the Breast Imaging-Reporting And Data System (BI-RADS) classification of the American College of Radiology

Table 1 Distribution of patient, tumour, and treatment characteristics of invasive breast cancer according to the mode of detection

	SP group (*n* = 878)		IC group (*n* = 181)		NSP group (*n* = 554)		SP vs. IC group	SP vs. NSP group	IC vs. NSP group
	n	(%)	n	(%)	n	(%)	p	p	p
Age							0.060	0.307	0.236
> 65 years	319	(36.3)	52	(28.7)	186	(33.6)			
≤ 65 years	559	(63.7)	129	(71.3)	368	(66.4)			
TNM stage							< 0.001	< 0.001	< 0.001
I	560	(63.8)	67	(37.0)	227	(41.0)			
II	255	(29.0)	86	(47.5)	174	(31.4)			
III	47	(5.4)	21	(11.6)	77	(13.9)			
IV	9	(1.0)	7	(3.9)	65	(11.7)			
Unknown	7	(0.8)	0	(0.0)	11	(2.0)			
Extent of disease							< 0.001	< 0.001	0.0014
Tumor with local spread (any T/N0/M0)	647	(73.7)	105	(58.0)	315	(56.9)			
T1	560	(86.6)	67	(63.8)	227	(72.1)			
T2	78	(12.1)	31	(29.5)	71	(22.5)			
T3	8	(1.2)	5	(4.8)	5	(1.6)			
T4	1	(0.1)	2	(1.9)	12	(3.8)			
Tumor with regional spread (any T/N+/M0)	215	(24.5)	69	(38.1)	163	(29.4)			
Advanced cancer (any T/any N/M+)	9	(1.0)	7	(3.9)	65	(11.7)			
Unknown	7	(0.8)	0	(0.0)	11	(2.0)			
SBR grade							0.001	0.022	0.147
1	228	(26.0)	27	(14.9)	115	(20.8)			
2	486	(55.4)	109	(60.2)	299	(54.0)			
3	136	(15.5)	42	(23.2)	110	(19.9)			
Unknown or not assessed	28	(3.2)	3	(1.7)	30	(5.4)			
Hormonal receptor status							< 0.001	< 0.001	0.171
OR+/PR+	633	(72.1)	105	(58.0)	331	(59.8)			
OR+/PR- or OR-/PR+	137	(15.6)	32	(17.7)	109	(19.7)			
OR-/PR-	88	(10.0)	41	(22.7)	89	(16.1)			
Unknown or not assessed	20	(2.3)	3	(1.7)	25	(4.5)			
Her-2 receptor status							0.350	0.068	0.899
Positive	86	(9.8)	23	(12.7)	72	(13.0)			
Negative	699	(79.6)	146	(80.7)	427	(77.1)			
Unknown or not assessed	93	(10.6)	12	(6.6)	55	(9.9)			
Type of treatment							< 0.001	< 0.001	< 0.001
Surgery ± RT	535	(60.9)	60	(33.2)	233	(42.1)			
Surgery + CT ± RT	293	(33.4)	91	(50.3)	199	(35.9)			
Neoadjuvant treatment *+ Surgery ± CT/RT	43	(4.9)	27	(14.9)	68	(12.3)			
No surgery (refusal, palliative treatment)	3	(0.3)	3	(1.7)	45	(8.1)			
Unknown	4	(0.5)	–	–	9	(1.6)			
Type of surgery							0.045	< 0.001	0.113
Breast-conserving surgery	692	(78.8)	129	(71.3)	327	(59.0)			
Mastectomy	178	(20.3)	49	(27.1)	171	(30.9)			
Unknown or no surgery	8	(0.9)	3	(1.7)	56	(10.1)			

Chi-square and Fisher's exact test did not including any missing values
SP group, patients detected by the screening programme; IC group, patients with interval cancer; NSP group, patients detected without participating in the screening programme; SBR, Scarff-Bloom-Richardson grade; OR, oestrogen receptor; PR, progesterone receptor; RT, adjuvant radiotherapy; CT, adjuvant chemotherapy
*Neoadjuvant treatment include neoadjuvant chemotherapy (38/43 in SP group, 25/27 in IC group and 57/68 in NSP group) or neoadjuvant hormonotherapy (5/43 in SP group, 2/27 in IC group and 11/68 in NSP group). One patient in NSP group receive both neoadjuvant hormonotherapy and adjuvant chemotherapy

Comparison of survival

The median follow-up for the study group was equal to 5.8 years (interquartile range, 5.3–6.4 years). One hundred and eighty-eight (11.7%) women died during the follow-up period, including 136 (72.3%) breast-cancer-related deaths. Kaplan-Meier survival curves are shown in Fig. 2. The 5-year specific survival rate was 92.0% in the IC group (95% CI, 89.9–94.0%), 96.4% in the SP group (95% CI, 95.8–97.1%) and 85.3% in the NSP (95% CI, 83.8–86.9%). Superior survival was observed in the IC group compared with the NSP group ($p = 0.015$). After 3 years, a significant survival difference emerged between the SP and IC groups ($p = 0.021$).

Univariate and multivariate Cox regression analyses are shown in Table 2. Due to collinearity between the treatment type and prognostic factors, treatment type was not included in the multivariate analysis. The final model retained three independent prognostic factors including an early stage at diagnosis ($p < 0.001$), hormone receptor-positive tumours ($p < 0.001$), and age ≤ 65 years ($p = 0.003$). Detection mode was not significant in survival analysis when taking into account TNM stage at diagnosis. A significant survival was observed in the IC group compared to the NSP group in the unadjusted model ($p < 0.001$). Analyses of overall survival confirmed these findings, except for the difference between the SP and IC groups, which was no longer significant ($p = 0.40$), with the 5-year overall survival rates being 93.7% in the SP group, 90.9% in the IC group, and 82.4% in the NSP group.

Discussion

Interval cancers are, as suggested by our study, diagnosed at a significantly less advanced stage compared with those in the NSP group. They show more aggressive features than screen-detected cancers, while undergoing more aggressive primary treatments with a higher rate of mastectomy and two-thirds undergoing chemotherapy. The individual data available for each patient confirm the linear gradient of TNM stage according to the mode of detection and provides interesting additional information on the initiated treatment. Most tumours in the SP group are localized cancers classically treated by breast-conserving surgery and radiotherapy. Tumours in the IC group are characterized by more local and regional spread justifying a more aggressive treatment with neoadjuvant therapy or adjuvant chemotherapy. Tumours in the NSP group show significantly more advanced cancers with metastastic and non-resectable cancers characterized by a greater proportion of palliative care.

As a majority of studies, we have found no significant differences in prognosis between women with interval cancer compared with an unscreened population, when taking into account the main prognostic factors [20–25]. The IC group had superior 5-year survival rates compared with the NSP group in univariate model, but this difference became reasonably non-significant after adjustment because of the strong survival advantage attributed to differences in the initial distribution of TNM stage at diagnosis [29–31]. The multivariate Cox model without adjustment on TNM stage confirmed the better and significant survival in the IC group than in the NSP group. Supplementary individual information on diagnosis and treatment provides essential results to properly understand the benefit on breast cancer survival.

We assume that the appropriate method for determining whether a cancer screening strategy works is the randomized controlled trial, with mortality as the endpoint. But the study of interval cancer on case diagnosis, treatment and survival is interesting, while the emphasis is now on evaluation of routine screening services for which

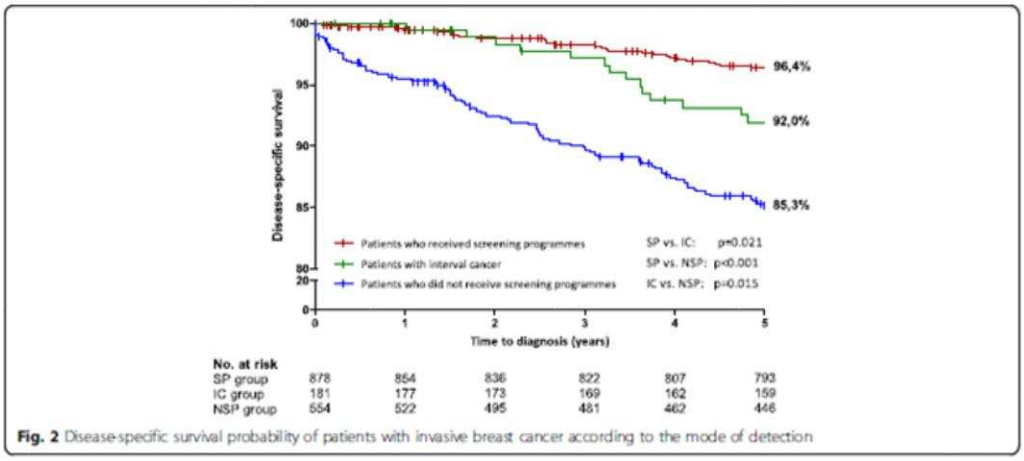


Fig. 2 Disease-specific survival probability of patients with invasive breast cancer according to the mode of detection

Table 2 Hazard ratios (HR) and 95% confidence intervals (CI) of prognostic factors in patients with invasive breast cancer by univariate and multivariate Cox regression analysis

	Univariate analysis			Multivariate analysis					
				With adjustment on TNM stage			Without adjustment on TNM stage		
	HR	95% CI	p	HR	95% CI	p	HR	95% CI	p
Mode of detection			< 0.001			0.325			< 0.001
NSP group	1			1			1		
SP group	0.25	0.17–0.37		0.75	0.48–1.15		0.28	0.19–0.42	
IC group	0.51	0.29–0.87		0.74	0.42–1.30		0.48	0.28–0.83	
Age			0.128			0.003			0.05
> 65 years	1			1			1		
≤ 65 years	0.77	0.54–1.08		0.58	0.41–0.84		0.70	0.49–0.99	
TNM stage			< 0.001			< 0.001			–
I	1			1			–		
II	2.62	1.40–4.90		2.59	1.33–5.02		–	–	
III	14.48	8.03–26.09		13.25	7.02–25.01		–	–	
IV	76.92	44.12–134.08		63.69	33.69–120.39		–	–	
SBR grade			< 0.001			–			–
1	1			–			–		
2	3.02	1.56–5.84		–	–		–	–	
3	5.82	2.93–11.59		–	–		–	–	
Hormonal receptor status			< 0.001			< 0.001			< 0.001
OR+/PR+	1			1			1		
OR+/PR- or OR-/PR+	2.78	1.82–4.25		1.97	1.28–3.02		2.49	1.63–3.81	
OR-/PR-	4.30	2.87–6.45		2.88	1.89–4.38		3.89	2.58–5.87	
Her-2 receptor status			0.050			–			–
Positive	1			–			–		
Negative	0.63	0.40–1.00		–	–		–	–	
Type of treatment			< 0.001			–			–
Surgery ± RT	1			–			–		
Surgery + CT ± RT	3.75	2.18–6.45		–	–		–	–	
Neoadjuvant treatment + Surgery ± RT/CT	11.36	6.33–20.39		–	–		–	–	
Absence of surgery	84.72	48.26–148.73		–	–		–	–	

SP group, patients detected by the screening programme; IC group, patients with interval cancer; NSP group, patients detected without participating in the screening programme; SBR, Scarff-Bloom-Richardson grade; OR, oestrogen receptor; PR, progesterone receptor; RT, adjuvant radiotherapy; CT, adjuvant chemotherapy

randomized trials may not be suitable [31]. The cancer registry provides robust data, ensuring the completeness of incident cancer cases and thereby avoiding selection bias. The use of individual data allowed the monitoring of a screening programme in a real setting, and with controlling for individual differences that might affect the primary outcome, particularly TNM stage at diagnosis [2]. Studies usually covered a period where the national screening programme was implemented gradually or referred to old data. Our study was initiated 4 years after the generalisation of the screening programme, which placed the analysis in a stable situation regarding diagnostic procedures, participation and care in breast cancer.

This study has some limitations. The present study did not allow to assess the occurrence of mammograms performed outside the screening programme (opportunistic screening) [32]. In a report issued by the French National Authority for Health [14], the participation for this mode of detection was estimated at 10% of the target population. The authors highlighted the difficulties in identifying these patients and the heterogeneity of practices regarding this type of screening. If individually screened patients could be distinguished from non-screen detected patients, the 5-year survival rate might have been even worse, and survival difference between non-screen detected patients and the others might have

been increased. Second, misclassification of the mode of detection cannot be excluded. Some interval cancers could be classified as screen-detected cancers if symptomatic women waited for the screening mammography instead of a consultation with the physician. Third, we did not interpret the results concerning survival differences between the SP group and the two others because of known biases such as lead-time and length-time which invariably provide a higher survival in screen-detected cancers.

Conclusions

In conclusion, more aggressive treatments were found in patients with interval cancers. Despite the aggressiveness of these cancers, women who participate in the screening process seem to benefit from early treatment. These results must be confirmed with long-term follow-up. Such a result could not be explained by overdiagnosis, but instead appeared to be the consequence of a reduction in late diagnosis due to participation in screening programme and an access to suitable and curative treatments. These findings reinforce the need to promote organised screening. Participation in a screening programme was important in facilitating early detection in these women. A survival benefit might be expected by increasing the participation rate in screening programmes if they are accessible to everyone at risk.

Abbreviations

BI-RADS: Breast Imaging-Reporting and Data System; CT: Adjuvant chemotherapy; HER-2: Human Epidermal Growth Factor Receptor 2; IC: Interval Cancer; NSP: Non-Screening Programme; OR: Oestrogen Receptor; PR: Progesterone Receptor; RT: Adjuvant Radiotherapy; SBR: Scarff-Bloom-Richardson grade; SP: Screening Programme

Acknowledgments

We thank the coordinating physicians of the screening facilities of the Poitou-Charentes region for their valuable contributions to this study: Dr. Françoise BOLVIN (ORCHIDEE 16), Dr. Anne FEYLER (LUCIDE 17), Dr. Sandrine ARLOT-LAMADE (ARCANDE 79), and Dr. Caroline TOURNOUX-FACON (DOCVIE 86).

Funding

The present study was funded by The French National Cancer Institute (grant number INCA_6955), which participated in data collection and analysis. The Poitou-Charentes cancer registry is sponsored by the Agence Régionale de Santé Nouvelle-Aquitaine.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Authors' contributions

GD and AQ performed the data collection, data analysis and data interpretation. GD and AQ wrote the manuscript. PI contributed to data interpretation and drafting of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

This study was approved by the French regulatory authorities (the "Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le Domaine de la Santé" and the "Commission Nationale Informatique et Libertés", authorisation number 907303). According to French law, patients were informed of their data registration and given the right to deny access or to rectify their personal data.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 27 July 2017 Accepted: 28 March 2018
Published online: 06 April 2018

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136:E359–86.
2. Sant M, Allemani C, Capocaccia R, Hakulinen T, Aareleid T, Coebergh JW, et al. Stage at diagnosis is a key explanation of differences in breast cancer survival across Europe. Int J Cancer. 2003;106:416–22.
3. Altobelli E, Lattanzi A. Breast cancer in European Union: an update of screening programmes as of march 2014 (review). Int J Oncol. 2014; 45:1785–92.
4. Lauby-Secretan B, Loomis D, Straif K. Breast-Cancer screening–viewpoint of the IARC working group. N Engl J Med. 2015;373:1479.
5. Paci E, EUROSCREEN Working Group. Summary of the evidence of breast cancer service screening outcomes in Europe and first estimate of the benefit and harm balance sheet. J Med Screen. 2012;19(Suppl 1):5–13.
6. Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. Lancet Lond. Engl. 2012; 380:1778–86.
7. Sarkeala T, Heinävaara S, Anttila A. Organised mammography screening reduces breast cancer mortality: a cohort study from Finland. Int J Cancer. 2008;122:614–9.
8. Weedon-Fekjær H, Romundstad PR, Vatten LJ. Modern mammography screening and breast cancer mortality: population study. BMJ. 2014; 348:g3701.
9. Moss SM, Nyström L, Jonsson H, Paci E, Lynge E, Njor S, et al. The impact of mammographic screening on breast cancer mortality in Europe: a review of trend studies. J Med Screen. 2012;19(Suppl 1):26–32.
10. Puliti D, Duffy SW, Miccinesi G, de Koning H, Lynge E, Zappa M, et al. Overdiagnosis in mammographic screening for breast cancer in Europe: a literature review. J Med Screen. 2012;19(Suppl 1):42–56.
11. Moss SM, Cuckle H, Evans A, Johns L, Waller M, Bobrow L, et al. Effect of mammographic screening from age 40 years on breast cancer mortality at 10 years' follow-up: a randomised controlled trial. Lancet Lond Engl. 2006; 368:2053–60.
12. Harding C, Pompei F, Burmistrov D, Welch HG, Abebe R, Wilson R. Breast Cancer screening, incidence, and mortality across US counties. JAMA Intern Med. 2015;175:1483–9.
13. Miller AB, Wall C, Baines CJ, Sun P, To T, Narod SA. Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. BMJ. 2014;348:g366.
14. La participation au dépistage du cancer du sein chez les femmes de 50 à 74 ans en France. Situation actuelle et perspectives d'évolution [Internet]. Haute Autorité de Santé; 2012. Available from: http://www.has-sante.fr/portail/upload/docs/application/pdf/2012-02/argumentaire_-_participation_depistage_cancer_du_sein_2012-02-02_15-27-14_245.pdf. Accessed 3 Apr 2018.
15. Le programme de dépistage organisé [Internet]. Institut National du Cancer; 2015. Available from: http://www.e-cancer.fr/Professionnels-de-sante/Depistage-et-detection-precoce/Depistage-du-cancer-du-sein/Le-programme-de-depistage-organise. Accessed 3 Apr 2018.
16. Domingo L, Blanch J, Servitja S, Corominas JM, Murta-Nascimento C, Rueda A, et al. Aggressiveness features and outcomes of true interval cancers: comparison between screen-detected and symptom-detected cancers. Eur J Cancer Prev Off J Eur Cancer Prev Organ ECP. 2013;22:21–8.
17. Caumo F, Vecchiato F, Strabbioli M, Zorzi M, Baracco S, Ciatto S. Interval cancers in breast cancer screening: comparison of stage and biological characteristics with screen-detected cancers or incident cancers in the absence of screening. Tumori. 2010;96:198–201.

18. Rayson D, Payne JI, Abdolell M, Barnes PJ, Macintosh RF, Foley T, et al. Comparison of clinical-pathologic characteristics and outcomes of true interval and screen-detected invasive breast cancer among participants of a Canadian breast screening program: a nested case-control study. Clin Breast Cancer. 2011;11:27–32.
19. Bordás P, Jonsson H, Nyström L, Lenner P. Survival from invasive breast cancer among interval cases in the mammography screening programmes of northern Sweden. Breast Edinb Scotl. 2007;16:47–54.
20. Holmberg LH, Tabar L, Adami HO, Bergström R. Survival in breast cancer diagnosed between mammographic screening examinations. Lancet Lond Engl. 1986;2:27–30.
21. Brekelmans CT, Peeters PH, Deurenberg JJ, Collette HJ. Survival in interval breast cancer in the DOM screening programme. Eur J Cancer Oxf Engl 1990. 1995;31A:1830–5.
22. Fracheboud J, de Koning HJ, Beemsterboer PM, Boer R, Verbeek AL, Hendriks JH, et al. Interval cancers in the Dutch breast cancer screening programme. Br J Cancer. 1999;81:912–7.
23. Zackrisson S, Janzon L, Manjer J, Andersson I. Improved survival rate for women with interval breast cancer - results from the breast cancer screening programme in Malmö, Sweden 1976-1999. J Med Screen. 2007;14:138–43.
24. Kalager M, Tamimi RM, Bretthauer M, Adami H-O. Prognosis in women with interval breast cancer: population based observational cohort study. BMJ. 2012;345:e7536.
25. Delacour-Billon S, Mathieu-Wacquant AL, Campone M, Auffret N, Amossé S, Allioux C, et al. Short-term and long-term survival of interval breast cancers taking into account prognostic features. Cancer Causes Control CCC. 2017; 28:69–76.
26. Jensen O, Parkin D, MacLennan R, Muir C, Skeet RG. Cancer Registration: Principles and Methods. International Agency for Research on Cancer; 1991.
27. Defossez G, Rollet A, Dameron O, Ingrand P. Temporal representation of care trajectories of cancer patients using data from a regional information system: an application in breast cancer. BMC Med Inform Decis Mak. 2014;14:24.
28. Quillet A, Defossez G, Ingrand P. Surveillance of waiting times for access to treatment: a registry-based computed approach in breast cancer care. Eur J Cancer Care (Engl). 2016;25:764–73.
29. Wishart GC, Greenberg DC, Britton PD, Chou P, Brown CH, Purushotham AD, et al. Screen-detected vs symptomatic breast cancer: is improved survival due to stage migration alone? Br J Cancer. 2008;98:1741–4.
30. Shen Y, Yang Y, Inoue LYT, Munsell MF, Miller AB, Berry DA. Role of detection method in predicting breast cancer survival: analysis of randomized screening trials. J Natl Cancer Inst. 2005;97:1195–203.
31. Duffy SW, Nagtegaal ID, Wallis M, Cafferty FH, Houssami N, Warwick J, et al. Correcting for lead time and length bias in estimating the effect of screen detection on cancer survival. Am J Epidemiol. 2008;168:98–104.
32. Vanier A, Leux C, Allioux C, Billon-Delacour S, Lomboll P, Molinié F. Are prognostic factors more favorable for breast cancer detected by organized screening than by opportunistic screening or clinical diagnosis? A study in Loire-Atlantique (France). Cancer Epidemiol. 2013;37:683–7.

3.2. Réutilisation des données de génétique moléculaire des cancers [38]

Un second domaine d'application a concerné l'accès des patients au diagnostic moléculaire des cancers en amont de leur traitement. Pour rappel, les premières thérapies ciblées ont été autorisées dans les années 2000 et sont venues compléter l'arsenal thérapeutique existant, représenté en grande partie par la classe des chimiothérapies conventionnelles (cytotoxiques). La prescription de ces thérapies ciblées étant guidée par les caractéristiques moléculaires de la tumeur (« médecine de précision »), des biomarqueurs peuvent alors être recherchés afin d'identifier les patients qui sont porteurs ou non d'une altération. Ainsi, plus de la moitié des thérapies ciblées autorisées disposent d'un biomarqueur conditionnant leur prescription, et l'accès à ces tests est dès lors indispensable pour permettre aux patients un accès à ces traitements [39].

Dans ce contexte, nous nous sommes intéressés à apprécier la proportion de patients atteints de cancer qui accédait en pratique à ces tests moléculaires, et de déterminer les facteurs qui influencaient leurs prescriptions. L'étude s'est focalisée sur le cancer colorectal (CCR), 3ème cancer le plus fréquent au niveau de l'ex-région Poitou-Charentes, en France et dans le monde. L'intérêt résidait dans la mise en relation des données du RGCPC (population exhaustive des patients atteints de CCR de type adénocarcinome en 2010, n=1 269) à celles de la plateforme régionale de génétique moléculaire des cancers, pour identifier les tests moléculaires mise en œuvre (codons KRAS 12/13, BRAF et instabilité microsatellite (MSI)) au regard des recommandations en vigueur.

Cette étude a été la première à fournir un état des lieux en population générale française. Parmi les 1 269 CCR inclus dans l'étude, les tests KRAS, BRAF et MSI étaient réalisés dans 35,1%, 10,5% et 10,9 % des cas, respectivement. La recherche de mutation KRAS était effectuée dans 65,5 % des cas de CCR métastatique, et 26,1 % des cas de CCR non métastatique (patients qui présentaient un risque élevé de récidive de la maladie). Parmi les CCR métastatiques, l'âge (<60 ans), le site de la tumeur primitive (côlon gauche) et le département du primo-traitement (taux de prescription les plus élevés en Deux-Sèvres > Vienne > Charente > Charente-Maritime) étaient des facteurs liés à la prescription des tests KRAS. Le test BRAF (indicateur pronostic) était contemporain du test KRAS pour 92,5 % des patients. Enfin les facteurs liés aux tests MSI étaient l'âge (<60 ans), le stade TNM (stade IV) et le lieu de traitement. Parmi les patients de moins de 60 ans, seuls 37,5 % avaient eu un test MSI.

Ces résultats ont permis d'identifier certains facteurs sur lesquels il était possible d'agir afin de réduire les disparités et renforcer la réalisation des tests moléculaires du CCR (tests MSI notamment).

Lire Article 5 *: Thiebault Q, Defossez G, Karayan-Tapon L, Ingrand P, Silvain C, Tougeron D. Analysis of factors influencing molecular testing at diagnostic of colorectal cancer.* ***BMC Cancer****. 2017 Nov 14;17(1):765. doi: 10.1186/s12885-017-3759-6. PMID: 29137623; PMCID: PMC5686889.*

Accéder à la suite « Réutilisation des données PMSI et RCP »

Thiebault et al. BMC Cancer (2017) 17:765
DOI 10.1186/s12885-017-3759-6

BMC Cancer

RESEARCH ARTICLE Open Access

Analysis of factors influencing molecular testing at diagnostic of colorectal cancer

Quentin Thiebault[1], Gautier Defossez[2,3], Lucie Karayan-Tapon[4], Pierre Ingrand[2,3], Christine Silvain[1,5] and David Tougeron[1,5*]

Abstract

Background: The aim of the study was to evaluate the current rate of molecular testing prescription (*KRAS* codons 12/13, *BRAF* and microsatellite instability (MSI)) in newly diagnosed colorectal cancer (CRC) patients and to determine which factors influence testing.

Methods: All incident CRC cases in 2010 were identified in the Poitou-Charentes General Cancer Registry. The exhaustive molecular testing performed was accessed in the French molecular genetics platform. Factors influencing prescription were analyzed using logistic regression.

Results: Among the 1269 CRCs included in the study, *KRAS*, *BRAF* and MSI testing accounted for 35.1%, 10.5% and 10.9%, respectively. *KRAS* testing was carried out in 65.5% of metastatic CRCs, and 26.1% of non-metastatic CRCs. Among metastatic CRCs, age (<60 years), site of primary tumour (left colon) and geographical area of treatment were factors related to *KRAS* testing. *BRAF* testing was contemporary to *KRAS* testing for 92.5% of patients. Factors related to MSI testing were age (<60 years), TNM stage (stage IV) and geographical area of treatment. Among CRC patients under 60 years old, only 37.5% had MSI testing.

Conclusion: These results underscore the need to reduce disparities in CRC molecular testing and highlight the limited application of the French guidelines, especially concerning MSI testing.

Keywords: Colorectal cancer, *KRAS*, Mutation, Molecular testing, *BRAF*, Microsatellite instability

Background

Colorectal cancer (CRC) is the third most common cancer worldwide [1]. To date, colorectal carcinogenesis has been classified in three distinct pathways: chromosomal instability (85%), microsatellite instability (MSI) (15%) and CpG island methylator phenotype (25%). MSI is related to a deficient DNA mismatch repair (dMMR) system due to germline mutation in a MMR gene in Lynch syndrome (LS), or more commonly to an epigenetic inactivation of *MLH1* in sporadic cases. Approximately 45% of CRC cases have a *KRAS* mutation [2] and only patients with wild-type (WT) *RAS* metastatic CRC (mCRC) may benefit from anti-epidermal growth factor receptor monoclonal antibody therapy (anti-EGFR mAbs) [3, 4]. A *BRAF* mutation (V600E) is present in approximately 12% of CRCs and confers a poor prognosis, especially in mCRCs [5–8]. In dMMR CRC, *BRAF* mutation is specific to a sporadic origin and eliminates a LS.

Since 2006, the French National Cancer Institute (INCa) has been supporting a national network of 28 hospital molecular genetics platforms throughout France, offering patients all essential molecular genetics techniques for all cancers. For CRC, *KRAS* (now complete *RAS*), *BRAF* and MSI testing are routinely performed. Since 2008, *KRAS* testing is supposed to be performed in all mCRC cases. Since *KRAS* and *BRAF* mutations are mutually exclusive [8], *BRAF* testing is performed only in *KRAS* WT tumours. In France, MSI testing is recommended in patients with a CRC at an age lower than 60 and/or if family history suggests a LS. Nevertheless, epidemiological data concerning these different testing procedures are lacking. A recent French retrospective study revealed that 81.1% of patients with a mCRC had *KRAS*

* Correspondence: davidtougeron@hotmail.fr
[1]Department of Gastroenterology, Poitiers University Hospital, 2 rue de la Milétrie, 86000 Poitiers Cedex, France
[5]Laboratory Inflammation, Tissus Epithéliaux et Cytokines, EA 4331, University of Poitiers, Poitiers, France
Full list of author information is available at the end of the article

testing [9]. This study has some limitations due the non-exhaustiveness of incident CRC cases included and patient recruitment based on physician willingness. The General Cancer Registry in the Poitou-Charentes region (GCRPC) covers an administrative region of 1.8 million people in south-western France (available at http://medphar.univ-poitiers.fr/registre-cancers-poitou-charentes/) and has been collecting all incident cancer cases, thereby enabling exhaustive analysis of the molecular analyses (using INCa molecular cancer genetics platform) performed in all incident CRC cases. The aim of the study was to analyze routine practice of *KRAS*, *BRAF* and MSI molecular testing among all the CRC patients in Poitou-Charentes diagnosed in 2010.

Methods

Study population

Since 2008, the GCRPC has included all incident cases of cancer, involving subjects regularly residing in the Poitou-Charentes region at the time of diagnosis, whatever the place of care. The Poitou-Charentes region comprises four departments: Charente, Charente-Maritime, Deux-Sèvres and Vienne. The minimum items recorded in the GCRPC were demographic data, tumour characteristics and treatment. According to the French law the data collected from the GCRPC was approved by the CCTIRS (Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé, approval n°07–374) and the CNIL (Commission Nationale de l'Informatique et des Libertés, approval n°907,303). Using the GCRPC 1375 incident CRC patients were identified in 2010 and after exclusion of non-relevant cases, 1269 patients were included in the study (Fig. 1).

Molecular testing

In 2010, *KRAS* mutational status (exon 2 codons 12 and 13) was determined at the specific request of a clinician. Regarding *BRAF* mutational status (V600E), analysis was mostly performed by the INCa hospital molecular genetics platforms in case of *KRAS* wild-type status. MSI was to be determined at the specific request of the clinician (suspicion of LS) or by the platforms for patients under 60 years old. All of the exhaustive molecular analyses ($n = 480$) from the different hospital molecular genetics platforms were itemized (Poitiers ($n = 401$) and other platforms ($n = 79$)).

Statistical analysis

The aim of the study was to evaluate the rate of prescription of molecular testing (*KRAS*, *BRAF* and MSI) regarding guidelines applicable in 2010 to newly diagnosed CRC patients. Secondary objectives were to analyze which criteria influenced *KRAS* molecular testing for metastatic and non-metastatic CRC patients respectively, and which characteristics influenced MSI molecular testing for all CRC patients.

The study was conducted in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement. The descriptive statistics used for quantitative parameters were mean and standard deviation; for qualitative parameters were frequency and percentage. A logistic regression was carried out on factors that could promote *KRAS* testing and

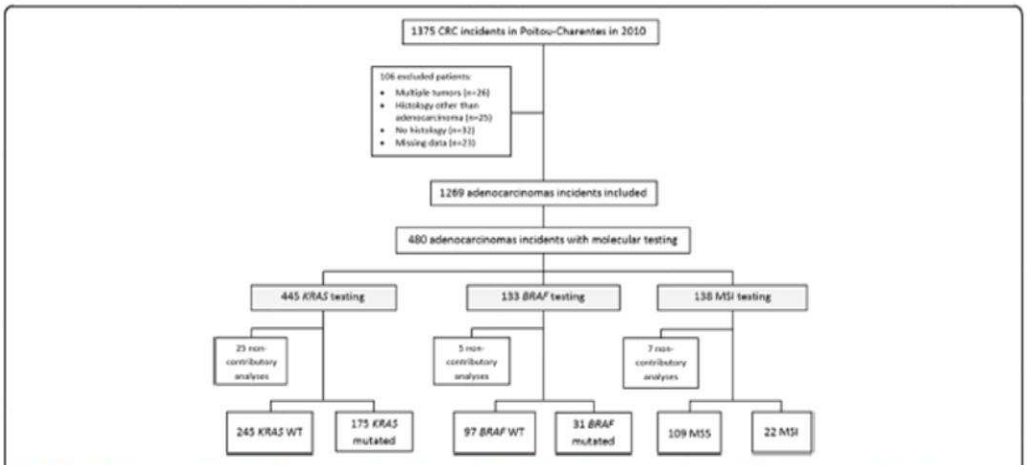


Fig. 1 Flowchart of the study. Abbreviations: CRC, colorectal cancer; WT, wild-type; MSI, microsatellite instability; MSS, microsatellite stable

MSI testing and determined odds ratios (OR) with a 95% confidence interval (CI). The geographical area of primary treatment was defined from the location of the center where the first treatment of CRC was performed. Status of the center was categorized as public, private or university hospital.

Statistically significant factors derived from univariate analysis (P values <0.25) were selected for multivariate analysis using a stepwise descending selection procedure with a significance threshold at 0.05. Possible interactions between independent risk factors were tested by including proper cross-product terms in the regression models, and likelihood ratio tests comparing models with and without the interaction term were used to estimate the significance of the interaction. Data management and statistical analyses were performed using SAS software version 9.4 (SAS Institute, Cary, NC, USA).

Results

Population

Between January 1st and December 31st 2010, 1269 incident cases of CRC were included in the study. The age-standardized incidence rates of CRC were respectively 38.3 per 100,000 in men and 26.9 per 100,000 in women. Mean age was 71.9 ± 11.8 years (Table 1). At diagnosis, 22.8% of CRCs were metastatic and 77.2% were non-metastatic.

Molecular testing

Overall, 480 CRCs (37.8% of the cohort) had at least one molecular test (*KRAS*, *BRAF* or MSI). *KRAS* was mutated in 41.7% of cases (n = 175/420), *BRAF* mutation in 24.2% (n = 31/128) and a dMMR phenotype was found in 16.8% (n = 22/131). Among the 480 molecular tests in Poitou-Charentes incident cases of CRC, 83.5% (n = 401)

Table 1 Patient and tumour characteristics

	All patients (n = 1269)	Patients without molecular test[a] (n = 789)	Patients with at least one molecular test[a] (n = 480)
Age (years, SD)	71.9 ± 11.8	74.4 ± 11.3	67.8 ± 11.4
Sex			
Women	578 (45.5%)	356 (45.1%)	222 (46.3%)
Men	691 (54.5%)	433 (54.9%)	258 (53.8%)
Site of the primary tumour			
Rectum	309 (24.4%)	220 (28.0%)	89 (18.8%)
Right colon	437 (34.4%)	268 (34.0%)	169 (35.0%)
Left colon	523 (41.2%)	301 (38.0%)	222 (46.3%)
TNM stage			
Stage I	219 (17.3%)	194 (24.6%)	25 (5.2%)
Stage II	380 (29.9%)	265 (33.6%)	115 (24.0%)
Stage III	380 (29.9%)	232 (29.4%)	148 (30.8%)
Stage IV	290 (22.8%)	98 (12.4%)	192 (40.0%)
Tumour grade (MD = 173)			
Well	436 (39.8%)	250 (37.7%)	186 (43.3%)
Moderate	576 (52.5%)	369 (55.4%)	207 (47.9%)
Poor	84 (7.7%)	46 (6.9%)	38 (8.8%)
Geographical area of primary treatment (MD = 4)			
Charente-Maritime	392 (31.0%)	286 (36.4%)	106 (22.1%)
Charente	226 (17.9%)	171 (21.6%)	55 (11.5%)
Deux-Sèvres	227 (17.9%)	110 (14.1%)	117 (24.4%)
Vienne	301 (23.8%)	137 (17.4%)	164 (34.2%)
Outside the region	119 (9.4%)	82 (10.4%)	37 (7.7%)
Status of the center (MD = 4)			
Public Hospital	551 (43.5%)	381 (48.6%)	170 (35.5%)
Private hospital	512 (40.5%)	292 (37.2%)	220 (45.7%)
University hospital	202 (16.0%)	112 (14.3%)	90 (18.8%)

MD missing data, *SD* standard deviation
[a]Molecular test defined as *KRAS*, *BRAF* and/or MSI testing

were performed in the platform of Poitiers and 16.5% ($n = 79$) outside the region.

The average time to obtain results of molecular tests, defined by the interval between the date of histological sampling and the date of molecular test results available in the platform, was 30.6 days for *KRAS* testing, 36.3 days for *BRAF* testing and 41.3 days for MSI testing.

KRAS testing

KRAS molecular testing was carried out in 35.1% ($n = 445/1269$), including 65.5% ($n = 190/290$) metastatic and 26.1% ($n = 255/979$) non-metastatic CRC patients. *KRAS* molecular testing was mainly requested by pathologists ($n = 174$, 39.1%), surgeons ($n = 105$, 23.6%) and oncologists ($n = 84$, 18.9%) (Table 2). Among mCRC patients, 68.6% ($n = 199/290$) received chemotherapy and among them 83.9% ($n = 167/199$) had *KRAS* molecular testing.

Among overall cohort, age at diagnosis, site of primary tumor, stage at diagnosis, geographical area of primary treatment and status of the center were the factors related to *KRAS* testing (data not shown). Secondly, analyses of metastatic and non-metastatic CRCs were performed separately, given that *KRAS* testing is recommended only in cases of mCRC.

Among mCRC patients, in multivariate analysis, age at diagnosis (<75 years; $p < 0.0001$), site of primary tumor (left colon; $p = 0.006$) and geographical area of primary treatment ($p = 0.01$) were factors related to *KRAS* molecular testing (Table 3). All mCRC patients treated with an anti-EGFR mAbs had *KRAS* molecular testing ($n = 42$). Among *KRAS* wild-type mCRC ($n = 101$), 40.6% were treated with anti-EGFR mAbs. More than half of *KRAS* molecular testing for mCRC patients was requested by pathologists ($n = 60$, 31.6%) and oncologists ($n = 51$, 26.3%).

Among non-metastatic CRC patients, in multivariate analysis, age at diagnosis (<75 years; $p < 0.0001$), site of primary tumor (right colon; $p = 0.026$), stage at diagnosis (stage II and III; $p < 0.0001$), geographical area of primary treatment ($p < 0.0001$) and status of the center (private hospital; $p < 0.0001$) were factors related to *KRAS* molecular testing (Table 4). *KRAS* molecular testing for non-metastatic CRC patients was mainly requested by pathologists ($n = 114$, 44.7%) and surgeons ($n = 72$, 28.2%).

Table 2 Specialty of physicians who order molecular testing

	KRAS ($n = 445$)	*BRAF* ($n = 133$)	MSI ($n = 138$)
Pathologists	174 (39.1%)	38 (28.6%)	34 (24.6%)
Surgeons	105 (23.6%)	22 (16.5%)	21 (15.2%)
Oncologists	84 (18.9%)	37 (27.8%)	43 (31.2%)
Gastroenterologists	9 (2.0%)	1 (0.7%)	1 (0.7%)
Others	4 (0.9%)	7 (5.3%)	5 (3.6%)
Non communicated/ unknown	69 (15.5%)	28 (21.1%)	34 (24.6%)

BRAF testing

BRAF molecular testing was performed in 10.5% ($n = 133/1269$), including 18.6% ($n = 54/290$) metastatic and 8.1% ($n = 79/979$) non-metastatic CRC patients. *BRAF* molecular testing was mainly requested by pathologists ($n = 38$, 28.6%), oncologists ($n = 37$, 27.8%) and surgeons ($n = 22$, 16.5%). *BRAF* molecular testing was contemporary to *KRAS* molecular testing for 92.5% of CRC patients ($n = 123/133$), of whom 93.5% ($n = 115/123$) were *KRAS* WT. Among the 101 *KRAS* WT mCRC patients, 47.5% ($n = 48$) had *BRAF* testing. Considering that *BRAF* testing should be performed in case of *KRAS* WT status, the factors associated with *BRAF* testing were not detailed as they were in fact similar to those for *KRAS* testing.

MSI testing

MSI molecular testing was performed in 10.9% ($n = 138/1269$), 39.4% ($n = 82/208$) in patients under 60 years and 5.3% ($n = 56/1061$) in patients over 60 years. MSI molecular testing was mainly requested by oncologists ($n = 43$, 31.2%) and pathologists ($n = 34$, 24.6%). Among the 138 patients with MSI testing, 58.0% ($n = 80/138$) had no *BRAF* testing. There was no significant difference in proportion of MSI testing between *BRAF*-mutated and *BRAF* WT CRC, respectively 38.7% ($n = 12/31$) and 43.3% ($n = 42/97$) ($p = 0.65$).

In multivariate analysis, age at diagnosis (<75 years; $p < 0.0001$), stage at diagnosis (stage II, III and IV; $p < 0.0001$) and geographical area of primary treatment ($p < 0.0001$) were factors related to MSI testing (Table 5). Among patients under 60 years old, 39.4% ($n = 82/208$) had MSI testing and 11.5% had an oncogenetic consultation ($n = 24/208$). Overall, among the 22 patients with dMMR CRC, we identified 6 *BRAF* wild-type CRCs (27.3%), 9 *BRAF*-mutated CRCs (40.9%) and 7 without *BRAF* testing (31.8%). Among patients with dMMR CRC and *BRAF* wild-type status or no *BRAF* testing, 61.5% had an oncogenetic consultation ($n = 8/13$).

Discussion

Our study is the first one to simultaneously evaluate three molecular testing procedures (*KRAS*, *BRAF* and MSI) in CRC. Rates for these molecular testing procedures were systematically linked to age at CRC diagnosis, site of primary tumour, stage at diagnosis, geographical area of primary treatment and status of the center.

KRAS testing was performed in 35.1% of CRCs and as expected was more frequent in patients with a metastatic disease (65.5%). Although *KRAS* status is required for the anti-EGFR mAbs used in mCRC, there are few data on *KRAS* testing rates. In a French retrospective

Table 3 Factors influencing *KRAS* testing in metastatic CRC patients

	KRAS testing *n* = 190/290 (65.5%)	Univariate analysis *P-Value*	Multivariate analysis		
			Odds ratio	95% CI	*P-Value*
Age (years)		< 0.0001			< 0.0001
> 75	42 / 107 (39.3%)		1	Ref	
60–75	97 / 126 (77.0%)		4.72	2.54–8.77	
< 60	51 / 57 (89.5%)		10.78	4.07–28.50	
Sex		0.016			0.33
Women	82 / 140 (58.6%)		1	Ref	
Men	108 / 150 (72.0%)		1.34	0.75–2.41	
Site of the primary tumour		0.0011			0.006
Rectum	35 / 64 (54.7%)		1	Ref	
Right colon	51 / 90 (56.7%)		1.44	0.67–3.07	
Left colon	104 / 136 (76.5%)		3.09	1.48–6.45	
Tumour grade (MD = 40)		0.52			
Well	63 / 89 (70.8%)				
Moderate	83 / 131 (63.4%)				
Poor	20 / 30 (66.7%)				
Geographical area of primary treatment (MD = 2)		0.0001			0.010
Charente-Maritime	42 / 87 (48.3%)		1	Ref	
Charente	32 / 53 (60.4%)		1.99	0.89–4.46	
Deux-Sèvres	37 / 45 (82.2%)		4.64	1.77–12.18	
Vienne	61 / 79 (77.2%)		2.88	1.36–6.13	
Outside the region	17 / 24 (70.8%)		2.02	0.67–6.12	
Status of the center (MD = 2)		0.026[a]			
Public Hospital	60 / 83 (72.3%)		–	–	–
Private hospital	83 / 143 (58.0%)				
University hospital	46 / 62 (74.2%)				

95% CI 95% confidence interval, *MD* missing data, *Ref* reference
[a]Not retained in the final multivariate model

study conducted in 2011 81.1% of mCRCs had *KRAS* testing [9] which is higher as compared our work. However, there are selection biases in Lièvre et al. study since patient recruitment was based on physician willingness. Finally, our rate is in accordance with that found in a large retrospective study published in 2011 concerning Europe, Latin America and Asia (69%) [10]. Moreover, in our study when limited to mCRC patients receiving first-line chemotherapy, *KRAS* molecular testing rate was higher (83.9%).

Among mCRC patients, in multivariate analysis young age at diagnosis, primary tumor located in left colon and geographical area of primary treatment were factors related to *KRAS* molecular testing. Frequent *KRAS* testing in young patients is probably explained by more "aggressive" treatment strategies in these patients, particularly anti-EGFR mAbs used. We have no explanation as to why *KRAS* testing was more frequent for left-sided tumors. *KRAS* testing was also significantly more frequent in the Vienne and Deux-Sèvres departments. In the Poitou-Charentes region there is only one university hospital located in the Vienne department. We can suppose that the higher rate of *KRAS* testing in Vienne department was linked to university hospital research programs and easier access to molecular testing. We observed that molecular testing procedures were mainly requested by pathologists and oncologists. An earlier request by gastroenterologists on initial biopsies should be encouraged to allow the availability of molecular tests results during the first oncological consultation in order to quickly define the optimal treatment for mCRC (RAS status and anti-EGFR treatment).

Our work showed that 26.1% of non-metastatic CRC cases had *KRAS* testing. The rate in the USA population is 5% [11]. Younger age, higher stage at diagnosis, geographical area of primary treatment and status of the

Table 4 Factors influencing *KRAS* testing in non-metastatic CRC patients

	KRAS testing *N* = 255/979 (26.1%)	Univariate analysis *P-Value*	Multivariate analysis		
			Odds ratio	95% CI	*P-Value*
Age (years)		0.0004			< 0.0001
> 75	90 / 446 (20.2%)		1	Ref	
60–75	123 / 382 (32.2%)		2.69	1.83–3.95	
< 60	42 / 151 (27.8%)		2.26	1.34–3.81	
Sex		0.25			0.15
Women	133 / 541 (24.6%)		1	Ref	
Men	122 / 438 (27.9%)		0.77	0.55–1.09	
Site of the primary tumour		0.010			0.026
Rectum	47 / 245 (19.2%)		1	Ref	
Right colon	105 / 347 (30.3%)		1.95	1.20–3.16	
Left colon	103 / 387 (26.6%)		1.47	0.92–2.35	
TNM stage		< 0.0001			< 0.0001
I	21 / 219 (9.6%)		1	Ref	
II	100 / 380 (26.3%)		5.24	2.98–9.21	
III	134 / 380 (35.3%)		9.62	5.47–16.90	
Tumour grade (MD = 133)		0.0082[a]			
Well	114 / 347 (32.8%)		–	–	–
Moderate	103 / 445 (23.2%)				
Poor	17 / 54 (31.5%)				
Geographical area of primary treatment (MD = 2)		< 0.0001			< 0.0001
Charente-Maritime	63 / 305 (20.7%)		1	Ref	
Charente	13 / 173 (7.5%)		0.19	0.10–0.37	
Deux-Sèvres	75 / 182 (41.2%)		3.74	2.35–5.95	
Vienne	94 / 222 (42.3%)		3.24	1.95–5.39	
Outside the region	10 / 95 (10.5%)		0.36	0.16–0.79	
Status of the center (MD = 2)		< 0.0001			< 0.0001
Public Hospital	77 / 408 (18.9%)		1	Ref	
Private hospital	149 / 429 (34.7%)		4.18	2.74–6.37	
University hospital	29 / 140 (20.7%)		0.88	0.44–1.75	

95% CI 95% confidence interval, *NA* not available, *MD* missing data, *Ref* reference
[a]Not retained in the final multivariate model

center were factors related to *KRAS* molecular testing in non-metastatic CRCs. We can suppose that it was conducted at the request of the clinician to quickly begin appropriate treatment in the event of development of metachronous metastases, especially in stage III patients. In addition, for some pathologists it was easier to address pathological samples to a molecular cancer genetics platform at the time of the first pathological examination rather than later, when the tumor blocs were archived.

To our knowledge there has been no previous study evaluating *BRAF* testing rates in CRC cases. In our study the rate of *BRAF* testing was 10.5% and the factors influencing *BRAF* testing are similar to those influencing *KRAS* testing. The rate of *BRAF*-mutated CRC (24.2%) was high as compared with the literature (approximately 12%) [12, 13]. *BRAF* testing was mostly performed directly by molecular cancer genetic platform in patients with *KRAS* wild-type CRC since the two mutations are mutually exclusive. This point explains the high rate of *BRAF*-mutated CRC since only *KRAS* WT CRCs were analyzed for *BRAF*.

Concerning MSI testing, the rate seems low (10.8%) but the dMMR CRC rate is in accordance with literature data [14, 15]. To our knowledge this is the first study that analyzing factors related to MSI testing rates. Like *KRAS* testing, MSI testing was associated in multivariate analysis with young age, higher tumor stage and geographical area of primary treatment. French guidelines

Table 5 Factors influencing MSI testing in all CRC patients

	MSI testing N = 138/1269 (10.9%)	Univariate analysis P-Value	Multivariate analysis		
			Odds ratio	IC 95%	P-Value
Age (years)		< 0.0001			< 0.0001
> 75	10 / 553 (1.8%)		1	Ref	
60–75	46 / 508 (9.1%)		5.77	2.80–11.89	
< 60	82 / 208 (39.4%)		59.6	27.82–127.85	
Sex		0.88			0.21
Women	62 / 578 (10.7%)		1	Ref	
Men	76 / 691 (11.0%)		0.75	0.47–1.18	
Site of the primary tumour		0.53			
Rectum	31 / 309 (10.0%)				
Right colon	44 / 437 (10.1%)				
Left colon	63 / 523 (12.1%)				
TNM stage		< 0.001			0.0001
I	6 / 219 (2.7%)		1	Ref	
II	37 / 380 (9.7%)		8.62	3.17–23.41	
III	49 / 380 (12.9%)		7.96	3.00–21.12	
IV	46 / 290 (15.9%)		8.79	3.29–23.48	
Tumour grade (MD = 173)		0.34			
Well	55 / 436 (12.6%)				
Moderate	56 / 576 (9.7%)				
Poor	9 / 84 (10.7%)				
Geographical area of primary treatment (MD = 4)		< 0.0001			< 0.0001
Charente-Maritime	3 / 392 (0.8%)		1	Ref	
Charente	21 / 226 (9.3%)		17.08	4.78–61.00	
Deux-Sèvres	19 / 227 (8.4%)		13.98	3.89–50.24	
Vienne	80 / 301 (26.6%)		69.59	20.51–236.01	
Outside the region	15 / 119 (12.6%)		13.53	3.64–50.29	
Status of the center (MD = 4)		< 0.0001[a]			
Public Hospital	36 / 551 (6.5%)		–	–	–
Private hospital	47 / 512 (9.2%)				
University hospital	55 / 202 (27.2%)				

95% CI 95% confidence interval, *NA* not available, *MD* missing data, *Ref* reference
[a]Not retained in the final multivariate model

recommended MSI testing for patients under 60 years old and/or *BRAF*-mutated CRC. Consequently, MSI testing was performed directly by the molecular cancer genetics platforms for patients under 60 years old and/or *BRAF*-mutated CRC when there was *KRAS*/*BRAF* testing. These points explain how it is that the factors influencing MSI testing are close to those influencing *KRAS*/*BRAF* testing.

Our study highlights the fact that guidelines for LS screening are not well-respected. Only 39.4% of CRC patients under 60 years old had MSI testing and some dMMR CRCs (31.8%) did not have *BRAF* testing to identify sporadic cases. Finally, most patients with a suspicion of LS (dMMR CRC with no *BRAF* mutation) did not have an oncogenetic consultation (38.5%). We were not able to determine if this was due to patient refusal or if patients had not been addressed to an oncogenetic consultant by their referring physician.

The average time to obtain results of *KRAS* tests in our study was 30.6 days (between histological sampling and the date when the molecular test results were available in the platform). Lièvre et al., calculated the median delay between physician prescription and reception of the results as 23.6 ± 28.2 days, a delay somewhat shorter because measured differently [9]. In addition, in contrast to the Lièvre et al. study, our study is reflective of real

life and exhaustive. To our knowledge, no previous study evaluated delays in *BRAF* and MSI testing.

The main strength of our study resides in the crossing of two reliable and exhaustive data banks, GCRPC and INCa molecular cancer genetics platforms. If none of the previous studies evaluating *KRAS* testing are as exhaustive, it is because they were based on incomplete database and/or on questionnaires sent to volunteer physicians. The main limitation of our work is the difficulty in extrapolating its results to other countries since CRC molecular tests are dependent on physicians' and pathologists' clinical practices. It is noteworthy that we accessed the molecular testing rates in 2010 since there is a delay of at least 2 years before obtaining high-quality CRC data from the GCRPC, a delay justified by the data collection process and the application of standards and requirements during case registration. Moreover, it is challenging to retrieve reliable and retrospective information on life-style and family history, but it would be interesting to complete this evaluation by including CCR risk factors which probably influences the choice of the clinician for ordering molecular testing. Finally, factors influencing these molecular testing procedures are relevant for countries which already performed these tests but also those who are implementing these tests in order to allow an optimal use, especially RAS testing for anti-EGFR therapy used in mCRC.

Conclusion

To conclude, this study is the first to provide a robust and exhaustive overview of molecular testing in CRC. As expected, we note a high level of *KRAS* testing in mCRC but also significant level in stage III CRC, which was probably undertaken in order to have *KRAS* results for patients with a high risk of disease recurrence. Moreover, MSI testing rate is low and not in accordance with French guidelines, which recommend systematic testing before the age of 60. In addition, these results highlighted on which factors it is possible to act to improve the molecular testing procedures essential to management of CRC patients, particularly MSI testing.

Abbreviations

CI: Confidence interval; CRC: Colorectal cancer; dMMR: Deficient DNA mismatch repair; GCRPC: General cancer registry in the poitou-charentes region; LS: Lynch syndrome; MSI: Microsatellite instability; OR: Odds ratios; WT: Wild-type

Acknowledgments

The authors wish to thank J. Arsham, an American translator, for having reviewed and revised the original English-language text. The authors thank V. Le Berre, a research secretary, for her help in editing and formatting the manuscript.

Funding

This work was supported in part by the Ligue contre le Cancer of Vienne, Deux-Sèvres, Charente and Charente-Maritime departments and the "Sport et Collection" foundations for the molecular MSI testing.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' contributions

Conception and design of the study: QT, GD, LK, PI, CS, DT. Data analysis and interpretation: QT, GD, LK, PI, CS, DT. Data acquisition, statistical analysis and writing the manuscript: QT, GD, DT. Final approval of the manuscript: QT, GD, LK, PI, CS, DT. All authors read and approved the final manuscript.

Ethics approval and consent to participate

This study was approved by the French regulatory authorities (the "Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le Domaine de la Santé" and the "Commission Nationale Informatique et Libertés", authorisation number 907303). According to French law, patients were informed of their data registration and given the right to deny access or to rectify their personal data. The informed consent was verbal as no biomedical intervention was performed.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Department of Gastroenterology, Poitiers University Hospital, 2 rue de la Milétrie, 86000 Poitiers Cedex, France. [2]Poitou-Charentes General Cancer Registry, Poitiers University Hospital, University of Poitiers, Poitiers, France. [3]INSERM, CIC 1402, Poitiers, France. [4]Department of Cancer Biology, Poitiers University Hospital, Poitiers, France. [5]Laboratory Inflammation, Tissus Epithéliaux et Cytokines, EA 4331, University of Poitiers, Poitiers, France.

Received: 14 June 2017 Accepted: 6 November 2017
Published online: 14 November 2017

References

1. Globocan. Estimated cancer incidence, mortality and prevalence worldwide in 2012. Lyon: International Agency for Research on Cancer; 2012. Available from: http://globocan.iarc.fr
2. Bos JL. The ras gene family and human carcinogenesis. Mutat Res. 1989;195:255–71.
3. Di Fiore F, Blanchard F, Charbonnier F, Le Pessot F, Lamy A, Galais MP, et al. Clinical relevance of KRAS mutation detection in metastatic colorectal cancer treated by Cetuximab plus chemotherapy. Br J Cancer. 2007;96:1166–9.
4. Bokemeyer C, Bondarenko I, Hartmann JT, de Braud F, Schuch G, Zubel A, et al. Efficacy according to biomarker status of cetuximab plus FOLFOX-4 as first-line treatment for metastatic colorectal cancer: the OPUS study. Ann Onco. 2011;22:1535–46.
5. Cantwell-Dorris ER, O'Leary JJ, Sheils OM. BRAFV600E: implications for carcinogenesis and molecular therapy. Mol Cancer Ther. 2011;10:385–94.
6. Weisenberger DJ, Siegmund KD, Campan M, Young J, Long TI, Faasse MA, et al. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat Genet. 2006;38:787–93.
7. Li WQ, Kawakami K, Ruszkiewicz A, Bennett G, Moore J, Iacopetta B. BRAF mutations are associated with distinctive clinical, pathological and molecular features of colorectal cancer independently of microsatellite instability status. Mol Cancer. 2006;5:2.
8. Rajagopalan H, Bardelli A, Lengauer C, Kinzler KW, Vogelstein B, Velculescu VE. Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status. Nature. 2002;418:934.
9. Lièvre A, Artru P, Guiu M, Laurent-Puig P, Merlin JL, Sabourin JC, et al. The KRAS mutation detection within the initial management of patients with metastatic colorectal cancer: a status report in France in 2011. Eur J Cancer. 2013;49:2126–33.

10. Cardiello F, Tejpar S, Normanno N, Mercadante D, Teague T, Wohlschlegel B, et al. Uptake of KRAS mutation testing in patients with metastatic colorectal cancer in Europe, Latin America and Asia. Target Oncol. 2011;6:133–45.
11. Chafton ME, Karlitz JJ, Schlichting JA, Chen VW, Lynch CF. Factors associated with guideline-recommended KRAS testing in colorectal cancer patients: a population-based study. Am J Clin Oncol. 2017;40(5):498–506.
12. Baldus SE, Schaefer KL, Engers R, Hartleb D, Stoecklein NH, Gabbert HE. Prevalence and heterogeneity of KRAS, BRAF, and PIK3CA mutations in primary colorectal adenocarcinomas and their corresponding metastases. Clin Cancer Res. 2010;16:790–9.
13. Tol J, Nagtegaal ID, Punt CJ. BRAF mutation in metastatic colorectal cancer. N Engl J Med. 2009;361:98–9.
14. Jung SB, Lee HI, Oh HK, Shin IH, Jeon CH. Clinico-pathologic parameters for prediction of microsatellite instability in colorectal cancer. Cancer Res Treat. 2012;44:179–86.
15. Sinicrope FA, Rego RL, Halling KC, Foster N, Sargent DJ, La Plant B, et al. Prognostic impact of microsatellite instability and DNA ploidy in human colon carcinoma patients. Gastroenterology. 2006;131:729–37.

3.3. Réutilisation des données PMSI et RCP (dans le contexte du SI-RGCPC)

Un troisième domaine d'application plus vaste a concerné la prise en charge des cancers et les déterminants de la qualité des soins : surveillance des délais d'accès aux traitements, exhaustivité de passage en réunion de concertation pluridisciplinaire (RCP), analyse des écarts aux recommandations. Ces évaluations se sont fondées sur l'exploitation des données produites en routine par le registre (modélisation du parcours de soins), complétées selon les cas par des retours aux dossiers médicaux spécifiques de l'objectif mis en œuvre.

3.3.1. Surveillance des délais d'accès aux traitements [40]

La mesure des délais est un élément essentiel de la qualité des soins et un traceur potentiel des inégalités d'accès aux soins. Malgré l'importance accordée à la question, élément central du Plan Cancer 2009-2013, peu d'études s'étaient intéressées à une approche évaluative globale en raison des difficultés à accéder à une information suffisamment complète.

Une étude importante avait été conduite dans 8 régions françaises conjointement par les réseaux de cancérologie et la fédération des observatoires régionaux de la santé (dont Poitou-Charentes), intitulée « *Étude sur les délais de prise en charge des cancers du sein et du poumon dans plusieurs régions de France en 2011* » [41]. Cette étude montrait que les délais d'accès au traitement variaient fortement selon les zones géographiques et les catégories socioéconomiques des patients. Aussi, cette même étude mettait en évidence une grande hétérogénéité des pratiques de recueil et de mise en commun des informations décrivant les prises en charge et les étapes clés du parcours de soins selon les fiches de RCP et les éléments du dossier médical. La méthodologie de cette étude, bien qu'elle apparaissait pertinente pour mesurer ces délais, s'était heurtée à des difficultés de mise en œuvre qui rendaient une application en routine difficile. Le rapport concluait à la nécessité de réfléchir à un système permettant de recueillir de façon précise les dates des évènements clés du parcours de soins, avec comme finalité de pouvoir évaluer de façon récurrente l'évolution de ces marqueurs.

Une autre étude, conduite sous l'égide du réseau Francim et intitulée « *Du diagnostic au premier traitement : délais de prise en charge des cancers enregistrés par les registres spécialisés du réseau Francim 1999-2008* » [42] avait permis de documenter précisément des variations dans les délais de prise en charge à partir des données du registre spécialisé des cancers du sein et des cancers gynécologiques de Côte d'Or. Cette approche présentait l'avantage de s'affranchir de l'exhaustivité de passage en RCP pour la sélection de la population d'étude et de reposer la production de ces indicateurs sur le savoir-faire et la connaissance fine des problématiques territoriales par les équipes des registres. En contrepartie, il n'apparaissait pas envisageable de généraliser ces indicateurs sans complément d'enquête systématique.

Le présent projet avait alors pour objectif d'analyser les parcours de soins sur un large échantillon de patients, et d'en déduire les délais d'accès aux traitements en utilisant l'algorithme de représentation des trajectoires de soins validé [35]. La démarche a été mise en œuvre sur une population représentative de 1 082 patientes atteintes de cancer du sein non métastatique diagnostiquées entre 2008 et 2010. L'originalité de cette étude résidait dans l'exploitation et la réutilisation secondaire des données rassemblées et reliées par les opérateurs du registre au moment de l'enregistrement des cas incidents de cancer. Le processus de collecte avait été enrichi pour les besoins de l'étude par les données individuelles d'activité de radiothérapie des centres de statut libéral, qui ne faisaient pas l'objet à l'époque d'une collecte structurée par le RGCPC. Un retour au dossier médical était réalisé pour documenter les cas inhabituels ou non conformes aux recommandations et compléter les trajectoires de soins (évènements traceurs et variables explicatives manquantes).

L'algorithme modélisait ainsi le parcours pour chaque patiente, selon une granularité définie à partir des référentiels HAS-INCa publiés. La séquence débutait à la date de diagnostic et était borné à 1 an afin de restreindre la séquence à la prise en charge initiale des patientes. Cette borne était modifiable afin de tenir compte des cas où la prise en charge initiale était supérieure à un an, ou faisait référence le cas échéant à une récidive ou une métastase précoce qui devait être ignorée. Chaque séquence était ensuite représentée sous la forme d'une chaîne de caractères ordonnés en vue d'en déduire automatiquement le calcul des délais. Les trajectoires réelles des patientes ont été alors confrontées aux parcours recommandés selon les référentiels nationaux et aux recommandations émises en RCP, et les limites de la méthode en termes de disponibilité et qualité de l'information ont été analysées.

Les résultats ont révélé une proportion de données manquantes très faibles (2,7% pour les évènements traceurs et 2,2% pour les variables explicatives), suggérant de pouvoir réduire de façon significative la part des retours aux dossiers médicaux systématique. La fréquente disponibilité des RCP pour cette localisation (> à 92%) et le contrôle individuel entrepris sur chaque dossier au sein du SI-RGCPC (via la confrontation de l'information au moment de la validation par les opérateurs du registre) étaient des éléments garants de la qualité des données analysées.

Cette étude a ainsi permis d'illustrer la prédisposition du registre à produire une évaluation chiffrée des délais d'accès au traitement, sans biais de recrutement et pour un coût marginal limité en routine. La généralisation de cette approche, basée sur des données classiquement utilisées par les registres des cancers en France, pourrait permettre de répondre plus facilement aux besoins exprimés par les institutions et les professionnels de santé en attente de ce type d'évaluation.

Lire Article 6 *: Quillet A, Defossez G, Ingrand P. Surveillance of waiting times for access to treatment: a registry-based computed approach in breast cancer care. Eur J Cancer Care (Engl). 2016 Sep;25(5):764-73. doi: 10.1111/ecc.12362. Epub 2015 Jul 30. PMID: 26223961.*

Accéder à la suite « Exhaustivité de passage en RCP »

European Journal of Cancer Care

Original Article

Surveillance of waiting times for access to treatment: a registry-based computed approach in breast cancer care

A. QUILLET, MD, RESEARCHER, *Registre général des cancers de Poitou-Charentes, Centre Hospitalier Universitaire de Poitiers, Université de Poitiers, Poitiers,* G. DEFOSSEZ, MD, RESEARCHER, *Registre général des cancers de Poitou-Charentes, Centre Hospitalier Universitaire de Poitiers, Université de Poitiers, Poitiers,* & P. INGRAND, MD, PHD, RESEARCHER, *Registre général des cancers de Poitou-Charentes, Centre Hospitalier Universitaire de Poitiers, Université de Poitiers, Poitiers and INSERM, CIC 1402, Poitiers, France*

QUILLET A., DEFOSSEZ G. & INGRAND P. (2015) *European Journal of Cancer Care*
Surveillance of waiting times for access to treatment: a registry-based computed approach in breast cancer care

The current study set out to automatically generate waiting times for access to surgery, chemotherapy and radiotherapy, and to analyse their determinants for non-metastatic breast cancer patients. We used data from the Poitou-Charentes regional cancer registry of women diagnosed with stages I–III breast carcinoma between 2008 and 2010. Waiting times were automatically computed from a previously validated algorithm modelling the care trajectory and then compared with national guidelines. The population of this study included 1082 patients. The compliance with guidelines ranged from 52.4% (access to adjuvant chemotherapy) to 89.2% (access to adjuvant radiotherapy). Younger age, a higher TNM stage, a lower grade, having a triple negative tumour, being the subject of multidisciplinary meetings and being a patient at a public hospital were associated with longer waiting times. The main result was the significant heterogeneity between geographical areas of treatment for all waiting times studied. The original, reproducible use of a registry-based automated algorithm to generate waiting times will help to follow these indicators routinely and efficiently.

Keywords: breast cancer, waiting times, automatic data processing, cancer registry.

INTRODUCTION

The quality of care for patients with cancer is conditioned by appropriate therapy delivery and optimal waiting times to ensure the best possible responses. In patients with breast cancer, the leading cause of cancer death in women (Malvezzi *et al.* 2014), delay in treatment is associated with increased mortality (Hershman *et al.* 2006; McLaughlin *et al.* 2012; Shin *et al.* 2013; Downing *et al.* 2014; Gagliato Dde *et al.* 2014). To meet of this major public health issue, controlling waiting times has been defined as a health policy priority in many countries (Cancer care Ontario, 2010; England Department of Health, Public Health England and NHS England 2013; New Zealand Ministry of Health, 2013; République Française 2014).

The monitoring of these waiting times is needed to highlight the associated determinants (Bouche *et al.* 2010; Institut National du Cancer, 2012; Mosunjac *et al.* 2012; Molinie *et al.* 2013; Plotogea *et al.* 2013; Vandergrift *et al.* 2013). During their care pathways, these patients undergo many diagnostic examinations and

Correspondence address: Defossez Gautier, Unité d'épidémiologie, biostatistique et registre des cancers de Poitou-Charentes, Centre Hospitalier Universitaire de Poitiers, Faculté de médecine, Université de Poitiers, 6, rue de la milétrie, TSA 51115 – 86073 POITIERS Cedex 9, France (e-mail: gautier.defossez@univ-poitiers.fr).

Accepted 29 June 2015
DOI: 10.1111/ecc.12362

European Journal of Cancer Care, 2015

therapeutic care which depend on tumour and patient characteristics. These different stages in the care pathway involve many points of contact with the health information system. To generate waiting times, few tools are able to integrate and mobilise all the information on care pathways distributed across many different health facilities and stored in heterogeneous formats. The monitoring of waiting times currently requires the implementation of specific surveys. The need to develop efficient methods has been pointed to by the French cancer plan (Institut National du Cancer, 2012).

During the creation of the Poitou-Charentes regional cancer registry, there was reflection and work on algorithmic developments to improve the recording capacity and efficiency of the registry (Jouhet *et al.* 2012, 2013). The cancer registry thus created a multi-source information system centred on the patient by providing an automated notification of incident tumours and a temporal representation of the care trajectory for patients from the available data. A computer algorithm producing temporal representations of breast cancer care trajectories was validated on a sample of 159 patients (Defossez *et al.* 2014). This study showed that 98% of these trajectories were correctly ordered and 94% of waiting times were accurately estimated. The aim of this study was to evaluate waiting times for access to surgery, chemotherapy and radiotherapy using a computer algorithm applied to data that is routinely integrated by a population cancer registry, and to analyse their determinants in non-metastatic breast cancer patients.

METHODS

Patients

In compliance with national and international recommendations, the general cancer registry of the Poitou-Charentes region (1.8 million inhabitants, south-western France) includes all incident cases of malignant tumours involving subjects regularly residing in the Poitou-Charentes region at the time of diagnosis, whatever the location of care. We randomly selected 1350 primary invasive breast carcinomas diagnosed between 1 January 2008 and 31 December 2010 in patients under 80 years of age from the cancer registry database. This sample amounted to 35% of eligible patients. After a review of medical records, metastatic tumours, male patients and patients concomitantly treated for another cancer or receiving all treatment outside of the region were excluded. This study was approved by the ethics committee of the French regulatory authorities in accordance with the provisions of the Declaration of Helsinki (the Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le Domaine de la Santé and the Commission Nationale Informatique et Libertés, authorization number 907303).

Recommendations

The French breast cancer guidelines (Haute Autorité de Santé, Institut National du Cancer, 2010) that are applicable to this study population specify the ordering of diagnostic and therapeutic acts to be carried out according to the severity and prognostic criteria of patients. The French guidelines are similar to the international guidelines (Senkus *et al.* 2013) and provide recommendations in terms of waiting times before access to treatments (Fig. 1). Thus, radiotherapy should be initiated within 12 weeks of surgery in patients not treated with adjuvant chemotherapy or within 6 months in patients treated with adjuvant chemotherapy. Chemotherapy should be initiated within 6 weeks after surgery. Finally, there is no specific recommended waiting time between biopsy and surgery. The evaluation of these waiting times should be carried out conditionally to the care pathway. We did not include patients with neoadjuvant treatment in the analysis because their small numbers would not have provided sufficient statistical power.

Data

The data used in this study were derived from the Poitou-Charentes multisource system cancer registry. This database is routinely augmented by establishments and organisations regularly caring for breast cancer patients living in Poitou-Charentes, including anatomical–pathological laboratories, hospital discharge reports from the French medical information programme, healthcare insurance services, cancer care centres and multidisciplinary oncology meetings.

The explanatory variables included the tumour characteristics (TNM stage, SBR grade, hormonal and HER2 status and age at diagnosis), care pathway characteristics (additional biopsy, secondary surgery and multidisciplinary meetings), and geographical and temporal determinants (geographical area of the patient's residence, geographical area of treatment, public vs. private hospital and year of diagnosis).

Automated computing of waiting times

In this study, we evaluated waiting times between biopsy and surgery (BS), between surgery and radiotherapy with adjuvant chemotherapy (SCR) or not (SR), and between

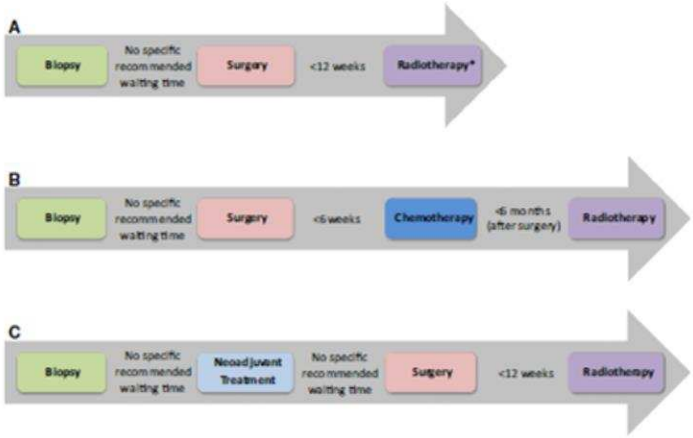


Figure 1. French guidelines for health care sequences for non-metastatic breast cancer. (a) In patients with good prognosis (excepting patients undergoing total mastectomy and with no relapse factors); (b) in patients with poor prognostic factors; (c) in patients with voluminous, infiltrating and/or inflammatory cancer.

surgery and adjuvant chemotherapy (SC) in patients without systemic neoadjuvant treatment. For each patient, 1–3 waiting times could therefore be computed.

To compute these waiting times, four tracer events corresponding to key steps in the management of patients with breast cancer were identified: biopsy (including fine needle aspiration), surgery, chemotherapy and radiotherapy. These tracer events were to be present and dated to the day in the data sources of the cancer registry. In cases of missing data, a systematic review of the medical records was performed to supplement this information.

After this control phase, we applied a previously validated algorithm (Defossez *et al.* 2014) to automatically generate a temporal linear representation of the care trajectories for each patient in the form of an ordered string of characters.

Every character in this representation corresponds to a day in care (B: biopsy, S: surgery, C: chemotherapy, R: radiotherapy and -: no tracer event). Waiting times were automatically computed from this representation by counting the number of characters between tracer events.

Statistical analysis

For each of the four waiting times studied, two types of analysis were performed. First, we performed logistic regressions to assess the conformity of the waiting times with the guidelines as the binary outcome. In the absence of specific recommendations for the BS waiting time, the median value of the sample was used as the threshold. The second approach was performed using analysis of variance on the logarithm of waiting times. A backward elimination procedure was performed to select variables based on the log-likelihood ratio between the models. The significance threshold to include variables in the models was set at 10%. Control for confounding factors and collinearity was performed at each stage in the modelling. The search for interactions was performed on the final models.

To help in the interpretation of the logarithmic parameter estimates, the percentage (P) corresponding to an increase (positive values) or reduction (negative values) in waiting times was calculated using the following formula: $P = (10^{\beta} - 1) \times 100$ (where β is the estimate of the parameter in the logarithmic model of waiting time). From this P, the corresponding number of days (D) was calculated on the basis of the median waiting time in the sample, using the following formula: $D = P \times M$ (where M is the median in the sample).

Data management and statistical analyses were performed using SAS software version 9.3 (SAS Institute, Cary, NC, USA).

RESULTS

In all, 1082 patients were included in this study. Among these patients, 91 received systemic neoadjuvant treatment (83 by chemotherapy and 8 by hormone therapy). For 17 patients, no waiting time could be computed because of undated tracer events or unusual care pathways.

Finally, 974 operated patients were included in the analysis of waiting times (Fig. 2). Among these patients, 54 (5.5%) were operated on without a prior positive biopsy (i.e. with prior negative biopsy, extemporaneous histological examination or no biopsy), 887 (91.1%) received adjuvant radiotherapy (among whom 407 also received adjuvant chemotherapy) and 429 (44.0%) received adjuvant chemotherapy. BS waiting time was computed for 893 patients (there were 27 undated biopsies) with a median of 32 days (IQR: 23–42 days). SR waiting time was computed for 480 patients with a median of 57 days (IQR: 47–72 days) and it complied with the guidelines in 89.2% of cases. SCR waiting time was computed for 405 patients (two undated radiotherapies) with a median of 179 days (range: 167–193 days) and it complied with the guidelines in 56.3% of cases. Finally, SC waiting time was computed for 426 patients (three undated chemotherapies) with a median of 42 days (IQR: 35–52 days) and it complied with the guidelines in 52.4% of cases. The medians of the BS, SR, SCR and SC waiting times were described for each of the determinants (Table 1).

Multivariate analyses were performed on the BS, SR, SCR and SC waiting times (Table 2). The geographical area of the patient's residence was not retained in the final models because of collinearity with the geographical area of treatment. A longer BS waiting time was associated with lower tumour grade (+3.8 days for tumour grade I vs. grade II and +5.5 days for tumour grade I vs. grade III, $P < 0.001$), younger age (+2.4 days, $P = 0.025$), additional biopsy (+8.4 days, $P = 0.022$) and being reviewed in a multidisciplinary meeting (+3.1 days, $P = 0.009$). Because of a significant interaction between the geographical area of surgical treatment and public vs. private surgery ($P = 0.002$), BS results are presented separately for the public and private sectors. Thus, an association between BS waiting time and the geographical area of surgery in both public ($P < 0.001$) and private ($P = 0.001$) facilities was found. A longer BS waiting time was found for patients treated outside the region (+8.0 days in public hospital and +14.2 days in private hospital).

SR waiting time was less compliant with the guidelines among patients with secondary surgery ($P < 0.001$) and those diagnosed in 2008 ($P = 0.019$). The analysis of variance showed an increase in SR waiting time for patients with larger tumour size (+6.5 days for 2–5 cm tumour vs. ≤2 cm tumour and +5.5 days for >5 cm tumour vs. ≤2 cm tumour, $P = 0.015$), secondary surgery (+27.7 days, $P < 0.001$) and diagnosis in 2008 (+4.5 days vs. 2010,

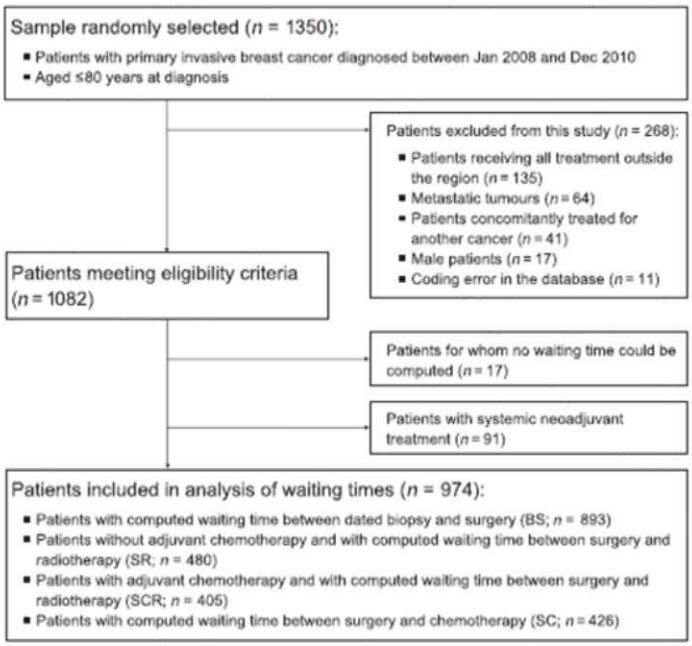


Figure 2. Study flow chart.

Table 1. BS, SR, SCR and SC waiting times based on tumour, care pathways, geographical and temporal determinants

	BS waiting time [days], n = 893		SR waiting time [days], n = 480		SCR waiting time [days], n = 405		SC waiting time [days], n = 426	
	n	Median [q1–q3]	n	Median [q1–q3]	n	Median [q1–q3]	n	Median [q1–q3]
Tumour characteristics								
Tumour size [cm]								
≤2	661	32 [24–42]	431	56 [46–71]	228	178 [166–195]	237	42 [35–54]
>2 and ≤5	198	32 [22–40]	41	64 [53–77]	150	180 [168–189]	161	41 [35–52]
>5	30	33 [22–43]	7	71 [59–98]	22	180 [169–194]	23	42 [35–55]
TNM stage*								
I	514	32 [24–42]	403	56 [46–71]	101	173 [148–184]	110	41 [36–49]
II	286	32 [22–41]	61	66 [53–78]	214	179 [170–194]	223	42 [35–54]
III	82	32 [25–39]	11	71 [65–83]	83	185 [168–199]	86	42 [34–54]
SBR grade*								
I	215	34 [26–43]	181	56 [47–69]	34	180 [169–196]	36	48 [36–56]
II	514	32 [22–41]	267	57 [46–72]	240	180 [166–193]	254	42 [35–54]
III	134	29 [22–39]	13	67 [53–69]	124	175 [166–187]	129	40 [34–49]
Triple negative tumour*								
Yes	71	28 [21–41]	11	71 [60–87]	62	179 [167–193]	64	43 [35–50]
No	726	32 [23–42]	403	56 [46–71]	323	177 [166–188]	341	42 [35–52]
Age at diagnosis [years]								
<65	603	30 [23–41]	276	57 [46–73]	325	178 [166–191]	338	42 [35–51]
≥65	290	34 [25–42]	204	58 [48–72]	80	182 [170–197]	88	44 [35–55]
Care pathway characteristics								
Additional biopsy*								
Yes	19	39 [28–54]	5	54 [41–66]	10	167 [163–180]	12	36 [28–53]
No	874	32 [23–42]	464	57 [47–72]	380	179 [167–193]	398	42 [35–52]
Secondary surgery								
Yes	126	30 [23–41]	57	82 [71–97]	83	188 [178–211]	84	50 [40–60]
No	766	32 [23–42]	422	56 [46–68]	322	175 [166–188]	342	41 [35–50]
Multidisciplinary meeting*,†								
Yes	188	33 [25–43]	443	59 [47–72]	376	179 [167]	382	42 [35–53]
No	705	32 [23–41]	37	52 [40–70]	29	177 [168–187]	44	41 [34–50]
Geographical and temporal characteristics								
Geographical area of patient's residence								
Vienne	224	32 [22–41]	113	68 [61–82]	104	187 [175–200]	105	45 [38–56]
Charente	171	31 [23–43]	90	63 [52–76]	73	175 [165–188]	82	39 [34–50]
Charente-Maritime North	194	32 [23–39]	102	46 [39–55]	98	174 [160–186]	99	39 [33–50]
Charente-Maritime South	140	32 [23–43]	86	48 [41–62]	58	170 [160–184]	60	39 [34–51]
Deux-Sèvres	164	34 [25–46]	89	62 [54–75]	72	180 [170–194]	80	45 [38–54]
Geographical area of treatment*,‡								
Vienne	245	32 [22–40]	136	68 [59–82]	126	187 [176–201]	125	48 [38–56]
Charente	147	30 [22–41]	82	63 [53–76]	61	175 [167–188]	64	39 [35–50]
Charente-Maritime North	196	32 [23–39]	124	44 [38–55]	110	173 [160–184]	106	38 [33–50]
Charente-Maritime South	83	28 [23–39]	69	51 [42–67]	52	172 [157–184]	49	42 [35–52]
Deux-Sèvres	130	34 [24–42]	69	62 [53–72	55	177 [167–190]	54	45 [35–54]
Outside the region	88	46 [33–56]	–	–	–	–	28	38 [34–43]
Hospital where surgery was performed*								
Public	375	36 [27–48]	199	60 [48–75]	176	182 [169–197]	184	45 [36–56]
Private	514	29 [21–37]	279	56 [46–71]	228	175 [166–187]	240	41 [35–50]
Year of diagnosis								
2008	290	32 [24–41]	163	59 [48–75]	132	179 [166–196]	142	43 [36–54]
2009	305	32 [23–41]	147	61 [47–75]	158	180 [167–191]	162	42 [35–53]
2010	298	32 [23–42]	170	56 [46–68]	115	178 [167–189]	122	40 [35–49]

BS, waiting time between biopsy and surgery; SC, waiting time between surgery and adjuvant chemotherapy; SCR, waiting time between surgery and radiotherapy with adjuvant chemotherapy; SR, waiting time between surgery and radiotherapy without adjuvant chemotherapy.

*Missing data are not presented in this table. They amounted to ≤1% for all variables apart from tumour grade [3%] and triple negative tumour [11%].

†The presence of a multidisciplinary meeting between biopsy and surgery for BS waiting time, between surgery and radiotherapy for SR and SCR waiting times and between surgery and chemotherapy for SC waiting time.

‡Geographical area of surgery for BS waiting time, geographical area of radiotherapy for SR and SCR waiting times and geographical area of chemotherapy for SC waiting time.

Table 2. Multivariate analysis of the compliance of BS, SR, SCR and SC with the waiting –time guidelines, and on the logarithm of BS, SR, SCR and SC waiting times

	Compliance of waiting times with French guidelines		Logarithm of waiting time			
	OR	95% CI	β	SE	P (%)*	D (days)†
Waiting time between biopsy and surgery without neoadjuvant treatment (BS)	*Threshold = 32 days (n = 859)*		*Median = 32 days (n = 859)*			
SBR grade						
II vs. I	1.92	1.33–2.76	−0.054	0.016	−11.7	−3.8
III vs. I	2.68	1.63–4.40	−0.082	0.021	−17.1	−5.5
Age (<65 years vs. ≥65 years)	0.47	0.34–0.65	0.031	0.014	7.5	2.4
Additional biopsy (Yes vs. No)	0.23	0.08–0.67	0.101	0.044	26.1	8.4
Multidisciplinary meeting in BS waiting time (No vs. Yes)	–	–	−0.045	0.017	−9.8	−3.1
Geographical area of surgery in public hospital						
Charente vs. Vienne	1.36	0.51–3.60	−0.033	0.042	−7.4	−2.4
Charente-Maritime North vs. Vienne	2.38	1.12–5.08	−0.082	0.035	−17.2	−5.5
Charente-Maritime South vs. Vienne	5.94	2.62–13.45	−0.118	0.036	−23.9	−7.6
Deux-Sèvres vs. Vienne	2.81	1.48–5.35	−0.076	0.029	−16.0	−5.1
Outside the region vs. Vienne	0.52	0.25–1.06	0.097	0.029	25.0	8.0
Geographical area of surgery in private hospital						
Charente vs. Vienne	0.44	0.24–0.78	0.048	0.024	11.7	3.8
Charente-Maritime North vs. Vienne	0.33	0.19–0.57	0.086	0.023	21.9	7.0
Charente-Maritime South vs. Vienne	0.46	0.21–1.00	0.020	0.034	4.8	1.5
Deux-Sèvres vs. Vienne	0.22	0.11–0.46	0.090	0.031	22.9	7.3
Outside the region vs. Vienne	0.20	0.05–0.76	0.160	0.062	44.4	14.2
Waiting time between surgery and radiotherapy without adjuvant chemotherapy (SR)	*Threshold = 84 days (n = 477)*		*Median = 57 days (n = 478)*			
Tumour size (cm)						
>2 and ≤5 vs. ≤2	–	–	0.047	0.017	11.5	6.5
>5 vs. ≤2	–	–	0.040	0.040	9.6	5.5
Secondary surgery (Yes vs. No)	0.07	0.03–0.16	0.172	0.015	48.6	27.7
Geographical area of radiotherapy						
Charente vs. Vienne	0.80	0.35–1.81	−0.027	0.015	−5.9	−3.4
Charente-Maritime North vs. Vienne	8.28	2.51–27.34	−0.167	0.013	−31.9	−18.2
Charente-Maritime South vs. Vienne	10.64	2.25–50.31	−0.104	0.015	−21.2	−12.1
Deux-Sèvres vs. Vienne	2.41	0.85–6.82	−0.036	0.015	−8.0	−4.6
Type of hospital where surgery was performed (Private vs. Public)	1.46	1.00–3.80	–	–	–	–
Year of diagnosis						
2009 vs. 2008	1.22	0.56–2.67	0.011	0.012	2.6	1.5
2010 vs. 2008	3.27	1.40–7.61	−0.035	0.011	−7.8	−4.5
Waiting time between Surgery and radiotherapy with adjuvant chemotherapy (SCR)	*Threshold = 180 days (n = 396)*		*Median = 179 days (n = 377)*			
TNM stage						
II vs. I	0.68	0.40–1.17	0.027	0.008	6.4	11.5
III vs. I	0.41	0.21–0.80	0.037	0.009	8.8	15.8
Triple negative tumour (Yes vs. No)	–	–	0.024	0.009	5.7	10.3
Age (<65 years vs. ≥65 years)	0.52	0.30–0.91	–	–	–	–
Secondary surgery (Yes vs. No)	0.30	0.17–0.53	0.038	0.008	9.2	16.4
Geographical area of radiotherapy						
Charente vs. Vienne	2.44	1.24–4.78	−0.027	0.010	−5.9	−10.6
Charente-Maritime North vs. Vienne	3.51	1.95–6.33	−0.050	0.009	−11.0	−19.6
Charente-Maritime South vs. Vienne	3.46	1.66–7.23	−0.044	0.011	−9.5	−17.1
Deux-Sèvres vs. Vienne	2.42	1.21–4.85	−0.029	0.011	−6.4	−11.5
Type of hospital where the surgery was performed (Private vs. Public)	2.04	1.29–3.22	−0.014	0.007	−3.1	−5.6

Table 2. *Continued*

	Compliance of waiting times with French guidelines		Logarithm of waiting time			
	OR	95% CI	β	SE	P (%)*	D (days)†
Waiting time between surgery and chemotherapy (SC)	*Threshold = 42 days (n = 417)*		*Median = 42 days (n = 417)*			
SBR grade						
II vs. I	1.41	0.65–3.04	−0.010	0.024	−2.2	−0.9
III vs. I	2.29	1.01–5.20	−0.048	0.026	−10.5	−4.4
Secondary surgery (Yes vs. No)	0.34	0.20–0.59	0.055	0.017	13.6	5.7
Multidisciplinary meeting between SC waiting time (No vs. Yes)	–	–	−0.045	0.022	−9.8	−4.1
Place where chemotherapy was administered						
Charente vs. Vienne	2.22	1.14–4.31	−0.040	0.022	−8.7	−3.7
Charente-Maritime North vs. Vienne	2.87	1.61–5.14	−0.072	0.019	−15.2	−6.4
Charente-Maritime South vs. Vienne	1.42	0.70–2.89	−0.020	0.023	−4.4	−1.8
Deux-Sèvres vs. Vienne	1.42	0.72–2.81	−0.035	0.022	−7.8	−3.3
Outside the region vs. Vienne	6.86	2.47–19.08	−0.087	0.029	−18.2	−7.7
Type of hospital where the surgery was performed (Private vs. Public)	1.72	1.10–2.69	−0.040	0.014	−8.7	−3.7

OR, odds ratio; CI, confidence interval; SE, standard error.
*Percentage of increase or reduction in waiting time. For example, patients with SBR grade II had a BS adjusted decrease of 11.7% compared with grade I.
†Number of days calculated on the basis of the median value in the sample. For example, BS adjusted waiting time in patients with SBR grade II would decrease by 3.8 days if that of patients with SBR grade I was 32 days.

$P < 0.001$). Significant heterogeneity between the geographical areas where radiotherapy was dispensed was found in both analyses ($P < 0.001$).

SCR waiting time was less compliant with the guidelines among patients with advanced TNM stage ($P = 0.031$), younger patients ($P = 0.022$), patients with secondary surgery ($P < 0.001$) and those treated in a public hospital ($P < 0.001$). The analysis of variance showed an increase in SCR waiting time among patients with advanced tumour stage (+11.5 days for stage II tumour and +15.8 days for stage III tumour, $P < 0.001$), triple negative tumours (+10.3 days, $P = 0.006$), secondary surgery (+16.4 days, $P < 0.001$) and among those treated in a public hospital (+5.6 days, $P = 0.040$). Significant heterogeneity between the geographical areas where radiotherapy was dispensed was found in both types of analysis ($P < 0.001$).

Finally, SC waiting time was less compliant with the guidelines among patients with secondary surgery ($P < 0.001$) and those treated in a public hospital ($P = 0.018$). From the results of the analysis of variance, increased SC waiting time was found for patients with lower tumour grade (+0.9 days for tumour grade I vs. grade II and +4.4 days for tumour grade I vs. grade III, $P = 0.028$), secondary surgery (+5.7 days, $P = 0.001$) and those treated in a public hospital (+3.7 days, $P = 0.006$). Significant heterogeneity between the geographical areas where chemotherapy was administered was found in the fit analysis ($P < 0.001$) and in the analysis of variance ($P = 0.002$).

DISCUSSION

In this study, waiting times to access treatments were automatically computed on a large sample of patients with breast cancer using an algorithm previously validated on data routinely collected by the cancer registry of Poitou-Charentes. The small proportion of initially missing data, for both tracer events (2.7%) and explanatory variables (2.2%), demonstrates the relevance of the use of this algorithm on this database, even without referring back to medical records, as well as reduced costs compared to survey protocols. Routine production of waiting time indicators is thus possible without separately implementing a collection of relevant information in medical records, which would be particularly resource-consuming in this setting. A recent French study presented an overview of the regional levels of waiting times for the various acts and key steps in the trajectories of patients with breast and lung cancer (Institut National du Cancer, 2012). This study underlined the time required to collect data, imposed by the scatter of information for a given care trajectory, and the difficulties linked to the heterogeneity of data collection methods and practices, which makes routine short-term follow-up of waiting times impossible.

The method proposed in this work responds to needs expressed by institutions and health professionals for routine indicators aiming to improve the quality of patient care. The routine production of these indicators will enable regular assessment of the match between care provided and official guidelines, an evaluation of waiting times to access treatment, and a comparison of results with those reported in the international literature. Although the data are initially fragmented across numerous sources in different territories, the information is collated by the Poitou-Charentes general cancer registry in a single base from which the data can be readily extracted and exploited. The crossing of the trajectories produced with certain clinical data will enable routine evaluation of the compliance of observed care provision trajectories with those set out in the guidelines, and an analysis of their determinants.

In the present study, the place of treatment is the main determinant impacting waiting times. This result is consistent with previous studies performed in the same region (Bouche *et al.* 2010), in France (Molinie *et al.* 2013) and worldwide (Plotogea *et al.* 2013; Vandergrift *et al.* 2013), and it highlights inequalities in access to treatment between geographical areas. These variations can be partly explained by internal organisational constraints in health establishments. Indeed, waiting times to access treatment depend on the availability of technical platforms and health professionals, and a lack of these services is generally perceived by patients as an event impeding the cancer care continuum (Bairati *et al.* 2006). We can, therefore, hypothesise that the therapeutic care access is faster in private hospitals because of patient overload in public hospitals. In France, establishments carrying out medical acts in oncology (surgery, chemotherapy or radiotherapy) must have an authorisation issued by the regional health agencies. To reduce waiting times for access to treatment, rather than increasing the number of authorised healthcare facilities, it seems more feasible to encourage collaboration between the different actors in cancer care, particularly between public and private hospitals. However, it should be borne in mind that patients have a role in the choice of place and options of treatment in their cancer care (Bouche *et al.* 2008; Rajan *et al.* 2013). The care trajectory of patients with breast cancer is mainly defined on the basis of prognostic and severity criteria. Taking account of these variables is therefore essential to understand organisational constraints linked to different care pathways. In our study, a longer waiting time between surgery and radiotherapy was found in patients with a more advanced tumour stage. This can be explained by the time required to perform complementary medical investigations in these women, to determine adjuvant chemotherapy indications. Also, when this treatment is complete, the different chemotherapy protocols proposed according to the tumour stage can also have an impact on the waiting time to access radiotherapy. An increase of 3–4 days was also found for patients for whom the medical record was presented in a multidisciplinary meeting. This result is similar to that described in a recent study (Molinie *et al.* 2013) and can be explained by organisational constraints of these meetings, usually held weekly.

Our results should be considered in view of certain limitations. The main limitation of the modelling algorithm used to represent the care trajectory is the need to access reliable and complete data sources (Defossez *et al.* 2014). To overcome this limitation, we systematically controlled medical records to search for undated tracer events. For the analysis of the determinants of access to treatment, the main limitation is the exclusion of patients receiving all their treatment outside the region (less than 10% of patients) on account of a risk of missing tracer events in their care pathways. Therefore, the interpretation of results on waiting times to access treatment outside the region should be taken with caution. In addition, we did not include patients who received neoadjuvant systemic treatment in the analysis. Indeed, this particular care pathway involved too few patients to assess the impact of determinants on waiting times. A complementary study will be implemented on a larger sample of this population to address this issue. Women over 80 years old were excluded from this study. Indeed, these patients usually have a specific care pathway because of the presence of numerous comorbidities and a lower benefit/risk ratio.

To achieve the aim of this study, we used two complementary indicators. The first relates to compliance with the French guidelines, which was a binary approach based on threshold values. To provide a quantitative approach, we used a linear analysis of variance of logarithms of waiting times. This logarithmic conversion was performed to correct the asymmetry of the data distribution as a result of minimum unavoidable waiting times related to organisational constraints and the wide spread of maximum waiting times attributable to individual variability. A study of waiting times to access treatment requires concomitant reflection on the consequences of delays for patients. Studies have shown that delays in access to treatment in breast cancer are associated with a greater risk of death, whether for access to surgery (McLaughlin *et al.* 2012; Shin *et al.* 2013), chemotherapy (Downing *et al.* 2014; Gagliato Dde *et al.* 2014) or radiotherapy (Hershman *et al.* 2006). However, the thresholds used in our study

were more stringent than those used in survival studies. It would be interesting to complete the vital status data of our patients to investigate the impact of these delays in terms of survival.

The main originality of this study is the reliance on standardised and unprocessed data sources that are routinely collected to automatically generate waiting time indicators. This method can be applied by other cancer registries, provided coded data sources using the international classification of disease (ICD-O-3, ICD-10) are available. In this particular form, the algorithm is based on the tracer events specific to breast cancer care pathways. However, it is possible to modify the scheduling and type of tracer event detected by this algorithm to apply it to other localisations, provided the key stages of these care pathways are identified. The results of this study show that the main determinant impacting waiting times is the place of treatment. The identification of determinants of this sort is essential to implement actions to reduce inequalities in access to treatment.

FUNDING

This study was granted by the French National Cancer Institute (Institut National du Cancer).

ACKNOWLEDGEMENTS

We thank the public and private hospitals, the Regional College of Medical Information, the CRISAP association of pathologists, healthcare insurance services, the Regional Health Agency, the Onco-Poitou-Charentes Network and the Cancer League for their valuable contributions to this study.

REFERENCES

Bairati I., Fillion L., Meyer F.A., Héry C. & Larochelle M. (2006) Women's perceptions of events impeding or facilitating the detection, investigation and treatment of breast cancer. *European Journal of Cancer Care (England)* **15**, 183–193.

Bouche G., Migeot V., Mathoulin-Pélissier S., Salamon R. & Ingrand P. (2008) Breast cancer surgery: do all patients want to go to high-volume hospitals? *Surgery* **143**, 699–705.

Bouche G., Ingrand I., Mathoulin-Pelissier S., Ingrand P., Breton-Callu C. & Migeot V. (2010) Determinants of variability in waiting times for radiotherapy in the treatment of breast cancer. *Radiotherapy and Oncology* **97**, 541–547.

Cancer care Ontario (2010) Ontario Cancer Plan 2011–2015. Available at: https://www.cancercare.on.ca/common/pages/UserFile.aspx?fileId=84204. Retrieved July 10, 2015

Defossez G., Rollet A., Dameron O. & Ingrand P. (2014) Temporal representation of care trajectories of cancer patients using data from a regional information system: an application in breast cancer. *BMC Medical Informatics and Decision Making* **14**, 24.

Downing A., Twelves C., Forman D., Lawrence G. & Gilthorpe M.S. (2014) Time to begin adjuvant chemotherapy and survival in breast cancer patients: a retrospective observational study using latent class analysis. *Breast Journal* **20**, 29–36.

England Department of Health, Public Health England and NHS England (2013) Improving Outcomes: A Strategy for Cancer. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/264511/IOSC_3rd_Annual_Report_-_Proof_version_-_9_December_2013_v2.pdf. (Retrieved July 10, 2015)

Gagliato Dde M, Gonzalez-Angulo AM, Lei X, Theriault RL, Giordano SH, Valero V, Hortobagyi GN & Chavez-Macgregor M. Clinical impact of delaying initiation of adjuvant chemotherapy in patients with breast cancer. *Journal of Clinical Oncology* 2014, **32**:735–744.

Haute Autorité de Santé, Institut National du Cancer. Guide – Affection longue durée. Tumeur maligne, affection maligne du tissu lymphatique ou hématopoïétique. Cancer du sein. 2010. Available at: http://www.has-sante.fr/portail/jcms/c_927251/fr/ald-n-30-cancer-du-sein. Retrieved July 10, 2015

Hershman D.L., Wang X., McBride R., Jacobson J.S., Grann V.R. & Neugut A.I. (2006) Delay in initiating adjuvant radiotherapy following breast conservation surgery and its impact on survival. *International Journal of Radiation Oncology Biology Physics* **65**, 1353–1360.

Institut National du Cancer. (2012) Étude sur les délais de prise en charge des cancers du sein et du poumon. Available at: http://www.e-cancer.fr/Expertises-et-publications/Catalogue-des-publications/Etude-sur-les-delais-de-prise-en-charge-des-cancers-du-sein-et-du-poumon. Retrieved July 10, 2015

Jouhet V., Defossez G., Burgun A., le Beux P., Levillain P., Ingrand P. & Claveau V. (2012) Automated classification of free-text pathology reports for registration of incident cases of cancer. *Methods of Information in Medicine* **51**, 242–251.

Jouhet V., Defossez G. & Ingrand P. (2013) Automated selection of relevant information for notification of incident cancer cases within a multisource cancer registry. *Methods of Information in Medicine* **52**, 411–421.

Malvezzi M., Bertuccio P., Levi F., La Vecchia C. & Negri E. (2014) European cancer mortality predictions for the year 2014. *Annals of Oncology* **25**, 1650–1656.

McLaughlin J.M., Anderson R.T., Ferketich A.K., Seiber E.E., Balkrishnan R. & Paskett E.D. (2012) Effect on survival of longer intervals between confirmed diagnosis and treatment initiation among low-income women with breast cancer. *Journal of Clinical Oncology* **30**, 4493–4500.

Molinie F., Leux C., Delafosse P., Ayrault-Piault S., Arveux P., Woronoff A.S., Guizard A.V., Velten M., Ganry O., Bara S., Daubisse-Marliac L. & Tretarre B. (2013) Waiting time disparities in breast cancer diagnosis and treatment: a population-based study in France. *Breast* **22**, 810–816.

Mosunjac M., Park J., Strauss A., Birdsong G., Du V., Rizzo M., Gabram S.G. & Lund M.J. (2012) Time to treatment for patients receiving BCS in a public and a private university hospital in Atlanta. *Breast Journal* **18**, 163–167.

New Zealand Ministry of Health (2013) National Cancer Programme: Work Plan 2012/13. Available at: http://www.health.govt.nz/system/files/documents/publications/national-cancer-programme-work-plan-2012-13.doc. Retrieved July 10, 2015

Plotogea A., Chiarelli A.M., Mirea L., Prummel M.V., Chong N., Shumak R.S., O'Malley F.P. & Holloway C.M.; Breast Screening Study Group. Factors associated with wait times across the breast cancer treatment pathway in Ontario. *SpringerPlus* 2013, **2**:388.

Rajan S., Foreman J., Wallis M.G., Caldas C. & Britton P. (2013) Multidisciplinary decisions in breast cancer: does the patient receive what the team has recommended? *British Journal of Cancer* **108**, 2442–2447.

République Française (2014) Plan Cancer 2014–2019. Available at: http://www.e-cancer.fr/Expertises-et-publications/Catalogue-des-publications/Plan-cancer-2014-2019

Senkus E., Kyriakides S., Penault-Llorca F., Poortmans P., Thompson A., Zackrisson S. & Cardoso F.; ESMO Guidelines Working Group. (2013) Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology* **24**(Suppl. 6), vi7–23.

Shin D.W., Cho J., Kim S.Y., Guallar E., Hwang S.S., Cho B., Oh J.H., Jung K.W., Seo H.G. & Park J.H. (2013) Delay to curative surgery greater than 12 weeks is associated with increased mortality in patients with colorectal and breast cancer but not lung or thyroid cancer. *Annals of Surgical Oncology* **20**, 2468–2476.

Vandergrift J.L., Niland J.C., Theriault R.L., Edge S.B., Wong Y.N., Loftus L.S., Breslin T.M., Hudis C.A., Javid S.H., Rugo H.S., Silver S.M., Lepisto E.M. & Weeks J.C. (2013) Time to adjuvant chemotherapy for breast cancer in National Comprehensive Cancer Network institutions. *Journal of the National Cancer Institute* **105**, 104–112.

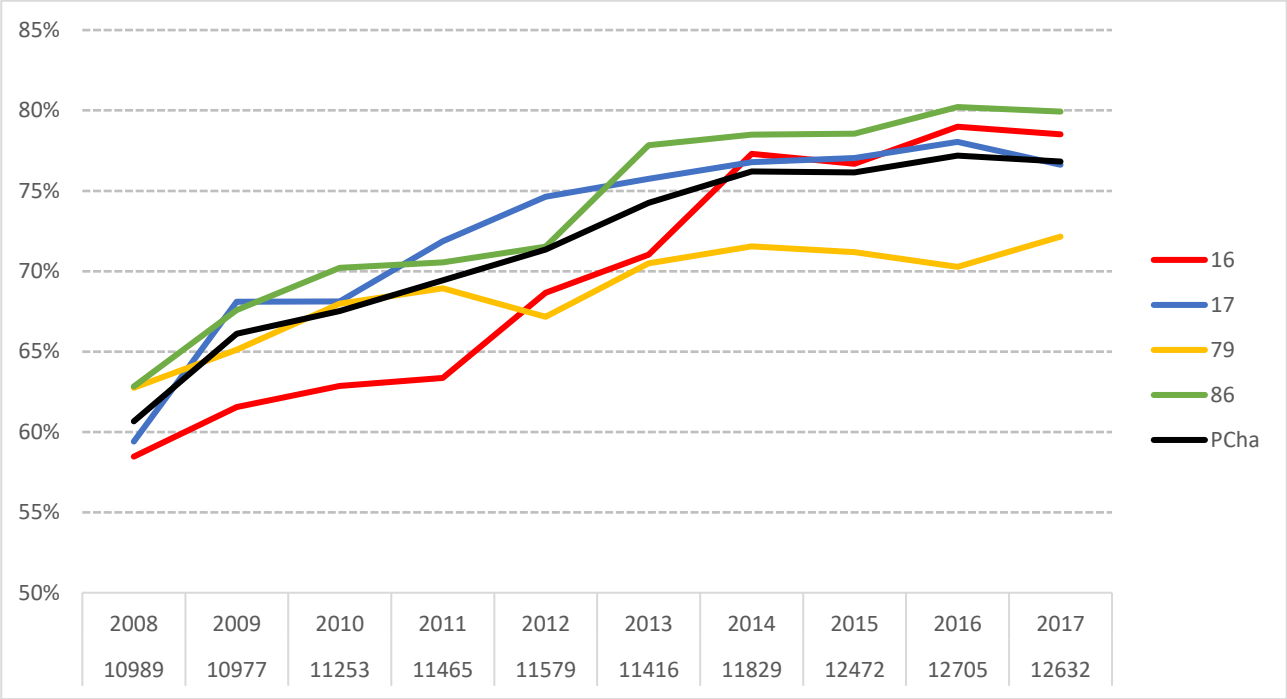
3.3.2. Exhaustivité de passage en RCP des nouveaux patients atteints de cancer

Un autre domaine d'application était celui des discussions pluridisciplinaires en oncologie, mises en œuvre pour répondre aux inégalités dans la pratique clinique, et au poids croissant des preuves suggérant que les patients traités dans des centres de soins spécialisés avaient de meilleurs résultats que ceux qui ne l'étaient pas [43–45]. En France, cette mission a été dévolue à partir de mars 2003 aux réseaux régionaux de cancérologie. Afin de s'assurer de l'accès des patients aux RCP, l'Institut national du cancer (INCa) et la Haute Autorité de santé (HAS) ont développé et généralisé un indicateur permettant d'évaluer le taux d'exhaustivité de passage en RCP, rendu obligatoire depuis septembre 2010 pour l'accréditation des établissements de santé ayant une activité de traitement du cancer [46,47]. Celui-ci est produit à partir de l'analyse d'un échantillon aléatoire annuel de 100 hospitalisations pour la prise en charge initiale d'un cancer. Si en pratique cet indicateur apporte des mesures de qualité intéressantes par établissement, il ne reflète pas en pratique la réalité du contexte de soins multidisciplinaires au niveau de la population. Les RCP examinent des patients de plus d'un hôpital, et certains membres individuels de l'équipe, en particulier les pathologistes et les oncologues, peuvent assister à plusieurs RCP dans différents endroits. La taille des échantillons proposés par l'indicateur est par ailleurs généralement insuffisante pour fournir des estimations par site de cancer.

Dans ce contexte, il était intéressant de chercher à en déduire cet indicateur à partir des données des registres de population. Le calcul du taux d'exhaustivité de passage en RCP au niveau du patient nécessite en effet l'identification de tous les nouveaux cas incidents de cancer – mission dévolue aux registres du cancer, et le suivi des RCP - mission spécifique des réseaux régionaux de cancérologie. Les RCP faisant l'objet d'une collecte systématique pour le fonctionnement de routine du RGCPC, celles-ci sont disponibles au moment de la validation, et sont systématiquement relues puis reliées par les opérateurs du registre aux tumeurs incidentes enregistrées. Cette relation permet alors d'évaluer pour chaque année consolidée la proportion de nouveaux patients atteints de cancer dont il est question lors des RCP pour tous les cas enregistrés (voir Figure 5). Les résultats peuvent être déclinés par âge, sexe, territoire de santé, localisation cancéreuse ou autres déterminants d'intérêt utiles pour justifier la mise en place de mesures correctives. Cette production trouve ainsi tout son intérêt dans l'accompagnement des réseaux de cancérologie et des 3C sur le déploiement et l'efficience des RCP.

Ce travail fait l'objet d'une publication en cours (à soumettre) portant sur les 10 premières années d'enregistrement du RGCPC : *Defossez G et al. Does new cancer patients benefit from systematic multidisciplinary team meeting? The reality of population practices in a French region. Revue ciblée: Int J Quality Health Care.*

Figure 5 : Evolution du taux d'exhaustivité de passage en RCP des nouveaux patients atteints de cancer sur la période 2008-2017, par département et pour l'ensemble de l'ex-région Poitou-Charentes (toutes localisations cancéreuses hors carcinomes cutanés, N= 117 314)



3.3.3. Evènements indésirables graves au décours d'une chimiothérapie [48]

Partant toujours de la modélisation du parcours de soins, la disponibilité d'informations sur la séquence thérapeutique au décours de l'enregistrement par le RGCPC présageait de pouvoir identifier de façon efficiente les chimiothérapies conventionnelles administrées en hospitalisation. Dans ce contexte et celui d'un nombre toujours croissant de malades traités par chimiothérapie, l'objectif a été d'estimer précisément l'incidence des effets indésirables graves (EIG) consécutifs à une chimiothérapie dans un territoire français (l'ex-région Poitou-Charentes, considérée *a priori* comme représentative des autres régions françaises concernant la prescription de chimiothérapie) grâce au SI-RGCPC, puis de chercher à en comprendre la sous-notification (en comparant le nombre de notifications spontanées d'EIG reçues par le Centre Régional de Pharmacovigilance (CRPV) du Poitou-Charentes par rapport au nombre d'EIG recueillis de façon exhaustive dans les dossiers médicaux d'un échantillon représentatif de patients traités par chimiothérapie). En second lieu, l'étude visait à caractériser ces EIG, après analyse et validation par le CRPV, et d'en identifier des déterminants de survenue selon leur typologie, la classe moléculaire, la localisation de la tumeur, le type de lieu d'administration, et les facteurs cliniques et contextuels.

Alors que les données disponibles proviennent d'essais cliniques, l'intérêt de cette étude reposait sur sa méthode d'échantillonnage, en utilisant la couverture d'un registre du cancer à l'échelle d'une région, garantissant ainsi une excellente représentation de la population concernée en vie réelle, ce qui était essentiel à la généralisation des résultats. En pratique, un échantillon représentatif de 1 023 tumeurs incidentes diagnostiquées en 2012 (hors hémopathies et cancers cutanés hors mélanome) a été constitué à partir de la base du SI-RGCPC. Pour chaque tumeur ayant fait l'objet d'une confirmation diagnostique en 2012 (4 000 tumeurs tirées au sort initialement sur l'ensemble de la base), la présence d'un traitement par chimiothérapie était recherchée dans les données sources du RGCPC. Une sélection de base repérait initialement :

i) Les individus traités par chimiothérapie en hospitalisation de jour ou en hospitalisation complète à partir des données PMSI (code CIM-10 Z511 « Séances de chimiothérapie pour tumeur »),
ii) Les individus ayant bénéficiés d'une chimiothérapie en ambulatoire à partir des actes CCAM traceurs de pose de chambre implantable réalisée en hospitalisation (acte CCAM EBLA003 « Pose de cathéter et de système diffuseur implantable sous-cutanée ») ou des codes CIM-10 d'ajustement et d'entretien des dispositifs d'accès vasculaire (code CIM-10 Z452 « Ajustement et entretien de dispositif d'accès vasculaire »).

La mention de traitement par chimiothérapie (IV ou per os) était ensuite recherchée via l'interrogation contextuelle et la consultation des comptes rendus de RCP intégrées dans le SI-RGCPC. En l'absence d'arguments évocateurs dans les données du RGCPC, un retour au dossier médical était mis en œuvre.

Sur la base des 4 000 tumeurs tirées au sort, six pour cent des individus « taggués » à partir des évènements traceurs du RGCPC comme « n'ayant pas été traités par chimiothérapie » ont été récupérés via le retour au dossier médical (à noter que les patients récupérés à partir des RCP n'étaient pas comptabilisés ici). Le motif le plus souvent rencontré était l'administration d'une chimiothérapie orale (dès lors non identifiable à partir des données PMSI). Plus rarement, il pouvait s'agir de la coexistence de tumeurs multiples pour lesquelles la chimiothérapie était à visée des 2 tumeurs et pointait vers l'une plutôt que l'autre dans le SI-RGCPC (une seule relation possible). A noter également que les actes de chimio embolisation (actes CCAM EDLF014-7), destinés à ralentir la progression des tumeurs du foie, n'avaient pas été retenus parmi les actes traceurs initiaux de chimiothérapie. A l'inverse, 15% des individus préalablement taggués comme « ayant reçu une chimiothérapie » ont été finalement exclus de l'échantillon. Le principal motif était un début d'administration d'une chimiothérapie à plus de 6 mois du diagnostic initial pour le traitement d'une récidive ou d'une évolution métastatique (ne faisant donc partie du traitement initial).

Les résultats ont montré que la chimiothérapie était associée à une incidence particulièrement élevée d'EIG, affectant environ 45% des patients. Les taux d'incidence les plus élevés étaient observés lorsque les patients étaient exposés à des inhibiteurs de la topo-isomérase II tels que l'étoposide et la bléomycine (69%), les vinca-alcaloïdes (67%), la topo-isomérase I (55 %) et les dérivés du platine (52 %). Le contexte clinique était corrélé à des taux d'incidence élevés, surtout en cas de métastases (53 %) et de comorbidités (51 %). Des différences substantielles étaient constatées selon la localisation, avec une incidence particulièrement élevée dans les cancers broncho-pulmonaires (59,0%).

Lire Article 7 : *Ingrand I, Defossez G, Lafay-Chebassier C, Chavant F, Ferru A, Ingrand P, Pérault-Pochat MC. Serious adverse effects occurring after chemotherapy: A general cancer registry-based incidence survey.* ***Br J Clin Pharmacol****. 2020 Apr;86(4):711-722. doi: 10.1111/bcp.14159. Epub 2020 Jan 16. PMID: 31658394; PMCID: PMC7098859.*

Par la suite, une étude ancillaire [49] a évalué le nombre de patients recevant des chimiothérapies à base de fluoropyrimidines en France (principal composant des protocoles de chimiothérapie combinée pour le traitement des tumeurs solides), ainsi que l'incidence réelle des effets indésirables graves liés aux fluoropyrimidines avant l'introduction récente du dépistage obligatoire de la dihydropyrimidine déshydrogénase (DPD) par les autorités sanitaires françaises (la DPD étant l'enzyme limitant le taux de catabolisme du 5-fluorouracile ou 5-FU). Après extrapolation au niveau national, le nombre de patients traités annuellement avec du 5-FU ou de la capecitabine était estimé à 76 200. Le taux d'EIG était de 32,2% sur les 6 premiers mois de traitement et 1,6% des patients ont

eu leur pronostic vital engagée ou une incapacité/invalidité significative liée aux fluoropyrimidines (1200 patients, dont 150 décès). Ces résultats ont permis de montrer l'importante toxicité des fluoropyrimidines, et l'espoir que le dépistage des carences en DPD puisse réduire ces événements iatrogènes et éradiquer les décès.

Barin-Le Guellec C, Lafay-Chebassier C, Ingrand I, Tournamille JF, Boudet A, Lanoue MC, Defossez G, Ingrand P, Perault-Pochat MC, Etienne-Grimaldi MC. Toxicities associated with chemotherapy regimens containing a fluoropyrimidine: A real-life evaluation in France. ***Eur J Cancer****. 2020 Jan;124:37-46. doi: 10.1016/j.ejca.2019.09.028. Epub 2019 Nov 9. PMID: 31715555.*

Une ré analyse des données est en cours sur les dérivés de platine.

Accéder à la suite « Parcours de soins des patients atteints d'hémopathies malignes »

Received: 27 May 2019 | Revised: 3 September 2019 | Accepted: 30 September 2019
DOI: 10.1111/bcp.14159

ORIGINAL ARTICLE

Serious adverse effects occurring after chemotherapy: A general cancer registry-based incidence survey

Isabelle Ingrand[1,2] | Gautier Defossez[1] | Claire Lafay-Chebassier[2,3] | François Chavant[2] | Aurélie Ferru[4] | Pierre Ingrand[1] | Marie-Christine Pérault-Pochat[2,3]

[1]Unité d'Epidémiologie et Biostatistique, Registre Général des Cancers Poitou-Charentes; INSERM CIC 1402; Université; CHU de Poitiers, France
[2]Service de Pharmacologie clinique et Vigilances; Université; CHU de Poitiers, France
[3]INSERM U1084-LNEC/INSERM CIC 1402; Université; CHU de Poitiers, France
[4]Pôle régional de cancérologie; CHU de Poitiers, France

Correspondence
Marie-Christine Pérault-Pochat (principal investigator), Pharmacologie Clinique et Vigilances, CHU – La Vie la Santé, 86021 POITIERS CEDEX, France.
Email: mc.perault-pochat@chu-poitiers.fr

Funding information
Agence Nationale de Sécurité du Médicament et des Produits de Santé, Grant/Award Number: AAP-2013-057; Institut National Du Cancer, Grant/Award Number: INCA AAP-SHS-E-SP-2013

Aims: Pharmaco-epidemiological surveys enable the frequency of serious adverse effects—and also the determining factors of their occurrence and seriousness—to be quantified. Few studies systematically gathering post-chemotherapy adverse effects data have been conducted. The objective was to assess the incidence of post-chemotherapy serious adverse effects on the basis of cancer registry data.

Methods: The population was composed of new invasive cancer cases, with the exception of haematopoietic tumours and cutaneous carcinomas. These cancers were identified in 2012 among patients living at the time of diagnosis in a region covered by a general cancer registry and by a French regional pharmacovigilance centre, and treated with neo-adjuvant and/or adjuvant first-intention chemotherapy, followed or not by radiotherapy. The study was based on a sample of 1000 patients from the registry, followed by the collection of serious adverse effects and the required information to constitute a pharmacovigilance file.

Results: Chemotherapy was associated with a particularly high incidence of serious adverse effects, affecting 44.5% (41.4–47.5%) of the patients. The highest incidence rates were observed when patients were exposed to topo-isomerase II inhibitors such as etoposide and bleomycin (69.2%), vinca-alkaloids (66.7%), topo-isomerase I inhibitors (54.5%) and platinum derivatives (52.0%). The clinical context was also linked to incidence, especially in case of metastases (53.3%) and comorbidities (51.3%). Substantial differences were found according to localisation, with a particularly high incidence in bronchial–pulmonary cancers (59.0%).

Conclusion: The high overall incidence rate of serious adverse effects should motivate a reinforcement of information about drug toxicities and improve knowledge by drawing on patient reporting.

KEYWORDS
chemotherapy, general cancer registry, incidence, serious adverse effects

The authors confirm that the Principal Investigator for this paper is Marie-Christine Pérault-Pochat and that she had direct clinical responsibility for the patients.

1 | INTRODUCTION

In France, the number of new cancer cases has increased by 20% in the last 10 years, from 320 000 in 2005 to 400 000 estimated cases in 2017.[1] The oncological therapeutic strategy is essentially based on local or loco-regional treatment (removal by surgery, radiotherapy) and/or on a systemic administration of chemotherapy or targeted therapy. The expectations for chemotherapy are *curative* (administered alone or as a complement to other so-called adjuvant or neoadjuvant treatments), or *palliative*, either prolonging life or improving quality of life. In the last few years, the number of patients treated with chemotherapy has increased twice as much as the number of new patients.[2] With 41 molecules that were granted marketing authorisation in France between 2004 and 2012, the field of oncology is particularly dynamic in the development and marketing of new drugs. Given the public health challenges, the health authorities ensure that drugs are made available rapidly (short duration of clinical trials, limited number of patients), which can contribute to masking potential adverse events (AEs). In oncology, the incidence of serious AEs (SAEs) is high,[3] with some being specific to chemotherapy. Antineoplastic agents are generally mentioned as being 1 of the first classes of drugs involved in iatrogenic pathologies, as much in terms of morbidity as mortality.[3,4] Some AEs are common to all types of chemotherapy due to their action mechanism, whilst others are specific to certain molecules with toxicity towards certain organs.

In the absence of treatment, cancer has a medium to long-term fatal evolution and the acceptable risk threshold is consequently much higher than in most pathologies. Taking into consideration this risk and the expected benefit of the treatments for the individual is an integral part of the strategy deployed by oncologists. The radiotherapy/chemotherapy combination or the associations of antineoplastic molecules could also expose treated patients to an increased risk of AE occurrences as well as increased risk of seriousness.

The current French pharmacovigilance system is based on health professionals' spontaneous notification of AEs, and notification was made compulsory for SAEs by the French public health code article R5144-19. Generally, substantial AE under-notification has been observed in France, even if the rate of notifications is regarded as 1 of the best in Europe. The frequency of AE notifications in France has been increasing over time. Indeed, after 18,907 notifications in 2002, 82,077 were filed in 2017, including 42,715 serious cases (French national pharmacovigilance database, URL: www.ansm.sante.fr). The frequency of spontaneous AE notification differs with the type of effect, its expected/unexpected nature and its seriousness.[5] In oncology, under-notification seems even higher, considering the relatively few notifications collected by the French regional pharmacovigilance networks. Generally speaking, oncologists do not automatically notify an AE. In oncology practice, AEs in cancer treatments have become almost synonymous with the treatments themselves.[6] It is therefore difficult, when informing patients of their treatment, to give each of them expert information on the level of risk of AE occurrences, their predictability, preventability or severity.

What is already known about this subject

- Knowledge on the incidence of serious adverse effects in chemotherapy is essentially derived from clinical trials.
- There is substantial evidence that clinicians under-report harm, and that there is a selection of patients.
- The results of clinical trials may not apply to patients treated in everyday clinical practice.

What this study adds

- In this study (on >1000 patients), the global incidence rate of serious adverse effects was 44.5% (range, 41.4–47.5%).
- This high percentage should lead to a reinforcement of information about drug toxicities among patients and oncologists.
- Knowledge could be improved by sensitisation of patients and motivation of medical doctors to report drugs toxicities.

Knowledge on incidence of SAEs in chemotherapy essentially comes from clinical trials, with substantial evidence that clinicians under-report harm, and that there is a selection of patients with good functional status in these trials, so that their results may not apply to patients treated in everyday clinical practice.[7] Pharmaco-epidemiology surveys are good tools to quantify the frequency and the context of SAE occurrences, as well as the determining factors of their occurrence and seriousness. However, when these surveys are conducted in the general population, they are confronted with substantial risks of bias, the selection bias in particular.[8] Cancer registries, on account of their exhaustive approach and their rigorous information validation criteria, constitute an optimal database to collect a sample that is representative of the population treated with anticancer chemotherapy.[9]

The objective of this study was to assess the incidence of postchemotherapy SAEs using cancer registry data.

2 | METHODS

2.1 | Patients and study

The study population was composed of new cases of infiltrating cancers, except haematopoietic tumours and cutaneous carcinomas (excepting melanoma), metastatic or not at diagnosis, identified in 2012 among patients living in 4 *départements* in the west of France covered by a general cancer registry[10] at the time of diagnosis and treated with first-intention adjuvant and/or neoadjuvant chemotherapy or radio-chemotherapy. Patients from these areas also come under the regional pharmacovigilance centre of the Poitou-Charentes area.

The study involved 2 tasks:

- identification, in the regional area, of all the new patients with cancer and treated with chemotherapy, associated or not with radiotherapy,
- identification of SAEs and collection of information required for the constitution of a pharmacovigilance file, complete with socio-demographic data.

Task 1. Identification of patients treated with chemotherapy (cancer registry)

Since 1 January 2008, the cancer registry has included any incident case of invasive malignant tumour (malignant haemopathies and solid tumours apart from basal-cell skin carcinomas) among patients residing regularly in the Poitou–Charentes area at the time of diagnosis, whatever the place of treatment. The gathering of source data for each patient (CNIL authorisation N°907303 to obtain nominative data) enables the identification of a given case of cancer from a number of independent data sources. Doctors from the registry conducted an individual validation procedure for every randomly selected tumour in the sample.

Phase 1. Each tumour was systematically checked for all the compulsory items of tumour registration according to international cancer registry standards (FRANCIM, IACR, IARC): date of diagnosis, topographic and ICD-03 morphology code, diagnostic basis, patient's place of residence, source of diagnostic evidence and eligibility criteria for this study. In case of synchronous multiple tumours, only the tumour targeted by chemotherapy was retained.

Phase 2. For each tumour for which a diagnosis was confirmed in 2012 (phase 1), the presence of treatment with chemotherapy within the 6 months following diagnosis was assessed.

Phase 3. For each tumour for which a diagnosis was confirmed in 2012 (phase 1), and treated with chemotherapy (phase 2), the place of chemotherapy treatment was identified.

Patient lists derived from prior identification steps were sent to the hospitals that provided patients with their chemotherapy.

Task 2. Detection of SAEs (pharmacovigilance centre)

Patient lists from each hospital, including only the patients from the hospital concerned, were provided by the hospitals to the pharmacovigilance centre (articles R5144-15 to 20—French public health code). Full perusal of medical files enabled identification of AEs occurring over a period of 6 months after initiation of chemotherapy, as present or not in the files, and defined as any type of harmful and non-intended manifestation occurring for a patient under treatment with 1 or several drugs, without assuming any link with the drug(s). An analysis conducted by pharmacists trained in pharmacovigilance enabled SAEs to be identified.

The AEs under consideration were: harmful and unwanted reactions resulting from authorised use of a medication at standard dosages, and those resulting from medication errors, or not in compliance with specifications in the marketing authorisation, including medication misuse and abuse.[11] The characteristic of seriousness was defined as any effect leading to hospitalisation (initial or prolonged), permanent disability, life-threatening situation or death, and congenital malformation (articles R5121-152—French public health code). Other SAEs were: cancellation or postponement of chemotherapy and a medically significant event assessed by the doctor or a pharmacovigilance investigator. This definition of the characteristic of seriousness is specific to pharmacovigilance (ICH Expert Working Group) but equivalent to the definition in oncology, which is expressed in grades from 0 to 5 (Common Terminology Criteria for Adverse Events), taking clinical and symptom-related aspects into consideration. A drug was considered as *suspect* according to the World Health Organisation imputability criteria.[12] Events with no link to the medication were not retained in the study.

2.2 | Analyses

Clinical and contextual factors associated with SAEs were determined by combining 2 sets of medical data:

- data from the registry (patient identification, sex, date of birth, age at diagnosis, date of incidence, basis for diagnosis, topographic and ICD-03 morphology of the tumour, morphological classification according to Berg,[13] differentiation grade, extension stage at diagnosis),
- data from the medical files, complemented by the pharmacovigilance investigators, socio-demographic data, clinical data, Charlson comorbidity index,[14] chemotherapy lines and any validated SAEs. Data management was carried out with a specially designed database using Microsoft Access. SAS Software version 9.4 was used for the statistical analysis.

The incidence rates of SAEs[15] per patient were calculated with a 95% confidence interval and subdivided into categories according to SAE typology, tumour topography and molecule class for chemotherapy. The global incidence rate was the number of patients for whom at least 1 SAE occurred, irrespective of its nature, divided by the total number of patients. The incidence rate according to SAE typology was the number of patients for whom a SAE occurred at least once divided by the total number of patients. The incidence rate according to tumour topography was the number of patients for whom at least 1 SAE occurred whatever its nature according to tumour topography divided by the total number of patients with that tumour topography. The incidence rate according to the class of chemotherapy molecule was the number of patients who received the treatment (grouped into classes) and for whom at least 1 SAE occurred, whatever its nature, divided by the total number of patients who received the treatment. Molecules associated with the SAEs were assessed on the basis of all

the molecules received between the beginning of treatment and the occurrence of a SAE, and from drugs identified as *suspect* in the patient's file. Aggregated incidence indicators were also calculated for the most frequent classes and topographies.

An epidemiological analysis was conducted to determine factors associated with SAEs. The effects analysed were divided into 2 categories: clinical and contextual variables on the one hand; and variables of exposure to classes of chemotherapy agents administered from the start of the treatment on the other hand. For the purpose of univariate and multivariate analyses, age groups were defined according to terciles. In univariate analysis, the significance of the effects was assessed using the logistic regression likelihood test, and any associations were quantified by the odds ratio and its 95% confidence interval. An increased risk of SAE associated with an effect yields an odds ratio >1, whilst a decreased risk yields an odds ratio <1. A multivariate analysis using a logistic regression model was performed to take conjointly into account the effects linked to chemotherapy molecules and the clinical and contextual effects. Given the variety of potential associations between the different types of chemotherapy, all binary variables associated with the different classes of chemotherapy were included in the multivariate analysis, independently from their univariate significance, except for the *other treatment* category, where the numbers were too small. Among the contextual clinical variables, only those for which the univariate significance was below 0.25 were included in the multivariate analysis. A selection of variables was conducted stepwise by a process of elimination according to the likelihood criteria. Variables for which the significance was <.05 according to Wald's test were retained for the final model. The adjusted odd-ratios and their confidence intervals were calculated. The quality of the fit of the logistic regression model was assessed using Hosmer–Lemeshow test.

The sample size was calculated to reach sufficient precision in estimates of the incidence of SAEs and to ensure statistical power for the analysis of its determining factors. A minimum number of 384 patients was required to ensure a precision of ±5% of the 95% confidence interval. To detect any doubling of the risk (RR = 2) with 80% power, with an unbalanced factor distribution, a sample size ranging from 200 to 500 patients was necessary. The final number was increased to at least 1000 patients to enable subgroup analyses to be conducted and constitute a reference database. In order to obtain a final sample size of 1000 patients treated with chemotherapy, 4000 records were randomly drawn from the registry. The study was approved by the French regulatory authorities; CCTIRS (10 September 2014, authorisation number 14.278bis), CNIL (23 December 2014, decision DR-2014-598).

3 | RESULTS

3.1 | Patient characteristics, tumour characteristics and its treatment

The survey was conducted on 1023 patients. Patient sociodemographics and the characteristics of the tumour and its treatment are shown in Table 1. Four patients under 18 years and 13 over 85 years were included in the study. No pregnant woman was included. The main tumour topographies were: lung, bronchi and trachea (20.0%); breast (18.9%); colon (9.1%); rectum and recto-sigmoid junction (7.7%); oral cavity, pharynx and larynx (9.8%); digestive tract (7.6%); and pancreas (5.3%). In 90.1% of cases, the diagnosis was based on a histological report on the primitive tumour. Metastases were present in 35.4% of the patients, and in more than half where the cancer localisation was lung, bronchi and trachea (62.9%), kidney and bladder (62.5%), pancreas (59.3%) or the digestive tract (stomach, oesophagus or small intestine; 51.3%).

In the 6 months that followed diagnosis, most patients received at least 1 platinum derivative (62.7%) and/or 1 fluoropyrimidine formulation (51.3%). 626 patients (61.2%) received 1–15 identical chemotherapy cycles over the 6-month period following diagnosis.

For the main topographies, the most frequently administered treatments were (Table 2):

- 5-fluorouracil (5-FU) and tegafur-uracil associated with molecules belonging to the following classes: alkylating agents other than platinum derivative, anthracyclines and taxanes for breast cancer,
- 5-FU/platinum derivative association in colorectal cancer,
- Platinum derivative in lung, bronchi and trachea cancers.

Fluoropyrimidine was administered for 87.6% of breast cancers (169/193 patients) and for 99.4% of colon, rectum and recto-sigmoid junction cancers (171/172 patients).

3.2 | Characteristics of SAEs

In total, 661 SAEs were identified, involving 455 patients, in the 6 months following the start of chemotherapy. However, 6 SAEs, observed on 2 occasions with identical molecules, were excluded from the analysis. Therefore, only the remaining 655 were described (Table 3). Over the same period, 4 of them were notified spontaneously by a clinician (0.6%) to the regional pharmacovigilance centre. No unexpected SAEs were observed. There was no report in the files suggesting that the SAEs described were attributable to medication other than chemotherapy or to herbal remedies. When co-medication was specified, there was no known risk of interaction.

More than half of the SAEs (57.3%) were blood and lymphatic system disorders. Half of them (49.8%) were observed during the first 2 cycles of chemotherapy. Nervous system effects occurred much later than the other SAEs. Most SAEs (70.8%) were resolved and the patient recovered. The consequences of the SAEs were hospitalisation or lengthening of a hospitalisation stay (33.7%), cancellation or delay of chemotherapy (27.2%), or another serious medical situation (36.2%). Fourteen SAEs led to a life-threatening situation or a disability; 8 patients died.

Five deaths were consecutive to SAEs:

TABLE 1 Characteristics of patients (n = 1023 patients). Data are n (%) or mean (standard deviation) [range]

Sociodemographic characteristics	
Sex	
Female	483 (47.2%)
Male	540 (52.8%)
Age at diagnosis (y)	62.5 (12.4) [0–92]
Marital status (n = 997)	
In a relationship	710/997 (71.2%)
Alone	287/997 (28.8%)
Body mass index (n = 985)	
Normal	546 (55.4%)
Overweight	300 (30.5%)
Obese	139 (14.1%)
Charlson comorbidity index	2.71 (1.87) [0–12]
Characteristics of the tumour	
Primary cancer location	
Lung, bronchi and trachea	205 (20.0%)
Breast	193 (18.9%)
Colorectum (colon, rectum and recto-sigmoid junction)	172 (16.8%)
Oral cavity, pharynx, larynx	100 (9.8%)
Digestive tract (stomach, oesophagus and small intestine)	78 (7.6%)
Pancreas	54 (5.3%)
Liver, biliary tracts	39 (3.8%)
Ovary, uterus	36 (3.5%)
Brain	28 (2.7%)
Female genital tract other than ovary and uterus	27 (3.5%)
Kidney and bladder	24 (2.3%)
Other and unclassifiable	67 (6.5%)
Basis of diagnosis	
Histology of a primary tumour	922 (90.1%)
Histology of a metastasis	78 (7.6%)
Cytology	12 (1.2%)
Specific tumour markers	2 (0.2%)
Paraclinical investigations	9 (0.9%)
Synchronous multiple tumours	32 (3.1%)
Tumour histological grade (n = 868)	
High grade	344 (39.6%)
Presence of metastases	362 (35.4%)
Morphological classification according to Berg	
Adenocarcinoma	652 (63.7%)
Squamous or transitional carcinoma	196 (19.2%)
Other	175 (17.1%)
Characteristics of chemotherapy care	
Place of chemotherapy	

(Continues)

TABLE 1 (Continued)

Sociodemographic characteristics	
Tertiary university	344 (33.6%)
Tertiary nonuniversity	485 (47.4%)
Secondary	194 (19.0%)
Number of patients who received at least once in the course of chemotherapy a molecule belonging to each class	
5-fluorouracil and tegafur-uracil	475 (46.4%)
Capecitabine	50 (4.9%)
Platinum derivatives (cisplatin, oxaliplatin, carboplatin)	641 (62.7%)
Taxanes (paditaxel, docetaxel)	283 (27.7%)
Other alkylating agents (cyclophosphamide, ifosfamide, dacarbazine, temozolomide, mitomycin C, actinomycin, trabectedin, busulfan, melphalan)	222 (21.7%)
Anthracyclines (epirubicin, doxorubicin)	191 (18.7%)
Other antimetabolites (gemcitabine, methotrexate, pemetrexed, raltitrexed, fludarabine)	177 (17.3%)
Targeted therapy or protein inhibitors (panitumumab, bevacizumab, cetuximab, trastuzumab, vemurafenib, crizotinib, temsirolimus, sunitinib, sorafenib)	171 (16.7%)
Other topo-isomerase II inhibitors (etoposide, bleomycin)	65 (6.4%)
Topo-isomerase I inhibitors (irinotecan, topotecan)	53 (5.2%)
Vinca-alkaloids (vincristine, vinorelbine)	50 (4.9%)
Other treatments (interferon, somatuline, MDM2 antagonist, cyclosporin)	10 (1.0%)

- Febrile neutropenia during cycle 2 of palliative treatment with cetuximab/carboplatin/5-FU (intravenous, IV) involving a 62-year-old patient (Charlson index 3) presenting cancer of the pharynx with bone metastases.
- Heart attack and cytopenia during cycle 1 of treatment with sunitinib (orally) involving a 78-year old female patient (Charlson index 3) presenting cancer of the kidney. Heart attack was the SAE that led to her death.
- Cardiopulmonary arrest during cycle 2 of treatment with paditaxel/carboplatin (IV) involving an 81-year old female patient (Charlson index 4) presenting cancer of the ovaries.
- Bicytopenia in cycle 2 of treatment with etoposide/carboplatin (IV) involving a 66-year old patient (Charlson index 3) presenting cancer of the lungs, bronchi and trachea.
- Bicytopenia during cycle 7 of treatment with pemetrexed (IV) involving a 66-year old patient (Charlson index 7) presenting cancer of the lungs, bronchi and trachea. The patient had received cisplatin/pemetrexed up to the 5th cycle included, and from cycle 6, pemetrexed alone. He had presented bicytopenia during cycle 3 and neutropenia during cycle 5.

TABLE 2 Treatments administered in the main cancer locations

	n (%)	*n* (%)	*n* (%)
Classes of molecules	Breast (*n* = 193)	Colorectum (*n* = 172)	Lung, bronchi and trachea (*n* = 205)
5-fluorouracil or tegafur-uracil	167 (86.5%)	138 (80.2%)	0
Capecitabine	2 (1.0%)	42 (24.2%)	0
Platinum derivatives	2	113 (65.7%)	189 (92.2%)
Other alkylating agents	176 (91.2%)	1	3
Other antimetabolites	0	2	97 (47.3%)
Topo-isomerase I inhibitors	0	20 (11.6%)	3
Anthracyclines	168 (87.0%)	0	2
Other topo-isomerase II inhibitors	0	0	48 (23.4%)
Vinca-alkaloids	1	0	44 (21.5%)
Taxanes	164 (85.0%)	0	31 (15.1%)
Targeted therapy/protein inhibitors	49 (25.4%)	36 (20.9%)	26 (12.7%)

Two SAEs could have been the cause of death:

- Pancytopenia during cycle 2 of treatment with pemetrexed/cisplatin (IV) involving a 54-year old patient (Charlson index 4) presenting cancer of the lungs, bronchi and trachea.
- Progressive multifocal leuko-encephalopathy during cycle 5 of treatment with carboplatin/etoposide (IV) involving a 58-year old patient (Charlson index 1) presenting cancer of the lungs, bronchi and trachea. The patient had received cisplatin/etoposide up to the second cycle included, then from cycle 3, carboplatin/etoposide. He had presented inappropriate antidiuretic hormone secretion during cycle 2.

One death was not linked to the SAE and resulted from kidney failure, which occurred during cycle 2 of treatment with vinorelbine/cisplatin (IV) involving a 76-year old female patient (Charlson index 3) presenting head and neck cancer. This was treated as a *never event*.

3.3 | Incidences of SAEs

The incidence rates for SAEs are presented in Table 4.

The global incidence rate of SAEs was 44.5% (95% confidence interval = 41.4–47.5%). It was 27.9% (25.1–30.6%) for the SAEs grouped under the system organ class *blood and lymphatic system disorders* in the 6 months following the start of chemotherapy. The incidence rate varied from 17.7% (12.0–23.5%) for patients presenting breast cancer and having received at least 1 cycle of fluropyrimidine, to 46.0% (38.9–53.1%) for patients presenting lung, bronchi and trachea cancer and treated with platinum derivative.

Incidence rates for lung, bronchi and trachea cancers (59.0% [52.3–65.8%]), colorectal cancers (50.0% [42.5–57.5%]), and cancers of the digestive tract (55.1% [44.1–66.2%]) were higher than the global incidence. Incidence rates for breast cancers (31.6% [25.0–38.2%]), cancers of the female genital tract (33.3% [15.5–51.1%]) and cancers of the liver and intrahepatic biliary tracts (25.6% [11.9–39.4%]) were lower than the global incidence.

3.4 | Factors and chemotherapy molecules associated with the occurrence of SAEs

The results of the analysis of clinical and contextual factors and chemotherapy molecules associated with the occurrence of SAEs are presented in Table 5. In univariate analysis, the incidence of SAEs was significantly higher among men (47.4 vs 41.2%), among patients presenting a Charlson comorbidity index of 3 or more (51.3 vs 40.4%), and with metastases (53.3 vs 39.6%). The incidence of SAEs was also higher for certain localisations, such as the lung, bronchi and trachea (59.0%), the digestive tract (55.1%), the colon, rectum and recto-sigmoid junction (50.0%).

The univariate analysis of the SAEs according to the chemotherapy therapeutic classes administered showed an excess incidence in presence of platinum derivatives (52.0%), topo-isomerase II inhibitors other than anthracyclines (69.2%), and vinca-alkaloids (66.7%). Conversely, a reduced incidence was observed with capecitabine (24.0%), alkylating agents other than platinum derivatives (32.6%), anthracyclines (31.6%) and taxanes (33.1%).

After the multivariate analysis using a logistic regression model, 8 variables were retained. Two were clinical context-related variables: Charlson's index and metastatic stage. The other 6 were therapeutic classes, 5 of which were already significant in the univariate analysis (platinum derivatives, capecitabine, topo-isomerase II inhibitors other than anthracyclines, vinca-alkaloids and taxanes). 5-FU therefore appeared as associated with excess risk, whilst anthracyclines and alkylating agents other than platinum derivatives no longer had any significant effect. The performance of this model in predicting the occurrence of a SAE was associated with a concordance index of 0.662.

The location of the tumour, highly significant in a univariate analysis, was not included in the final model because of the strong relationship with the therapeutic strategy. It is worth noting, however.

TABLE 3 Description of the typology, evolution and gravity of serious adverse effects (SAEs; n = 655 in 455 patients)

MedDRA system organ class	n (n = 655)	%
Blood and lymphatic system disorders	375	57.3%
Nervous system disorders	56	8.5%
Gastrointestinal disorders	41	6.3%
General disorders and administration site conditions (asthenia, hyperthermia, inflammation of the mucous membranes, delayed healing, thrombosis)	40	6.1%
Renal and urinary disorders	38	5.8%
Skin and tissue disorders	33	5.0%
Immune system disorders	15	2.3%
Ear and labyrinth disorders	13	2.0%
Cardiac disorders + vascular disorders	12	1.8%
Metabolism and nutrition disorders	7	1.1%
Investigations	6	0.9%
Vascular disorders	5	0.8%
Hepatobiliary disorders	3	0.5%
Musculoskeletal and connective tissue disorders	4	0.6%
Eye disorders	2	0.3%
Respiratory, thoracic and mediastinal disorders	2	0.3%
Endocrine disorders	1	0.1%
Psychiatric disorders	1	0.1%
Infections and infestations	1	0.1%
Evolution		
Recovered/resolved	464	70.8%
Unknown	83	12.7%
Not recovered/not resolved	53	8.1%
In the process of recovery/regression	46	7.0%
Deaths linked to SAEs*	5	0.8%
SAEs that could have been the cause of death*	2	0.3%
Death with no link to the SAE*	2	0.3%
Gravity		
Another serious medical situation**	237	36.2%
Hospitalisation or lengthening of a hospitalisation stay	221	33.7%
Cancellation or the delay of chemotherapy	178	27.2%
Life-threatening situation	9	1.4%
Death***	5	0.8%
Incapacity or disability	5	0.8%

*these 9 SAEs correspond to 8 deceased patients.
**a medically significant effect, at the discretion of the physician.
***of which 4 SAEs resulted in death and 1 SAE may have resulted in death.

TABLE 4 Incidence rates of serious adverse effects (SAEs; by decreasing incidence rates)

Global incidence rate	SAEs	Rate (95%CI)
No SAE	568	55.5% (52.5–58.6%)
At least 1 SAE	455	44.5% (41.4–47.5%)
1 SAE	314	30.7%
2 SAEs	94	9.2%
3 SAEs	37	3.6%
4 SAEs	8	0.8%
5 SAEs	2	0.2%
Incidence rate according to SAE typology (system organ class)		
Blood and lymphatic system disorders	285	27.9% (25.1–30.6%)
Nervous system disorders	55	5.4% (4.0–6.8%)
Gastrointestinal disorders	41	4.0% (2.8–5.2%)
General disorders and administration site conditions	38	3.7% (2.6–4.9%)
Renal and urinary disorders	37	3.6% (2.5–4.8%)
Skin and tissue disorders	33	3.2% (2.1–4.3%)
Cardiac disorders + vascular disorders	17	1.7% (0.9–2.5%)
Immune system disorders	15	1.5% (0.7–2.2%)
Ear and labyrinth disorders	13	1.3% (0.6–2.0%)
Incidence rate according to primary cancer location		
Lung, bronchi and trachea	121/205	59.0% (52.3–65.8%)
Digestive tract (stomach, oesophagus and the small intestine)	43/78	55.1% (44.1–66.2%)
Colorectum (colon, rectum and recto-sigmoid junction)	86/172	50.0% (42.5–57.5%)
Pancreas	25/54	46.3% (33.0–59.6%)
Kidney and bladder	11/24	45.8% (25.9–65.8%)
Brain	11/28	39.3% (21.2–57.4%)
Ovary, uterus	14/36	38.9% (23.0–54.8%)
Oral cavity, pharynx, larynx	37/100	37.0% (27.5–46.5%)
Female genital tract other than ovary and uterus	9/27	33.3% (15.5–51.1%)
Breast	61/193	31.6% (25.0–38.2%)
Liver, biliary tracts	10/39	25.6% (11.9–39.4%)
Other and unclassifiable	27/67	40.3% (28.5–52.0%)
Incidence rate according to the class of chemotherapy molecule (all the molecules received between the beginning of treatment and the occurrence of a SAE)		
Topo-isomerase II inhibitors other than anthracyclines	45/65	69.2% (68.0–80.5%)
Vinca-alkaloids	32/48	66.7% (53.3–80.0%)
Topo-isomerase I inhibitors	24/44	54.5% (39.8–69.3%)
Platinum derivatives	331/637	52.0% (48.1–55.8%)

(Continues)

TABLE 4 (Continued)

Global incidence rate	SAEs	Rate (95%CI)
Antimetabolites other than 5-fluorouracil or capecitabine	86/175	49.1% (41.7–56.6%)
5-fluorouracil	211/470	44.9% (40.4–49.4%)
Targeted therapy or protein inhibitors	64/155	41.3% (33.5–49.0%)
Taxanes	84/254	33.1% (27.3–38.9%)
Other alkylating agents	72/221	32.6% (26.4–38.8%)
Anthracyclines	60/190	31.6% (25.0–38.2%)
Capecitabine	12/50	24.0% (12.2–35.8%)
Other treatments	1/9	11.1% (0.3–48.3%)*

*95% exact CI.

that an alternative multivariate statistical model could be proposed, retaining the location of the tumour, comorbidities and 3 classes of chemotherapy: platinum derivatives, topo-isomerase II inhibitors other than anthracyclines, and capecitabine. Despite a lesser statistical performance on the likelihood criterion, this last model had a similar concordance index of 0.662.

The search for interactions between factors in the multivariate logistic model did not evidence any significant interaction that could alter the formulation and interpretation of the model.

4 | DISCUSSION

Cancer treatments involving chemotherapy have been associated with a particularly high incidence of SAEs, affecting 44.5% of the patients, without unexpected SAE. This study was the first to be carried out on a large representative sample of over 1000 patients.

Few pharmaco-epidemiological studies have been conducted, and existing studies involved smaller numbers of patients, older on the whole, treated with different chemotherapy protocols for different types of non-haematological cancers at different stages. The occurrence of at least 1 SAE is reported for 42–60% of patients: these SAEs thus occurred in 49% of a population of 110 patients older than 75 years[16] or 53% of 500 patients older than 65 years.[17] In an Australian cohort that followed 449 patients, aged under 65 years for 72.5%, over a median period of 5.64 months, 60% of the patients reported at least 1 SAE.[18] More recently, among 151 Thai patients aged 70 years and over, 42% experienced SAEs.[19] Only 1 French single-centre study assessed the occurrence of SAEs to be 45% among 397 patients.[20]

The incidence of SAEs is linked to many determinants, dependent at the same time on the tumour, the clinical context and the molecules administered. However, Pearce *et al*. did not find any difference in AE incidence, whether serious or not, according to the type of cancer (breast, colorectal, non-small cell pulmonary cancers), but this result could have been influenced by the method used in this study for the collection of AEs among patients.[18]

The highest incidence rates for SAEs were observed when patients were exposed to topo-isomerase II inhibitors other than anthracyclines (69.2%), vinca-alkaloids (66.7%), topo-isomerase I inhibitors (54.5%) or platinum derivatives (52.0%). The risk for the most severe SAEs (grades III and IV) is indeed linked to the greater toxicity of certain chemotherapy protocols.[16] The association of several chemotherapy treatments is a factor associated with the occurrence of an AE among older patients.[19] Two studies have identified the molecules most frequently associated with AEs whatever the level of seriousness.[21,22] Le Beller *et al*. mainly found cisplatin–carboplatine, taxanes, fluorouracil and gemcitabine.[20]

The clinical context influences the incidence of SAEs, especially the presence of metastases (53.3%) and comorbidities (51.3% for a Charlson index ≥3). These factors have also been found in previous studies.[16,17,19] Substantial differences have also been detected depending on the primary cancer location, with a particularly high incidence in bronchial–pulmonary cancers (59.0%). Wahlang *et al*. found an AE incidence higher in men than in women, a difference they explained by a higher incidence of lung cancers (22.7%) in their study on an Indian population.[23] The multivariate analysis showed that these determinants interact, since the molecules used, and the associations in chemotherapy protocols, are adapted to cancer location, stage and the patient's clinical state.

Certain limitations linked to the inclusion criteria should be considered in the interpretation of the results. The most frequent cancer locations (breast, colon, lung) were included. Indeed, resort to chemotherapy is not so widely used for malignant haemopathies (leukaemia and lymphoma), nor for cancers of the nervous system, where different protocols are used, so that these results are not applicable. Patient inclusion in the initial stages of treatment, defined as a period of 6 months following diagnosis, made it possible to take account of the effects of confusion related to past medical history. However, the procedure used puts the emphasis on the first lines of treatment to the detriment of secondary treatment lines or the treatment of recurrences. Finally, there is still a lack of information on SAEs among patients younger than 18 years or older than 85 years, and also on SAEs following targeted therapies or immunotherapy. We therefore suggest that further specific studies are required.

Nevertheless, a strength of this study is that it included subjects irrespective of age, and provided for a more or less exhaustive collection of comorbidities. In addition, patient record review is the most widely used method, accepted worldwide, to measure incidence rates of AEs and it was used here on a population representatively sampled from cancer registry clinical data.[24]

The main interest in this study is its original sampling method, using a general cancer registry on a regional scale, which ensures exhaustive monitoring, thus guaranteeing an excellent representation of the population concerned, which is essential to any generalisation of the results. Systematic checking of the patients' medical files was nevertheless necessary to ensure that the information on the treatments received and on SAE occurrences was precise and complete, in addition to information routinely collected by the registry. This mode of survey is standard in pharmacovigilance, but the difficulties

TABLE 5 Factors associated with the occurrence of serious adverse effects (SAEs; univariate and multivariate analysis)

		Univariate analysis		Multivariate analysis	
	SAEs *n* (%)	*P*	Odds ratio	*P*	Odds ratio
Sex (*n* = 1023)		.046		NS	
Female (*n* = 483)	199 (41.2%)		1		
Male (*n* = 540)	256 (47.4%)		1.29 (1.00–1.65)		
Age at diagnosis (*n* = 1023)		.17		NS	
<60 years (*n* = 386)	159 (41.2%)		1		
60–69 years (*n* = 332)	160 (48.2%)		1.33 (0.99–1.79)		
≥70 years (*n* = 305)	136 (44.6%)		1.15 (0.85–1.56)		
Body mass index (*n* = 985)		.70		-	
Normal (*n* = 546)	245 (44.9%)		1		
Overweight (*n* = 300)	128 (42.7%)		0.91 (0.69–1.22)		
Obese (*n* = 139)	65 (46.8%)		1.08(0.74–1.57)		
Marital status (*n* = 997)		.46		-	
In a relationship (*n* = 710)	325 (45.8%)		1		
Alone (*n* = 287)	124 (43.2%)		0.90 (0.68–1.89)		
Charlson comorbidity index (*n* = 1023)		.010		.0071	1.44 (1.10–1.87)
0–2 (*n* = 498)	201 (40.4%)		1		
≥3 (*n* = 525)	254 (51.3%)		1.39 (1.08–1.77)		
100% social insurance coverage (*n* = 1023)		.42		-	
Yes (*n* = 941)	422 (44.8%)		1		
No (*n* = 82)	33 (40.2%)		0.83 (0.52–1.31)		
Place of chemotherapy (*n* = 1023)		.12		NS	
Tertiary university (*n* = 344)	150 (43.6%)		1		
Tertiary nonuniversity (*n* = 485)	206 (42.5%)		0.96 (0.72–1.26)		
Secondary (*n* = 194)	99 (51.0%)		1.35 (0.95–1.92)		
Metastases (*n* = 1023)		<.0001		.013	
No (*n* = 661)	262 (39.6%)		1		
Yes (*n* = 362)	193 (53.3%)		1.74 (1.34–2.25)		1.42 (1.08–1.88)
Synchronous multiple tumours (*n* = 1023)		.66			
No (*n* = 991)	442 (44.6%)		1		
Yes (*n* = 32)	13 (40.6%)		0.85 (0.41–1.74)		
Morphological classification according to Berg (*n* = 1023)		.29		-	
Adenocarcinoma (*n* = 652)	278 (42.6%)		1		
Squamous and transitional carcinoma (*n* = 196)	93 (47.5%)		1.22 (0.88–1.67)		
Other (*n* = 175)	84 (48.0%)		1.24 (0.89–1.74)		
Cancer location (*n* = 1023)		<.0001		NS	
Breast (*n* = 193)	61 (31.6%)		1		
Lung, bronchi and trachea (*n* = 205)	121 (59.0%)		3.12 (2.06–4.70)		
Colorectum (*n* = 172)	86 (50.0%)		2.16 (1.41–3.31)		
Oral cavity, pharynx, larynx (*n* = 100)	37 (37.0%)		1.27 (0.76–2.11)		
Pancreas, liver, biliary tracts (*n* = 93)	35 (37.6%)		1.31 (0.78–2.19)		
Digestive tract (*n* = 78)	43 (55.1%)		2.66 (1.55–4.56)		
Female genital tract (other than breast; *n* = 63)	23 (36.5%)		1.24 (0.69–2.26)		
Other and unclassifiable (*n* = 119)	49 (41.2%)		1.52 (0.94–2.44)		
Molecules administered between the beginning of chemotherapy and the onset of SAE					
5-fluorouracil or tegafur-uracil (*n* = 470)		.80		.0067	

(Continues)

TABLE 5 (Continued)

		Univariate analysis	Multivariate analysis
No	244 (44.1%)	1	1
Yes	211 (44.9%)	1.03 (0.81–1.32)	1.46 (1.11–1.93)
Capecitabine (*n* = 50)		.0028	.045
No	443 (45.5%)	1	1
Yes	12 (24.0%)	0.38 (0.19–0.73)	0.49 (0.24–0.98)
Platinum derivatives (*n* = 637)		<.0001	<.0001
No	124 (32.1%)	1	
Yes	331 (52.0%)	2.28 (1.75–2.98)	1.77 (1.33–2.35)
Other alkylating agents (*n* = 221)		<.0001	NS
No	383 (47.8%)	1	
Yes	72 (32.6%)	0.53 (0.39–0.72)	
Antimetabolites (*n* = 175)		.17	NS
No	369 (43.5%)	1	
Yes	86 (49.1%)	1.25 (0.90–1.74)	
Topo-isomerase I inhibitors (*n* = 44)		.17	NS
No	431 (44.0%)	1	
Yes	24 (54.5%)	1.53 (0.83–2.80)	
Anthracyclines (*n* = 190)		<.0001	NS
No	395 (47.4%)	1	
Yes	60 (31.6%)	0.51 (0.37–0.72)	
Other topo-isomerase II inhibitors (*n* = 65)		<.0001	.0019
No	410 (42.8%)	1	
Yes	45 (69.2%)	3.01 (1.75–5.17)	2.48 (1.40–4.40)
Vinca-alkaloids (*n* = 48)		.0015	.0043
No	423 (43.4%)	1	
Yes	32 (66.7%)	2.61 (1.41–4.82)	2.54 (1.34–4.83)
Taxanes (*n* = 254)		<.0001	.019
No	371 (48.2%)	1	
Yes	84 (33.1%)	0.53 (0.39–0.71)	0.68 (0.49–0.94)
Targeted therapy or protein inhibitors (*n* = 155)		.39	NS
No	391 (45.0%)	1	
Yes	64 (41.3%)	0.86 (0.61–1.21)	

involved in its implementation and the great amount of time staff need to devote to the survey and the analysis are real obstacles, leading to long delays in implementation and limiting the possibilities for repetition in time and space for comparative purposes.

To improve the notification of SAEs in oncology, the first lever on which action could be taken is the sensitisation of physicians towards reporting of SAEs, in the form of support and back-up rather than an accusatory approach.[25]

5 | CONCLUSION

The overall incidence rate for SAEs was 44.5%. This high percentage should lead the medical community to reinforce information about drug toxicities for patients and oncologists. In this way, knowledge could be improved by raising awareness among patients on the need to report on quality of life and among medical doctors on the need to report on drug toxicities.

ACKNOWLEDGEMENTS

The authors would like to thank:

- the pharmacovigilant pharmacists: Marion Allouchery, Vincent Delpech, Inès Miladi;
- the cancer registry technicians for their valuable contributions to the study: Aurélie Delzor, Tiffanie Girault, Soizic Lellouch, Nicolas Mériau;
- the oncologists, medical practitioners and directors of the participating centres described below for their useful contributions in access to medical records: Poitiers, La Rochelle, Niort, Cognac,

Saintes, Royan, Jonzac, Bordeaux, Angoulême, Rochefort, Angers, Limoges, Nord Deux-Sèvres, Nantes, Cholet, Tours, Montmorillon and Châtellerault;

- the pathologists, departments of medical data processing in public and private hospitals, and health insurance services for their assistance for routine collection of data;
- and Angela Verdier for the translation.

COMPETING INTERESTS

Isabelle Ingrand, Gautier Defossez, Claire Lafay-Chebassier, François Chavant, Aurélie Ferru, Pierre Ingrand andMarie-Christine Perault-Pochat have no conflicts of interest directly relevant to the content of this article.

This work was supported by grants from the *Institut national du cancer* (INCa) and the *Agence nationale de sécurité du médicament et des produits de santé* (ANSM).

CONTRIBUTORS

Isabelle Ingrand, Claire Lafay-Chebassier, Pierre Ingrand and Marie-Christine Perault-Pochat wrote the manuscript; Isabelle Ingrand, Gautier Defossez, François Chavant, Pierre Ingrand and Marie-Christine Perault-Pochat designed the research; Isabelle Ingrand, Gautier Defossez, François Chavant, Pierre Ingrand and Marie-Christine Perault-Pochat performed the research; Isabelle Ingrand and Pierre Ingrand analysed the data; Aurélie Ferru and Marie-Christine Perault-Pochat helped to interpret the adverse effects.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Isabelle Ingrand https://orcid.org/0000-0002-2709-1585
Gautier Defossez https://orcid.org/0000-0002-7615-954X
Claire Lafay-Chebassier https://orcid.org/0000-0001-8970-5514
Aurélie Ferru https://orcid.org/0000-0003-3981-6640
Pierre Ingrand https://orcid.org/0000-0002-7765-2520
Marie-Christine Pérault-Pochat https://orcid.org/0000-0002-3289-5502

REFERENCES

1. Jéhannin-Ligier K, Dantony E, Bossard N, et al. *Cancer incidence and mortality projections in metropolitan France in 2017*. Technical report. Saint-Maurice: Santé publique France; 2017
2. Situation de la chimiothérapie des cancers - Rapport 2012. *Boulogne-Billancourt: INCa*; 2013:1-102.
3. Pouyanne P, Haramburu F, Imbs JL, Bégaud B. Admissions to hospital caused by adverse drug reactions: cross sectional incidence study. *French Pharmacovigilance Centres BMJ*. 2000; 320(7241):1036.
4. Bénard-Laribière A, Miremont-Salamé G, Pérault-Pochat MC, Noize P, Haramburu F. EMIR study group on behalf of the French network of pharmacovigilance centres. Incidence of hospital admissions due to adverse drug reactions in France: the EMIR study. *Fundam Clin Pharmacol*. 2015;29(1):106-111. https://doi.org/10.1111/fcp.12088
5. Moride Y, Haramburu F, Requejo AA, Bégaud B. Under-reporting of adverse drug reactions in general practice. *Br J Clin Pharmacol*. 1997; 43(2):177-181.
6. Lau PM, Stewart K, Dooley M. The ten most common adverse drug reactions (ADRs) in oncology patients: do they matter to you? *Support Care Cancer*. 2004;12(9):626-633.
7. Seruga B, Templeton AJ, Badillo FE, Ocana A, Amir E, Tannock IF. Under-reporting of harm in clinical trials. *Lancet Oncol*. 2016;17(5): e209-e219. https://doi.org/10.1016/S1470-2045(16)00152-2
8. Csizmadi I, Collet JP, Boivin JF, Hoffman BB, Lefkowitz RJ. Bias and confounding. In: Strom BL, ed. *Pharmacoepidemiology*. Fourth ed. Chichester: John Wiley & Sons; 2007:791-809.
9. Sankila R, Black R, Coebergh JW, et al. *eds. Evaluation of Clinical Care by Cancer Registries. IARC Technical Publication No. 37*. International Agency for Research on Cancer: Lyon; 2003.
10. Registre des cancers Poitou-Charentes. 2018 http://medphar.univ-poitiers.fr/registre-cancers-poitou-charentes/ Accessed March 4, 2019.
11. Bonnes pratiques de pharmacovigilance. Saint-Denis: ANSM; 2018.
12. World Health Organization. Safety monitoring of medicinal products: guidelines for setting up and running a pharmacovigilance Centre. Uppsala: the Uppsala Monitoring Centre, WHO Collaborating Centre for International Drug Monitoring; 2000.
13. Berg JW. Morphologic classification of human cancer. In: Shottenfeld D, Fraumeni J Jr, eds. *Cancer Epidemiology and Prevention*. 2nd ed. New York: Oxford University Press; 1996:28-44.
14. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987;40(5):373-383.
15. Bégaud B. *Dictionary of Pharmacoepidemiology*. Chichester: John Wiley & Sons; 2000.
16. Marinello R, Marenco D, Roglia D, et al. Predictors of treatment failures during chemotherapy: a prospective study on 110 older cancer patients. *Arch Gerontol Geriatr*. 2009;48(2):222-226. https://doi.org/10.1016/j.archger.2008.01.011
17. Hurria A, Togawa K, Mohile SG, et al. Predicting chemotherapy toxicity in older adults with cancer: a prospective multicenter study. *J Clin Oncol*. 2011;29(5):3457-3465. https://doi.org/10.1200/JCO.2011.34.7625
18. Pearce A, Haas M, Viney R, et al. Incidence and severity of self-reported chemotherapy side effects in routine care: A prospective cohort study. PLoS ONE 2017; 12(10). e0184360. doi: https://doi.org/10.1371/journal.pone.0184360.
19. Phaibulvatanapong E, Srinonprasert V, Ithimakin S. Risk factors for chemotherapy-related toxicity and adverse events in elderly Thai cancer patients: a prospective study. *Oncology*. 2018;94(3):149-160. https://doi.org/10.1159/000485078
20. Le Beller C, Henry S, Levée M, et al. Assessment after 12-months of systematic report of cancer treatment adverse drug reaction. *Fundamental Clin Pharmacol*. 2003;17(2):228.
21. Sharma A, Kumari KM, Manohar HD, Bairy KL, Thomas J. Pattern of adverse drug reactions due to cancer chemotherapy in a tertiary care hospital in South India. *Perspect Clin Res*. 2015 Apr-Jun;6(2):109-115. https://doi.org/10.4103/2229-3485.154014
22. Mallik S, Palaian S, Ojha P, Mishra P, Pak J. Pattern of adverse drug reactions due to cancer chemotherapy in a tertiary care teaching hospital in Nepal. *Park J Pharm Sci*. 2007 Jul;20(3):214-8.
23. Wahlang JB, Laishram PD, Brahma DK, Sarkar C, Lahon J, Nongkynrih BS. Adverse drug reactions due to cancer chemotherapy in a tertiary care teaching hospital. *Ther Adv Drug Saf*. 2017;8(2):61-66. https://doi.org/10.1177/2042098616672572
24. Hanskamp-Sebregts M, Zegers M, Vincent C, van Gurp PJ, de Vet HC, Wollersheim H. Measurement of patient safety: a systematic review of the reliability and validity of adverse event detection with record review. *BMJ Open*. 2016;6:e011078. https://doi.org/10.1136/bmjopen-2016-011078

25. Hardy AC, Canevet JP, Girard C, Jourdain M, Morel S, Bournot MC, Buyck JF, Dalichampt M, Tallec A, Ingrand I et al. Sircade-Volet 2 Étude sociologique des dynamiques de qualification et de gestion des EI médicamenteux. [Rapport de recherche] INca; ANSM. 2018. https://halshs.archives-ouvertes.fr/halshs-01923896/document]

How to cite this article: Ingrand I, Defossez G, Lafay-Chebassier C, et al. Serious adverse effects occurring after chemotherapy: A general cancer registry-based incidence survey. *Br J Clin Pharmacol.* 2020;86:711–722. https://doi.org/10.1111/bcp.14159

3.3.4. Parcours de soins des patients atteints d'hémopathies malignes [50–52]

Une série de travaux appliqués aux hémopathies malignes s'est intéressée aux parcours de soins de deux hémopathies sensiblement différentes : le myélome multiple et le lymphome B diffus à grandes cellules. L'intérêt de la démarche résidait dans l'expertise et la vision multidisciplinaire, territoriale et exhaustive du RGCPC en vue d'analyser la qualité des prises en charge et ses déterminants, et leurs impacts notamment en termes de survie globale.

- Exemple du myélome multiple [50,51]

Le pronostic du myélome a été transformé par l'arrivée de nouvelles thérapeutiques et dont l'usage a été rapidement intégré dans les recommandations de traitement (à partir de 2008). Cette tumeur, dont la prise en charge est souvent morcelée entre plusieurs spécialités, était une maladie modèle pour tenter de dégager des déterminants d'inégalités d'accès aux soins.

Une première publication a permis d'identifier des facteurs géographiques et organisationnels associés à une moins bonne qualité des soins, sur une série de 367 patients atteints de myélome multiple incident entre 2008 et 2010 en Poitou-Charentes [50] :

i) L'éloignement géographique par rapport au centre de référence et l'absence de passage en RCP étaient les 2 principaux déterminants qui impactaient sur la qualité du bilan diagnostic et pronostic (bilans conformes aux recommandations dans 98% et 58% des cas, respectivement) ;
ii) La prise en charge dans un centre secondaire (non spécialisé) était le principal déterminant d'une moins bonne compliance au traitement (bilan conforme aux recommandations dans 89% des cas).

Cette étude soulignait ainsi l'importance de l'organisation territoriale de l'offre de soins et les points d'amélioration à mettre en œuvre de sorte de parvenir au meilleur équilibre entre la proximité géographique et l'expertise disponible pour les décisions thérapeutiques, en vue que les patients puissent recevoir les meilleurs soins possibles.

Un second papier s'était intéressé plus précisément à l'impact de l'âge sur l'adhésion aux recommandations de prise en charge [51]. Les résultats montraient des disparités liées à l'âge sur la qualité du bilan pronostic et le traitement de première ligne, avec un impact en terme de survie globale.

La principale limite résidait dans la non prise en compte du statut socio-économique des individus comme variable d'ajustement sur ces questions d'inégalités, désormais disponible au sein des registres de population français. A la différence des démarches précédentes, un retour au dossier médical systématique était rendu ici nécessaire pour colliger les informations relatives au bilan d'extension (*e. g.* bilan radiologique osseux), pronostic (cytogénétique notamment) et thérapeutique

(molécules administrées), absentes le plus souvent du SI-RGCPC. A noter que le faible taux de passage en RCP pour cette pathologie sur la période d'étude (< 50%) était un élément prédisposant à un faible niveau d'informations initiales au sein du SI-RGCPC.

Puyade M, Defossez G, Guilhot F, Leleu X, Ingrand P. Age-related health care disparities in multiple myeloma. ***Hematol Oncol****. 2018 Feb;36(1):224-231. doi: 10.1002/hon.2422. Epub 2017 Apr 21. PMID: 28429426.*

Puyade M, Defossez G, Guilhot F, Ingrand P. Multiple myeloma: the quality of care is linked to geographical and organisational determinants. A study in a French registry. ***Eur J Cancer Care*** *(Engl). 2016 Sep;25(5):855-63. doi: 10.1111/ecc.12414. Epub 2015 Nov 25. PMID: 26603508.*

- Exemple du lymphome B diffus à grandes cellules [52]

Si le R-CHOP est le traitement standard de premier recours pour le lymphome B diffus à grandes cellules (DLBCL) de novo, le choix de traitement optimal d'un DLBCL secondaire (*i. e.* second cancer primitif) est moins conventionnel du fait du pronostic et du traitement du premier cancer. Partant du constat que les DLBCL secondaires avaient un pronostic moins bon que les DLBCL de novo, l'étude s'est intéressée à la prise en charge des seconds cancers de type DLBCL et à leurs impacts en termes de survie global au regard de l'utilisation ou non de R-CHOP.

En pratique, tous les DLBCL diagnostiqués entre 2008 et 2015 en Poitou-Charentes et ayant un ou plusieurs cancers primitifs antérieurs (solides et hématologiques) ont été inclus (132 DLBCL secondaires parmi 1 251 DLBCL de novo). Les patients non traités par R-CHOP (22%) avaient une survie globale moins bonne en analyse uni et multivariée, avec un état de santé général plus souvent altéré et un traitement antérieur par chimiothérapie plus fréquent (60% vs. 26%) que ceux traités par R-CHOP. Ces résultats suggéraient que ces deux facteurs pouvaient en pratique influencer sur la décision du clinicien à ne pas utiliser de R-CHOP. Toutefois, la notion de chimiothérapie antérieure n'avait pas d'impact en termes pronostic, soulevant la question du rapport risque-bénéfice de la non-utilisation du R-CHOP pour prévenir la cardiotoxicité liée aux anthracyclines.

Systchenko T, Defossez G, Guidez S, Laurent C, Puyade M, Debiais-Delpech C, Dreyfus B, Machet A, Leleu X, Delwail V, Ingrand P. R-CHOP appears to be the best first-line treatment for second primary diffuse large B cell lymphoma: a cancer registry study. ***Ann Hematol****. 2020 Jul;99(7):1605-1613. doi: 10.1007/s00277-020-04100-8. Epub 2020 May 25. PMID: 32451709.*

3.4. Réutilisation des données massives en santé : exemple de la plateforme INSHARE [53,54]

Au global, des échanges structurés entre le registre et d'autres gestionnaires de bases de données en santé permettent d'envisager de réduire les coûts, d'intensifier la recherche centrée sur le patient et d'accélérer la découverte de nouvelles connaissances. C'est le constat également établi plus largement par les institutions et la communauté scientifique, qui soutiennent depuis quelques années une politique d'incitation à l'ouverture des données massives en santé (DMS). Le partage et l'exploitation efficiente des DMS impliquent toutefois de résoudre en amont un certain nombre de verrous, parmi lesquels figurent le caractère sensible des données (règles de gouvernance et de sécurité), leur manque d'interopérabilité (normalisation sur le plan sémantique et syntaxique) et leur qualité variable pour une analyse et une gestion efficace.

Dans ce contexte, la plateforme expérimentale INSHARE (pour INtegrating and Sharing Health dAta for Research) s'est positionnée rapidement comme une preuve de concept visant à explorer comment les technologies récentes pouvaient répondre à ces défis. Sous l'impulsion du Pr Marc Cuggia et de l'équipe projet « Données massives en Santé » du CHU de Rennes, la démarche d'INSHARE a été de réunir les acteurs (fournisseurs et utilisateurs des données) afin de définir, à partir de scénarios réels, les conditions favorables à la ré exploitation efficiente des DMS pour la recherche médicale. Le consortium de ce projet sur 3 ans a regroupé des équipes spécialisées en informatique médicale, biostatistique, épidémiologie, protection des données, technologies du « big data », et plusieurs fournisseurs de données : CHU et registres régionaux ou nationaux (dont le RGCPC). L'objectif était de démontrer la faisabilité et l'intérêt d'une plateforme technologique basée sur les entrepôts de données biomédicales (EDBM), dédiée au partage collaboratif des DMS.

Le programme scientifique s'est articulé autour de 3 axes :

1) La définition de la gouvernance et de la politique de protection des données,
2) La description des spécifications de l'environnement de la plateforme (intégration, qualité, traitement, sécurité, crypto-tatouage des données),
3) La définition de cas d'usage et des besoins des utilisateurs de la plateforme.

Au-delà de la gouvernance et du développement du prototype de la plateforme [53], l'enrichissement des registres épidémiologiques constituait un cas d'usage expérimental pour l'exploitation des DMS. L'objectif était de travailler à l'appariement des données des registres du cancer (preuve de concept établie sur le territoire du Poitou-Charentes) aux données du registre R.E.I.N. (registre national de l'insuffisance rénale chronique terminale déployé depuis 2001 sous l'égide de l'Agence de la Biomédecine), en vue de chercher à alimenter de façon efficiente le statut des patients en insuffisance rénale chronique terminale (IRCT) et traitée par dialyse face à la maladie cancéreuse. Le registre

R.E.I.N. ne renseignant pas obligatoirement les données relatives au cancer, le croisement des registres suggérait de récupérer l'information de façon fiable et stable à partir de l'expertise des registres du cancer (sans retour au dossier médical) tout en renforçant la volumétrie des effectifs disponibles pour les questions de recherche sous-jacentes.

Tous les patients en IRCT qui ont commencé une dialyse en Poitou-Charentes entre 2008 et 2015 ont été inclus (n=1 634). Les données sur le(s) cancer(s) ont été extraites parmi tous les cas incidents de cancer répertoriés sur la même période au sein du RGCPC et appariées aux cas d'IRCT au sein de la plateforme INSHARE. L'étude s'est intéressée à évaluer l'association entre le diabète et le risque de développer un cancer après le début de la dialyse en utilisant le modèle de Fine & Gray et le risque compétitif de décès. Plusieurs études reportaient en effet une augmentation du risque de cancer chez les sujets diabétiques, tandis que certains auteurs observaient un risque de décès par cancer moins élevé chez les patients diabétiques en IRCT que chez ceux qui étaient non-diabétiques [55,56].

Les résultats ont finalement montré que le risque de développer un cancer après le début de la dialyse était plus faible chez les patients diabétiques dialysés que chez les patients non diabétiques. En outre, par rapport à la population générale, le risque de cancer était plus élevé chez les patients dialysés non diabétiques, mais pas chez ceux qui étaient diabétiques [54].

D'un point de vue technique, cette étude a permis de fournir l'un des cas d'usage pour tester le prototype de la plateforme INSHARE. Cette plateforme a été la première autorisée en France à intégrer et à partager en toute sécurité des données multi sources et des données de santé multi-échelles à des fins de recherche. En termes de perspectives, cette plateforme ouvrait la possibilité d'être utilisé pour réaliser des études similaires dans d'autres régions ou pour enrichir les questions de recherche, avec des informations supplémentaires telles que les consommations individuelles de médicaments ou les demandes de remboursement provenant du Système National des Données de Santé (SNDS) auxquelles les registres du cancer sont particulièrement demandeurs.

Bouzillé G, Westerlynck R, Defossez G, Bouslimi D, Bayat S, Riou C, Busnel Y, Le Guillou C, Cauvin JM, Jacquelinet C, Pladys P, Oger E, Stindel E, Ingrand P, Coatrieux G, Cuggia M. Sharing Health Big Data for Research - A Design by Use Cases: The INSHARE Platform Approach. ***Stud Health Technol Inform****. 2017;245:303-307. PMID: 29295104.*

Lire Article 8 *: Pladys A, Defossez G, Lemordant P, Lassalle M, Ingrand P, Jacquelinet C, Riou C, Bouzillé G, Van Hille P, Vigneau C, Cuggia M, Bayat S. Cancer risk in dialyzed patients with and without diabetes.* ***Cancer Epidemiol****. 2020 Apr;65:101689. doi: 10.1016/j.canep.2020.101689. Epub 2020 Feb 29. PMID: 32126508.*

Accéder à la suite « Discussion générale »

Cancer Epidemiology 65 (2020) 101689

Contents lists available at ScienceDirect

Cancer Epidemiology

journal homepage: www.elsevier.com/locate/canep

Cancer risk in dialyzed patients with and without diabetes

Adélaïde Pladys[a], Gautier Defossez[b,c], Pierre Lemordant[d], Mathilde Lassalle[e], Pierre Ingrand[b,c], Christian Jacquelinet[e,f,*], Christine Riou[d], Guillaume Bouzillé[d], Pascal Van Hille[d], Cécile Vigneau[g,h], Marc Cuggia[d], Sahar Bayat[a]

[a] Univ Rennes, EHESP, REPERIS (Recherche en pharmaco-épidémiologie et recours aux soins) – EA 7449, F-35000 Rennes, France
[b] Poitou-Charentes General Cancer Registry, Poitiers University Hospital, University of Poitiers, Poitiers, France
[c] INSERM, CIC1402, Poitiers, France
[d] Univ Rennes, CHU Rennes, INSERM, LTSI – UMR 1099, F-35000 Rennes, France
[e] Renal Epidemiology and Information Network (REIN), Biomedicine Agency, La Plaine Saint-Denis, France
[f] CESP Centre for Research in Epidemiology and Population Health, Inserm UMRS 1018, Univ Versailles-Saint Quentin, Univ Paris-Saclay, Univ Paris Sud, Villejuif, France
[g] University of Rennes 1, INSERM U1085-IRSET, Rennes, France
[h] CHU Pontchaillou, Department of Nephrology, Rennes, France

ARTICLE INFO

Keywords:
Cancer occurrence
Cancer registry
Diabetes
End stage renal disease
INSHARE platform
REIN registry

ABSTRACT

Background: The risk of cancer is higher in patients with renal diseases and diabetes compared with the general population. The aim of this study was to assess in dialyzed patients, the association between diabetes and the risk to develop a cancer after dialysis start.

Methods: All patients who started dialysis in the French region of Poitou-Charentes between 2008 and 2015 were included. Their baseline characteristics were extracted from the French Renal Epidemiology and Information Network and were linked to data relative to cancer occurrence from the Poitou-Charentes General Cancer Registry using a procedure developed by the INSHARE platform. The association between diabetes and the risk of cancer was assessed using the Fine & Gray model that takes into account the competing risk of death.

Results: Among the 1634 patients included, 591 (36.2 %) had diabetes and 91 (5.6 %) patients developed a cancer (n = 24 before or at dialysis start, and n = 67 after dialysis start). The risk to develop a cancer after dialysis initiation was lower in dialyzed patients with diabetes than without diabetes (SHR = 0.54; 95 %CI: 0.32-0.91). Moreover, compared with the general population, the cancer risk was higher in dialyzed patients without diabetes, but not in those with diabetes.

Conclusion: The risk of developing a cancer in the region of Poitou-Charentes is higher in dialyzed patients without diabetes than with diabetes.

1. Introduction

Type 2 diabetes is an important risk factor for End Stage Renal Disease (ESRD). In France, 47 % of patients who started a Renal Replacement Therapy (RRT) in 2017 were diabetic, and diabetes was the second cause of ESRD [1]. A recent French study showed that the current increase in the standardized incidence rate of RRT is mainly due to type 2 diabetes-related ESRD [2].

Beside diabetes, patients with ESRD often present comorbidities, particularly cardiovascular diseases, and are at higher risk to develop cancer compared with the general population [3–8]. The higher incidence of cancer in patients with ESRD is explained by pro-tumor factors directly or indirectly associated with ESRD and the treatment regimens. In dialyzed patients, cancer concerns frequently the urinary system [4,5,7–10]. For instance, Lin et al., recently reported that the risk of kidney (SHR = 40.3; 95 %CI: 13.4–121.8), bladder (SHR = 41.95; 95 %CI: 25.1–70.1) and upper urinary tract (SHR = 61.3; 95 %CI: 18.6–202.5) cancer is higher in dialyzed patients than in controls who never received dialysis [7]. Indeed, urinary system organs are specifically affected by the nature of the disease that led to ESRD [primary kidney disease, acquired cystic kidney disease or associated urological abnormalities] and this could contribute to cancer development [3,5,10]. In addition, altered DNA repair, impaired immune system function, chronic infection or inflammation, and also chronic immunosuppressive medication intake are factors associated with cancer development [3,6,10,11].

* Corresponding author at: Agence de la biomédecine, 1 avenue du Stade de France. 93212 Saint Denis La Plaine Cedex, France.
E-mail address: christian.jacquelinet@biomedecine.fr (C. Jacquelinet).

https://doi.org/10.1016/j.canep.2020.101689
Received 13 December 2019; Received in revised form 14 February 2020; Accepted 17 February 2020
Available online 29 February 2020
1877-7821/

In the general population, diabetes increases the cancer risk [11–14]. A Sweden study showed that type 2 diabetes is associated with higher standardized incidence rates of cancer in pancreas and liver [14], two organs affected by the metabolic alterations linked to diabetes [15]. Hyperinsulinemia, increased oxidative stress and insulin resistance as well as stimulation of insulin-like growth factors, due to pancreas function alteration, have carcinogenic effects [12–15]. Several factors common to both diseases could also explain the association between diabetes and cancer, such as obesity, hyperglycemia [14,15] and sedentary lifestyle [14]. However, the association between antidiabetic drugs and cancer is still debated. Indeed, it has been observed that insulin is associated with an increased cancer risk, whereas metformin may reduce cancer incidence [14–17], by decreasing tumor cell proliferation.

Although ESRD and diabetes are independently associated with higher cancer risk, their combined presence does not further increase the overall risk of cancer in comparison to each disease on its own [11]. On the other hand, in France, a large national cohort study observed that among dialyzed patients with ESRD, the risk of death by cancer was 30 % lower in patients with diabetes than without diabetes [18]. Consequently, a pilot study analyzed the medical records of dialyzed patients in Bretagne, a French region, during the 2002–2007 period to evaluate the potential association between diabetes and the incidence of specific malignancies in ESRD [19]. Unfortunately, there were too few cancer events to detect a significant association between diabetes and cancer incidence among dialyzed patients. Therefore, the present study analyzed data from a larger population of dialyzed patients and followed for a longer period (data extracted from the French Renal Epidemiology and Information Network; REIN) after linkage to a regional cancer registry. The objective was to assess the association between diabetes and cancer incidence after dialysis start among patients with ESRD in the Poitou-Charentes region.

2. Material and methods

2.1. Study population

For the present study, all ≥18-year-old incident patients who started dialysis (hemodialysis or peritoneal dialysis) in the French region of Poitou-Charentes between 2008 and 2015 were included. Poitou-Charentes is located in south-west France and had 1 783 991 inhabitants in 2012 [20]. The patient data were extracted from two registries:

- The REIN registry was established in 2002 and since 2011 covers all French regions. REIN includes all patients with ESRD who start RRT (either dialysis or kidney transplantation) and live in France [21]. Data on patients in the Poitou-Charentes region started to be included in REIN from 2007.
- The Poitou-Charentes General Cancer Registry was put in place at the beginning of 2008, and includes all incident cases of malignant tumor (hematological malignancies and solid tumors except for non-melanoma skin cancers) in subjects who reside regularly in the Poitou-Charentes region at the time of diagnosis, and in compliance with national and international guidelines [22].

To study cancer incidence among patients who initiated dialysis in the Poitou-Charentes region, data from the REIN registry for this region were linked to and integrated with data from the regional cancer registry. The matching procedure was automatically performed by a determinist algorithm using nominative variables that are common to both databases (family name, maiden name, name, sex, and birth date) using the INSHARE (INtegrating and Sharing Health dAta for Research) technologic platform which was developed during this project to facilitate access to health big data and foster collaborative research [23].

Ethical approval was granted by the French National Research Agency. Subjects involved in our study were extracted from the French REIN registry that received the agreement by the CNIL Commission Nationale de l'Information et des Libertés in 2010 agreement number: 903,188 Version 3. All involved subjects received an information leaflet before giving their verbal consent to participate. The ethics committee approved this procedure.

2.2. Data collection

For this study, patient-related data and cancer-related data were extracted from their respective databases.

Patient-related data were from the REIN registry. Three categories of variables were collected: i) sociodemographic data: sex and age; ii) clinical data at dialysis start: primary kidney disease, several comorbidities and biological parameters; and iii) ESRD management: dialysis regimen, vascular access, date of first dialysis, of renal transplantation, and of death.

Data on cancer in dialyzed patients were extracted from the regional cancer registry. For the study, the date of diagnosis, the anatomical site of origin, and the morphology (histology) of the cancer, according to the International Classification of Diseases for Oncology (ICD-O) third edition, were collected.

2.3. Statistical analyses

Baseline characteristics were first described. Then, the characteristics of dialyzed patients with and without diabetes were compared using the Chi-square test.

Time to outcome (i.e., cancer occurrence after dialysis start) was assessed from dialysis start to the date of cancer diagnosis. The date of renal transplantation, date of death, or the endpoint (December 31, 2015) were used for patients without cancer occurrence. History of cancer was defined as a cancer that occurred before dialysis start, an active cancer at baseline (i.e., dialysis start), or a cancer as the origin of the primary renal disease. The data about cancers occurred before 2008 were issued from the REIN registry.

As death was considered a competing event, the association between patient-related data and the outcome of interest (i.e., cancer occurrence after dialysis start) was assessed by using univariate and multivariable Fine and Gray models that take into account the competing event of death. Missing data were handled using the Multiple Imputation by Chained Equations (MICE) approach with ten imputations and five cycles. To assess the association between diabetes and *de novo* cancer and to take into account the confounding effect of history of cancer, a second analysis was performed only among patients who did not have history of cancer.

Standardized Incidence Ratios (SIR) were used to compare the cancer incidence in the study population and in the French general population. SIR were calculated by dividing the observed number of cancer cases in the study population by the expected number of cases that would occur if the cancer incidence rate in the French general population [24] was applied to the study population, using person-years at risk for a given age and sex. Then, 95 % Confidence Intervals (95 %CI) were calculated. The SIR was calculated in function of the sex and diabetes status (yes/no).

Variables with a p-value < 0.20 in univariate models were included in the multivariate models. A p-value < 0.05 was considered statistically significant. Results were reported as Subdistribution Hazard Ratios (SHR) with 95 % CI. All statistical analyses were performed with the STATA 13.1 software.

3. Results

3.1. Patients on dialysis

Between 2008 and 2015, 1648 patients initiated dialysis in the

Table 1
Patients' characteristics at baseline by diabetes status.

	With diabetes n = 591 n (%)	Without diabetes n = 1043 n (%)	p
Sex			0.883
Men	390 (66.3)	692 (66)	
Women	201 (33.7)	351 (34)	
Age (mean ± sd)	71.2 ± 11.2	66.9 ± 16.2	< 0.001
Hemoglobin (g/dl)			0.015
< 10	286 (48.4)	474 (45.4)	
10-12	236 (39.9)	389 (37.3)	
> 12	45 (7.6)	132 (12.7)	
Missing	24 (4.1)	48 (4.6)	
Albumin (g/dl)			0.727
< 30	115 (19.5)	192 (18.4)	
≥ 30	290 (49)	504 (48.3)	
Missing	186 (31.5)	347 (33.3)	
BMI (kg/m²)			< 0.001
< 18.5	6 (1.0)	53 (5.1)	
18.5-23	87 (14.7)	302 (29)	
23-25	69 (11.7)	165 (15.8)	
≥ 25	380 (64.3)	411 (39.4)	
Missing	49 (8.3)	112 (10.7)	
Tobacco			< 0.001
Current/former smoker	257 (43.5)	483 (46.3)	
Non-smoker	257 (43.5)	484 (46.4)	
Missing	77 (13)	76 (7.3)	
Number of cardiovascular disease[a]			< 0.001
0	142 (24)	504 (48.3)	
1	120 (20.3)	217 (20.8)	
≥ 2	329 (55.7)	322 (30.9)	
Respiratory insufficiency			< 0.001
Yes	121 (20.5)	137 (13.1)	
No	465 (78.7)	901 (86.4)	
Missing	5 (0.8)	5 (0.5)	
Hepatic disease			0.198
Yes	14 (2.4)	15 (1.4)	
No	573 (97)	1025 (98.3)	
Missing	4 (0.6)	3 (0.3)	
History of cancer[b]			0.079
Yes	96 (16.2)	206 (19.8)	
No	495 (83.8)	837 (80.2)	
Physical impairment[c]			< 0.001
Yes	103 (17.4)	103 (9.9)	
No	479 (81.1)	936 (89.7)	
Missing	9 (1.5)	4 (0.4)	
First dialysis modality			0.177
Peritoneal Dialysis	72 (12.2)	152 (14.6)	
Hemodialysis	519 (87.8)	891 (85.4)	
Vascular access			0.390
Catheter	219 (37.1)	385 (36.9)	
Arteriovenous fistula	236 (39.9)	411 (39.4)	
Other	64 (10.8)	94 (9.0)	
Missing	72 (12.2)	153 (14.7)	

[a] Cardiovascular diseases: coronary artery disease, peripheral vascular disease, congestive heart failure, arrhythmia, aortic aneurism, and cerebrovascular disease.

[b] History of cancer: cancer occurred before dialysis start, active cancer at baseline, cancer as the cause of primary renal disease.

[c] Physical impairment: physical impairment of ambulation, para- or hemiplegia, blindness, member amputation and mental disability. BMI: Body Mass Index.

Poitou-Charentes region. After the exclusion of patients aged < 18 years and patients with unknown diabetes status (n = 14), 1634 patients were finally included in the study among whom 591 (36.2 %) had diabetes (only 4.6 % had type 1 diabetes) (Supporting Information Figure S1). The median follow up duration was 19,3 months (IQR: 7.7–36.3 months). During the follow up, 243 diabetic patients (41 %) and 330 non-diabetic ones (32 %) died. Forty-three diabetic (7%) and 253 non-diabetic patients (24 %) were transplanted.

The patients' characteristics according to their diabetes status are summarized in Table 1. Compared with patients without diabetes (n = 1043), patients with diabetes (n = 591) were older (71.2 ± 11.2 vs 66.9 ± 16.2 years; p < 0.001) and had more often comorbidities.

3.2. Cancer occurrence

Between 2008 and 2015, 245 patients included in the study had at least one cancer among which 154 were diagnosed before or at dialysis start. Among the 91 (5.6 %) patients who developed an incident cancer after dialysis initiation (Supporting Information Fig. S1), 24 patients had another cancer before or at dialysis start and 67 were considered as having a *de novo* cancer. Among the 91 patients, only 30 had diabetes.

Overall, the most common cancer sites were lung and kidney (13.2 %), followed by multiple myeloma and bladder (9.9 %). The median interval between dialysis start and cancer diagnosis was 13.2 (IQR: 5.9–26.8) months. Cancer sites differed according to the patients' diabetes status (Table 2).

Analysis of the SIR values by sex and diabetes status (Table 3) showed that the risk of cancer occurrence was significantly higher in non-diabetic men (SIR = 1.48; 95 %CI: 1.08–1.89) and non-diabetic women (SIR = 1.81; 95 %CI: 1.03–2.77) compared with men and women in the general population. Conversely, the risk of cancer was comparable in patients with diabetes and in the general population.

3.3. Factors associated with the risk of cancer

Analysis of the association between patient-related data and cancer occurrence after dialysis initiation using univariate and multivariate Fine & Gray models showed that in the unadjusted model (Table 4, left panel), diabetes (SHR = 0.79; 95 %CI: 0.51–1.22) and history of cancer (SHR = 1.51; 95 %CI: 0.94–2.43) were not significantly associated with the risk to develop a cancer during RRT.

In the adjusted model (Table 4, right panel), diabetes, but not history of cancer was associated with a lower risk of cancer after dialysis start (SHR = 0.54; 95 %CI: 0.32-0.91). Conversely, the interaction between diabetes and history of cancer was associated with higher risk to develop a cancer (SHR = 3.29; 95 %CI: 1.24–8.73).

3.4. Factors associated with the risk of de novo cancers

The association between patient-related factors and *de novo* cancer events after dialysis start (n = 67) was assessed using univariate and multivariate Fine & Gray models in patients without history of cancer (n = 1332; Supporting Information Table S1). After adjustment for sex, age and tobacco use, diabetes was associated with a lower risk of *de novo* cancer after dialysis start (SHR = 0.52; 95 %CI: 0.31-0.89).

4. Discussion

This study shows that among patients with ESRD who started dialysis in the French region of Poitou-Charentes between 2008 and 2015, cancer incidence after dialysis initiation was higher in patients without than with diabetes. This result was confirmed also when assessing the occurrence of *de novo* cancers in the subpopulation without history of cancer (i.e., tumors diagnosed before or at dialysis start). Moreover, compared with the general French population, cancer risk was higher in dialyzed patients without diabetes. Conversely, the risk to develop a cancer was similar in patients with ESRD and diabetes and in the general French population.

Our results are close to those by Wong et al., who showed that mild to moderate chronic kidney disease does not increase the risk of cancer in patients with type 2 diabetes [11]. The authors suggested that renal disease and diabetes lead to common pro-tumor factors (chronic inflammation, DNA mutation…) that contribute to the association with cancer, but without additive effects. A previous study showed no significant association between diabetic nephropathy and cancer [6].

Table 2
Sites of cancers occurring after dialysis start.

	Entire n (%)	Median interval between dialysis start and cancer (months)*	With diabetes n (%)	Without diabetes n (%)
Lung	12 (13.2)	17.5 (9.3–34.9)	4 (13.3)	8 (13.1)
Kidney	12 (13.2)	8.4 (2.1–22.8)	5 (16.7)	7 (11.5)
Multiple myeloma	9 (9.9)	9.6 (4.3–20.3)	0	9 (14.7)
Bladder	9 (9.9)	7.8 (4.2–9.4)	6 (20.0)	3 (4.9)
Colon	8 (8.8)	17.1 (11.4-25.8)	4 (13.3)	4 (6.6)
Prostate	6 (6.6)	11.7 (8.8–27.3)	1 (3.3)	5 (8.2)
Stomach	5 (5.5)	5.9 (2.9–6.3)	3 (10.0)	2 (3.3)
Thyroid	5 (5.5)	14.2 (10.5-17.5)	0	5 (8.2)
Rectum	4 (4.4)		1 (3.3)	3 (4.9)
Lymphomas	3 (3.3)		1 (3.3)	2 (3.3)
Breast	2 (2.2)		1 (3.3)	1 (1.6)
Esophagus	2 (2.2)		0	2 (3.3)
Pancreas	2 (2.2)		0	2 (3.3)
Others	12 (13.2)		4 (13.3)	8 (13.1)
Total	**91**	**13.2 (5.9–26.8)**	**10.5 (4.2–26.4)**	**16.5 (6.0–26.8)**

* Median duration and interquartile ranges in months.

Table 3
Standardized incidence ratios for all malignancies, stratified by sex and diabetes status.

	Person-years	O/E	SIR (95 %CI)
Men			
All	2438.5	69/50.72	1.36 (1.06–6.22)
Diabetics	814.5	24/20.35	1.18 (0.76–1.60)
Non-diabetics	1524.0	45/30.37	1.48 (1.08–1.89)
Women			
All	1341.5	22/15.01	1.47 (0.92–6.86)
Diabetics	500.0	6/6.16	0.97 (0.36–1.69)
Non-diabetics	841.5	16/8.85	1.81 (1.03–2.77)

O: Observed number of cancer cases; E: expected number of cancer cases, based on the general population in France; SIR: Standardized Incidence Ratio (ratio of observed number of cancer cases in our study by the expected number of cases that would occur if the cancer incidence rate in the French general population).

Conversely, another work found that the incidence of any cancer is higher in patients with primary ESRD not caused by diabetes [25]. A recent study in Taiwan also reported that the risk of new cancers after dialysis start is lower in patients with ESRD and diabetes (HR = 0.74; 95 %CI: 0.67-0.81) than in those without diabetes [26].

Our analysis also showed that the cancer site distribution differed according to the diabetes status. Indeed, cancers of the urinary system (kidney: 16.7 %, and bladder: 20 %) were more common in patients with diabetes, whereas multiple myeloma was more frequent in patients without diabetes. In our study, only one patient with diabetes (3.3 %) developed prostate cancer (n = 5, 8.2 %, in the group without diabetes). A previous meta-analysis showed that the risk of developing prostate cancer is significantly lower in diabetic men [27], possibly due to hormonal changes (decreased insulin or testosterone levels) that may have growth inhibitory effects on prostate cancer cells. Moreover, as diabetic men are more likely to be screened for prostate-specific antigen, this cancer should be more easily diagnosed compared with non-diabetic men.

Previous studies showed that metformin therapy is associated with a reduced risk of cancer [17,28], including colon-rectal cancer, among patients without diabetes [29]. Metformin activates the AMP-activated protein kinase (AMPK) signaling pathway that induces cell cycle arrest and apoptosis of myeloma cells [17,28,29]. We could hypothesize that patients with diabetes in our study could have been treated with metformin before reaching stage 4 kidney disease, and this might have had a protective role against cancer cell proliferation/growth. As recent KDIGO guidelines recommend that metformin should not be used in patients on dialysis [30], dialyzed patients with diabetes are often treated with insulin that is associated with higher cancer risk [31,32].

Table 4
Association of patient-related data with cancer occurrence after dialysis start (univariate and multivariate Fine & Gray models).

	Univariate Fine & Gray model SHR (95 %CI)	Multivariate Fine & Gray model SHR (95 %CI)
Sex (vs Men)		
Women	0.61 (0.38-0.98)	0.78 (0.47–1.31)
Age (vs 40–59 years)		
18-39	0.69 (0.16–2.99)	0.63 (0.15–2.70)
60-79	1.37 (0.75–2.50)	1.46 (0.79–2.73)
≥80	0.76 (0.37–1.55)	0.88 (0.42–1.87)
Tobacco (vs Non-smoker)		
Current/former smoker	1.93 (1.22–3.05)	1.79 (1.07–3.0)
Hemoglobin (vs 10 – 12 g/dl)		
< 10	1.24 (0.77–1.98)	n/a
> 12	0.82 (0.37–1.80)	n/a
BMI (vs 23 – 25 kg/m²)		
< 18.5	1.49 (0.59–3.73)	n/a
18.5-23	0.91 (0.47–1.75)	n/a
≥25	0.86 (0.48–1.53)	n/a
Diabetes (vs No)		
Yes	0.79 (0.51–1.22)	0.54 (0.32-0.91)
History of cancer* (vs No)		
Yes	1.51 (0.94–2.43)	0.98 (0.53–1.82)
Diabetes + History of cancer (vs No, No)		
Yes, yes	n/a	3.29 (1.24-8.73)
Hepatic disease (vs No)		
Yes	0.53 (0.07–3.80)	n/a
Respiratory insufficiency (vs No)		
Yes	1.41 (0.86–2.31)	n/a
Physical impairment (vs No)		
Yes	0.35 (0.14-0.86)	n/a
Cardiovascular disease (vs 0)		
1	1.03 (0.58–1.83)	n/a
≥2	1.19 (0.74–1.91)	n/a

* History of cancer: Cancer before or at dialysis start; BMI: Body Mass Index; SHR: Subdistribution Hazard Ratio; 95 %CI: 95 % Confidence Interval.

However, as renal function impairment leads to reduced insulin resistance and insulin clearance [33], patients with diabetes on dialysis need less insulin than patients with normal kidney function. Consequently, lower insulin doses might also contribute to reduce the cancer risk among dialyzed patients with diabetes.

Moreover, the French recommendations suggest that all patients with ESRD and hypertension or albuminuria should be treated with Angiotensin-Receptor Blockers (ARBs), particularly if they have diabetes [34]. It has been shown that renin-angiotensin system inhibitors do not increase the risk of cancer development [35], or have a

protective effect against cancer development [36–38]. In our population, patients with diabetes also had frequently hypertension, therefore they might have been taking ARBs, with an additional anti-tumor effect. Nevertheless, as medications are not recorded in the REIN registry, we could not include them in our analysis.

Finally, patients with diabetes, which is often characterized by the presence of several comorbidities (i.e., cardiovascular diseases, respiratory insufficiency, obesity…), might have had a closer medical follow-up than patients with fewer comorbidities, and consequently more targeted preventive treatments and medical examinations. Otherwise, diabetic patients with chronic kidney disease could develop cancer and die before the terminal stage of the kidney disease; or the association of cancer and diabetes could be considered as a barrier to dialysis initiation. This may explain the lower rate of cancer in patients with diabetes than in those without diabetes at the moment of their inclusion in the REIN registry.

This study has several strengths. We studied cancer incidence by taking into account the competing risk of death to avoid over-evaluating the risk of death for patients with diabetes before cancer development. From a technical point of view, this study provided one of the use cases to develop the INSHARE platform. This platform is the first authorized in France to securely integrate and share multisource and multiscale health data for research purposes. In our study, INSHARE facilitated the linkage procedure to merge REIN data to data relative to cancer occurrence and characteristics and made easier the data integration for statistics analysis. In terms of perspectives, the platform can be used to perform similar studies in other regions or to enrich the study with supplementary information such as individual drug consumptions or reimbursed claims coming from the French National Health Data System [39].

Our study has also limitations. As the cancer registry of the Poitou-Charentes region started only in 2008, we could not assess the cancer history of patients with the same method before and after 2008. Because our study was conducted on a specific French region and relatively low numbers of patients during a limited time period, our results can't be generalized to the entire population. Further studied are needed in order to confirm our results. Moreover, unmeasured confounding variables might influence the results. For example, we did not have data on treatments and therefore we could not study the association between drugs and cancer development. Finally, the results of the Fine and Gray model might be influenced by unmeasured confounders of the effect of the diabetes on cancer and on death.

5. Conclusions

Our study showed that the risk of developing a cancer, including *de novo* cancer, is higher in dialyzed patients without diabetes than with diabetes. Moreover, in comparison to the general population, the risk of developing a cancer was higher in non-diabetic than in diabetic dialyzed patients.

6. Authorship contribution

AP, GD, PI, CJ, MC, SB contributed to the study conception and design, and provided general support to the study. GD, MS contributed to data acquisition. GD, PL, MS, GB, PVH controlled the data and algorithms qualities. PL performed the linkage procedure to merge data from the REIN and national Cancer registry. AP performed statistical analyses and AP, GD, PI, CJ, CR, MC, SB interpreted results. AP wrote the main body of this original article. BS and SG: contributed to data collection and provided general support to the study. And finally, all authors helped to revise the manuscript and approved the final manuscript for publication.

Funding sources

French National Research Agency funded this work inside the Integrating and Sharing Health Data for Research Project (Grant No. ANR-15-CE19-0024).

CRediT authorship contribution statement

Adélaïde Pladys: Conceptualization, Project administration, Formal analysis, Supervision, Validation, Resources, Writing - original draft, Writing - review and editing. **Gautier Defossez:** Data curation, Supervision, Validation, Resources, Funding acquisition, Writing - original draft, Writing - review and editing. **Pierre Lemordant:** Data curation, Software. **Mathilde Lassalle:** Data curation. **Pierre Ingrand:** Supervision, Validation, Resources, Writing - original draft, Writing - review and editing. **Christian Jacquelinet:** Conceptualization, Project administration, Supervision, Validation, Resources. **Christine Riou:** Conceptualization, Project administration, Supervision, Validation, Resources. **Guillaume Bouzillé:** Data curation. **Pascal Van Hille:** Data curation. **Cécile Vigneau:** Writing - original draft, Writing - review and editing. **Marc Cuggia:** Conceptualization, Project administration, Supervision, Validation, Resources, Funding acquisition. **Sahar Bayat:** Conceptualization, Project administration, Supervision, Validation, Resources, Funding acquisition, Writing - original draft, Writing - review and editing.

Declaration of Competing Interest

All authors declare that there is no conflict of interest regarding the publication of this article.

Acknowledgments

We thank Elisabetta Andermarcher for English revisions. We thank the French National Research Agency for funding this work inside the Integrating and Sharing Health Data for Research Project (Grant No. ANR-15-CE19-0024). We gratefully acknowledge all REIN registry and Poitou-Charentes cancer registry participants. The centers participating in the registry are listed in the REIN annual report (2017): https://www.agence-biomedecine.fr/IMG/pdf/rapport_rein_2017_v3.pdf.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.canep.2020.101689.

References

[1] REIN annual report 2017. https://www.agence-biomedecine.fr/IMG/pdf/rapport_rein_2017_v3.pdf.
[2] C. Vigneau, A. Kolko, B. Stengel, et al., Ten-years trends in renal replacement therapy for end-stage renal disease in mainland France: lessons from the French Renal Epidemiology and Information Network (REIN) registry, Nephrol. Ther. 13 (2017) 228–235.
[3] P. Maisonneuve, L. Agodoa, R. Gellert, et al., Cancer in patients on dialysis for end-stage renal disease: an international collaborative study, Lancet 354 (1999) 93–99.
[4] J.H. Stewart, G. Buccianti, L. Agodoa, et al., Cancers of the kidney and urinary tract in patients on Dialysis for end-stage renal disease: analysis of data from the United States, Europe, and Australia and New Zealand, J. Am. Soc. Nephrol. 14 (2003) 197–207.
[5] J.H. Stewart, C.M. Vajdic, L.T. Van Leeuwen, et al., The pattern of excess cancer in dialysis and transplantation, Nephrol. Dial. Transplant. 24 (2009) 3225–3231.
[6] C.Y.C. Cheung, G.C.W. Chan, S.K. Chan, et al., Cancer incidence and mortality in chronic Dialysis population: a multicenter cohort study, Am. J. Nephrol. 43 (2016) 153–159.
[7] M.Y. Lin, M.C. Kuo, C.C. Hung, et al., Association of Dialysis with the risks of cancers, PLoS One 10 (4) (2015) e0122856.
[8] H.F. Lin, Y.H. Li, C.H. Wang, C.L. Chou, D.J. Kuo, T.C. Fang, Increased risk of cancer in chronic dialysis patients: a population-based cohort study in Taiwan, Nephrol. Dial. Transplant. 27 (2012) 1585–1590.
[9] Y.C. Lee, S.Y. Hung, H.K. Wang, et al., Is there different risk of cancer among end-

stage renal disease patients undergoing hemodialysis and peritoneal dialysis? Cancer Med. 7 (2) (2018) 485–498.
[10] S. Vamvakas, U. Bahnerb, A. Heidland, Cancer in end-stage renal disease: potential factors involved, Am. J. Nephrol. 18 (1998) 89–95.
[11] G. Wong, S. Zoungas, S. Lo, et al., The risk of cancer in people with diabetes and chronic kidney disease, Nephrol. Dial. Transplant. 27 (2012) 3337–3344.
[12] V. Donadon, M. Balbi, P. Casarin, A. Vario, A. Alberti, Association between hepatocellular carcinoma and type 2 diabetes mellitus in Italy: potential role of insulin, World J. Gastroenterol. 14 (37) (2008) 5695–5700.
[13] S.S. Coughlin, E.E. Calle, L.R. Teras, J. Petrelli, M.J. Thun, Diabetes mellitus as a predictor of cancer mortality in a large cohort of US adults, Am. J. Epidemiol. 159 (2004) 1160–1167.
[14] X. Liu, K. Hemminki, A. Försti, K. Sundquist, J. Sundquist, J. Ji, Cancer risk in patients with type 2 diabetes mellitus and their relatives, Int. J. Cancer 137 (2015) 903–910.
[15] P. Vigneri, F. Frasca, L. Sciacca, G. Pandini, Vigneri R. Diabetes and cancer, Endocr. Relat. Cancer 16 (2009) 1103–1123.
[16] G. Shlomai, B. Neel, D. Le Roith, E.J. Gallagher, Type 2 diabetes mellitus and Cancer: the role of pharmacotherapy, J. Clin. Oncol. 34 (2016) 4261–4269.
[17] C.J. Currie, C.D. Poole, E.A.M. Gale, The influence of glucose-lowering therapies on cancer risk in type 2 diabetes, Diabetologia 52 (2009) 1766–1777.
[18] A. Pladys, C. Couchoud, A. LeGuillou, M. Siebert, C. Vigneau, S. Bayat, Type 1 and Type 2 Diabetes and Cancer mortality in the 2002-2009 Cohort of 39 811 French Dialyzed Patients, PLoS One 10 (5) (2015) e0125089.
[19] A. Le Guillou, A. Pladys, W. Kihal, et al., Is cancer incidence different between type 2 diabetes patients compared to nondiabetics in hemodialysis? A study from the REIN registry? Nephrol. Ther. 14 (2018) 142–147.
[20] Institut national de la statistique et des études économiques (Insee). Insee Analyses Poitou-Charentes 2014, n°08. www.insee.fr.
[21] C. Couchoud, B. Stengel, P. Landais, et al., The renal epidemiology and information network (REIN): new registry for endstage renal disease in France, Nephrol. Dial. Transplant. 21 (2006) 411.
[22] O.M. Jensen, D.M. Parkin, R. MacLennan, C.S. Muir, R.G. Skeet, Cancer Registration: Principles and Methods, IARC, Lyon, 1991.
[23] G. Bouzillé, R. Westerlynck, G. Defossez, et al., Sharing health big data for research - a design by use cases: the INSHARE platform approach, Stud. Health Technol. Inform. 245 (2017) 303–307.
[24] F. Binder-Foucard, A. Belot, P. Delafosse, L. Remontet, A.S. Woronoff, N. Bossard, Estimation Nationale De l'incidence Et De La Mortalité Par Cancer En France Entre 1980 Et 2012. Partie 1 – Tumeurs Solides, Institut de veille sanitaire, Saint-Maurice (Fra), 2013 122p.
[25] A.M. Butler, A.F. Olshan, A.V. Kshirsagar, et al., Cancer incidence among US medicare ESRD patients receiving hemodialysis, 1996-2009, Am. J. Kidney Dis. 65 (5) (2015) 763–772.
[26] C.C. Chien, M.M. Han, Y.H. Chiu, et al., Epidemiology of cancer in end-stage renal disease dialysis patients: a national cohort study in Taiwan, J. Cancer 8 (1) (2017) 9–18.
[27] J.S. Kasper, E. Giovannucci, A meta-analysis of diabetes mellitus and the risk of prostate Cancer, Cancer Epidemiol. Biomarkers Prev. 15 (11) (2006) 2056–2062.
[28] G. Libby, L.A. Donnelly, P.T. Donnan, D.R. Alessi, A.D. Morris, J.M. Evans, New users of metformin are at low risk of incident cancer: a cohort study among people with type 2 diabetes, Diabetes Care 32 (9) (2009) 1620–1625.
[29] K. Hosono, H. Endo, H. Takahashi, et al., Metformin suppresses colorectal aberrant crypt foci in a short-term clinical trial, Cancer Prev. Res. Phila. (Phila) 3 (9) (2010) 1077–1083.
[30] KDIGO, Clinical practice guideline for the evaluation and management of chronic kidney disease, Kidney Int. Suppl. 2013 3 (1) (2012) 1–163.
[31] V. Donadon, M. Balb, P. Casarin, A. Vario, A. Alberti, Association between hepatocellular carcinoma and type 2 diabetes mellitus in Italy: potential role of insulin, World J. Gastroenterol. 14 (37) (2008) 5695–5700.
[32] S.S. Coughlin, E.E. Calle, L.R. Teras, J. Jennifer Petrelli, M.J. Thun, Diabetes mellitus as a predictor of Cancer mortality in a large cohort of US adults, Am. J. Epidemiol. 159 (2004) 1160–1167.
[33] C.M. Rhee, A.M. Leung, C.P. Kovesdy, K.E. Lynch, G.A. Brent, K. Kalantar-Zadeh, Updates on the management of diabetes in dialysis patients, Semin. Dial. 27 (2) (2014) 135–145.
[34] Haute Autorité de Santé, Guide Du Parcours De Soins – Maladie Rénale Chronique De l'adulte, HAS, Saint-Denis La Plaine, 2012https://www.has-sante.fr/portail/upload/docs/application/pdf/2012-04/guide_parcours_de_soins_mrc_web.pdf.
[35] T. Datzmann, S. Fuchs, D. Andree, B. Hohensteinc, J. Schmitte, C. Schindler, Systematic review and meta-analysis of randomised controlled clinical trial evidence refutes relationship between pharmacotherapy with angiotensin receptor blockers and an increased risk of cancer, Eur. J. Intern. Med. (2019), https://doi.org/10.1016/j.ejim.2019.04.019.
[36] C.C. Huang, W.L. Chan, Y.C. Chen, et al., Angiotensin II receptor blockers and risk of cancer in patients with systemic hypertension, Am. J. Cardiol. 107 (7) (2011) 1028–1033.
[37] C.H. Chang, J.W. Lin, L.C. Wu, M.S. Lai, Angiotensin receptor blockade and risk of cancer in type 2 diabetes mellitus: a nationwide case-control study, J. Clin. Oncol. 29 (22) (2011) 3001–3007.
[38] G. Verhoest, T. Dolley-Hitze, F. Jouan, et al., Sunitinib combined with Angiotensin-2 Type-1 receptor antagonists induces more necrosis: a murine xenograft model of renal cell carcinoma, Biomed Res. Int. (2014), https://doi.org/10.1155/2014/901371.
[39] P. Tuppin, J. Rudant, P. Constantinou, C. Gastaldi-Ménager, et al., Value of a national administrative database to guide public decisions: from the système national d'information interrégimes de l'Assurance maladie (SNIIRAM) to the système national Des Données De Santé (SNDS) in France, Rev. Epidemiol. Sante Publique 65 (Suppl. 4) (2017) S149–67, https://doi.org/10.1016/j.respe.2017.05.004.

4. Discussion générale

Ce travail présente le développement et les nombreuses applications d'un modèle de registre fondé sur la réutilisation de données médicales standardisées, et visant à répondre de façon optimisée aux objectifs de surveillance épidémiologique et d'évaluation des parcours de soins. L'enjeu était de soutenir un modèle de registre de nouvelle génération, proposant des gains d'efficience au regard des avancées technologiques récentes, et ayant vocation à accroître les domaines d'application et possibilités d'utilisation des données.

Le RGCPC est un registre récent, qui s'est développé dans un contexte local favorable et une excellente participation et adhésion des partenaires. L'informatisation des données médicales et l'interopérabilité croissante des systèmes d'information étaient propices à de tels développements au moment de la création du RGCPC en 2007. Le développement du SI a été un processus long qui s'est étalé sur plusieurs années, et qui s'est basé sur une réflexion préalable sur les avantages et les contraintes des différentes méthodes développées à l'international, et adapté au contexte politique et juridique français.

Il était important de trouver un équilibre entre les ressources disponibles et la productivité du registre, tout en satisfaisant un niveau très élevé d'exhaustivité, de qualité, de comparabilité et de disponibilité des données. Si le rôle fondamental des registres de population au niveau international est acquis et n'est plus à remettre en cause, avec un vaste éventail d'actions proposées [1], c'est le type de données que les registres de population peuvent fournir qui va déterminer en premier lieu l'étendue de leur rôle et de leur fonction. Ainsi mettre en place un recueil très large sur des données de faible qualité rendait leur utilisation limitée. En contrepartie, avoir un ensemble minimal de données pour faire de l'incidence et de la survie revenait à mobiliser en amont, à des fins d'exhaustivité, un maximum de données [9], et il devenait dès lors implicite de vouloir élargir les données colligées et enregistrées pour garantir cette première finalité. Dans ce contexte, l'efficience du registre se trouvait intimement liée à la disponibilité des données et à la façon de les traiter.

4.1. Automatisation du traitement de l'information

La première partie de ce travail démontre l'intérêt d'une automatisation généralisée à toute la chaîne de production des données, nécessaire pour limiter les coûts en routine et raccourcir les délais de production. Le paradigme global était de parvenir à décloisonner et à mettre en relation les bases de données de santé appropriées pour servir les objectifs de surveillance, d'évaluation et de recherche du registre.

4.1.1. Standardisation des sources de données

La standardisation des sources de données, impliquant pour certaines d'entre elles la conversion des données vers un format standard commun et normalisé, a été un prérequis indispensable à la mise en œuvre du traitement des données. Les sources d'information produisent en effet le plus souvent des données riches et structurées, mais très hétérogènes et utilisant différentes terminologies médicales (ADICAP, SNOMED, CIM-O3, CIM-10). Les efforts initiaux consentis auprès de chaque source de données (partenaires du registre) pour circonscrire les différents systèmes et logiciels métiers ont été un gage de réussite et de pérennité. Les exports mis en place ont permis de recueillir des données hautement structurées et selon un format uniforme selon les sources, facilitant dès lors grandement leur intégration dans le SI du registre. Notons toutefois l'importance dans ce contexte de maintenir une veille et une maintenance adaptative aux évolutions de classification et des logiciels métiers des structures partenaires.

Le déploiement des modules d'exports auprès des pathologistes a contribué à accéder efficacement aux comptes rendus d'anatomopathologie (CRAP) sans mettre en œuvre incidemment un retour au dossier médical, représentant un gain d'efficience majeur. Cet objectif était rendu possible toutefois par l'effectivité et la qualité du codage prodigué en amont par les pathologistes, et soulevait l'intérêt bipartite de mettre en œuvre une politique « qualité » impliquant des démarches continues d'évaluation et d'amélioration des pratiques professionnelles. Le codage du (ou des) diagnostic(s) lésionnel(s) au sein de chaque structure ACP étant un préalable nécessaire à l'exploitation des données, des audits internes ou externes doivent permettre de s'assurer continuellement de la qualité des données ACP produites, leur mise à jour et la conformité au référentiel rendant possible leur transmission (pour la coordination des soins, la santé publique et la recherche). Il y aurait tout intérêt à ce que les registres du cancer et les pathologistes travaillent de façon concertée à améliorer la saisie et l'exploitation des informations et la qualité du rendu diagnostique, pour optimiser la prise en charge et le service rendu aux patients, et renforcer leur rôle synergique dans l'évaluation des politiques de lutte contre le cancer. C'est le rôle engagé notamment dans certaines régions de France par les CRISAP (Centre de Regroupement Informatique et Statistique en Anatomie Pathologique) et l'AFAQAP (Association Française d'Assurance Qualité en Anatomie et Cytologie Pathologiques) [57–59]. Au-delà, les pathologistes sont également des acteurs obligatoires des campagnes de dépistage (le codage prenant une part importante pour faciliter le travail en aval des structures de gestion), et leurs informations sont de plus en plus fréquemment versées dans le dossier communicant de cancérologie (dans le cadre du dossier médical partagé), questionnant l'efficience du modèle au regard de leurs sollicitations répétées et l'intérêt qu'il y aurait

à structurer les flux et le partage des données médicales au sein de plateformes de données, en vue d'économiser les circuits d'informations et d'améliorer l'efficience des données exploitées.

Le contexte local peut parfois révéler aussi certains freins à la mise en œuvre de ce type d'export, avec la réticence de certains praticiens à transmettre les données nominatives, ou des difficultés techniques qui peuvent être liées à la structuration des données (comme l'utilisation partielle du codage ADICAP ou l'absence de certaines variables indicatives comme le code postal de résidence). Dans ce contexte, des solutions alternatives existent, avec des initiatives croissantes de machine learning en France [32,60] comme dans de nombreux autres pays [61–63], et qui demeurent sous-exploitées. La recherche des cas et des informations représente un coût important pour les registres, et ces initiatives méritent d'être renforcées afin de substituer autant que possible ces méthodes opératoires au travail manuel de récupération de l'information.

La collecte des données PMSI était aussi un enjeu majeur à la mise en œuvre des approches algorithmiques engagées sur les données. Parce que les patients, les actes qu'ils subissent, ainsi que les structures médicales visitées y sont recensées, le PMSI permet l'étude exhaustive de la fréquentation des structures de soins publiques et privées et propose des « marqueurs » des déplacements géographiques de chaque patient. De plus, l'enregistrement des actes médicotechniques (chirurgical ou non) dans le résumé d'unité médicale vient renforcer la qualification du motif d'hospitalisation et permet de mieux appréhender les limites inhérentes aux pratiques et aux disparités de codage. L'adaptation récente de la méthode de quantification de l'activité soumise à seuil dans le cadre des dispositions d'autorisations spécifiques au traitement du cancer[11] atteste de l'intérêt d'une méthode fondée sur les actes [64]. L'avantage pour la collecte des données PMSI était que le RGCPC pouvait s'appuyer sur le format national structuré des fichiers RSS[12] (résumé de sortie standardisée) produits par le logiciel groupeur de chaque établissement de santé public et privé (fichiers RSS utilisés pour le calcul des ressources des établissements de santé depuis 2004 dans le cadre de la tarification à l'activité). La procédure visait ensuite à rendre nominatif les RSS (ajout du nom, prénom et nom de jeune fille) par le biais des services d'information médicale, complétés des données de résidence et des communes de naissances. Ce format d'export unique a permis ainsi de garantir l'accès aux données d'hospitalisation complètes incluant les actes CCAM, et simplifiant nettement leur implémentation (un programme unique) dans le SI-RGCPC. A noter que les données PMSI qui alimentent aujourd'hui les plateformes de données de santé (Health Data Hub, SNDS) sont les sous-

[11] CIRCULAIRE N° DHOS/O/INCa/2008/101 du 26 mars 2008 relative à la méthodologie de mesure des seuils de certaines activités de soins de traitement du cancer

[12] « Un résumé de sortie standardisé (RSS) est constitué de l'ensemble des résumé d'unité médicale (RUM) relatifs au même séjour hospitalier d'un malade dans le secteur MCO (médecine, chirurgie, obstétrique). Un RUM est produit à la fin de chaque séjour de malade dans une unité médicale assurant des soins de MCO, quel que soit le mode de sortie de cette unité. Le RUM contient un nombre limité d'informations d'ordre administratif et médical, qui doivent être systématiquement renseignées et codées selon des nomenclatures et des classifications standardisées, afin de bénéficier d'un traitement automatisé » (https://www.atih.sante.fr/glossaire)

produits (RSA : Résumé de sortie anonymisé) d'un processus automatique de pseudonymisation des RSS réalisé par le logiciel national Genrsa, questionnant à nouveau l'efficience du modèle au regard des efforts disproportionnés engagés par le RGCPC pour accéder régulièrement à ces données.

4.1.2. Interopérabilité

Au-delà de la collecte, l'enjeu était de rendre ces sources et leurs terminologies interopérables sur le plan sémantique pour souscrire aux traitements algorithmiques ultérieurs. La gestion de grands volumes de données réclamait un système d'interconnexion (couplage) des enregistrements performant en l'absence de clé d'identification commune. Le couplage déterministe, associé à une standardisation en amont des données sur les champs d'identification clés (nom, prénom, sexe, date de naissance, nom de jeune fille) [65] était propice à rassembler efficacement de grands volumes de données, tout en restant vigilant sur le risque de regroupement abusif en cas d'homonymie (le taux d'erreur d'homonymie augmentant lui-même de façon linéaire avec le nombre d'enregistrements [66]). Toutefois, une intervention humaine étant établie sur chaque dossier au moment de la visualisation et de l'enregistrement des tumeurs, elle permet en toute vraisemblance de repérer d'éventuelles discordances et de les corriger[13]. Au-delà du couplage déterministe très sensible à la qualité des données qui lui sont soumises, les méthodes de couplage probabiliste implémentées dans le SI-RGCPC se fondent sur des méthodes publiées (algorithme phonétique, mesures de dissimilarité) [67,68]. L'enjeu était de trouver le meilleur compromis entre les parts des doublons détectés et des doublons avérés, ceux-ci étant tous soumis à une revue manuelle (pas d'appariement probabiliste automatique). S'agissant de 13% des patients qui font l'objet d'un regroupement manuel[14], et considérant la part des doublons résiduels extrêmement faible (< à 1%) à l'issue de l'intervention humaine et des contrôles effectuées en bout de chaîne sur les tumeurs incidentes, le système d'interconnexion des enregistrements est fiable et performant, même s'il reste toujours perfectible [69].

4.1.3. Approches algorithmiques

La plus-value majeure du SI-RGCPC repose surtout sur le gain apporté par les approches algorithmiques. Une fois en effet que le registre s'est attaché à avoir accès à un nombre suffisant de sources d'informations et qu'il a chaîné efficacement toutes les données utiles concernant un individu, l'enjeu était de rendre l'information accessible et de mettre le personnel du registre en configuration optimale pour l'enregistrement. Les choix opérés étaient guidés par le besoin de simplifier, autant

[13] Trois homonymes dégroupés pour 350.000 individus dans le SI-RGCPC
[14] Données calculées sur 10 ans d'enregistrement (2008-2017), non fournies dans le rapport

que possible, le processus manuel d'interprétation, d'exploitation et de saisie engagé sur chaque dossier par le personnel du registre.

- **Notification des tumeurs**

L'algorithme de notification des tumeurs répond à un premier niveau de simplification des tâches. La notification assiste l'utilisateur en l'affranchissant de la saisie des variables obligatoires et en lui faisant la(e) meilleure(s) proposition(s) de tumeur(s) au regard des données disponibles. Elle permet en outre une planification intelligente des tâches opérationnelles selon les règles et prérequis internes et la base des variables notifiées (*e. g.* sélection des dossiers par utilisateur, lieu de résidence, établissement, territoire de santé). La qualité des résultats produits par l'algorithme dépendant singulièrement de la collecte et de la complétude des données, le point majeur est le maintien d'un flux élevé et régulier d'informations en provenance des partenaires. Ceci est d'autant plus désiré que l'approche sélective propose une approche combinée (tumeur composite) pour améliorer la précision de l'algorithme. Le choix a été toutefois de garantir la qualité des données enregistrées par une revue manuelle systématique des dossiers notifiés, tandis que certaines initiatives internationales proposent de réaliser un enregistrement automatique [22–25,27,28]. Le but était de se préserver des difficultés liées à certains cas d'usage particuliers et d'augmenter l'amplitude d'enregistrement du set minimal de données (grade, stade d'extension, traitement) pour un temps passé limité. Le temps dévolu à chaque dossier va toutefois varier de façon importante selon la (ou les) localisation(s) cancéreuse(s). Il est fréquent par exemple de réunir un niveau d'information élevé (disponibilité des pièces diagnostiques, des RCP, de l'intégralité du plan de traitement, des marqueurs de surveillance ou de récidives) pour des localisations dont la prise en charge est standardisée et le plan de traitement conforme au mode de présentation au diagnostic (cancer du sein, prostate, côlon-rectum, thyroïde). Ce type de dossier n'implique dès lors aucune modification, sinon minime, du contenu des variables obligatoires, mais demande à ce que les variables relatives au pronostic et au traitement soient complétées. Il est fréquent dans cette situation de s'affranchir de retours aux dossiers lorsque l'analyse du parcours de soins est confortée par la disponibilité et la convergence des données issues des RCP. A l'inverse, il est illusoire de colliger un dossier pour des tumeurs confirmées cliniquement ou radiologiquement ou pour lesquelles les informations nécessaires à l'enregistrement sont rendues partielles compte tenu des spécificités liées à la localisation (foie, voies biliaires, pancréas, hémopathies malignes).

➢ Typologie des trajectoires de soins

De manière générale, c'est l'association de la dématérialisation de la notification des tumeurs à la typologie des trajectoires de soins qui rend l'approche performante. Le cancer étant une maladie relevant d'une prise en charge multidisciplinaire, on comprend rapidement l'hétérogénéité des trajectoires individuelles, et leurs descriptions se retrouvent ardues voire impossibles intellectuellement, si l'on ne rend pas explicite l'enchaînement des évènements rencontrés. La démarche sous la forme d'une séquence d'événements, horodatés et organisés de façon chronologique, était appropriée du point de vue de l' « utilisateur registre » pour reconstituer les parcours individuels des patients parmi les nombreuses approches disponibles [70–73]. Cette représentation présente l'avantage de mettre en lumière les interactions des patients avec les équipes médicales et les structures institutionnelles sur l'évolution de leur trajectoire, et de mieux comprendre ainsi les logiques d'orientation et de prise en charge territoriales. Si la pratique au sein des registres français est le retour au dossier médical, la recherche et l'organisation des informations sont trop souvent chronophages en raison du manque d'outils adaptés à cette fin et parce que les données sur les patients sont souvent dispersées dans plusieurs dossiers. La visualisation de ces séquences apporte une aide précieuse pour apprécier le contexte des données et de repérer immédiatement des séquences classiques comme des séquences inhabituelles. La difficulté va résider dans le choix des évènements et des états selon la ligne de temps chronologique [74]. Un nombre de classes trop élevé sera difficile à décrire et interpréter. A l'inverse s'il est trop faible, l'hétérogénéité des classes risque d'être importante et l'on ne pourra guère dégager de parcours-types. Le choix de la typologie supposait donc un arbitrage, le critère principal étant que la typologie sélectionnée soit cohérente et porteuse d'enseignements du point de vue de la mission du registre. Une seule représentation ne pouvant être efficace pour tous les types de données utiles dans le domaine du cancer, il serait bénéfique de spécialiser les représentations des trajectoires de soins selon les localisations ou groupes de localisations cancéreuses.

4.1.4. Qualité des données

Le RGCPC, comme tout registre du cancer en France et dans le monde, fournit régulièrement des indications objectives sur la qualité des données qu'il produit selon quatre dimensions bien définies : comparabilité, validité, ponctualité et exhaustivité [10,11]. Cette évaluation répond en France à une certification sous l'égide du CER[15] (Comité d'Evaluation des Registres) dont la mission est de s'assurer tous les 5 ans du remplissage des critères d'excellence et de la qualité de chaque registre

[15] Mis en place par Santé publique France (SpF), l'Institut national du cancer (INCa) et l'Institut national de la santé et de la recherche médicale (Inserm) : https://www.santepubliquefrance.fr/comite-d-evaluation-des-registres

(dernière labellisation pour le RGCPC le 01/11/2015, classé « A »). Le RGCPC rapporte dans ce cadre des données favorables et comparables à celles des autres registres français du cancer sur la base des indicateurs classiquement utilisés : contrôle de la stabilité des données d'incidence dans le temps, exploration du ratio mortalité/incidence, évaluation du nombre de sources de notification par cas (exhaustivité), pourcentage de cas vérifiés histologiquement, proportion de cas sans primitif connu, évaluation de la proportion de données manquantes (validité), standardisation des pratiques de classification et de codage (comparabilité), respect des calendriers de chargement de la base nationale commune et mise en ligne des données d'incidence (ponctualité).

La configuration du SI-RGCPC joue un rôle important dans la démarche d'amélioration continue de la qualité des données. Le système dispose d'une interface ergonomique qui assiste en permanence l'utilisateur du registre (règles de sémiologie graphique, accès simplifié aux données, pré-remplissage des items obligatoires, contrôle à la saisie, priorisation des tâches opératoires) et le positionne dans une configuration optimale pour la validation comme pour le contrôle des tâches accomplies. Les données sont contrôlées quotidiennement (toutes les nuits) selon les règles internationales éditées par l'IARC et l'IACR (programme implémentée dans le SI-RGCPC) et complétées de procédures internes spécifiques, permettant de sensibiliser chaque utilisateur à la qualité d'exécution des tâches effectuées la veille.

L'intérêt réside également dans l'autonomie relative à l'ajout de nouvelles fonctionnalités déduites des besoins des utilisateurs. Tout processus supplémentaire qui vise à améliorer la qualité des données contribue à améliorer l'efficacité du registre. Dans ce contexte, la disponibilité exponentielle des données électroniques contribue à cette opportunité. Trois développements récents au sein du SI-RGCPC ont concouru à améliorer significativement la qualité des données enregistrées : implémentation des listes électorales[16] pour la qualité de l'enregistrement des lieux de résidence et des lieux de naissance ; implémentation de la BD Adresse®[17] pour la qualité du géocodage des tumeurs ; implémentation des fichiers INSEE[18] des personnes décédées pour la qualité de la mise à jour du statut vital.

4.1.5. Efficience, coût

Au-delà de la qualité des données se pose nécessairement la question de l'efficience et des retombées attendues en termes de productions. L'efficacité opérationnelle du RGCPC peut se juger au travers de la consommation des ressources utilisées pour la production du résultat. D'une manière générale, cela revient à estimer le coût moyen par cas à partir des informations sur les dépenses et le nombre

[16] Répertoire électoral unique : https://www.insee.fr/fr/information/3539086
[17] BD Adresse : https://geo.data.gouv.fr/fr/datasets/3a96ad3eb9f3fad6bb8ee2dcddd76b01ae3baff5
[18] Fichier des personnes décédées depuis 1970 : https://www.insee.fr/fr/information/4190491

de cas signalés. En pratique, la tâche est plus ardue car il importe d'identifier les postes clés et sources de variations susceptibles d'influencer le coût réel de l'enregistrement des cancers. Plusieurs travaux de comptabilité analytique ont été publiés dans ce contexte aux USA, afin de permettre aux registres et aux décideurs politiques de prendre les décisions optimales sur l'allocation des ressources [75–78]. Ces travaux ont l'intérêt d'avoir délimité les facteurs influençant le coût de l'enregistrement, et d'avoir été appliqués à l'ensemble des registres financés par le NPCR (National Program of Cancer Registries). Ces travaux sont riches d'enseignement et montrent d'une manière générale que le coût moyen par cas diminue à mesure que le nombre de cas déclarés augmente. Les registres du cancer à faible volume ont dépensé en moyenne 93,1 dollars pour signaler un cas contre 27,7 dollars pour les registres à volume élevé. A l'inverse, le coût moyen augmentait à mesure que le territoire géographique couvert était plus étendu. Quand bien même les registres privilégient les flux électroniques de données, il n'en demeure pas moins que des déplacements sur sites sont nécessaires et engagent des frais de personnel et de déplacement. Le coût n'était pas plus élevé pour les États dont les données répondaient aux critères de certification de données de haute qualité, et n'était pas associé au pourcentage de la population vivant en zone urbaine. En revanche, le coût augmentait en fonction du coût de la vie au niveau du territoire surveillé (frais généraux de fonctionnement, y compris la rémunération des employés), et il était plus important pour le déploiement de nouveaux registres que pour le renforcement des registres existants. Les registres à faible volume étaient finalement moins assujettis à utiliser les formats de données électroniques et avaient des coûts plus élevés que les registres à volume élevé pour toutes les activités clés : collecte et extraction de données, gestion et administration.

Au global, ces résultats suggéraient des économies d'échelle en suscitant le partage des coûts fixes entre plusieurs unités de production (qui étaient globalement les mêmes qu'ils soient amenés à enregistrer 1 000 ou 100 000 nouveaux cas annuels). Le partage des ressources de gestion des bases de données permettait également de favoriser l'uniformité de la collecte des données et des pratiques de contrôle de la qualité, tout en permettant aux registres de partager plus facilement les informations sur les résidents qui bénéficiaient de services de diagnostic et de traitement sur les territoires limitrophes.

Partant de ce constat, il était délicat d'apprécier l'efficience du RGCPC sans disposer d'une évaluation comparative aux autres registres français. Le RGCPC fait état toutefois de signaux très favorables au regard de sa production. Le RGCPC recense annuellement 12 600 tumeurs malignes pour une population d'1,8 millions d'habitants, le positionnant parmi les registres à volume d'activité élevé. Il fonctionne pour son activité de surveillance avec une équipe de 7 équivalents temps-plein[19].

[19] 1 temps médical, 1 temps de data manager et 5 temps d'enquêteurs

Le RGCPC contribue à 13% de la population surveillée au sein de la zone registre (3% du territoire métropolitain). Ses données ont été intégrées dans la base de données commune aux registres des cancers en 2015 et sont utilisées depuis 2018 pour la surveillance nationale dans le cadre du Programme de Travail Partenarial (PTP) qui associe le réseau FRANCIM des registres du cancer, le service de biostatistique-bioinformatique des Hospices Civils de Lyon, Santé publique France et l'Institut national du cancer. Leur prise en compte a été l'un des arguments validés par le Conseil scientifique du PTP cancer pour le changement de méthode d'estimation de l'incidence nationale (*i. e.* la zone registre étant considérée comme représentative de la France métropolitaine en termes d'incidence des cancers). Cette prise en compte a permis de se passer du corrélat de la mortalité, avec comme conséquences majeures : 1/ la production d'estimations d'incidence par sous-types topographiques et histologiques pour la première fois en France [79] ; 2/ l'utilisation d'une même méthode pour les tumeurs solides et les hémopathies malignes [80] ; 3/ l'amélioration *in fine* de la qualité des estimations pour les localisations à faible létalité. Cette étude phare du PTP, mise à jour tous les 5 ans, revêt une importance majeure puisque sa coordination générale nous a été confiée et a abouti à un certain nombre de livrables et de valorisation scientifique et médiatique[20].

Defossez G, Uhry Z, Delafosse P, Dantony E, d'Almeida T, Plouvier S, Bossard N, Bouvier AM, Molinié F, Woronoff AS, Colonna M, Grosclaude P, Remontet L, Monnereau A and the French Network of Cancer Registries (FRANCIM). Cancer incidence and mortality trends in France over 1990-2018 for solid tumors: the sex gap is narrowing. BMC Cancer. 2021 Jun 24;21(1):726.

Au-delà de l'apport du RGCPC pour les estimations nationales, la prise en compte des données du RGCPC a permis de produire également des estimations régionales et départementales d'incidence pour un nombre plus élevé de localisations cancéreuses (gain de puissance statistique) [81].

D'une manière générale, l'efficience du RGCPC se résume en partie à la réduction de la part du temps humain passé sur les processus de notification et de saisie. Le RGCPC a produit un investissement rentable permettant de concentrer le travail de l'équipe aujourd'hui sur l'activité d'enregistrement et l'investigation des dossiers médicaux complexes ou pour lesquels des compléments d'informations sont inévitables dans le cadre de l'activité de veille sanitaire, sinon motivés et convoités dans le cadre des questions de recherche.

[20] https://www.santepubliquefrance.fr/maladies-et-traumatismes/cancers/cancer-du-sein/documents/rapport-synthese/estimations-nationales-de-l-incidence-et-de-la-mortalite-par-cancer-en-france-metropolitaine-entre-1990-et-2018-volume-1-tumeurs-solides-etud

4.1.6. Transférabilité

Le modèle privilégie la réutilisation de données médicales standardisées disponibles à l'échelle du territoire national, ce qui le rend parfaitement reproductible pour répondre aux objectifs de production des indicateurs d'incidence et de survie souhaités. Le SI-RGCPC a été transposé en 2016 en région Corse dans le cadre du déploiement d'un nouveau registre des cancers soutenu par la Collectivité Territoriale de Corse. Les choix opérés par ses responsables étaient guidés par l'expertise du RGCPC sur l'interopérabilité des sources d'information en cancérologie et le traitement des données, et la stratégie de pouvoir disposer d'un environnement logiciel clé en main pour l'enregistrement des cas incidents de cancer. Le SI-RGCPC a fait l'objet d'une protection numérique en date du 27 juillet 2016 auprès de l'Agence de Protection des Programmes[21]. Au terme de la signature d'une convention de partenariat et d'un accord de confidentialité protégeant intellectuellement le savoir-faire du RGCPC, un livrable complet du logiciel a été diffusé aux responsables, associant les codes sources (import, identitovigilance, notification des tumeurs, modélisation du parcours de soins, monitorage des données), les tables de nomenclatures, l'interface de validation et un document de synthèse. L'installation a été mise en œuvre par le personnel qualifié du registre Corse. Le fonctionnement du logiciel a été rapidement effectif, et le registre Corse a pu débuter sa mission dans les meilleures conditions. Des échanges réguliers et amicaux avec ses responsables perdurent et nous attendons avec eux aujourd'hui le premier retour du Comité d'Evaluation des Registres (1ère demande de labellisation en septembre 2020).

4.2. Réutilisation secondaire des données

La seconde partie de ce travail illustre les bénéfices à réutiliser les données générées au fil du parcours de soins des patients. Les domaines d'application sont nombreux et variées et relèvent aussi bien de l'évaluation des programmes de dépistage [36], que du diagnostic moléculaire [38], du traitement et du suivi du cancer [40,50–52], ou encore de la pharmaco-épidémiologie [48,49]. Les registres du cancer sont les seuls à compiler des informations validées sur les diagnostics et la gravité des pathologies (type histologique, stades d'extension, grades de différenciation …). Le travail de standardisation des critères d'inclusion et du codage des items selon des normes internationales confère aux données finales une qualité spécifique et une comparabilité des données dans le temps et dans l'espace au niveau international. En accédant ainsi à toutes les données importantes concernant les actes et prescription ayant eu lieu depuis le diagnostic jusqu'à l'issue de la maladie, en passant par les différentes étapes de traitement et les récidives, les registres sont stratégiquement les plus à

[21] Référencé sous le certificat IDDN.FR.001.310017.000.S.P.2016.000.1000

même de fournir des informations sur la qualité et l'efficacité des soins prodigués aux patients dans la « vraie vie ». Les données de vie réelle soulèvent des aspects qui ne peuvent pas être capturés lors d'essais contrôlés randomisés, rappelant aussi la valeur des données observationnelles pour décrire l'usage d'un traitement dans un environnement non contrôlé (effets indésirables, pratiques de prescription, efficacité des traitements) [82].

Les registres du cancer répondent déjà de façon active à l'évaluation des pratiques de soins par le biais des études dites hautes résolutions [83–85]. Cependant, ces études sont ponctuelles car elles requièrent la mise en place d'enquêtes ad-hoc et sont établies le plus souvent sur des échantillons représentatifs de la population des cas incidents afin de limiter le coût de production de ce type d'évaluation. La mise en œuvre d'un système intégré d'évaluation des pratiques de soins permettrait de répondre plus avantageusement à ces attentes, tout en conservant l'expertise avancée du registre sur les problématiques territoriales, d'environnement sociodémographique, économique et d'organisation des soins. L'avantage tient notamment dans l'interprétation, l'harmonisation et le travail de comparaison des données brutes mis en œuvre spécifiquement par le registre au moment de l'enregistrement pour s'assurer de la véracité des données [86]. Le fait de croiser les bases de données médico-administratives et cliniques procure la possibilité au registre d'enrichir le sens et la portée des données brutes rassemblées, et d'offrir un outil suffisamment puissant pour répondre aux problématiques d'évaluation des soins.

Il importe ainsi de promouvoir le déploiement de systèmes d'informations intégrés d'évaluation des pratiques de soins en cancérologie fondés sur les registres de population. C'est une stratégie embrassée par de nombreux pays qui ont mis en place, à la fois du point de vue technique et opérationnel et du point de vue de la gouvernance, les moyens et conditions nécessaires à la réutilisation des données de santé [87–92]. Une étude comparative récente menée par l'OCDE (Organisation de Coopération et de Développement Economiques) rappelle la situation assez défavorable de la France en termes de maturité sur la réutilisation des données (couverture, interopérabilité, normalisation, gouvernance), positionnant la France au 22ème rang des 30 pays évalués et dans le groupe des 7 pays les moins avancés en termes de gouvernance [93]. La France a toutefois engagé un tournant ces dernières années pour exploiter le potentiel des données massives en santé, en revoyant notamment son dispositif d'accès aux données.

4.3. Perspectives à l'ère des données massives en santé

De multiples bénéfices peuvent être attendus de l'exploitation des données massives en santé. La France a, de ce point de vue, un socle de données administratives qui constitue une réelle richesse, le système national de données de santé (SNDS). Le SNDS est un entrepôt de données médico-

administratives pseudonymisées contenant l'ensemble des soins présentés au remboursement de la population française. Constitué initialement des données de l'assurance maladie (base SNIIRAM : Système National d'Information Inter-régimes de l'Assurance Maladie) et enrichi des données des hôpitaux (base PMSI), le SNDS contient aujourd'hui les causes médicales de décès (base du CépiDC de l'Inserm) et vise à intégrer deux jeux de données supplémentaires (les données relatives au handicap et les données de remboursement d'organismes d'assurance complémentaires pour un échantillon de la population).

Jusqu'alors confus et complexe, la loi de modernisation du système de santé du 26 janvier 2016[22] a revu le dispositif d'accès aux données et a posé le principe d'une ouverture large aux données pour la mise en œuvre de traitements à des fins de recherche, d'étude ou d'évaluation présentant un caractère d'intérêt public. Concrètement, les jeux de données extraits de ces bases médico-administratives sont rendus totalement anonymes pour être réutilisés sans nécessité d'autorisation préalable (*open data*). Pour les données présentant un risque de ré-identification, la loi a prévu des accès permanents pour des organismes publics ou chargés d'une mission de service public. Pour les autres utilisateurs, des accès sont autorisés par la CNIL pour des projets spécifiques.

Le SNDS attire aujourd'hui une vaste communauté de chercheurs grâce à sa quasi-exhaustivité à l'échelle de la population française et grâce à son décloisonnement ville - hôpital qui permet de travailler sur le parcours de soin complet des patients. Cette base contribue à connaître efficacement les éléments constitutifs du parcours de soins et donne lieu à de nombreuses publications, notamment dans le domaine du cancer [94–98]. La réutilisation de ses données constitue ainsi un enjeu stratégique majeur de santé publique pour les registres de population. Sa principale faiblesse réside en revanche dans sa complexité d'utilisation et dans l'absence des résultats d'examens cliniques ou paracliniques qui confère le besoin d'enrichir le SNDS de bases de données externes [99]. C'est sur ces éléments que le réseau ReDSiam (Réseau pour mieux utiliser les Données du Système national des données de santé) a été constitué en 2013 en vue d'élaborer des algorithmes par domaine pathologique et de promouvoir les méthodes d'analyse des BDMA [100]. C'est le cas par exemple pour l'identification des pathologies cancéreuses, lorsqu'il s'agit d'approcher au plus juste le nombre des cas pour estimer l'incidence ou lorsqu'il s'agit de repérer chaque cas incident de façon à pouvoir reconstituer sa trajectoire de soins [101]. De façon générale, « big data » n'est pas synonyme de « good data » et les algorithmes d'intelligence artificielle ont besoin de données valides, rappelant le rôle des registres du cancer en tant qu'expert pour valider et calibrer les données du SNDS.

Il existe donc une logique forte de réciprocité qui consiste d'une part à ce que les registres puissent accéder aux données du SNDS pour répondre intelligemment à leurs missions, et que d'autre part les

[22] https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000031912641/

données du SNDS disposent de l'expertise et de la connaissance approfondie des registres pour mieux cartographier et utiliser les données. De façon plus générale, ces échanges doivent s'intégrer aujourd'hui au niveau du Health Data Hub (HDH ou Plateforme des données de santé) qui, basée sur les recommandations du rapport Villani de 2018 sur l'intelligence artificielle [102], constitue la plateforme d'accès et de mutualisation des données pour la recherche et l'innovation en santé. Fondé d'abord sur les données médico-administratives avec le SNDS, le HDH vise à réunir des sources très variées (génomiques, cliniques, hospitalières ...) pour répondre au défi de l'usage des traitements algorithmiques dits d'« intelligence artificielle ».

Il conviendrait dans ce contexte que les registres du cancer puissent définir le cadre de gouvernance et le contenu du(es) jeu(x) de données (catalogue) qu'ils sont susceptibles de partager au sein du HDH. D'un côté, les données de santé créées sont considérées comme étant financées par la solidarité nationale et comme faisant partie du patrimoine collectif. D'un autre côté, le partage de la valeur créée et les propriétés intellectuelles liées à la mise à disposition de ces données soulèvent des interrogations tout à fait légitimes. La question de la gouvernance est une question majeure qui représente aujourd'hui le principal frein à l'extension du HDH, et à laquelle les registres doivent y répondre de façon structurée.

A l'inverse, ces questions pointent aussi l'absence de réciprocité. L'accès aux données du HDH (SNDS) constitue une opportunité de moderniser le circuit des données des registres de morbidité. Pourtant, cette question soulève encore de nombreux défis techniques et juridiques, puisque l'accès aux données du HDH ne permet pas de répondre en l'état à une simplification du mode opérationnel des registres du cancer pour leurs missions de surveillance. Le premier défi implique de pouvoir apparier les données des registres à celles du HDH, ces dernières étant cryptées afin de garantir leur confidentialité. Chaque patient est repéré dans le SNDS par un pseudonyme obtenu par l'application d'un procédé cryptographique irréversible (appelé FOIN) au NIR (Numéro d'Inscription au Répertoire[23]). Un appariement direct via un tiers de confiance ne peut pas être envisagé aujourd'hui pour les registres, qui n'ont pas autorité à disposer du NIR (le décret n° 2019-341 du 19 avril 2019[24] relatif à la mise en œuvre de traitements comportant l'usage du NIR n'étend pas l'autorisation aux registres du cancer). Cette phase implique dès lors de mettre en œuvre un appariement indirect déterministe (si possibilité de reconstruire le NIR à partir du nom de naissance, prénom, sexe, date et lieu de naissance) ou probabiliste (sur diverses variables en commun entre le registre et le SNDS). La

[23] Le NIR, ou plus communément « numéro de sécurité sociale », est le numéro d'inscription au répertoire national d'identification des personnes physiques (RNIPP) de l'INSEE, délivré à la naissance et utilisé notamment par les organismes d'assurance maladie pour la délivrance des « cartes vitales ».

[24] Décret n° 2019-341 du 19 avril 2019 relatif à la mise en œuvre de traitements comportant l'usage du numéro d'inscription au répertoire national d'identification des personnes physiques ou nécessitant la consultation de ce répertoire : https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000038396526/

question de l'usage du NIR par les registres demeure cependant la solution (juridique) à privilégier. Deuxièmement, l'accès aux données du SNDS et leur analyse ne peut se faire que dans un cadre d'hébergement très restrictif respectant le référentiel de sécurité du SNDS (traçabilité des accès et des traitements, confidentialité des données, intégrité des données). Les registres du cancer doivent donc être « SNDS compatibles » et apporter la preuve du respect du référentiel de sécurité et du règlement européen sur la protection des données à caractère personnel (RGPD). Il est important que les registres mettent en œuvre des solutions sécurisées d'hébergement, et prônent même au-delà un alignement progressif de leurs systèmes d'information qui veilleraient à simplifier les démarches techniques et réglementaires à l'échelle nationale et en réduiraient les coûts. Enfin, le catalogue des données intégrées dans le HDH ne détient pas les sources classiquement utilisées par les registres du cancer (laboratoires d'anatomopathologie, dossier communicant de cancérologie, données des laboratoires d'hématologie biologique …). Ce dernier point est majeur et limite *de facto* la portée du HDH aujourd'hui pour répondre aux problématiques de surveillance des cancers. Le concept est toutefois essentiel à défendre et rappelle l'intérêt à ce que chaque acteur « potentiel » du HDH œuvre aujourd'hui pour définir les apports et les bénéfices mutuels qui émaneront du partage et du croisement ultérieur de leurs données. C'est l'objectif convoité aujourd'hui par l'Institut National du Cancer qui, au travers de sa plateforme de données en oncologie, souhaite impulser des travaux visant à identifier et délimiter le partage des données de santé entre les différents professionnels et institutions impliquées dans le domaine du cancer.

4.4. Extension des travaux du RGCPC avec la Plateforme Cancer (INCa)

Dans ce contexte et celui des travaux instaurés avec la plateforme INSHARE [53,54], le RGCPC prévoit de croiser à titre expérimental ses données à celles de la plateforme de données en oncologie de l'INCa, dont le socle est la « cohorte cancer ». Basée sur le SNDS et développée depuis 2010, la cohorte cancer inclut les personnes diagnostiquées, traitées ou suivies pour une pathologie cancéreuse, ainsi que les personnes ayant recours aux soins en raison d'un risque élevé de cancer (elle comprend à la fois les cas incidents et les cas prévalents de cancer).

Ce projet expérimental, nommé A3CANSER[25], visera sur 3 ans (2021-2023) à identifier et définir les méthodes d'appariement optimales, dictionnaires de variables, enjeux d'interopérabilité et complémentarité entre les données des registres (jeux de données socle et/ou complémentaires) et celles de la plateforme cancer, en prévision de croisements ultérieurs. L'enjeu est de définir les contours de la mise à disposition d'une source de données à haute valeur ajoutée (base registres) à la

[25] A3CANSER : Appariement, Alignement et Accessibilité des données de santé en CANcérologie pour la Surveillance, l'Evaluation et la Recherche en population (coordonnateur du projet : Gautier DEFOSSEZ, financement INCa)

plateforme cancer ou tout autre plateforme de données de santé. Un second axe de travail cherchera à substituer les exports PMSI du registre à l'interrogation du SNDS et tester ainsi le remplacement des extractions locales de lots de données par des interrogations ciblées du SNDS, dans un objectif d'amélioration de l'efficience des registres et en prévision d'un fonctionnement qui pourrait se fonder à l'avenir sur les données des plateformes de données de santé (preuve de concept). L'enjeu de ce deuxième axe est de promouvoir le déploiement d'une démarche simplificatrice de réutilisation des données de santé pour les missions de surveillance et d'évaluation des registres de morbidité, et de lutter dans le même temps contre une dispersion des moyens collectifs engagés sur des évaluations similaires et des populations le plus souvent sélectionnées (gain d'efficience, économie d'échelle).

Conclusion

Ce travail documente un modèle de registre du cancer de génération récente, proposant des méthodes innovantes pour optimiser la surveillance épidémiologique du cancer et assimiler le rôle stratégique d'évaluation des pratiques de soins en vie réelle par les registres du cancer. Poser les fondements d'un système de surveillance optimisé au moment de la création du RGCPC était opportun face à la disponibilité exponentielle des données de santé et les initiatives analogues mises en œuvre à l'international. De nombreux verrous opérationnels ont laissé place progressivement à des méthodes algorithmiques de traitement de l'information, nous mettant en capacité de rassembler des volumes élevés de données et de satisfaire un premier niveau d'enregistrement sur les indicateurs de surveillance épidémiologique classiques (incidence, survie).

Les méthodes algorithmiques développées ont consenti à des gains d'efficience majeurs en proposant une simplification des tâches opérationnelles en amont de l'enregistrement des tumeurs. Le souci d'une harmonisation des formats d'exports (*i. e.* un format commun et universel par source à l'échelle nationale) a été le garant d'une simplification du degré d'intégration et du traitement des données dans le SI-RGCPC. Cette configuration nous a permis d'instaurer un rythme d'export soutenu (pluri annuel), étant attentif à proposer un niveau optimal d'informations pour les opérateurs du registre au moment de la validation. L'expérience montre que le SI-RGCPC est plus réactif que le SNDS qui récupère des données à N+1 lorsque le RGCPC intègre les données de l'année en cours.

La dématérialisation de la notification a été un second palier nous ayant permis de réduire le temps humain investi dans la confrontation et l'analyse de la concordance des données sources. L'expérience acquise des travaux méthodologiques initiés sur les actes CCAM et celle issue des premières années d'enregistrement du RGCPC a affecté de façon très positive les choix stratégiques opérés dans les processus de traitement de l'information. La typologie des enregistrements initiée pour la représentation chronologique des parcours de soins nous a confortée dans l'idée de contextualiser méthodiquement l'information disponible à l'échelle individuelle, en vue de limiter le bruit induit par les imprécisions ou les erreurs de codage. En pratique, le système n'aspire pas à proposer de l'enregistrement automatique, mais la question mériterait toutefois d'être approfondie au regard des performances acquises pour certaines localisations sur la base des items obligatoires pour l'incidence et la survie (*e. g.* les cancers du sein, les plus fréquents chez la femme, qui regroupent un niveau d'information souvent très élevé).

Enfin, la mise en œuvre d'une modélisation des parcours de soins a été le troisième palier, qui, d'une manière générale, a été le garant principal des performances de l'approche proposée, en plaçant l'opérateur du registre dans une configuration optimale pour mener à bien ses tâches

d'enregistrement. La visualisation des séquences offre à l'utilisateur une vision décloisonnée des données qui permet de renforcer l'effet synergique et additif des multiples données sources chaînées et ordonnées dans le temps. L'utilisateur se retrouve en capacité de produire une analyse intelligible du parcours de soins, dont les bénéfices sont tirés de la confrontation des résultats issus des documents textuels (ACP, RCP) aux évènements issus de la représentation des données médico-administratives (PMSI). Cela permet à l'utilisateur de qualifier en pratique le mode de présentation au diagnostic de la tumeur (stade d'extension notamment) à la lumière du traitement mis en œuvre, et inversement, et d'apprécier ainsi les éventuelles déviances ou incohérences du parcours (adéquation de prise en charge, respect des bonnes pratiques, pertes de chances, inégalités). Le croisement opéré contribue au final à enrichir le sens et la portée des données brutes rassemblées, et de souscrire à un second niveau d'enregistrement qui répond, au-delà des finalités épidémiologiques classiques, aux problématiques d'évaluation des soins et des services de santé.

Le système d'information regroupe ainsi un ensemble organisé de ressources qui permet de traiter, distribuer l'information et coordonner les activités selon l'alignement stratégique et les missions clés du RGCPC : surveillance, évaluation et recherche. L'environnement de développement du SI-RGCPC repose sur des outils et langages informatiques familiers et les méthodes algorithmiques développées sont fondées sur la réutilisation de données standardisées qui concourent à leur transférabilité. Sa configuration contribue à promouvoir la contextualisation et l'enrichissement des parcours de soins des patients atteints de cancer, en réponse à des évaluations non biaisées et pour des domaines d'applications variées, aussi bien dans le champ de la prévention et du dépistage, que celui du diagnostic, du traitement et du suivi du cancer. Le travail préalable de standardisation des critères d'inclusion et du codage des items selon des normes internationales conforte ainsi la mise en œuvre d'une stratégie collective qui veillerait à renforcer les missions des registres de population, en soutenant le déploiement de systèmes d'informations intégrés répondant à la double finalité de surveillance et d'évaluation des pratiques.

L'ère des plateformes des données de santé constitue une réelle opportunité pour y répondre mais aspire en même temps à une redéfinition claire de la stratégie nationale de surveillance pour en lever les obstacles juridiques. Si d'un côté les registres doivent s'acquitter d'une définition du contenu et du mode de gouvernance des jeux de données susceptibles d'y être déposées, il est vital qu'ils disposent en toute réciprocité d'un accès sécurisé aux données, disponibles et futures, indispensables à leurs missions. L'accès au NIR serait un premier élément de réponse. Au-delà de ce point d'orgue juridique, il est utile que les registres engagent un alignement progressif de leurs systèmes d'informations qui veillerait à mutualiser les méthodes et les outils de traitement de l'information (solutions d'hébergement, module d'interconnexion des enregistrements, gestion commune des

terminologies, choix et implémentation des méthodes algorithmiques). Ces choix opérationnels et stratégiques dégageraient des gains d'efficience collectifs majeurs et permettraient de réduire les coûts de production, susceptibles d'être réinvestis dans une plus grande valorisation, dans de nouvelles missions et dans une couverture plus complète de la population.

Références bibliographiques

1. Saracci R, Wild C, International Agency for Research on Cancer. International Agency for Research on Cancer: the first 50 years, 1965-2015 [Internet]. 2015 [cited 2020 Jul 17]. Available from: http://www.iarc.fr/en/publications/books/iarc50/

2. Bray F, Znaor A, Cueva P. Planification et développement des registres du cancer basés sur la population dans les pays à revenu faible et intermédiaire. Lyon: CIRC; 2015. (Publications techniques du CIRC). Report No.: 43.

3. Doll R, Payne PM, Waterhouse JAH. Cancer Incidence in Five Continents Vol. I. A technical report. IARC Publ. 1966;

4. Jensen O, Parkin D, MacLennan R, Muir C, Skeet RG. Cancer Registration: Principles and Methods. International Agency for Research on Cancer; 1991.

5. Forman D, Bray F, Brewster DH, Gombe Mbalawa C, Kohler B, Pineros M, et al. Cancer incidence in five continents. Volume X [Internet]. 2014 [cited 2018 Sep 20]. Available from: http://www.iarc.fr/en/publications/pdfs-online/epi/sp164/

6. Bray F, Ferlay J, Laversanne M, Brewster DH, Gombe Mbalawa C, Kohler B, et al. Cancer Incidence in Five Continents: Inclusion criteria, highlights from Volume X and the global status of cancer registration. Int J Cancer. 2015 Nov 1;137(9):2060–71.

7. Bray F, Colombet M, Mery L, Piñeros M, Znaor A, Zanetti R and Ferlay J, editors (2017). Cancer Incidence in Five Continents, Vol. XI (electronic version). Lyon: International Agency for Research on Cancer.

8. Siesling S, Louwman WJ, Kwast A, van den Hurk C, O'Callaghan M, Rosso S, et al. Uses of cancer registries for public health and clinical research in Europe: Results of the European Network of Cancer Registries survey among 161 population-based cancer registries during 2010–2012. Eur J Cancer. 2015 Jun;51(9):1039–49.

9. Forsea A-M. Cancer registries in Europe—going forward is the only option. ecancermedicalscience [Internet]. 2016 May 12 [cited 2020 Jun 25];10. Available from: http://www.ecancer.org/journal/10/full/641-cancer-registries-in-europe-going-forward-is-the-only-option.php

10. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. Eur J Cancer Oxf Engl 1990. 2009 Mar;45(5):747–55.

11. Parkin DM, Bray F. Evaluation of data quality in the cancer registry: principles and methods Part II. Completeness. Eur J Cancer Oxf Engl 1990. 2009 Mar;45(5):756–64.

12. ENCR. Recommendations for a Standard Dataset for the European Network of Cancer Registries [Internet]. 2005. Available from: https://www.encr.eu/recommendations-and-working-groups

13. Engholm G, Ferlay J, Christensen N, Bray F, Gjerstorff ML, Klint A, et al. NORDCAN--a Nordic tool for cancer information, planning, quality control and research. Acta Oncol Stockh Swed. 2010 Jun;49(5):725–36.

14. Gjerstorff ML. The Danish Cancer Registry. Scand J Public Health. 2011 Jul;39(7 Suppl):42–5.

15. Lynge E, Sandegaard JL, Rebolj M. The Danish National Patient Register. Scand J Public Health. 2011 Jul;39(7 Suppl):30–3.

16. Cancer Registry of Norway. All cancer cases are to be reported to the Cancer Registry of Norway [Internet]. Available from: https://www.kreftregisteret.no/en/General/About-the-Cancer-Registry/

17. The Swedish Cancer Register. Reporting procedures [Internet]. Available from: https://www.socialstyrelsen.se/en/statistics-and-data/registers/register-information/swedish-cancer-register/

18. The Icelandic Cancer Registry. About the Icelandic Cancer Registry [Internet]. Available from: https://www.krabb.is/krabbameinsskra/en/activities/about-icr/

19. Finish Cancer Registry. Cancer information notification (data collection) [Internet]. Available from: https://cancerregistry.fi/data-collection/cancer-information-notification/

20. Glaser SL, Clarke CA, Gomez SL, O'Malley CD, Purdie DM, West DW. Cancer surveillance research: a vital subdiscipline of cancer epidemiology. Cancer Causes Control CCC. 2005 Nov;16(9):1009–19.

21. White MC, Babcock F, Hayes NS, Mariotto AB, Wong FL, Kohler BA, et al. The history and use of cancer registry data by public health cancer control programs in the United States. Cancer. 2017 Dec 15;123 Suppl 24:4969–76.

22. Black RJ, International Agency for Research on Cancer, editors. Automated data collection in cancer registration. Lyon; 1998. 52 p. (IARC technical report).

23. Tagliabue G, Maghini A, Fabiano S, Tittarelli A, Frassoldi E, Costa E, et al. Consistency and accuracy of diagnostic cancer codes generated by automated registration: comparison with manual registration. Popul Health Metr. 2006 Sep 28;4:10.

24. Contiero P, Tittarelli A, Maghini A, Fabiano S, Frassoldi E, Costa E, et al. Comparison with manual registration reveals satisfactory completeness and efficiency of a computerized cancer registration system. J Biomed Inform. 2008 Feb;41(1):24–32.

25. Tognazzo S, Emanuela B, Rita FA, Stefano G, Daniele M, Fiorella SC, et al. Probabilistic classifiers and automated cancer registration: an exploratory application. J Biomed Inform. 2009 Feb;42(1):1–10.

26. Marrett LD, Clarke EA, Hatcher J, Weir HK. Epidemiologic research using the Ontario Cancer Registry. Can J Public Health Rev Can Sante Publique. 1986 Jun;77 Suppl 1:79–85.

27. Clarke E, Marrett LD, Kreiger N. Cancer registration in Ontario: a computer approach. In: Cancer Registration: Principles and Methods. Lyon; 1991. (IARC scientific publications).

28. Simonato L, Zambon P, Rodella S, Giordano R, Guzzinati S, Stocco C, et al. A computerised cancer registration network in the Veneto region, north-east of Italy: a pilot study. Br J Cancer. 1996 Jun;73(11):1436–9.

29. Tognazzo S, Andolfo A, Bovo E, Fiore AR, Greco A, Guzzinati S, et al. Quality control of automatically defined cancer cases by the automated registration system of the Venetian Tumour Registry. Quality control of cancer cases automatically registered. Eur J Public Health. 2005 Dec;15(6):657–64.

30. Francis F, Terroba C, Persoz C, Gagliolo J-M, Alla F. Quelle place pour les registres de morbidité à l'ère des données massives de santé ? Rev DÉpidémiologie Santé Publique. 2020 Apr;68(2):117–23.

31. Evaluation du plan cancer 2003-2007. Rapport final [Internet]. Haut Conseil de la Santé Publique; 2009 Jan. Available from: file:///C:/Users/GDEFOS~1/AppData/Local/Temp/hcspr20090131_EvaluationCancer.pdf

32. Jouhet V, Defossez G, Burgun A, le Beux P, Levillain P, Ingrand P, et al. Automated classification of free-text pathology reports for registration of incident cases of cancer. Methods Inf Med. 2012;51(3):242–51.

33. Jouhet V, Defossez G, CRISAP, CoRIM, Ingrand P. Automated selection of relevant information for notification of incident cancer cases within a multisource cancer registry. Methods Inf Med. 2013;52(5):411–21.

34. Working Group Report&NA; International rules for multiple primary cancers (ICD-0 third edition): Eur J Cancer Prev. 2005 Aug;14(4):307–8.

35. Defossez G, Rollet A, Dameron O, Ingrand P. Temporal representation of care trajectories of cancer patients using data from a regional information system: an application in breast cancer. BMC Med Inform Decis Mak. 2014 Apr 2;14:24.

36. Defossez G, Quillet A, Ingrand P. Aggressive primary treatments with favourable 5-year survival for screen-interval breast cancers. BMC Cancer. 2018 Apr 6;18(1):393.

37. Miller AB, Wall C, Baines CJ, Sun P, To T, Narod SA. Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. BMJ. 2014 Feb 11;348:g366.

38. Thiebault Q, Defossez G, Karayan-Tapon L, Ingrand P, Silvain C, Tougeron D. Analysis of factors influencing molecular testing at diagnostic of colorectal cancer. BMC Cancer. 2017 Nov 14;17(1):765.

39. Les thérapies ciblées dans le traitement du cancer en 2015. États des lieux et enjeux. Boulogne-Billancourt: Institut National du Cancer; 2016 Jul p. 82.

40. Quillet A, Defossez G, Ingrand P. Surveillance of waiting times for access to treatment: a registry-based computed approach in breast cancer care. Eur J Cancer Care (Engl). 2016 Sep;25(5):764–73.

41. Institut National du Cancer (INCa). Étude sur les délais de prise en charge des cancers du sein et du poumon dans plusieurs régions de France en 2011. Boulogne-Billancourt; 2012. (Collection Etudes et Perspectives).

42. Institut National du Cancer (INCa). Du diagnostic au premier traitement : délais de prise en charge des cancers enregistrés par les registres spécialisés du réseau Francim 1999-2008. Boulogne-Billancourt; 2012. (Collection Etat des lieux & des connaissances).

43. Fleissig A, Jenkins V, Catt S, Fallowfield L. Multidisciplinary teams in cancer care: are they effective in the UK? Lancet Oncol. 2006 Nov;7(11):935–43.

44. Sainsbury R, Haward B, Rider L, Johnston C, Round C. Influence of clinician workload and patterns of treatment on survival from breast cancer. Lancet Lond Engl. 1995 May 20;345(8960):1265–70.

45. Stiller CA. Centralisation of treatment and survival rates for cancer. Arch Dis Child. 1988 Jan;63(1):23–30.

46. Haute Autorité de Santé. Réunion de concertation pluridisciplinaire en cancérologie (RCP) [Internet]. [cited 2017 Jun 16]. Available from: https://www.has-sante.fr/portail/jcms/c_2676637/fr/reunion-de-concertation-pluridisciplinaire-en-cancerologie-rcp

47. Circulaire n◦ DHOS/SDO/2005/101 du 22 février 2005 relative à l'Organisation des soins en cancérologie. [Internet]. Available from: http://solidarites-sante.gouv.fr/fichiers/bo/2005/05-03/a0030034.htm

48. Ingrand I, Defossez G, Lafay-Chebassier C, Chavant F, Ferru A, Ingrand P, et al. Serious adverse effects occurring after chemotherapy: A general cancer registry-based incidence survey. Br J Clin Pharmacol. 2020;86(4):711–22.

49. Barin-Le Guellec C, Lafay-Chebassier C, Ingrand I, Tournamille J-F, Boudet A, Lanoue M-C, et al. Toxicities associated with chemotherapy regimens containing a fluoropyrimidine: A real-life evaluation in France. Eur J Cancer Oxf Engl 1990. 2020 Jan;124:37–46.

50. Puyade M, Defossez G, Guilhot F, Ingrand P. Multiple myeloma: the quality of care is linked to geographical and organisational determinants. A study in a French registry. Eur J Cancer Care (Engl). 2016 Sep;25(5):855–63.

51. Puyade M, Defossez G, Guilhot F, Leleu X, Ingrand P. Age-related health care disparities in multiple myeloma. Hematol Oncol. 2018 Feb;36(1):224–31.

52. Systchenko T, Defossez G, Guidez S, Laurent C, Puyade M, Debiais-Delpech C, et al. R-CHOP appears to be the best first-line treatment for second primary diffuse large B cell lymphoma: a cancer registry study. Ann Hematol. 2020 Jul;99(7):1605–13.

53. Bouzillé G, Westerlynck R, Defossez G, Bouslimi D, Bayat S, Riou C, et al. Sharing Health Big Data for Research - A Design by Use Cases: The INSHARE Platform Approach. Stud Health Technol Inform. 2017;245:303–7.

54. Pladys A, Defossez G, Lemordant P, Lassalle M, Ingrand P, Jacquelinet C, et al. Cancer risk in dialyzed patients with and without diabetes. Cancer Epidemiol. 2020;65:101689.

55. Le Guillou A, Pladys A, Kihal W, Siebert M, Haddj-Elmrabet A, Cernon C, et al. [Is cancer incidence different between type 2 diabetes patients compared to non-diabetics in hemodialysis? A study from the REIN registry]. Nephrol Ther. 2018 May;14(3):142–7.

56. Pladys A, Couchoud C, LeGuillou A, Siebert M, Vigneau C, Bayat S. Type 1 and type 2 diabetes and cancer mortality in the 2002-2009 cohort of 39,811 French dialyzed patients. PloS One. 2015;10(5):e0125089.

57. Barrès D, Bergeron C, Cartier I, Denis P, Doumecq-Lacoste J-M, Feutry C, et al. Reproductibilité du diagnostic cytologique : étude du CRISAP Ile-de-France. Rev Fr Lab. 1999 Dec;1999(318):59–63.

58. Bergeron C, Cartier I, Guldner L, Lassalle M, Savignoni A, Asselain B. Lésions précancéreuses et cancers du col de l'utérus diagnostiqués par le frottis cervical, Ile-de-France, enquête Crisap, 2002. Bull Epidémiologique Hebd. 2005;

59. Bailly L, Mariné-Barjoan E, Ambrosetti D, Roussel J-F, Caissotti C, Ettore F, et al. [Data quality of cancer registration by Adicap codes, used by French pathologists from Paca, 2005-2006]. Ann Pathol. 2009 Apr;29(2):74–9.

60. Fabacher T, Godet J, Klein D, Velten M, Jegu J. Machine learning application for incident prostate adenocarcinomas automatic registration in a French regional cancer registry. Int J Med Inf. 2020 Jul;139:104139.

61. Löpprich M, Krauss F, Ganzinger M, Senghas K, Riezler S, Knaup P. Automated Classification of Selected Data Elements from Free-text Diagnostic Reports for Clinical Research. Methods Inf Med. 2016;55(04):373–80.

62. Oleynik M, Patrão DFC, Finger M. Automated Classification of Semi-Structured Pathology Reports into ICD-O Using SVM in Portuguese. Stud Health Technol Inform. 2017;235:256–60.

63. Kavuluru R, Hands I, Durbin EB, Witt L. Automatic Extraction of ICD-O-3 Primary Sites from Cancer Pathology Reports. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2013;2013:112–6.

64. Le Bihan-Benjamin C, Rochhi M, Bousquet PJ, Simonnet JA, Ferrari C, De Luze S. Autorisation en chirurgie du cancer. Adapter le méthode de quantification de l'activité soumise à seuil [Internet]. INCa; 2019. Available from: file:///C:/Users/GDEFOS~1/AppData/Local/Temp/Autorisation_en_chirugie_du_%20cancer_adapter_methode_quantification_activite_soumise_seuil_mel_20190618-1.pdf

65. Quantin C, Binquet C, Bourquard K, Pattisina R, Gouyon-Cornet B, Ferdynus C, et al. Which are the best identifiers for record linkage? Med Inform Internet Med. 2004 Dec;29(3–4):221–7.

66. Schmidtmann I, Sariyar M, Borg A, Gerold-Ay A, Heidinger O, Hense H-W, et al. Quality of record linkage in a highly automated cancer registry that relies on encrypted identity data. GMS Med Inform. 2016 Jun 13;Biometrie und Epidemiologie; 12(1):Doc02.

67. Jaro MA. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. J Am Stat Assoc. 1989 Jun 1;84(406):414–20.

68. Winkler WE. Overview of Record Linkage and Current Research Directions (Statistics #2006-2). 2006 Feb 8; Available from: http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf

69. Nicholson N, Perego A. Interoperability of population-based patient registries. J Biomed Inform X. 2020 Sep;6–7:100074.

70. Nguena Nguefack HL, Pagé MG, Katz J, Choinière M, Vanasse A, Dorais M, et al. Trajectory Modelling Techniques Useful to Epidemiological Research: A Comparative Narrative Review of Approaches. Clin Epidemiol. 2020;12:1205–22.

71. Hsu W, Taira RK, El-Saden S, Kangarloo H, Bui AAT. Context-based electronic health record: toward patient specific healthcare. IEEE Trans Inf Technol Biomed Publ IEEE Eng Med Biol Soc. 2012 Mar;16(2):228–34.

72. Thiessard F, Mougin F, Diallo G, Jouhet V, Cossin S, Garcelon N, et al. RAVEL: retrieval and visualization in ELectronic health records. Stud Health Technol Inform. 2012;180:194–8.

73. Ledieu T, Bouzillé G, Thiessard F, Berquet K, Van Hille P, Renault E, et al. Timeline representation of clinical data: usability and added value for pharmacovigilance. BMC Med Inform Decis Mak. 2018 Oct 19;18(1):86.

74. Monroe M, Lan R, Lee H, Plaisant C, Shneiderman B. Temporal event sequence simplification. IEEE Trans Vis Comput Graph. 2013 Dec;19(12):2227–36.

75. Weir HK, Berg GD, Mansley EC, Belloni KA. The National Program of Cancer Registries: explaining state variations in average cost per case reported. Prev Chronic Dis. 2005 Jul;2(3):A10.

76. Subramanian S, Tangka FKL, Beebe MC, Trebino D, Weir HK, Babcock F. The cost of cancer registry operations: Impact of volume on cost per case for core and enhanced registry activities. Eval Program Plann. 2016 Apr;55:1–8.

77. Tangka FKL, Subramanian S, Beebe MC, Weir HK, Trebino D, Babcock F, et al. Cost of Operating Central Cancer Registries and Factors That Affect Cost: Findings From an Economic Evaluation of Centers for Disease Control and Prevention National Program of Cancer Registries. J Public Health Manag Pract JPHMP. 2016 Oct;22(5):452–60.

78. Beebe MC, Subramanian S, Tangka FK, Weir HK, Babcock F, Trebino D. An Analysis of Cancer Registry Cost Data: Methodology and Results. J Regist Manag. 2018;45(2):58–64.

79. Defossez G, Le Guyader-Peyrou S, Uhry Z, Grosclaude P, Colonna M, Dantony E, et al. National estimates of cancer incidence and mortality in metropolitan France between 1990 and 2018 [Internet]. Saint-Maurice (Fra): Santé publique France; 2019 [cited 2019 Oct 10] p. 372. Available from: https://www.santepubliquefrance.fr/maladies-et-traumatismes/cancers/cancer-du-sein/documents/rapport-synthese/estimations-nationales-de-l-incidence-et-de-la-mortalite-par-cancer-en-france-metropolitaine-entre-1990-et-2018-volume-1-tumeurs-solides-etud

80. Uhry Z, Chatignoux E, Dantony E, Colonna M, Roche L, Fauvernier M, et al. Multidimensional penalized splines for incidence and mortality trend analyses and validation of national cancer incidence estimates (In press). Int J Epidemiol. 2020.

81. Chatignoux E, Uhry Z, Grosclaude P, Colonna M, Remontet L. How to produce sound predictions of incidence at a district level using either health care or mortality data in the absence of a national registry: the example of cancer in France. Int J Epidemiol. 2020 Nov 24.

82. Bégaud B, Polton D, Von Lennep F. Les données de vie réelle, un enjeu majeur pour la qualité des soins et la régulation du système de santé. L'exemple du médicament. Boulogne-Billancourt; 2017. (Rapport réalisé à la demande de Madame La Ministre de la santé Marisol Touraine).

83. Minicozzi P, Van Eycken L, Molinie F, Innos K, Guevara M, Marcos-Gragera R, et al. Comorbidities, age and period of diagnosis influence treatment and outcomes in early breast cancer. Int J Cancer. 2019 May 1;144(9):2118–27.

84. Cowppli-Bony A, Uhry Z, Remontet L, Voirin N, Guizard A-V, Trétarre B, et al. Survival of solid cancer patients in France, 1989-2013: a population-based study. Eur J Cancer Prev Off J Eur Cancer Prev Organ ECP. 2017;26(6):461–8.

85. Sant M, Meneghini E, Bastos J, Rossi PG, Guevara M, Innos K, et al. Endocrine treatment and incidence of relapse in women with oestrogen receptor-positive breast cancer in Europe: a population-based study. Breast Cancer Res Treat. 2020 Sep;183(2):439–50.

86. Bellazzi R. Big data and biomedical informatics: a challenging opportunity. Yearb Med Inform. 2014 May 22;9:8–13.

87. Dewdney SB, Lachance J. Electronic Records, Registries, and the Development of "Big Data": Crowd-Sourcing Quality toward Knowledge. Front Oncol [Internet]. 2017 Jan 30 [cited 2021 Jan 23];6. Available from: http://journal.frontiersin.org/article/10.3389/fonc.2016.00268/full

88. Prodhan S, King MJ, De P, Gilbert J. Health Services Data: The Ontario Cancer Registry (a Unique, Linked, and Automated Population-Based Registry). In: Sobolev B, Levy A, Goring S, editors. Data and Measures in Health Services Research [Internet]. Boston, MA: Springer US; 2016 [cited 2020 Jul 20]. p. 1–27. Available from: http://link.springer.com/10.1007/978-1-4899-7673-4_18-1

89. Laschkolnig A, Habl C, Renner A-T, Bobek J, European Commission, Directorate-General for Health and Food Safety, et al. Study on Big Data in public health, telemedicine and healthcare: final report. [Internet]. Luxembourg: Publications Office; 2016 [cited 2020 Dec 23]. Available from: http://dx.publications.europa.eu/10.2875/734795

90. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. J Am Med Inform Assoc. 2007 Jan 1;14(1):1–9.

91. Sledge GW, Miller RS, Hauser R. CancerLinQ and the future of cancer care. Am Soc Clin Oncol Educ Book Am Soc Clin Oncol Annu Meet. 2013;430–4.

92. Health Registries for Research (HRR) – Facilitating the use and security of Norwegian health registries in research [Internet]. Available from: https://hrr.w.uib.no/

93. Readiness of electronic health record systems to contribute to national health information and research [Internet]. 2017 Dec [cited 2021 Jan 23]. (OECD Health Working Papers; vol. 99). Report No.: 99. Available from: https://www.oecd-ilibrary.org/social-issues-migration-health/readiness-of-electronic-health-record-systems-to-contribute-to-national-health-information-and-research_9e296bf3-en

94. Lefeuvre D, Le Bihan-Benjamin C, Pauporté I, Medioni J, Bousquet P-J. French Medico-Administrative Data to Identify the Care Pathways of Women With Breast Cancer. Clin Breast Cancer. 2017 Jul;17(4):e191–7.

95. Lefeuvre D, Catajar N, Le Bihan Benjamin C, Ifrah N, De Bels F, Viguier J, et al. Breast cancer screening: Impact on care pathways. Cancer Med. 2019 Jul;8(8):4070–8.

96. Janah A, Gauthier LR, Morin L, Bousquet PJ, Le Bihan C, Tuppin P, et al. Access to palliative care for cancer patients between diagnosis and death: a national cohort study. Clin Epidemiol. 2019;11:443–55.

97. Janah A, Le Bihan-Benjamin C, Mancini J, Bouhnik A-D, Bousquet P-J, Bendiane M-K. Access to inpatient palliative care among cancer patients in France: an analysis based on the national cancer cohort. BMC Health Serv Res. 2020 Aug 26;20(1):798.

98. Didier R, Gouysse M, Eltchaninoff H, Le Breton H, Commeau P, Cayla G, et al. Successful linkage of French large-scale national registry populations to national reimbursement data: Improved data completeness and minimized loss to follow-up. Arch Cardiovasc Dis. 2020 Aug;113(8–9):534–41.

99. Scailteux L-M, Droitcourt C, Balusson F, Nowak E, Kerbrat S, Dupuy A, et al. French administrative health care database (SNDS): The value of its enrichment. Therapie. 2019 Apr;74(2):215–23.

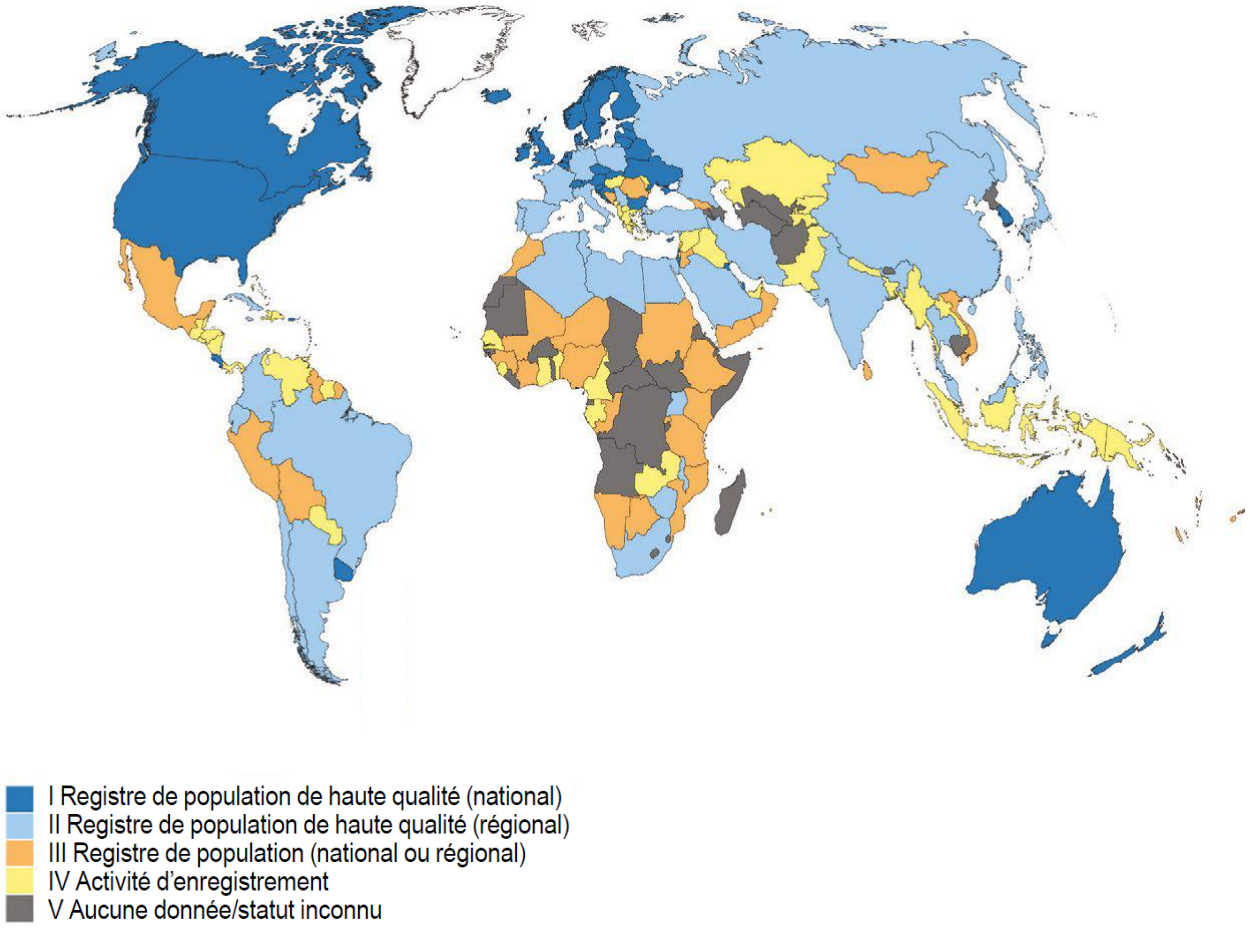
100. Goldberg M, Carton M, Doussin A, Fagot-Campagna A, Heyndrickx E, Lemaitre M, et al. [The REDSIAM network]. Rev Epidemiol Sante Publique. 2017 Oct;65 Suppl 4:S144–8.

101. Bousquet P-J, Caillet P, Coeuret-Pellicer M, Goulard H, Kudjawu YC, Le Bihan C, et al. [Using cancer case identification algorithms in medico-administrative databases: Literature review and first results from the REDSIAM Tumors group based on breast, colon, and lung cancer]. Rev Epidemiol Sante Publique. 2017 Oct;65 Suppl 4:S236–42.

102. Rapport Villani : donner un sens à l'intelligence artificielle (IA) [Internet]. Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation; 2018 Mar. Available from: https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf

Annexes

Annexe 1 : Situation et couverture mondiales en 2013 (source : CIRC / OMS) pour l'enregistrement des cancers dans la population, selon l'étendue de la couverture (nationale ou régionale) et la qualité (haute qualité vs simple enregistrement ou juste certaines activités d'enregistrement).



Annexe 2 : Liste des 28 registres du cancer de France métropolitaine et d'outre-mer appartenant au réseau FRANCIM au 1er Janvier 2021.

	Libellé du Registre	Département(s)	Population Insee[1]
1	Général Calvados	14	658 262
2	Général Poitou-Charentes (Charente, Charente-Maritime, Deux-Sèvres, Vienne)	16, 17, 79, 86	1 777 773
3	Général Doubs et territoire de Belfort	25, 90	672 451
4	Général Gironde	33	1 463 662
5	Général Hérault	34	1 062 036
6	Général Isère	38	1 215 512
7	Général Loire-Atlantique Vendée (Loire-Atlantique, Vendée)	44, 85	1 938 021
8	Général Manche	50	499 531
9	Général Lille et sa région[2]	59	785 256
10	Général Bas Rhin	67	1 099 269
11	Général Haut Rhin	68	753 056
12	Général Somme	80	571 211
13	Général Tarn	81	377 675
14	Général Haute-Vienne	87	376 058
		TOTAL	**13 276 473 (21.2%)**
15	Spécialisé Digestif Calvados	14	685 262
16	Spécialisé Digestif Bourguignon (Côte d'Or, Saône-et-Loire)	21, 71	1 081 930
17	Spécialisé Digestif Finistère	29	899 870
18	Spécialisé Gynéco Côte d'Or	21	525 931
19	Spécialisé Hémato Basse Normandie (Calvados, Manche, Orne)	14, 50, 61	1 475 684
20	Spécialisé Hémato Côte d'Or	21	525 931
21	Spécialisé Hémato Gironde	33	1 463 662
22	Spécialisé SNC Gironde	33	1 463 662
23	Spécialisé Thyroïde Marne Ardennes	08, 51	849 681
24	Spécialisé Thyroïde Rhône	69	1 744 236
		TOTAL	**7 515 063 (11.9%)**
	ZR	Départements couverts par les registres généraux + Côte d'Or	**13 802 404 (21.9%)**
	HZR	Tous les autres départements	**49 267 940 (78.1%)**
	FRANCE		**63 070 344**
25	Général Guadeloupe	971	404 635
26	Général Martinique	972	392 291
27	Général Guyane	973	237 549
28	National Pédiatrique (tumeurs solides et hémopathies malignes)	FRANCE	11 572 872[3]

[1] Population légale en 2011 (https://www.insee.fr/fr/statistiques/2132415/?geo=METRO-1)

[2] Le Registre général des cancers de Lille et sa région couvre partiellement le département du Nord (59) avec l'agglomération de Lille et sa métropole, soit 785 256 habitants (30.0%) sur 2 579 208 pour l'ensemble du département

[3] Le Registre national des tumeurs de l'enfant recouvre la population âgée de 0 à 18 ans. L'effectif présenté ici est restreint par souci de simplicité à la population âgée de 0 à 14 ans (effectif disponible par défaut sur le site de l'Insee)

Légende : ZR= Zone registre ; HZR= Hors zone registre ; La zone registre (ZR) est arbitrairement définie ici à partir des 19 départements disposant d'un registre général et le département de la Côte d'Or qui dispose de trois registres spécialisés (digestif, sein, gynécologique et hémopathies malignes).