Conséquences de la crise Covid-19 en oncologie médicale : promesses, réalisations et défis d'un entrepôt de données de santé

Sonia Priou, Guillaume Lamé, Marija Jankovic, Romain Bey, Christel Daniel, Gilles Chatellier, Christophe Tournigand, Emmanuelle Kempf

Conséquences de la crise Covid-19 en oncologie médicale : promesses, réalisations et défis d'un entrepôt de données de santé

Sonia Priou [1], Guillaume Lamé [1], Marija Jankovic [1], Romain Bey [2], Christel Daniel [2,3], Gilles Chatellier [4], Christophe Tournigand [5], Emmanuelle Kempf [3,5]

[1] Laboratoire de Génie Industriel, CentraleSupélec, Université Paris Saclay
[2] Département SI, Innovation et Données, Assistance Publique—Hôpitaux de Paris
[3] Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé, LIMICS, Sorbonne Université, Inserm, Université Sorbonne Paris Nord
[4] Département d'Informatique Médicale, Centre-Université de Paris (APHP-CUP), Assistance Publique-Hôpitaux de Paris
[5] Département d'Oncologie Médicale, CHU Henri Mondor, Assistance Publique—Hôpitaux de Paris

Résumé. Les Entrepôts de Données de Santé (EDS) regroupent de nombreuses informations disponibles sur les patients traités au sein d'un établissement hospitalier. Ces données ont été largement utilisées pendant la pandémie de Covid-19 pour informer rapidement des conséquences de la crise Covid-19 sur l'état de santé des patients, notamment en cancérologie. Néanmoins, ces analyses nécessitent une grande prudence car différents types de biais peuvent survenir lors de l'exploitation de ces données. Nous présentons un bilan réflexif du travail réalisé sur l'EDS de l'Assistance Publique – Hôpitaux de Paris (AP-HP) pour analyser l'impact de la Covid19 sur la prise en charge des patients atteints de cancer. Nous analysons différents biais constatés sur les données structurées et non-structurées, et leur impact sur l'étude. Une fois ces points pris en compte, les EDS présentent un potentiel indéniable. Pour l'illustrer, nous présentons brièvement les résultats obtenus sur l'impact du Covid19 sur la prise en charge en cancérologie à l'AP-HP.

Mots clés : Données de santé, Qualité des données, Système complexe

1 Introduction

L'épidémie de Covid-19 a bousculé la recherche en santé. Une étude récente [Horbach, 2020] a indiqué que depuis le début de l'épidémie, les revues médicales ont drastiquement diminué leurs délais entre la soumission et la publication pour les articles concernant la Covid-19. Alors que la course à la publication est dénoncée par une partie de la communauté scientifique depuis des années, cette pression à publier a été importante lors de la crise Covid-19 et a parfois entrainé de la négligence sur la qualité des recherches, voire des suspicions de fraude. Les recherches utilisant les bases de données médicales ont été particulièrement sujettes à polémique après que deux articles publiés dans des revues prestigieuses ont été rétractés quelques semaines après leur publication car leurs auteurs ne pouvaient produire les bases de données supposément utilisées [Mehra, Ruschitzka and Patel, 2020] [Mehra *et al.*, 2020].

Ces faits invitent à la prudence sur la qualité des publications disponibles sur la Covid-19 et la validité des résultats utilisant des données médicales. Cependant, les dossiers électroniques restent une source d'information riche pour la recherche clinique : tous les sujets ne peuvent pas être traités par des études expérimentales interventionnelles (pour des raisons éthiques par exemple), favorisant ainsi les études observationnelles dans certains cas. L'utilisation des données de vie réelle pour la réalisation d'études observationnelles est de plus en plus courante, mais ces données sont difficiles à exploiter et leur utilisation est sujette à divers types de biais [Agniel *et al.*, 2018]. Des démarches commencent à être proposées pour identifier et limiter les erreurs et les fraudes liées à l'exploitation de données médicales [Boetto *et al.*, 2021].

Il convient de renforcer ces initiatives et de proposer des cadres de travail adaptés à ces données, afin de garantir la fiabilité et la reproductibilité des résultats. Les actions à mettre en place se situent à différents niveaux : fournisseurs de données de santé, investigateurs, éditeurs…

Nous contribuons à cet effort à travers l'analyse critique d'un projet mené au sein de l'Assistance Publique – Hôpitaux de Paris (AP-HP) et visant à analyser les conséquences de la crise liée à la Covid-19 sur les patients nouvellement diagnostiqués avec un cancer. Cet article constitue un retour d'expérience sur l'exploitation d'un entrepôt de données de santé (EDS) et un bilan sur les difficultés rencontrées lors de ce projet. Nous présentons les potentiels d'exploitation d'un EDS pour l'analyse de parcours de soins ainsi que les défis liés à l'étude de données de vie réelle dans le cadre d'études cliniques et du pilotage des trajectoires de soins des patients. Un résumé rapide des résultats cliniques tirés de ces travaux est également présenté.

2 Contexte et méthodes

Pendant la pandémie, la France et les établissements de santé ont pris des mesures drastiques afin de répondre au mieux aux besoins des patients infectés par la Covid-19 tout en maintenant un accès aux soins aux patients hors Covid-19. Cette restructuration des soins pose la question des éventuelles conséquences sur la survie des patients nouvellement diagnostiqués avec un cancer. De premières études ont été publiées suggérant des conséquences indirectes néfastes de la pandémie et plus particulièrement des mesures exceptionnelles mis en place sur le parcoures de soins des patients, avec davantage de morts du fait des retards diagnostics [Sud *et al.*, 2020] [Lai *et al.*, 2020]. Néanmoins, les résultats de ces études de modélisation étaient marqués par une forte incertitude.

Dès début 2021, des données de santé de vie réelle ont été disponibles au sein des EDS pour effectuer des analyses précises sur le devenir des patients nouvellement diagnostiqués avec un cancer pendant la première vague. Une étude réalisée aux États-Unis sur les patients nouvellement diagnostiqués avec un cancer solide métastatique montrait que la pandémie n'avait pas eu d'impact sur les délais de traitements [Parikh *et al.*, 2021]. En Angleterre, une étude montrait un allongement des délais de prises en charge des patients, sans pour autant conclure à une détérioration de la qualité des soins [Fox *et al.*, 2022].

L'AP-HP comprend 39 hôpitaux répartis sur la région Île-de-France. Depuis 2015, l'AP-HP centralise les données de tous ses hôpitaux et les met à disposition de chercheurs au sein de son EDS [Daniel, 2020]. Les informations pseudonymisées de 11 millions de patients remontant jusqu'en 2012 sont ainsi disponibles pour la recherche. Les données proviennent de différents logiciels spécialisés et sont regroupées dans un EDS commun. Ce processus est complexe, car si certains logiciels sont communs à tous les hôpitaux de l'AP-HP, pour d'autres, chaque hôpital a pu choisir une solution informatique différente (e.g., logiciel de prise de rendez-vous). Pour le moment, seule une partie des logiciels ont un flux de données consolidé entre le logiciel source et l'EDS. L'EDS intègre également des données publiques, comme les fichiers nationaux de décès mis à disposition par l'INSEE (Institut National de la Statistique et des Études Économiques) (Fig. 1).

La force de l'EDS de l'AP-HP repose sur la possibilité de croiser plusieurs catégories de données :

- Des informations structurées (données dont la valeur est prédéfinie et dont la structure est formatée), dont celles du PMSI (Programme de Médicalisation des Systèmes d'Information, qui intègre l'activité médicale codée selon la Classification Internationale des Maladies 10eme révision (CIM-10) pour les diagnostics et selon la Classification Commune des Actes Médicaux (CCAM) pour les actes)
- Des données non structurées comme du texte libre ou de l'imagerie. En effet, l'EDS de l'AP-HP dispose de plus de 70 millions de comptes-rendus médicaux et de 33 millions d'examens d'imagerie.

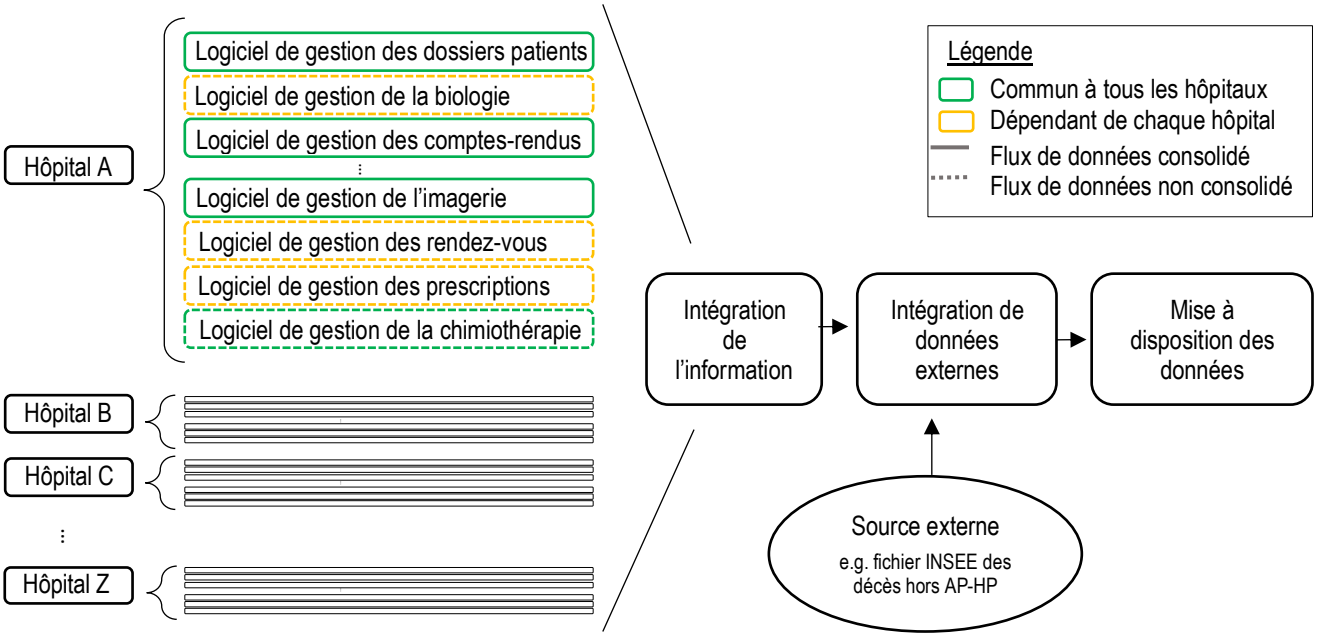

Figure 1. Processus d'intégration des données à l'EDS de l'AP-HP

Conjointement, ces données permettent d'obtenir une vision plus robuste de la réalité : les données non structurées sont complémentaires aux données structurées du PMSI. L'envergure de cet entrepôt en fait une source d'information unique dans le domaine médical européen.

Pour répondre à la problématique de l'impact de la crise Covid-19 sur les patients nouvellement diagnostiqués avec un cancer, un groupe de travail constitué d'oncologues médicaux, d'épidémiologistes, d'ingénieurs données et de *data scientists* a été mis en place au sein de l'EDS de l'AP-HP. Les questions auxquelles souhaitait répondre ce groupe de travail étaient les suivantes : quelles sont les conséquences de l'épidémie de Covid-19 sur les patients nouvellement diagnostiqués avec un cancer à l'AP-HP ? Les trajectoires de soins ont-elles été modifiées ? Les résultats cliniques ont-ils été affectés ? Nous allons décrire les étapes clés des travaux publiés, les défis rencontrés et comment ils ont pu être surmontés.

3 Résultats

3.1 Construction des parcours patient à analyser

L'analyse du parcours des patients atteints de cancer comporte quelques étapes clés : les modalités des procédures et le type de diagnostic, les différentes phases de traitement (type, dosage, durée) et l'issue du traitement (devenir clinique, consommation de soins). La structure de ce parcours diffère d'un type de cancer à un autre, et nombre des trajectoires de patients se déroulent en partie « en ville » (médecine générale, spécialistes, imagerie et biologie de ville). Pour analyser différents types de cancer à partir de l'EDS de l'AP-HP, il est donc tout d'abord nécessaire d'identifier un parcours de référence intra-hospitalier qui sera commun à un grand type de cancers. Ce parcours ne pourra pas être détaillé mais comportera les principales étapes et les traitements en oncologie.

Ensuite, pour analyser de façon fiable et exhaustive les parcours de soins, il est nécessaire de se concentrer sur les portions des parcours se déroulant au sein de l'AP-HP. En effet, l'EDS ne couvre que les informations générées lors de séjours ou consultations à l'AP-HP : toute procédure diagnostique ou thérapeutique réalisée dans un hôpital hors AP-HP ou en ville n'est pas disponible. La radiothérapie, par exemple, est couramment

réalisée dans des établissements privés partenaires de l'AP-HP. Les informations sur ces actes réalisés hors AP-HP ne sont pas contenues dans l'EDS, ce qui ne permet pas d'analyser les parcours centrés sur ce mode de traitement. Il serait impossible, par exemple, de calculer de façon fiable un indicateur de « délai avant traitement » si la date du traitement est inconnue. Certains parcours de soins ne sont ainsi pas identifiables.

3.2 Identification des sources de données disponibles

Une fois qu'un parcours reconstituable est identifié, les données permettant de calculer les délais entre chaque étape peuvent être collectées. Les patients sont ensuite répartis dans différents sous-parcours (e.g., chirurgie ou exclusivement de la chimiothérapie). Un premier dilemme se pose sur la détection de la date de diagnostic du cancer. Différentes dates sont envisageables, chacune avec ses difficultés :

- Le patient peut arriver à l'hôpital après un diagnostic de cancer effectué en ville, mais l'EDS ne contient aucune information structurée sur les procédures réalisées en ville.
- La date de la première rencontre entre le patient et le médecin hospitalier pourrait constituer une date de diagnostic approximative. Les consultations (service hospitalier et date) sont disponibles dans l'EDS, mais aucune codification de diagnostic ne leurs sont associées : le PMSI analyse l'activité médicale (classifications CIM-10 et CCAM) uniquement pour les hospitalisations (soit de jour, soit complète). On ne connaît donc pas la cause d'une consultation. Or un patient peut consulter un gastro-entérologue pour une suspicion de cancer comme pour d'autres troubles digestifs. Il n'est donc pas possible de déterminer si une consultation dans un service donné a eu lieu pour un cancer.
- Une autre piste repose sur la classification des tumeurs malignes comme une affection de longue durée (ALD30). Cette prise en charge permet l'exonération de l'ensemble des soins à partir d'une date de diagnostic précisée lors de la demande d'ALD30, même si le diagnostic a été fait en ville. Néanmoins, ces données ne sont pas disponibles dans l'EDS. Par ailleurs, un patient peut bénéficier d'ALD30 pour d'autres causes, ou ne pas en bénéficier, ce qui rend cette information peu fiable.
- Ainsi, pour ce projet, nous avons choisi d'utiliser comme date de diagnostic la première hospitalisation du patient à l'AP-HP avec un codage CIM-10 de cancer. Dans la majorité des cas, cette date est certainement ultérieure au début réel du parcours de soins pour cancer, mais elle apparaît comme la seule donnée fiable dont nous disposons.

De même, pour le moment, les données structurées de l'EDS sont essentiellement constituées des données du PMSI, mais tous les traitements perçus par les patients ne sont pas codés via ce système. Par exemple, les séances de chimiothérapie sont codées dans les tables du PMSI, mais les informations concernant le dosage et les substances ne le sont pas. Il existe un système de prescription de la chimiothérapie qui permettrait d'obtenir ces informations, mais les données de ce logiciel ne sont pas encore intégrées à l'EDS.

Enfin, la validation du contenu des tables utilisées est indispensable pour l'étude. Des différences dans la collecte des données peuvent induire des biais : la même information n'est pas collectée uniformément pour tous les individus au cours du temps. Par exemple, l'analyse des tables de l'EDS de l'AP-HP nous a permis de mettre en évidence une profondeur historique des données différentes pour chaque hôpital, du fait du déploiement successif du logiciel commun de gestion des dossiers patients. La mise en place de ce logiciel s'est faite progressivement. Ainsi, pour certains hôpitaux les données disponibles remontent à 2012, alors que pour d'autres elles commencent mi-2019. Un hôpital de l'AP-HP ne dispose pas encore de ce logiciel commun, ses données ne sont donc pas disponibles dans l'EDS de l'AP-HP. Ce biais pourrait amener à constater une augmentation de l'incidence au fil du temps, simplement parce que le périmètre de collecte des données s'étend progressivement. Un compromis entre la quantité et la qualité des données disponibles doit être décidé. Ici, l'étude cherche à comparer la période pré-pandémie et la période de la pandémie, ce qui nécessite une complétude de l'information similaire dans les périodes. Nous nous sommes donc restreints aux hôpitaux avec un historique suffisant pour notre analyse, diminuant la cohorte de patients d'environ 30%.

3.3 Mettre à profit les données textuelles

En complément des données structurées, les données textuelles ont une double utilité :

- Valider le codage de certaines variables, en croisant avec les données structurées ;
- Fournir des informations indisponibles dans les données structurées.

Les données structurées utilisées dans les études rétrospectives ont une visée médico-administrative et ne sont pas dédiées à la recherche. Il est donc indispensable de traiter ces données avec attention, car des biais liés au codage peuvent se répercuter sur les résultats. Pour certains types d'actes, plusieurs systèmes de codage peuvent être pertinents. Certains actes de radiothérapie peuvent ainsi être codés à l'aide d'un code CIM-10 et/ou d'un code CCAM. En analysant l'ensemble des patients avec un codage de radiothérapie par CIM-10 ou CCAM, nous observons que 10% de patients ont un code CCAM et un code CIM-10, 90% des patients ont uniquement un code CIM-10 et <0.5% des patients ont uniquement un code CCAM. Afin de trancher sur le(s) système(s) de codage à utiliser, une vérification de la justesse du codage des séances de radiothérapie via l'analyse des comptes-rendus textuels a été réalisée. On observe un sous-codage des actes CCAM et une justesse du codage CIM-10. Cette vérification manuelle nous a également permis d'identifier que la date de la séance n'était pas toujours exacte. En hospitalisation de jour, la séance de radiothérapie est codée à la date de la séance. En hospitalisation complète, toutes les séances sont codées le 1er jour de l'hospitalisation. Ces observations sont indispensables à l'évaluation des délais entre les différentes phases de traitement.

Les documents textuels permettent également de mener des études plus approfondies, en extrayant des informations supplémentaires. Nous avons par exemple mis en place des méthodes de Traitement Automatique des Langues (TAL) pour extraire les scores TNM des nouvelles tumeurs opérées. Le score TNM permet une classification de dissémination cancéreuse selon la taille de la tumeur principale, l'envahissement des ganglions régionaux et la présence de métastases. Obtenir ces scores permet d'identifier les stades d'avancement pronostique des tumeurs lors des diagnostics effectués avant la pandémie et pendant la pandémie. Pour extraire le score TNM, nous avons utilisé les comptes-rendus d'anatomopathologie.

Comme pour les données structurées, il est d'abord indispensable de qualifier la complétude des documents afin de ne pas introduire de biais de sélection. Par exemple, il est possible que les documents de certains hôpitaux ne se soient pas bien intégrés pendant une période, laissant ainsi un défaut de documents. Il est donc indispensable de vérifier la stabilité du nombre de comptes-rendus identifiés dans le temps avant d'entreprendre une étude sur ces documents. Pour les patients identifiés avec un cancer unique du côlon résécable, nous avons identifié un taux constant de 3% des patients sans comptes-rendus d'anatomopathologie depuis 2018. Comme nous nous intéressons dans cette étude à la comparaison de la période pré-pandémie à la période de la pandémie, ce taux stable de documents manquants n'a pas introduit de biais dans l'étude. Une validation de la date structurée du document en comparaison à la date extraite du texte a également été mis en place, pour vérifier que les comptes-rendus sont correctement datés dans la base. En effet, l'ajout en masse de documents non intégrés dans l'EDS pourrait engendrer des dates d'enregistrement dans la table dédiée erronées et ainsi fausser les analyses. La validation de la qualité et de la quantité des informations disponibles est indispensable pour ne pas introduire de biais dans l'étude.

Les performances du modèle de TAL développé ici ont été validées selon une procédure standardisée :

1. Annotation « à la main » de la tâche souhaitée par des experts du domaine.
2. Séparation des annotations en un jeu d'entrainement et un jeu de validation.
3. Utilisation du jeu d'entrainement pour améliorer les performances de l'algorithme.
4. Calcul des performances finales sur le jeu de validation.

Il est primordial travailler en aveugle du jeu de validation. La validation du modèle sur un jeu de données nouveau permet de garantir une estimation non biaisée de ses performances. Bien que très chronophage, la

validation des algorithmes de TAL par annotation manuelle des documents est indispensable. Pour l'algorithme d'extraction du score TNM, un oncologue médical a annoté plusieurs centaines de comptes-rendus d'anatomopathologie afin que nous puissions confirmer les performances de l'algorithme développé.

Enfin, au-delà de l'analyse du contenu, l'existence de données non structurées dans l'EDS peut permettre de collecter indirectement et rapidement des données structurées. Ainsi, les Réunions de Concertation Pluridisciplinaires (RCPs) en cancérologie ne sont pas codées dans le PMSI. Néanmoins, leurs comptes-rendus sont intégrés à l'EDS, ce qui permet de dater les RCPs à partir de la date du compte-rendu.

3.4 Un environnement en mouvement

L'analyse des données médicales suppose une bonne connaissance de leurs conditions de production (codage notamment) et de mise à disposition. Le codage structuré des actes et des diagnostics n'est ni instantané ni définitif. En effet, le codage PMSI est utilisé pour le remboursement des dépenses hospitalières par l'Assurance Maladie, aussi des codes peuvent être ajoutés, modifiés ou supprimés lors de la sortie du patient si le codage ne peut être réalisé avant ou lors de la validation du codage par la Direction de l'Information Médicale (DIM) de l'hôpital (parfois plusieurs mois après la fin de l'hospitalisation). De ce fait, en fonction de la date de réalisation de l'analyse, l'information pour un même patient et un même séjour peut varier. Une étude sur l'ampleur de ce phénomène est en cours. Afin de rendre les études reproductibles, il est donc indispensable de faire des sauvegardes des tables utilisées pour produire l'analyse finale. Ceci permet d'avoir une traçabilité et une validation des analyses lors d'éventuelles publications. A l'AP-HP, les données mises à disposition des chercheurs sont des captures à un instant t de la base pseudonymisée. Il est ainsi toujours possible d'exploiter la même table sans subir les effets de cet environnement mouvant.

L'EDS de l'AP-HP est une structure jeune. De nouvelles infrastructures sont en cours de développement afin de remédier à certains biais et de faciliter l'accès aux données, et de nouveaux flux de données sont en cours d'intégration, selon un système de priorisation. Les données disponibles évoluent donc au fil du temps, ainsi que l'infrastructure pour leur analyse. Il nous est apparu indispensable de garder un lien très étroit avec les responsables de cette infrastructure pour anticiper ces changements et évaluer leurs impacts sur nos résultats.

Afin de mener à bien ce projet et de réaliser des études rigoureuses, de nombreuses expertises ont été nécessaires. Tout projet mené au sein de l'EDS de l'AP-HP doit être soutenu devant un Comité Scientifique & Éthique. Un encadrement de la procédure à suivre pour obtenir l'approbation du comité a été indispensable. Ensuite, la création de la cohorte d'intérêt et la mise à disposition des données pseudonymisées dans un espace protégé a été possible grâce à des équipes spécialisées dans le traitement massif de données. Un partenariat entre des professionnels de santé (médecins cliniciens et DIMs) et des experts d'analyses statistiques et de développement d'algorithmes de *machine learning* a permis l'aboutissement de l'étude. La spécificité des données cliniques rend impossible leur exploitation sans expertise médicale. Au total, plus d'une dizaine de personnes sont intervenues tout au long du projet.

3.5 Résultats cliniques

Notre projet nous a permis de décrire l'évolution de l'incidence de différents types de cancers, de leur mode de prise en charge thérapeutique, des délais entre RCP et traitement et de la mortalité (y compris en tenant compte de l'infection ou non par la Covid19) en 2018-19 et 2020-21. Une première étude [Kempf *et al.*, 2021] portant sur l'ensemble des cancers a montré une diminution du nombre de nouveaux diagnostics de cancer à l'AP-HP pendant le premier confinement. Comparé à la moyenne de 2018 et 2019, l'AP-HP a connu une baisse de 40% de nouveaux diagnostics pendant le premier confinement français (17 mars 2020 – 11 mai 2020). Aucun effet de rattrapage (hospitalisations après la crise des patients non vus pendant la crise) n'est visible dans les données jusqu'à présent. Une analyse plus détaillée [Kempf *et al.*, 2022] a été conduite sur le cancer du côlon afin de comprendre comment les modifications des circuits de soins ont affecté le stade

d'avancement des cancers au diagnostic, les délais de prise en charge et les traitements des nouveaux diagnostics de cancer. Les patients diagnostiqués après le début de la pandémie ne semblent pas être diagnostiqués dans des stades plus avancés de leur cancer. De même, les délais de prise en charge intra-hospitaliers sont stables si on compare la période pré-pandémie à la période de la pandémie. On observe que les patients diagnostiqués et traités à l'AP-HP après janvier 2020 et ayant été sujets à une infection à la Covid-19 ont un taux de survie globale à un an plus faible que ceux diagnostiqués pré-pandémie ou ceux de diagnostiqués après 2020 sans contraction du virus.

Les résultats obtenus ont été restitués sous forme de publications académiques. Ils ont également été présentés à la direction de l'AP-HP de façon fréquente pendant la crise. Aucune donnée brute n'a pu être communiqué. Seules des données agrégées ont été présentées.

4 Discussion

Les EDS constituent une ressource majeure dans l'analyse de dossiers informatisés. La force de l'EDS de l'AP-HP réside dans sa taille, avec l'agrégation multi-sites d'informations variées. L'exploitation des données de tous les hôpitaux de l'AP-HP permet d'avoir une représentation large du milieu hospitalier en Ile-de-France.

Dans le domaine de la cancérologie, d'autres structures de données sont également disponibles. L'Institut Nationale du Cancer (INCa) dispose d'un sous-ensemble du système national de santé (SNDS) spécifique à la cancérologie. Cette table regroupe le codage de toutes les structures médicales à un niveau national (PMSI hospitalier, SNIIRAM (Système National d'Information Inter-Régimes de l'Assurance Maladie), causes médicales de décès), pour tous les patients traités en France, permettant ainsi des études statistiques sur de grandes cohortes. Par rapport à l'EDS de l'AP-HP, la force de cette table repose dans la présence des informations de ville et hospitalières. Cependant, il n'y a aucun lien entre le codage et d'autres informations complémentaires sur le patient (imagerie, biologie, comptes-rendus…). En termes de spécificités, par rapport à une base nationale de données structurées (SNDS/PMSI), les EDS hospitaliers peuvent présenter les avantages suivants :

- Richesse des données : comptes-rendus, images, données de génomique…
- Possibilité de croiser les informations entre données structurées et non structurées
- Rapidité de mise à disposition des données.

Mais ils ont également des faiblesses :

- Périmètre restreint : un EDS comprend une cohorte de patients (ceux vus dans l'établissement) nécessairement inférieure à une base nationale, ce qui peut affaiblir la puissance de l'étude.
- Périmètre restreint également car un EDS couvre par définition seulement les soins prodigués au sein de l'établissement, ce qui peut ne représenter qu'une part limitée de la trajectoire de soins.
- Chaque EDS est unique, à la différence du PMSI où la validation et la diffusion de bonnes pratiques est possible au niveau national [Quantin *et al.*, 2014] .

Les registres de cancérologie sont une autre structure de données cliniques. Leur force repose sur la complétude, la quantité et la qualité de l'information à disposition pour une population sur une zone géographique définie. Cependant, la constitution des registres est un processus long, ce qui ne permet pas d'études en temps « quasi-réel », pour lesquelles les EDS peuvent être plus adaptés.

De nombreuses études proposent des recommandations pour l'étude de dossiers patients informatisés, e.g. [Kohane *et al.*, 2021][Penberthy *et al.*, 2021]. Il convient maintenant de définir les modalités pratiques d'implémentation de ces recommandations pour garantir la fiabilité des résultats et l'efficacité des projets.

Pour cela, il est nécessaire de spécifier les rôles, responsabilités et outils applicables à différentes échelles : l'infrastructure, l'organisation des projets, et les méthodes d'exploitation des données. L'exploitation des données de santé est une opportunité majeure pour l'analyse de parcours de patients à des fins de recherche, de pilotage et d'évaluation. Pour remplir cette promesse, il faut nous donner les moyens de la rigueur et de la reproductibilité indispensables pour garantir la crédibilité des résultats.

Références

Agniel, D., Kohane, I.S. and Weber, G.M. (2018) 'Biases in electronic health record data due to processes within the healthcare system: retrospective observational study', BMJ, p. k1479. doi:10.1136/bmj.k1479.

Boetto, E. et al. (2021) 'Frauds in scientific research and how to possibly overcome them', Journal of Medical Ethics, 47(12), pp. e19–e19. doi:10.1136/medethics-2020-106639.

Daniel, C. (2020) 'La recherche clinique à partir d'entrepôts de données. L'expérience de l'Assistance Publique – Hôpitaux de Paris (AP–HP) à l'épreuve de la pandémie de Covid-19', La Revue de Médecine Interne, 41(5), pp. 303–307. doi:10.1016/j.revmed.2020.04.005.

Fox, L. et al. (2022) 'Association between COVID-19 burden and delays to diagnosis and treatment of cancer patients in England', Journal of Cancer Policy, 31, p. 100316. doi:10.1016/j.jcpo.2021.100316.

Horbach, S.P.J.M. (2020) 'Pandemic publishing: Medical journals strongly speed up their publication process for COVID-19', Quantitative Science Studies, 1(3), pp. 1056–1067. doi:10.1162/qss_a_00076.

Kempf, E. et al. (2021) 'New cancer cases at the time of SARS-Cov2 pandemic and related public health policies: A persistent and concerning decrease long after the end of the national lockdown', European Journal of Cancer, 150, pp. 260–267. doi:10.1016/j.ejca.2021.02.015.

Kempf, E. et al. (no date) 'Impact of two waves of Sars-Cov2 outbreak on the number, clinical presentation, care trajectories and survival of patients newly referred for a colorectal cancer: A French multicentric cohort study from a large group of university hospitals', International Journal of Cancer, n/a(n/a). doi:10.1002/ijc.33928.

Kohane, I.S. et al. (2021) 'What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask', Journal of Medical Internet Research, 23(3), p. e22219. doi:10.2196/22219.

Lai, A.G. et al. (2020) 'Estimated impact of the COVID-19 pandemic on cancer services and excess 1-year mortality in people with cancer and multimorbidity: near real-time data on cancer care, cancer deaths and a population-based cohort study', BMJ Open, 10(11), p. e043828. doi:10.1136/bmjopen-2020-043828.

Mehra, M.R. et al. (2020) 'Retraction: Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. N Engl J Med. DOI: 10.1056/NEJMoa2007621.', New England Journal of Medicine, 382(26), pp. 2582–2582. doi:10.1056/NEJMc2021225.

Mehra, M.R., Ruschitzka, F. and Patel, A.N. (2020) 'Retraction—Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis', The Lancet, 395(10240), p. 1820. doi:10.1016/S0140-6736(20)31324-6.

Parikh, R.B. et al. (2021) 'Impact of the COVID-19 Pandemic on Treatment Patterns for Patients with Metastatic Solid Cancer in the United States', JNCI: Journal of the National Cancer Institute, p. djab225. doi:10.1093/jnci/djab225.

Penberthy, L.T. et al. (no date) 'An overview of real-world data sources for oncology and considerations for research', CA: A Cancer Journal for Clinicians, n/a(n/a). doi:10.3322/caac.21714.

Quantin, C. et al. (2014) 'Qualité des données périnatales issues du PMSI : comparaison avec l'état civil et l'enquête nationale périnatale 2010', Journal de Gynécologie Obstétrique et Biologie de la Reproduction, 43(9), pp. 680–690. doi:10.1016/j.jgyn.2013.09.004.

Sud, A. et al. (2020) 'Effect of delays in the 2-week-wait cancer referral pathway during the COVID-19 pandemic on cancer survival in the UK: a modelling study', The Lancet Oncology, 21(8), pp. 1035–1044. doi:10.1016/S1470-2045(20)30392-2.