Structuration, standardisation et enrichissement par traitement automatique du langage des données relatives au cancer au sein de l'entrepôt de données de santé de l'Assistance Publique -Hôpitaux de Paris

Emmanuelle Kempf

Sorbonne Université

Ecole doctorale Pierre Louis de Santé Publique

Laboratoire d'informatique médicale et d'ingénierie en connaissances en santé

Structuration, standardisation et enrichissement par traitement automatique du langage des données relatives au cancer au sein de l'entrepôt de données de santé de l'Assistance Publique – Hôpitaux de Paris

Par Emmanuelle Kempf

Thèse de doctorat de Sciences des données

Dirigée par Xavier Tannier, co-encadrée par Christel Daniel

Présentée et soutenue publiquement le 9 octobre 2023

Devant un jury composé de :

Rougé Bugat, Marie-Eve, PU-PH, Présidente
Cuggia, Marc, PU-PH, Rapporteur
Jouhet, Vianney, PU-PH, Rapporteur
Le Bihan, Christine, PH, Examinatrice
Tannier, Xavier, Professeur, Directeur
Daniel, Christel, PH, Co-encadrante

Table des matières

Contexte de la thèse

I. L'oncologie médicale, un enjeu de santé publique

1. Etat des lieux de la discipline

Le cancer est à l'origine d'une mort sur six, et entraîne chaque année environ 10 millions de décès dans le monde. Ce nombre est amené à croître significativement dans le futur, en raison de l'allongement de la durée de la vie et de l'augmentation de notre exposition aux carcinogènes, environnementaux notamment (1,2). La précocité du diagnostic de la maladie, le développement d'innovations thérapeutiques efficaces ajustées au profil moléculaire de la tumeur et aux préférences du patient, la personnalisation des parcours de soins des patients constituent ainsi autant d'enjeux de prise en charge, favorisés par le déploiement des innovations des traitements anti-tumoraux et de support. La prise en charge thérapeutique des patients atteints de cancer constitue un défi majeur pour les cliniciens concernés, aussi aguerris soient-ils, compte tenu de l'engagement du pronostic vital des patients, de l'hétérogénéité clinico-biologique de la maladie, de l'efficacité encore insuffisante des thérapeutiques impliquées et de la cinétique rapide d'actualisation des données scientifiques sous-jacentes à la pratique clinique. Le profilage moléculaire des tumeurs est permis par la démocratisation des techniques de séquençage à haut débit du génome humain (3). Ce faisant, chaque type de tumeur se retrouve classé en autant de groupes de maladies définies par les altérations moléculaires ciblables par ces nouvelles molécules, créant de plus en plus de niches thérapeutiques (4). Pluridisciplinaire par essence, l'oncologie médicale constitue une discipline médicale propre à des améliorations concernant plusieurs pistes dynamiques de recherche.

2. Leviers de l'amélioration de la prise en charge des patients

a. Pilotage et système de santé apprenant

Les autorités sanitaires développent des programmes d'assurance qualité (AQ) fondés sur de bonnes pratiques cliniques et des recommandations de prise en charge. Ces programmes d'AQ peuvent accompagner des dispositifs de certifications de professionnels de santé et d'accréditations d'établissements de santé. Les indicateurs de qualité et de sécurité des soins (IQSS) contribuent au pilotage des structures de soins au cours du temps et comparativement aux structures du même type. La mesure d'IQSS peut contribuer à améliorer les prises en charge, réduire l'hétérogénéité de la pratique médicale et réduire les coûts (5). En France, l'institut national du cancer (INCa) développe, en partenariat avec la haute autorité de santé (HAS), des IQSS spécifiques au cancer. La production d'IQSS requiert souvent une saisie manuelle fastidieuse de données de sorte que l'évaluation de la qualité et de la sécurité des soins repose souvent sur des campagnes ponctuelles et ne porte donc que sur un très faible échantillon des pratiques faisant l'objet de l'évaluation. Les données nécessaires à la production d'IQSS sont pour partie collectées par ailleurs au sein du dossier patient dans le cadre du soin. Ainsi, cette double saisie dans le cadre du soin d'une part et du contrôle qualité d'autre part est une perte de temps et source d'incohérence.

Face à ce problème, la stratégie nationale de santé 2018-22 posait comme ambition de « développer des indicateurs de résultat, de vigilance et d'alerte pour les trois secteurs de l'offre de soins » dont le recueil « devra être automatisé sans surcharge de travail pour les professionnels » (6). Dans ce cadre, le présent cas d'usage s'est intéressé à la faisabilité du recueil automatique d'IQSS pour la prise en charge du cancer du sein, cancer le plus fréquent chez les femmes. Différentes études ont tenté de calculer automatiquement des IQSS pour le cancer du sein à partir de données médico-administratives structurées. Parmi les indicateurs d'intérêt pour la communauté médicale, seule une faible part est calculable à partir de données médico-administratives : 9 indicateurs sur 46 dans une première étude, 9 sur 367 dans une étude plus récente (7,8). Le jeu d'indicateurs développé par l'INCa pour évaluer la qualité des soins du cancer du sein suit aussi cette stratégie d'exploitation des données structurées : le processus de choix des IQSS excluait explicitement les indicateurs non calculables à partir des données structurées des bases médico-administratives (9).

Si ce choix est compréhensible car il permet un calcul à l'échelle nationale, les études ont largement montré que le type des indicateurs de performance mesurés à l'échelle d'un établissement affecte le comportement des acteurs de soins dans leur périmètre (10). Dans le cas du cancer du sein, les indicateurs non calculables à partir de données médico-administratives couvrent des dimensions importantes du processus de soins, et sont pour certains demandés dans les référentiels d'accréditation comme celui de l'organisation des instituts contre le cancer européens (OECI - oeci.eu) (11,12). Il est donc important d'essayer de couvrir au mieux les indicateurs pertinents, et pas seulement ceux facilement obtenables. Pour cela, il faut aller au-delà des données médico-administratives et utiliser un éventail de données plus riche pour obtenir les indicateurs. Le développement des dossiers patients informatisés présente une opportunité pour dépasser le périmètre des données accessibles via les seules sources médico-administratives (13,14). Cependant, un des verrous principaux

concernant le calcul d'IQSS réside dans le niveau de disponibilité des données au sein des dossiers patients informatisés. Ainsi, dans une étude américaine récente sur la qualité des soins en oncologie, le taux moyen de disponibilité des variables nécessaires au calcul des IQSS n'était que de 23% (15). Sur 19 indicateurs de qualité choisis, seuls deux étaient calculables pour plus de 1% des patients étudiés. En effet, la plupart des informations dans les dossiers patients informatisés sont contenus dans le texte libre des comptes rendus (16). L'exploitation des dossiers patients informatisés à grande échelle suppose donc de pouvoir extraire automatiquement les données du texte libre, et notamment à l'aide d'algorithmes de traitement automatique du langage naturel dont la fiabilité devra être précisément évaluée.

>> Cas d'usage 1 relatif à l'automatisation du calcul des indicateurs de qualité et de sécurité des soins EUSOMA du cancer du sein

Le cas d'usage d'application des algorithmes de structuration des items du jeu de données minimales liées au cancer cherchait à étudier la faisabilité du calcul automatisé d'IQSS pour le cancer du sein, à partir de données des dossiers patients informatisés. Les IQSS concernés choisis étaient ceux proposés par l'European Society of Breast Cancer Specialists (EUSOMA), et mesurés à partir du PMSI et du contenu textuel des comptes rendus d'anatomopathologie contenus dans l'entrepôt de données de santé (EDS) de l'Assistance Publique – Hôpitaux de Paris (AP-HP) (17,18). La calculabilité des IQSS a été établie à partir de la combinaison de la disponibilité des comptes rendus au sein de l'EDS de l'AP-HP, de la présence des variables élémentaires dans les comptes rendus, et des performances des algorithmes d'extraction des variables élémentaires. La méthodologie du cas d'usage 1 relatif à l'automatisation du calcul des IQSS EUSOMA se trouve dans l'Annexe 1a. Les principaux résultats du projet se trouvent dans la Partie III du mémoire de thèse, à la section 'Portabilité' des pages 88-89 et dans l'Annexe 1b.

b. Recherche clinique expérimentale

Depuis la seconde guerre mondiale, la survie globale médiane des patients atteints de cancer a doublé tout site primitif tumoral confondu, principalement en raison de la mise en œuvre de politiques de santé de prévention et de la mise à disposition sur le marché de médicaments antitumoraux innovants (19). Volet incontournable de cette spécialité médicale, le développement pharmaceutique permet, via la conduite d'études expérimentales randomisées, d'évaluer le développement de nouvelles molécules anti-tumorales ajustées au profilage moléculaire des tumeurs, en vue de l'amélioration du pronostic des patients atteints de cancer. L'un des obstacles récurrents à la réussite des essais cliniques interventionnels consiste en la faiblesse de participation des patients. L'inclusion des patients dans des essais thérapeutiques, condition incontournable au développement pharmaceutique ou de nouvelles modalités de prise en charge en oncologie, s'en retrouve complexifiée, avec un gâchis et un taux d'échecs non négligeables (20). Seulement jusqu'à 10% des patients atteints de cancer semblent être inclus au moins une fois dans un essai clinique (21). Une étude a montré que 18% des 789 essais cliniques en oncologie lancés entre 2000 et 2011 par l'institut

national du cancer américain (NCI) ont clôturé avec moins de 50% du nombre cible de patients atteints après 3 ans ou plus (22). Plus inquiétant encore, environ 20% des essais cliniques de phase II à IV en oncologie ne parviennent pas à se terminer 7 ans après leur lancement et malgré le recrutement cumulé de 48 000 patients (23).

Parmi les facteurs incriminés, la littérature a, par le passé, pointé du doigt le possible impact d'une représentation négative de la recherche clinique chez certains oncologues sur le taux d'inclusion des patients (24). D'autres études ont révélé qu'un manque d'information médicale peut entraîner une compréhension sous-optimale de l'intérêt de la recherche expérimentale et une crainte concernant le phénomène de randomisation chez une partie des patients atteints de cancer solide, mais la majorité de la population générale semblerait incline à participer à de telles études si l'occasion se présentait (25–27). La littérature plus récente montre en effet que le taux de retrait de consentement de la part des patients inclus dans des essais cliniques en oncologie ne dépasse pas 10%, que ces derniers semblent percevoir les bénéfices de leur participation à de telles études et qu'en revanche leur sortie de protocole est associée à un vécu émotionnel négatif (28–30). En parallèle des facteurs liés aux patients, une récente revue de la littérature des 30 dernières années a révélé que la part substantielle des obstacles au recrutement des patients atteints de cancer dans des essais cliniques avait part au processus de mise en correspondance des critères d'éligibilité avec les caractéristiques des patients pressentis (31). En effet, le nombre et la complexité des critères d'éligibilité des essais cliniques rendent leur adéquation avec les caractéristiques cliniques, biologiques et radiologiques des patients un défi quotidien pour les cliniciens impliqués dans l'inclusion des patients (32,33). Lorsque réalisée manuellement, la mise en correspondance des caractéristiques d'un patient avec les critères d'éligibilité d'un essai clinique est une tâche longue qui peut se laisser déborder par la progression tumorale du patient concerné, et l'échec définitif de son inclusion (20).

>> Cas d'usage 2 relatif à l'aide au recrutement (projet PENELOPE)

Le travail de thèse fut l'occasion de la création et du lancement de l'initiative française '*Patients ENrollment in clinical trials from rEal-Life databases using Common Data ModEls*' (PENELOPE) visant à améliorer le recrutement des patients au sein d'essais cliniques en tirant parti des dossiers patients informatisés en abordant au niveau national les défis d'interopérabilité et de qualité des données, en se concentrant sur un ensemble de données minimales telles que défini, pour le cancer, lors de la première partie du travail de thèse. Le groupe de travail français, associant l'AP-HP, l'institut Curie et les hôpitaux d'Angers, Brest, Nantes, Reims, Nancy, Strasbourg et Tours s'est constitué autour de la dynamique de thèse afin de concevoir l'interopérabilité de bases de données contenant des données de santé de vie réelle dédiées au cancer et en premier lieu à l'aide au recrutement de patients dans des essais cliniques. La démarche s'appuie sur la caractérisation de l'évolution de la maladie tumorale avec un niveau de qualité suffisant pour être réutilisable à des fins de préscreening à une échelle nationale en s'appuyant sur des terminologies standards. Les objectifs de l'étude préalable OMOP du projet PENELOPE étaient de comparer l'expressivité de deux versions différentes du modèle de données OMOP (v5.3 et 5.4) concernant la représentation du jeu de données minimales développées dans la première partie de la thèse pour le cas d'usage de

l'aide au recrutement en oncologie (cas d'usage 2) (34). En effet, alors que la version actuellement implémentée au sein de l'EDS de l'AP-HP est la v5.3, l'extension oncologie spécifiquement développée par OHDSI est contenue dans la version ultérieure (v5.4). La méthodologie et les résultats du cas d'usage 2 PENELOPE sont décrits au sein de la Partie II du manuscrit de thèse. Un glossaire dédié au projet se trouve au sein de l'Annexe 2.

c. Recherche clinique observationnelle

La recherche observationnelle permet, quant à elle, de valider en population réelle les résultats issus des études pivotales permettant l'enregistrement des innovations thérapeutiques par les agences régulatrices américaine, européenne et autres (35). Les études épidémiologiques sont précieuses, ainsi, pour des raisons de pharmacovigilance liées aux thérapies antitumorales, mais également concernant l'identification de facteurs de risque de cancer (36,37). Enfin, elles représentent autant d'outils précieux pour évaluer l'impact de mesures de santé publique sur les parcours de soins des patients, comme, p. ex., lors de la récente pandémie de SARS-CoV-2 qui a structurellement modifié l'organisation des soins (38). L'épidémie de SARS-CoV-2 et les décisions prises par les tutelles pour y faire face, dont deux périodes de confinement national, ont largement perturbé l'organisation des soins en oncologie en France. Au début de la pandémie de SARS-CoV-2, les systèmes de santé ont été fortement touchés par l'afflux de patients contaminés. La pandémie de SARS-Cov2 a conduit à des politiques sanitaires spécifiques visant à réduire à la fois sa dissémination et l'encombrement des hôpitaux. Parmi eux, des confinements ont eu lieu dans le monde entier, et en France pendant les trois périodes suivantes : 17 mars au 11 mai 2020 ; du 30 octobre au 15 décembre 2020 et du 3 avril au 3 mai 2021. De nombreux gouvernements ont instauré des mesures de confinement et de distanciation sociale, qui ont également affecté les services de santé. Par exemple, les centres de soins de santé spécialisés ont adopté des stratégies de triage pour les patients présentant des dysfonctionnements aigus d'organes (39). En France, Les consultations des médecins généralistes et des spécialistes ont chuté respectivement de 40 % et 50 % au cours de la période du premier mois de confinement (40).

Ces perturbations peuvent avoir altéré les trajectoires de soins des patients atteints de cancer dans le monde entier, et particulièrement les mesures de dépistage (41). La réaffectation de médecins de diverses spécialités dans des unités temporaires dédiées aux patients atteints du SARS-CoV-2 et les modifications apportées aux voies de traitement auraient pu perturber les délais des soins contre le cancer et le traitement choisi pour les patients (42–44). Afin de désengorger les établissements de santé de patients atteints de pathologie non aiguë, les sociétés savantes d'oncologie telles que l'American Society of Clinical Oncology ont publié des recommandations pour adapter les stratégies de soins à la situation de la pandémie, comme le report de toute procédure de dépistage du cancer pendant l'épidémie (45–47). En France, le 2 avril 2020, l'INCa a décidé d'interrompre les programmes nationaux de dépistage du cancer (48,49). Le nombre d'interventions diagnostiques du cancer a considérablement diminué pendant le confinement en France, avec l'utilisation de consommables tels que les contrastes iodés pour le scanner, les contrastes de gadolinium pour l'imagerie par résonance magnétique (IRM) et les liquides laxatifs pour la préparation des coloscopies diminuant respectivement de 500 000 (jusqu'à -72%), 280 000 (jusqu'à -72%) et 250 000 (jusqu'à -86%), entre le 16 mars et le 13 septembre 2020 (50). De premières données en France et ailleurs

firent rapidement état d'un nombre de consultations médicales et de nouveaux patients récemment diagnostiqués avec un cancer en net repli pendant le premier confinement du printemps 2020, par rapport aux années précédentes.

>> Cas d'usage 3 relatif aux études observationnelles sur données (projet CovOnco)

Les soignants ont craint, dans ce contexte, une rupture irréversible dans les filières diagnostiques et thérapeutiques des patients atteints de cancer. Cela signifierait que les patients atteints d'un nouveau cancer soient identifiés plus tard, et pourraient donc potentiellement : 1/ souffrir d'un retard dans le début de leur traitement antitumoral, 2/ se présenter au diagnostic avec des cancers plus avancés en termes de dissémination tumorale que de coutume. Par conséquent, il sembla important d'évaluer l'importance d'une telle baisse, car les retards de traitement peuvent avoir un impact considérable sur le pronostic des patients atteints de cancer, les études de modélisation anticipant des dizaines de milliers de morts supplémentaires liées à la gestion de l'épidémie (51–53). Le projet CovOnco a ainsi tenté d'analyser au plus proche l'impact de l'épidémie sur le système de soins en oncologie, pour informer le pilotage de l'AP-HP et la communauté scientifique, du déploiement éventuel de nouvelles modalités de prise en charge des patients atteints de cancer - incluant d'éventuels retards diagnostiques et aggravations des pronostics des patients. La méthodologie et les principaux résultats du cas d'usage 3 CovOnco se trouvent dans les Annexes 3a - 3c.

d. Aide à la décision médicale par vision par ordinateur

L'intelligence artificielle (IA) est une discipline mathématique et informatique dont l'objectif est le développement d'algorithmes capables d'apprendre à partir des données et d'outiller l'humain lors de la réalisation d'activités variées. Le domaine de la santé est un des secteurs où l'IA suscite le plus d'espoir avec en perspective des médecins « augmentés » disposant d'assistants numériques leur proposant automatiquement des diagnostics, des pronostics et des options thérapeutiques tenant compte de la variabilité individuelle chère à la médecine personnalisée (54). Les images médicales constituent des données essentielles à l'évaluation diagnostique et pronostique des patients et l'imagerie médicale est un des premiers domaines d'émergence de l'IA en santé. Si des systèmes d'aide au diagnostic en imagerie médicale existent depuis plus de vingt ans dans le domaine de l'analyse automatique d'images médicales, les nouvelles technologies de l'IA, portées par l'augmentation de la puissance de calcul disponible et la sophistication des algorithmes, ont permis, autour de 2010, l'émergence d'une discipline récente, la « imagomique », consistant à intégrer les données nécessaires au développement et à l'évaluation d'algorithmes d'intelligence artificielle diagnostiques et pronostiques basés sur l'extraction automatique de données physiopathologiques à partir d'images médicales (scanner, IRM, PET, lames digitales, etc) (55). Le développement des modèles d'IA en imagerie médicale dépend de la disponibilité de jeux de données de qualité, en termes de de données d'entrée (variables explicatives) mais également de sortie (variables expliquées).

>> Cas d'usage 4 relatif à l'IA en imagerie (projet 'Challenge AI for health')

En décembre 2020, à la jointure du plan IA2021 et de la stratégie Smart Santé annoncés respectivement en octobre 2018 et en septembre 2020, la Région Île-de-France a annoncé le lancement du « AI for Health Challenge 2020 » : deuxième Challenge international sur l'oncologie en partenariat avec l'AP-HP, l'Institut Curie et le soutien de Startup Inside, Medicen Paris Région et OVH, doté d'un grand prix d'un million d'euros (56–58). L'initiative a ciblé des startups et PME européennes, associées entre elles ou avec des laboratoires de recherche, avec implantation à court terme dans la région d'Île-de-France. « Convaincue qu'un accès facilité aux données de santé constitue aujourd'hui un facteur clé de succès pour le développement et la validation des futures innovations en santé ainsi que pour la structuration de la filière de la médecine du futur en Ile-de-France, la Région vise également à contribuer à l'émergence des futurs champions européens en IA appliquée à la santé » (58). L'objectif du projet visait à développer des algorithmes d'identification de biomarqueurs en oncologie solide. A l'Institut Curie, il s'agissait d'informer la décision thérapeutique dans le cancer du poumon, via la constitution d'une classification des cancers broncho-pulmonaire à partir de données multiples multimodales (cliniques, radiologiques, anatomopathologiques et génétiques) issues d'une large cohorte de patients. A l'AP-HP, il s'agissait de prédire des facteurs histologiques pronostiques dans le cholangiocarcinome intra-hépatique, un cancer rare, dont l'incidence augmente et le pronostic demeure sombre (59,60). En s'appuyant sur l'expertise des centres de référence que sont Paul Brousse, Beaujon et Henri Mondor, le défi proposé par l'AP-HP consistait à développer un modèle de vision par ordinateur à partir de scanners standards afin de mieux prédire l'agressivité des tumeurs opérées. L'objectif du partenariat développé avec la start-up lauréate serait de développer ultérieurement des outils d'intelligence artificielle pour l'analyse d'imageries (scanners standards et lames d'anatomopathologie), afin de prédire les facteurs pronostiques, comme la fibrose et l'infiltrat immunitaire, mais aussi des altérations moléculaires portées par la tumeur, et ciblables par des thérapies antitumorales. Un descriptif du rationnel scientifique et de la méthodologie du projet est disponible au sein de l'Annexe 4.

II. Les bases de données de santé à grande échelle, autant d'opportunités pour la discipline, nonobstant quelques verrous

Le développement actuel des EDS suscite un enthousiasme certain au sein de la communauté médicale et scientifique qui y voit une mine d'informations précieuses liées à l'expérience dite « en vie réelle » des patients et ce, dans le cadre d'une collecte massive d'informations liée aux soins (61). De tels entrepôts ont pu voir le jour grâce au développement des techniques de stockage de données numériques à grande échelle, à la mise en œuvre croissante de l'informatisation des dossiers médicaux et à l'interfaçage des différents logiciels encadrant les parcours de soins des patients. Ainsi, de plus en plus d'établissements de santé s'engagent dans la mise en œuvre d'EDS afin d'améliorer le pilotage de l'activité hospitalière et de faire avancer la recherche et l'innovation en favorisant la réalisation d'études observationnelles, la mise en place d'outils d'optimisation d'essais cliniques et le développement d'algorithmes

d'aide à la décision, basés notamment sur les nouvelles technologies d'intelligence artificielle. Les EDS intègrent les informations collectées en routine à l'occasion de consultations médicales, séjours d'hospitalisations, réalisation d'examens paracliniques diagnostiques ou de suivi, etc. Les données concernées sont très variées : données démographiques, signes vitaux, résultats d'analyses biologiques, comptes rendus et images de radiologie ou d'anatomopathologie, comptes rendus et données de séquençage génétique, de prescriptions, dispensations ou administrations médicamenteuses, de notes ou comptes rendus cliniques de médecins ou de personnels paramédicaux, de données générées par des dispositifs médicaux (électrocardiogrammes, courbes pondérales, etc.).

1. Des promesses …

Dans le domaine du cancer particulièrement, l'hétérogénéité de la pathologie, des parcours de soins et des patients eux-mêmes, les biais de sélection significatifs des études cliniques interventionnelles et le développement abyssal des sous-catégories nosographiques n'en augmentent que d'autant leur attrait (62–65). La mise à disposition d'une information médicale, paramédicale, médico-sociale et environnementale standardisée et numérisée concernant des dizaines de milliers de cas permet aux cliniciens et chercheurs de constituer des cohortes de patients partageant un ensemble de caractéristiques similaires avec un minimum de biais de sélection, et capturant l'information constitutive de leurs « génome », « phénome » et « exposome » (66). Ces technologies sont susceptibles d'améliorer la prise en charge des patients atteints de cancer, en embrassant les quatre leviers d'améliorations présentés dans la partie précédente, et étayées par les quatre cas d'usage de la thèse correspondants et développés ci-après :

- Concernant le *pilotage de l'activité* et le calcul des IQSS (cas d'usage 1), les systèmes d'informations de santé à grande échelle pourraient devenir « apprenants », au sens où les données cliniques colligées en leur sein permettraient, via leur analyse en pilotage de l'activité hospitalière, l'intégration de pratiques soignantes optimisées et, de ce fait, améliorant les soins prodigués aux patients.

- Les dossiers patients informatisés sont susceptibles de devenir un précieux soutien des différentes phases d'un *essai clinique* (cas d'usage 2) : conception du protocole, étude de faisabilité, recrutement de patients, collecte de données d'efficacité et de tolérance (67–78). Par exemple, l'utilisation des données de soins de routine semble prometteuse pour évaluer semi-automatiquement l'éligibilité des patients atteints de cancer aux différents essais cliniques en cours, et ce, pour accélérer leur accès à l'innovation thérapeutique (67,79). L'étape de préscreening peut se définir comme « une sélection initiale de candidats potentiels à partir d'une population pertinente basée sur un sous-ensemble de critères d'admissibilité prioritaires » (67). Un outil de préscreening numérique n'est qu'une composante d'un système d'aide au recrutement d'essais cliniques (SAREC) dédié au recrutement de patients atteints de cancer. Cette phase initiale pourrait bénéficier du secours des dossiers patients informatisés en vue de l'identification automatique de correspondance entre les caractéristiques d'un patient et celles d'un essai clinique, même si l'éligibilité finale des patients est moins susceptible d'être automatisée (67,68,80,81).

- Le rôle des EDS n'est plus à démontrer concernant la tenue d'*études épidémiologiques* dédiées au cancer, et ce, potentiellement à partir de bases préexistantes ou interopérées (cas d'usage 3) (82–85). Plus récemment, le développement et la mise à disposition des données -omics qui sous-tendent la nosographie de la discipline et l'individualisation des prises en charges en oncologie trouvent une cible idéale de stockage dans les EDS. Accessibles en masse, ces informations permettent d'étayer la recherche translationnelle qui ne cesse de raffiner le développement thérapeutique ciblé du cancer, suivant le paradigme de la médecine dite de précision (86).

- Enfin, l'automatisation de l'*aide à la décision médicale* en oncologie (cas d'usage 4) connaît un regain actuel d'intérêt, à l'heure où la cinétique de renouvellement des recommandations de bonnes pratiques émises par les sociétés savantes ainsi que l'éclatement nosographique précédemment exposé permettent difficilement aux praticiens un alignement en temps réel entre la littérature et la caractérisation « personnalisée » d'un patient atteint de cancer (87,88). Toutefois, la plus grande prudence demeure de mise concernant cette démarche qui ne vise pas à substituer l'art médical du processus décisionnel partagé avec le patient par la machine, si performante fût-elle (89).

2. … et des verrous

Malgré ces séduisantes promesses, les EDS présentent des obstacles de taille quant à leur exploitation systématique en vue d'applications translationnelles, cliniques, épidémiologiques, de pilotage et interventionnelles. En France, la HAS a ainsi ressenti la nécessité de mettre récemment à disposition des recommandations concernant les modalités du développement d'EDS (90). Parmi les obstacles non spécifiques à la discipline de l'oncologie, les enjeux éthiques et particulièrement ceux liés au respect de la vie privée des patients et dont les données sont exploitées représente un sujet de débat intense (91–94). La question du coût financier et humain concernant la mise en place, le développement et la maintenance des EDS constitue un frein à la pérennisation de leur exploitation (95).

En rapport avec leurs applications directes, l'exploitation massive des données de santé liées au cancer requiert que ces données soient accessibles et réutilisables dans le contexte précis des différents cas d'usage d'intérêt. D'une part, l'accessibilité des données comporte un premier volet organisationnel lié à leurs règles de gouvernance et un second volet technologique de leur gestion respectant la démarche *findable, accessible, interoperable, reusable* (FAIR) promue à l'heure du partage des données massives en santé (96). D'autre part, la réutilisation des données dépend de leur niveau de qualité et de pertinence eu égard au cas d'usage en question. A cet endroit, cette démarche, qualifiante pour les EDS, s'inscrit dans la dialectique de la recherche clinique dite « pragmatique » développée depuis plusieurs décennies déjà, et qui valorise l'effectivité réelle d'une donnée issue de la recherche, plutôt que la qualité de son élaboration originelle (97). Ainsi, trois niveaux de défis majeurs semblent d'emblée identifiables à l'heure de l'exploitation des données massives en santé :

- Leur qualité et utilité par rapport aux cas d'usages d'intérêt ;

- Leur niveau de standardisation ;
- Leur niveau de structuration (98,99).

a. Verrou 1 : la qualité et l'utilité des données massives rapportée aux cas d'usage (ou la nécessaire identification d'un jeu de données minimales en oncologie)

Les problèmes de qualité des données de « vie réelle » des EDS hospitaliers sont de plus en plus documentés dans la littérature, ainsi que les biais d'analyses susceptibles d'en découler (100,101). La question de la qualité peut se décliner selon différentes catégories parmi lesquelles les suivantes sont les plus discutées dans la littérature : accessibilité, précision, exhaustivité, conformité, cohérence et fiabilité (102). La qualité des données sous-jacentes aux études semble encore sous-évaluée par leurs auteurs (103). Si l'accès aux EDS demeure séduisant pour l'utilisation secondaire des données, les données structurées disponibles proviennent majoritairement de sources médico-administratives. En France, ces données utilisées pour le programme de médicalisation des systèmes d'information (PMSI) sont insuffisantes à caractériser les patients atteints de cancer, selon les jeux de données minimales attendus, et particulièrement dans le cadre de la médecine dite de précision. Ainsi, le stade tumoral est connu de longue date pour être sous-évalué dans les données issues du codage, alors que le statut métastatique est une notion complexe à extraire des données administratives des dossiers patients (104,105).

Les données clés d'oncologie mises à disposition des utilisateurs des entrepôts de données sont souvent manquantes dans les espaces de travail dédiés (106). Une étude réalisée à partir de la base de données nationale du cancer américaine a montré qu'entre 40% et 71% des 63 variables enregistrées dans le registre national du cancer étaient manquantes au sein d'une cohorte de plus de 4 millions de patients atteints des 3 cancers les plus fréquents, la survie globale de cette sous-population étant significativement moindre que celle des patients sans donnée manquante (107). Par exemple, le type de diagnostic du cancer, information clé de la caractérisation d'un patient concerné, manque souvent d'exhaustivité et de précision dans les EDS. Ainsi, une étude américaine a montré que cette catégorie de variable était significativement influencée par le type de parcours de soins des patients, et les personnels responsables du codage du diagnostic (108). Parallèlement, l'identification des données correspondant au traitement du cancer, comme l'administration de chimiothérapie, sont associées à des métriques sous-optimales dans les EDS, comparées aux données sources des dossiers patients ou à celles des registres d'oncologie (109,110). Le renseignement du statut vital des patients - qui constitue le critère de jugement étalon de la majorité des études cliniques observationnelles et interventionnelles en oncologie – peut également manquer de fiabilité, malgré le chainage avec les registres nationaux des décès (111).

Ainsi, la formalisation du type de variables déterminantes pour la caractérisation d'un patient atteint de cancer – type de cancer, historique des traitements antitumoraux reçus, résultats d'imagerie médicale, analyse histologique et de biologie moléculaire de la pièce tumorale, résultats biologiques, etc., permettrait leur hiérarchisation et leur priorisation au sein de l'ensemble des données de soin à intégrer en routine dans les EDS, et faciliterait le contrôle de leur qualité.

b. *Verrou 2 : le niveau de standardisation des données (ou l'optimisation nécessaire de leur modélisation)*

Une des limites majeures des EDS actuels réside en l'absence de leur standardisation commune dans le domaine de l'oncologie. Cette caractéristique freine ainsi leur exploitation à large échelle, que ce soit au sein d'un même établissement de santé que dans le cadre de projets de recherche multicentriques internationaux sur données interopérées (112). Pour permettre une interopérabilité sémantique et syntaxique des EDS, il semble nécessaire de mettre en place un modèle commun à toutes les institutions impliquées, impliquant un alignement terminologique pérenne et actualisé dans le temps (113). L'intégration des données éparpillées au sein de différents systèmes, développés en silo et non communicants, nécessite de mettre en place une chaîne de structuration et de standardisation des données reposant sur l'adoption de modèles et de terminologies standards. Il est question, ici, d'adopter des contraintes sémantiques par la formalisation de jeux de valeurs, formellement associés à des attributs d'un modèle de données. Ces travaux dans le domaine de l'interopérabilité des données doivent permettre, à terme, d'intégrer des données hospitalières et de médecine de ville, des données générées au domicile des patients et des données environnementales. Pareille démarche permettrait de disposer d'une vision plus complète du parcours de santé des patients atteints de cancer, ainsi qu'un support efficient pour les études de recherche translationnelles en oncologie (114). Actuellement, les données nécessaires à la caractérisation des patients atteints de cancer sont aujourd'hui contenues dans une compilation de données hétérogènes, structurées ou non structurées, peu hiérarchisées et non croisées avec d'autres sources d'information.

A l'heure où se développent les plateformes de partage et d'exploitation de données à plus grande échelle qu'elles soient nationales (Health Data Hub en France, Medical Informatics Initiative (MII) en Allemagne, Shared Health Research Information Network (SHRINE) et National Patient-Centered Clinical Research Network (PCORnet) aux Etats-Unis) ou internationales (espace de données de santé européen (EHDS), Observational Health Data Sciences and Informatics (OHDSI), European Health Data Evidence Network (EHDEN), TriNetX), les EDS hospitaliers ont vocation à constituer un composant majeur de ces initiatives (115–122). Pour cela, les EDS sont appelés à standardiser leurs données selon des référentiels internationaux tels que p. ex. le standard Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR), le modèle de données commun Observational Medical Outcomes Partnership (OMOP), le modèle Digital Imaging and Communications in Medicine (DICOM) pour l'imagerie et les terminologies de référence (classification internationale des maladies, 10e révision (CIM10), Logical Observation Identifiers Names and Codes (LOINC), Anatomical Therapeutic Chemical (ATC), etc.) (122–128). Au-delà d'un choix purement théorique, chacun de ces standards, non mutuellement exclusifs, entraine avec lui une certaine capacité à répondre aux différentes spécificités d'une spécialité médicale et/ou de chacun des cas d'usage concernés. Ainsi, le modèle OMOP a développé une extension génomique en 2019 et dédiée au cancer en 2021 et l'initiative mCODE vise à développer un guide d'implémentation lié à l'oncologie (129–133).

c. *Verrou 3 : le niveau de structuration des données (ou l'apport du traitement automatique du langage naturel)*

Un autre obstacle, et non des moindres, réside dans le stockage sous format textuel et non structuré d'informations de santé clés (134). Certaines initiatives cherchent à développer des algorithmes de traitement automatique du langage naturel afin de constituer, en temps réel, des jeux de données de patients atteints de cancer disponibles pour la constitution de cohortes (135–141). Ces techniques à l'intersection de l'intelligence artificielle, de l'informatique et de la linguistique permettent la modélisation, la hiérarchisation et la mise en perspective des données issues de comptes rendus textuels, correspondant au support largement majoritaire de l'information relative au diagnostic, au traitement et au suivi des patients atteints de cancer (142). La caractérisation d'une maladie à partir des EDS peut ainsi s'appuyer sur la reconnaissance d'entités nommées, une sous-tâche du traitement automatique du langage naturel consistant à identifier et caractériser dans les textes les mentions d'entités ou de concepts pertinents pour un objectif donné. Ce domaine a tiré pleinement parti du développement récent des modèles d'apprentissage machine (143). Ces techniques trouvent un essor particulier en oncologie avec la prise en charge de la grande dimension et du manque de structuration des données concernées (144).

Alors que les systèmes d'extraction textuelle traditionnels basés sur des règles nécessitent le développement par les experts de terminologies exhaustives et de plusieurs itérations de règles, l'apprentissage machine pourrait être plus résilient face à l'évolution de la variabilité linguistique des multiples entités médicales, ainsi que face à l'hétérogénéité formelle des textes libres sous-jacents aux comptes rendus médicaux (145). Les algorithmes d'apprentissage classiques se sont ainsi vus, depuis quelques années, remplacés par les réseaux de neurones artificiels, en raison du bond sans précédent des puissances de calcul informatique et à l'augmentation massive des données collectées disponibles pour leur entraînement (146). Par ailleurs, des innovations méthodologiques récentes ont permis une amélioration des méthodes neuronales pour les principales tâches du domaine du traitement du langage naturel : reconnaissance d'entités nommées, *question-answering*, et classification. Ainsi, les plongements lexicaux (*text embeddings*) permettent la modélisation d'une représentation de la langue. Les modèles de langues ainsi construits pour un usage particulier peuvent être transférés dans d'autres domaines ou pour d'autres applications, ce d'autant que différentes architectures neuronales ont pu améliorer significativement leurs performances (147). Néanmoins, la langue française reste encore trop peu le sujet d'applications efficientes de ces innovations, y compris les modèles de langue de grande taille (*large language models)* (148–150).

L'adaptation des modèles d'apprentissage machine à de nouveaux jeux de données en oncologie est une tâche chronophage qui limite l'extension de leurs applications, notamment en raison d'habitus linguistiques conjecturaux (151). Les algorithmes d'apprentissage machine sont des outils précieux pour des tâches telles que l'extraction de relations qui pourraient éventuellement raccourcir le temps d'extraction d'informations, mais leur entraînement nécessite de larges jeux de données manuellement annotées ce qui constitue une limite à leur déploiement (152,153). Les tâches d'annotations demeurent des obstacles majeurs au développement de modèles d'apprentissage machine, aboutissant à une majorité de corpus

de textes annotés de taille moyenne allant de quelques centaines à quelques milliers de documents seulement (154). Plusieurs techniques d'optimisation des tâches d'annotations manuelles ont été développées à des fins d'automatisation. Certains auteurs ont recours à des systèmes de règles comme outils de pré-annotation automatique en vue de l'entraînement ultérieur de modèles d'apprentissage machine (155–157).

III. Contributions

L'objectif de cette thèse était de démontrer qu'il est possible de lever les verrous d'utilisation secondaire des données de l'EDS de l'AP-HP concernant des patients atteints de cancer solide à diverses finalités telles que le pilotage de la sécurité et de la qualité des soins, et les projets de recherche clinique observationnelle et expérimentale.

Dans cette perspective, il s'est agi de :

- Proposer une liste d'informations clés de caractérisation d'un patient atteint de cancer solide, en vue de la constitution d'un jeu de données minimales (objectif 1) ;
- Identifier un modèle de données commun de représentation de ce jeu de données minimal, et optimiser la structuration et la standardisation des données selon ce modèle dans le domaine de l'oncologie (objectif 2) ;
- Identifier et extraire les informations clés non structurées par traitement automatique du langage naturel à partir des comptes rendus textuels contenus dans l'EDS de l'AP-HP, tout en discutant la façon d'optimiser les techniques d'extraction textuelle (objectif 3).

1. Premier objectif : spécification d'un jeu de données minimales polyvalent en oncologie

L'étape préliminaire du projet de thèse a consisté à définir les items constitutifs d'un jeu de données minimales liées au cancer, et ce, en vue de diverses finalités. Il s'est agi de définir les données nécessaires pour le pilotage, l'aide au recrutement dans les études observationnelles et interventionnelles, permettant dans le même temps la constitution d'un jeu de données propre à servir de référence pour le développement algorithmique en intelligence artificielle.

2. Deuxième objectif : identification d'un modèle de données commun optimal en oncologie

Dans une démarche de recherche d'interopérabilité d'EDS en lien avec le cancer, le travail a cherché à optimiser la standardisation des données d'intérêt identifiées lors de la première partie, selon le modèle de données commun international OMOP. Cette étape de la thèse a évalué l'expressivité de la version 5.3 d'OMOP, telle qu'actuellement déployée au Health Data Hub ainsi qu'à l'AP-HP, puis estimé la plus-value théorique apportée par son extension oncologie, à savoir la version 5.4 du modèle.

3. Troisième objectif : extraction de données du jeu de données minimales par traitement automatique du langage naturel

Des travaux de structuration et d'enrichissement du jeu de données minimales identifié lors de la première partie du travail ont été réalisés. Ceci a permis de développer et valider des algorithmes de traitement automatique des langues relatifs à l'extraction de cette information médicale clinique, histologique et radiologique, en vue d'une application immédiate au sein d'études de pilotage, épidémiologiques ou de développement d'algorithmes de vision par ordinateur dans le domaine de l'oncologie. En parallèle, les options méthodologiques des tâches d'extraction textuelle ont été discutées, dans un contexte de limites des ressources d'annotations.

Chacun des trois objectifs sera déployé au sein de trois parties thématiques dédiées, elles-mêmes décomposées en Introduction, Méthodes, Résultats et Discussion.

Partie I : Spécification d'un jeu de données minimales polyvalent en oncologie

I. Introduction

L'utilisation secondaire des données de "vie réelle" en oncologie pose la question de l'adéquation au cas d'usage envisagé de données collectées dans le cadre du soin et donc *a priori* non adaptées au contexte de l'utilisation secondaire. Quel que soit le cas d'usage envisagé, il convient dans un premier temps d'identifier au sein des données de "vie réelle", celles qui sont nécessaires à la réalisation du cas d'usage. La première étape de la thèse consista en un inventaire médical des données minimales d'intérêt en lien avec le cancer. Il s'agissait d'identifier les informations clés cliniques, biologiques et radiologiques permettant de caractériser pertinemment un patient atteint de cancer solide, en vue d'applications très diverses. L'originalité du travail réalisé dans le cadre de la thèse résida dans la perspective ascendante du choix des items d'intérêt, s'inscrivant dans la veine de la recherche dite pragmatique en oncologie (158). Il s'agissait de pouvoir exploiter de façon efficiente un EDS en l'état, c'est-à-dire de définir les items permettant de vérifier et valoriser l'applicabilité de l'EDS à des cas d'usages divers en oncologie, plutôt que de définir les items permettant la modélisation exhaustive de la pathologie tumorale (97).

La recherche d'un jeu de données minimales pragmatique telle qu'optée dans le travail de thèse s'inscrit dans une démarche déployée depuis quelques années, et particulièrement aux Etats-Unis. Dès 2009, une collaboration entre l'ASCO, l'institut national du cancer américain et le National Community Cancer Center Program conduit au projet Clinical Oncology

Requirements for the EHR (CORE) qui cherchait à aider les cliniciens dans le développement et l'adoption de dossiers patients informatisés, en précisant les fonctionnalités, données clés minimales et conditions d'interopérabilité attendues en oncologie, et ce, dans une approche purement orientée par la clinique de routine (159). Cette démarche s'enracine dans la pratique clinique de routine médicale qui vise à synthétiser autour d'un nombre minimal d'informations les caractérisations d'un patient atteint de cancer, et ce, particulièrement lors des réunions de concertation pluridisciplinaires (RCP) permettant d'instruire les cas et proposer les trajectoires de soins optimaux aux patients concernés (160,161).

L'objectif de cette première étape de la thèse était de proposer une méthodologie de définition et une première validation de contenu d'une liste minimale de données de santé de "vie réelle" à des fins de repérage et de caractérisation de patients atteints de cancer, et ce, dans le cadre de la réalisation de quatre cas d'usage dans le domaine de la cancérologie : le pilotage de la qualité et de la sécurité des soins, l'aide au recrutement, les recherches épidémiologiques et le développement de l'IA en oncologie. Les sources de données de "vie réelle" ont été tirées des dossiers patients informatisés.

II. Méthodes

1. Constitution d'un jeu de données minimales polyvalent en oncologie

Depuis des décennies, des jeux de données de santé minimum ne cessent de se développer. Parmi eux, le Minimum Basic Data Set a tenté de formaliser la pratique en médecine générale au tout début des années 1990's (162). S'il est associé à une sensibilité optimale pour l'identification de cas de cancer à partir de données d'entrepôts, sa valeur prédictive positive ne permet pas d'en faire un outil robuste de constitution de cohortes en vue de la tenue d'études épidémiologiques en oncologie (163).

La méthodologie de spécification de listes minimales de données comporte différentes étapes: i) constitution du domaine, ii) état de l'art et état des lieux, iii) spécification d'une première version de la liste minimale de données par un groupe d'experts, iv) révision de la liste par le groupe d'experts, v) validation de la liste (164). Lorsque la liste minimale de données doit être gérée par un système d'information de santé interopérable, une étape supplémentaire de structuration et de standardisation de cette liste selon des standards internationaux doit être réalisée. Une étape complémentaire consiste à évaluer, dans le cadre de preuves de concept, la liste minimale de données selon deux critères : la faisabilité de la constitution d'un jeu de données de qualité d'une part, et l'utilisabilité de ces données dans la réalisation du cas d'usage d'intérêt. La contribution du travail de thèse a porté sur l'étape méthodologique de spécification de données d'intérêt du dossier patient informatisé à partir de cas d'usage variés et sur une première évaluation de la qualité de ces données, de leur pertinence et de leur utilité dans le cadre des cas d'usages considérés. La Figure 1 illustre les étapes de la méthodologie adoptée pour la spécification de la première version du jeu de

données minimales polyvalent en oncologie - et le périmètre couvert dans le cadre de la thèse. Dans cette perspective, le jeu de données minimales issu de la connaissance experte et des référentiels métiers a été confronté et informé par les quatre cas d'usages du travail de thèse.

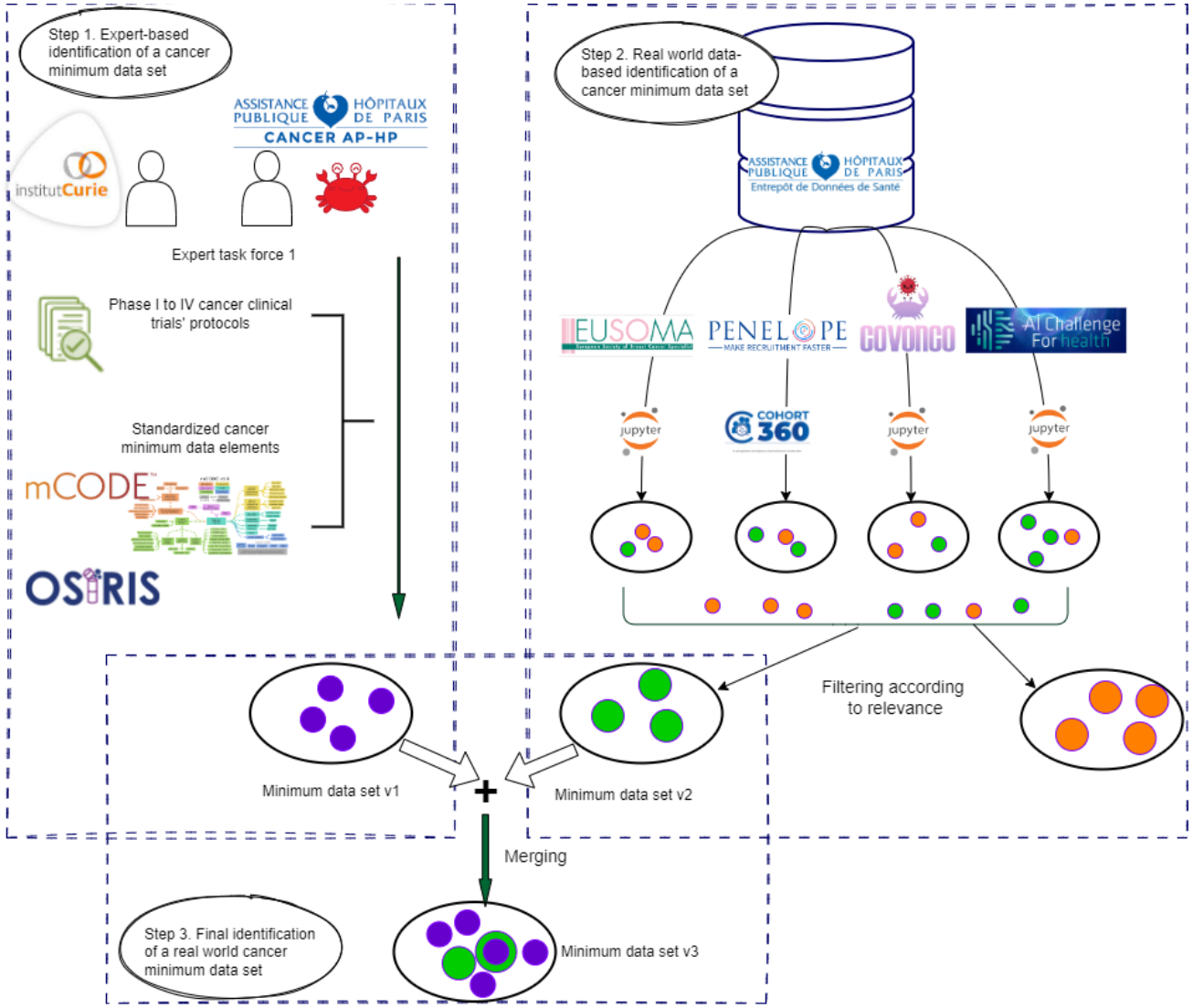


Figure 1. Méthodologie adoptée pour la spécification de la première version du jeu de données minimales polyvalent en oncologie, en articulation avec les cas d'usages de la thèse

2. Matériel : sources de données, environnements et espaces de travail utilisés

L'évaluation de l'utilité du jeu de données minimales polyvalent en oncologie a été réalisée, comme tout le travail de thèse, à partir des données de patients atteints de cancer solide contenues dans l'EDS de l'AP-HP, et mise en œuvre au sein des environnements et espaces de travail dédiés aux différents cas d'usages.

a. L'EDS de l'AP-HP : données et règles d'accès

L'AP-HP, plus grand centre hospitalier universitaire européen, s'est dotée depuis 2017 d'un EDS qui rassemble les données administratives et médicales de près de 11 millions de patients prise en charge dans les 38 hôpitaux qui constituent le groupe hospitalier (165,166). La population de patients atteints d'un cancer solide et dont l'information médicale est contenue et exploitable selon les conditions réglementaires dans l'EDS de l'AP-HP s'élève à 147 870 patients, entre juillet 2017 et juin 2023, soit environ 2 500 nouveaux cas par mois. A l'AP-HP, les efforts de standardisation permettent d'exploiter des données à l'échelle nationale dans le cadre de projets en lien avec le Health Data Hub et d'intégrer des consortiums internationaux tels que le Consortium for Clinical Characterization of SARS-CoV-2 by EHR (4CE). Dans le cadre de leur intégration dans l'EDS de l'AP-HP, les données sources issues des dossiers patients informatisés sont transformées, pseudonymisées et normalisées à l'aide de terminologies internationales (CIM-10, CIM-O, LOINC, ATC, etc.). Il existe par ailleurs une remontée biannuelle des données de mortalité intra- et extrahospitalière, issue du registre national des décès supervisé par l'institut national de la statistique et des études économiques (INSEE). Les contrôles de qualité des données sont effectués pendant la phase d'intégration des données ainsi que pendant la conduite des différents projets de recherche ou de pilotage.

La constitution de cet EDS hospitalier a été autorisée par la Commission Nationale de l'Informatique et des Libertés (CNIL) le 19 janvier 2017 à des fins de pilotage hospitalier et de recherche (agrément n°19800120) (165,166). Dans un engagement unique auprès de la CNIL, l'AP-HP a déclaré respecter la méthodologie nationale de référence MR-004 régissant les traitements de données à caractère personnel à des fins de recherche d'intérêt public n'impliquant pas la personne humaine, en particulier la recherche réutilisant des données précédemment enregistrées. La sécurité de l'EDS de l'AP-HP est assurée aux niveaux matériel, logiciel et organisationnel. En effet, une matrice d'autorisation conforme aux règles d'accès et d'exploitation des données, ainsi que des solutions garantissant la confidentialité des données, ont été mises en place. En particulier, la pseudonymisation des données est mise en œuvre sur les données structurées et non structurées permettant leur utilisation dans la recherche de données, et ce, conformément à la CNIL, à la direction générale de la protection des données (dont la méthodologie de référence MR004) et au Règlement Général sur la Protection des Données (RGPD).

b. *L'EDS de l'AP-HP : environnements (solutions et espaces de travail)*

L'utilisation secondaire des données du dossier patient informatisé repose sur des applications dédiées aux différents contextes d'utilisation. Le pilotage (cas d'usage 1) repose sur l'application Pilote basée sur la solution Cognos d'IBM. A noter cependant que des IQSS peuvent être développés et évalués en environnement de recherche avant d'être mis en œuvre dans la solution Pilote. Pour la recherche et l'innovation, un ensemble de solutions a été développé à l'AP-HP dont l'application Cohort360, qui permet de requêter et d'explorer les données de son EDS, en consommant des données structurées sous forme de ressources FHIR à partir d'une base de données dont le schéma est dérivé du modèle OMOP (167). Cohort360 permet de caractériser des cohortes d'intérêt pour des projets d'épidémiologie (cas d'usage 3) ou de développement et d'évaluation d'algorithmes d'IA (cas d'usage 4), et de spécifier les données d'intérêt à mettre à disposition dans le cadre de ces projets.

Les données nécessaires aux projets sont transférées au sein de ces espaces de travail, accessibles aux équipes projets, au sein desquels sont mises également à disposition les capacités de calcul (clusters de calcul et cartes graphiques) et de stockage adaptées à l'analyse des données - dont certaines peuvent être massives : documents cliniques, images médicales...) ainsi que des outils et bibliothèques de code. Les environnements d'évaluation de la qualité et de l'utilité du jeu de données minimales en oncologie ont donc été l'application cohort360 pour le cas d'usage 2 (projet PENELOPE) et les espaces de travail projets dédiés pour les cas d'usage 1, 3 (CSE 20-0055_COVONCO_AP) et 4 (CSE 20-0088_Challenge IA-FIAVOBI).

Ouverture des espaces de travail

L'accord du comité scientifique et éthique (CSE) de cet EDS fut obtenu dans le cadre du projet épidémiologique COVONCO regroupant l'ensemble de la population atteinte de cancer solide le 15 mai 2020 (IRB00011591, approbation CSE 20-0055_COVONCO_AP). Dans le cadre de l'édition 2020 du projet 'Challenge AI for Health' financé par la Région Ile-de-France, un accord parallèle du CSE permit l'accès aux données de patients atteints d'un cholangiocarcinome et pris en charge à l'AP-HP, en vue de la structuration des comptes rendus textuels d'anatomopathologie post-opératoires pour les stades localisés réséqués (IRB00011591, approbation CSE 20-0088_Challenge IA-FIAVOBI) le 12 mars 2021. Un amendement de ces deux protocoles fut obtenu du CSE les 30 juin 2023 et 30 août 2022, respectivement, afin d'ouvrir les droits d'exploitation du modèle d'apprentissage profond Bidirectional Encoder Representations from Transformers (BERT) entraîné sur les données de l'EDS de l'AP-HP. Seules les données strictement nécessaires et pertinentes au regard des objectifs de la recherche ont été collectées et traitées au cours des différents projets de la thèse (168).

Mise à disposition des données de recherche et partage du code

Les données mises à disposition proviennent principalement du dossier de santé électronique ORBIS de l'AP-HP et d'autres applications cliniques. Le déploiement du dossier électronique ORBIS s'est échelonné dans le temps depuis le début de son implémentation en 2014, et est toujours en cours pour certains services cliniques de l'AP-HP. Les données cliniques sont interopérables, normalisées à l'aide de terminologies internationales (CIM-10, CIM-O, LOINC, ATC, etc.) et accessibles aux utilisateurs autorisés via un portail Jupyter (Python, R, Scala...). A

l'occasion de l'amendement de juin 2023 validé par le comité scientifique et éthique de l'EDS de l'AP-HP, les données initialement fournies au format Informatics for Integrating Biology & the Bedside (i2b2) furent chargées au format OMOP dans l'espace projet CSE 20-0055_COVONCO_AP. Le code utilisé pour sélectionner et analyser les données des projets de la thèse est en libre accès avec possible partage des différents algorithmes (AP-HP Github, Zenodo, License open source BSD 3-clause).

Les variables suivantes ont été intégrées dans les espaces projets Jupyter :

- Données démographiques et données de parcours de soins au moment de l'admission, du transfert et de la sortie de l'hôpital ;
- Antécédents médicaux, codés selon la 10e édition de la classification internationale des maladies (CIM-10) ;
- Actes médicaux pratiqués pendant l'hospitalisation, codés selon la classification commune des actes médicaux (CCAM, 11e édition) ;
- Documents cliniques, en particulier comptes rendus médicaux de consultations, d'anatomopathologie et d'imagerie médicale, comptes rendus d'hospitalisation et de réunions de concertations pluridisciplinaires ;
- Statut vital avec données de mortalité intra- et extra-hospitalière.

III. Résultats

1. Constitution du groupe de travail *Cancer Research Application on Bigdata* (CRAB)

Le traitement des quatre cas d'usages a impliqué une équipe pluridisciplinaire constituée d'épidémiologistes, ingénieurs en génie industriel, oncologues cliniciens, médecins des départements d'information médicale (DIM), data scientists, ingénieurs et informaticiens, des services cliniques et techniques de l'AP-HP, des laboratoires de Sorbonne Université et de Centrale Supélec. Ce groupe de travail baptisé *Cancer Research Application on Bigdata (CRAB)* est né de la dynamique de recherche impulsée par la thèse, et la volonté conjointe d'exploiter les données de l'EDS de l'AP-HP de façon efficiente, tout en faisant avancer les thématiques méthodologiques propres aux sciences des données et améliorer la qualité de l'entrepôt dans le contexte d'une boucle rétroactive étroite et vertueuse avec ses équipes techniques.

Plus particulièrement, la réponse à un appel à projets lancé par la Fondation ARC pour la recherche contre le cancer dédié à la pandémie de SARS-CoV2 a permis le recrutement d'une scientifique des données formée aux données de santé au sein de l'unité de recherche clinique du groupe hospitalo-universitaire Henri Mondor - Albert Chenevier, et ce, en partenariat étroit avec l'équipe Innovations et Données de la Direction des Services Numériques de l'AP-HP, elle-même en charge du déploiement de son EDS (référence de subvention COVID202001343). En fonction de la typologie du cas d'usage traité, des spécialistes titulaires de l'AP-HP du cas d'usage considéré ont été conviés aux réunions hebdomadaires du groupe de travail ainsi

constitué et invités à participer aux publications connexes : pneumologues, gastro-entérologues, chirurgiens, anatomopathologistes, etc. Une interaction étroite avec les co-porteuses de la *mission IA & cancer* de l'établissement hospitalier francilien a permis de faire converger les réflexions concernant les modalités d'optimisation d'exploitation des données de son entrepôt, et ce, dans un dialogue étroit avec les équipes de la Direction de la Recherche Clinique et de l'Innovation (DRCI) et de la Direction de la Stratégie et de la Transformation (DST).

2. Etat de l'art et état des lieux

Différentes initiatives de spécifications de listes minimales de données de "vie réelle" dans le domaine de l'oncologie à diverses finalités ont été mises en œuvre par des sociétés savantes, agences sanitaires institutionnelles ou organismes de standardisation.

a. Initiatives des sociétés savantes et agences sanitaires

L'American Society of Clinical Oncology (ASCO), en collaboration avec MITRE Corporation, a spécifié une liste minimale de données d'oncologie du dossier patient informatisé. Cette liste appelée Minimal Common Oncology Data Elements (mCODE) comporte 90 items ou éléments de données se répartissant entre les six domaines suivants : patient, maladie, signes vitaux et résultats biologiques, génomique, thérapeutique et devenir clinique (131,169). Cette liste a été spécifiée afin d'améliorer l'interopérabilité des dossiers patients informatisés et ainsi faciliter le partage de données de base dans le domaine de l'oncologie dans le cadre du soin et de la recherche. Dans le domaine du cancer, les anatomopathologistes ont produit les spécifications les plus abouties de jeux de données minimales dans leur domaine, décliné par type de cancer (170–175). En France, les données d'anatomopathologie ont été sélectionnées à l'aide des référentiels de structuration des comptes rendus correspondants émis par l'INCa en 2011 dans le cadre de la mesure 20 du plan Cancer 2009 – 2013, et repris par l'agence du numérique en santé en 2018 (176,177).

En oncologie clinique, le premier défi réside dans le difficile consensus sur le choix des données clés du cancer, et le second (qui en est le corollaire) dans l'hétérogénéité de formalisation des informations finalement identifiées comme prioritaires (178–180). En effet, en fonction de la finalité attribuée au jeu de données, son contenu peut être particulièrement orienté vers le cas d'usage considéré (181,182). En France, l'INCa pilote l'initiative francophone du grOupe inter-SIric sur le paRtage et l'Intégration des donnéeS clinico-biologiques en oncologie (OSIRIS) relative à 67 items cliniques et 65 items -omiques, déployé autour du concept de médecine de précision (183).

b. Initiatives des organismes impliqués dans la standardisation des données massives

Les organismes de standardisation HL7 et DICOM ainsi que l'initiative OHDSI spécifient des modèles de données pertinents dans le cadre de l'utilisation secondaire des données du dossier patient informatisé en oncologie. HL7 est une organisation à but non lucratif dédiée au développement de l'interopérabilité des données de santé et la standardisation du

protocole d'échange informatisé de données cliniques, financières et administratives entre systèmes d'information hospitaliers. HL7 définit un ensemble de spécifications techniques qui sont diversement intégrées au corpus des normes formelles américaines (ANSI) et internationales (ISO). Initialement américaines, ces spécifications s'exportent et tendent à devenir un standard international pour ce type d'application. FHIR est un standard développé par HL7 décrivant des formats de données et des éléments (appelés « ressources ») ainsi qu'une interface de programmation applicative (API) pour les échanges des informations dans le domaine de la santé. Les spécifications HL7 FHIR représentent un modèle de messages que s'échangent différents acteurs, qui, une fois respectées, permettent à une application FHIR de requêter et consommer des ressources en provenance de différents serveurs FHIR.

L'initiative HL7 Vulcan vise à faciliter l'intégration des activités de soins et de recherche pour améliorer la qualité de vie et la survie des patients, et réduire les coûts par l'utilisation de standards d'interopérabilité HL7 FHIR (184). Ce programme d'accélération FHIR développe des ressources de recherche FHIR nécessaires à la réalisation de cas d'utilisation secondaire de données de "vie réelle" identifiés et classés par ordre de priorité, et a défini un profil FHIR dédié à cet usage. L'objectif principal de ce guide d'implémentation HL7 FHIR « recherche sur données de vie réelle » est d'aider à définir un ensemble minimal de ressources et d'éléments FHIR d'intérêt dans le cadre de la réutilisation secondaire des données des dossiers patients informatisés dans un contexte de recherche lors de deux étapes bien identifiées : i) déterminer les patients d'intérêt en fonction de critères d'inclusion et d'exclusion et ii) récupérer les données de soins de ces patients nécessaires à la réalisation des objectifs de recherche. Les informations clés nécessaires à l'étape de construction de la cohorte ont été spécifiées : données démographiques, caractéristiques de la visite, diagnostics, tests de laboratoire, actes, médicaments. Le guide d'implémentation est développé en utilisant une approche itérative de couverture de cas d'usages qui finiront par permettre la définition d'un ensemble minimal de ressources et d'éléments communs.

Des initiatives complémentaires apportent des spécifications additionnelles spécifiques à des domaines (ex : mCODE, DiaMorph, DaVinci projects, etc). Ainsi, le guide d'implémentation correspondant à mCODE et promu par HL7 US précise un ensemble d'informations minimales pour permettre la réalisation de travaux de recherche en oncologie de qualité, ainsi que les terminologies de santé à utiliser pour décrire effectivement un patient atteint de cancer. L'initiative CodeX met en œuvre et évalue des implémentations de mCODE dans des cas d'usages d'amélioration des soins ou de la recherche dans le domaine du cancer (185).

Le modèle OMOP développé par le consortium OHDSI fut en effet la cible du choix stratégique réalisé par les directions du Health Data Hub et de l'EDS de l'AP-HP (186). Depuis quelques années déjà, le consortium OHDSI développe et promeut le modèle de données commun OMOP pour la réalisation de recherches observationnelles (telles que la construction de cohortes de patients homogènes par rapport à des expositions ou à des maladies). Au-delà d'un modèle physique de base de données, le modèle de données commun OMOP repose aussi sur une approche terminologique forte. Ainsi, pour chaque concept médical, un terme d'une terminologie de santé (Medical Subject Headings (MeSH), Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT), LOINC, Medical Dictionary for Regulatory Activities (MedDRA), …) est reconnu comme étant le terme à utiliser dans une base de données OMOP. Récemment, une extension du modèle OMOP dédiée à l'oncologie a été développée afin de

fournir un meilleur cadre de représentation commune de l'évolution naturelle de la maladie tumorale (34,130,187).

3. Spécification et évaluation d'un jeu de données minimales polyvalent de données de dossiers patients informatisés dans le domaine de l'oncologie

a. Spécification d'une version initiale de jeu de données minimales en oncologie

Comme illustré par la Figure 1, l'identification des items d'intérêt fut réalisée par un binôme d'oncologues médicaux (Emmanuelle Kempf et Alain Livartowski), experts en recherche clinique expérimentale et sur données de vie réelle. Emmanuelle Kempf (EK) coordonne le groupe de travail CRAB à l'AP-HP et Alain Livartowski (AL) est coordonnateur de Consore, outil de requêtage au sein des bases de données fédérées d'UNICANCER, et animateur du groupe de travail OSIRIS – RWD (188). La délibération concernant la spécification du jeu de données minimales eut lieu lors d'une simulation de recherche d'inclusion d'un patient atteint de cancer dans des essais cliniques aléatoirement identifiés parmi ceux en cours au sein d'UNICANCER et de l'AP-HP. Les protocoles de 10 essais cliniques en oncologie solide académiques et industriels, de phase II à IV, ont été identifiés à partir de la base de données virtuelle ClinicalTrials.gov. AL et EK ont analysé les critères d'éligibilité de chaque essai clinique et défini une liste de critères d'inclusion et de non-inclusion pertinents pour l'étape préliminaire automatisable de recherche d'adéquation entre un patient et un protocole d'essai clinique donnés (préscreening) (67,189). AL et EK ont défini une liste théorique de 15 éléments minimaux pertinents permettant la confrontation des caractéristiques d'un patient atteint de cancer solide avec celles des critères d'éligibilité d'un essai clinique et de 5 autres éléments pour identifier la source de ces informations (Tableau 1, colonne 1).

AL et EK ont discuté dans quelle mesure cette liste de 15 éléments minimaux établie dans le cadre du cas d'usage d'aide au recrutement (cas d'usage n°2) était de nature à satisfaire un premier niveau de réalisation des autres cas d'usage : cas d'usage n°1 (pilotage), n°3 (étude de l'impact de la pandémie de SARS-CoV2 sur la prise en charge de patients atteints de cancer) et n°4 (développement/évaluation d'IA en imagerie oncologique). Ils ont, de la sorte, convenu du caractère polyvalent de cette première liste.

Tableau 1. Spécification des informations constitutives du jeu de données minimales en oncologie, à partir de 10 protocoles d'essais cliniques académiques et industriels de phase I à IV et utilité dans les quatre cas d'usages concernés par la thèse

Items constitutifs du jeu de données minimales en oncologie solide (n=15)	**Cas d'usage 1** Pilotage EUSOMA	**Cas d'usage 2** Etudes expérimentales PENELOPE	**Cas d'usage 3** Etudes observationnelles CovOnco	**Cas d'usage 4** Aide à la décision 'Challenge AI for health'
Patient				
• Date de naissance (1)	x	x	x	x
• Sexe (2)	x	x	x	x
Tumeur primitive				
• Localisation anatomique de la tumeur (topographie) (3)	x	x	x	x
• ***Type anatomopathologique de la tumeur (morphologie)*** (4)	x	x		x
• Stade tumoral				
○ ***Stade clinique de la tumeur (cTNM)*** (5)	x	x	x	x
○ ***Stade anatomopathologique de la tumeur (pTNM)*** (6)	x	x	x	x
• ***Biomarqueurs tumoraux*** (7)	x	x		x
• Nombre de sites métastatiques (8)	x	x	x	x
• Localisation des sites métastatiques (9)		x		
• Évolutivité tumorale (10)		x		
Traitement antitumoral				
• Type de traitement antitumoral (chirurgie, radiothérapie, chimiothérapie, ...) (11)	x	x	x	x
• Nombre de lignes de traitement antitumoral systémique (1 à n) (12)		x	x	
• Noms et type de médicaments antitumoraux (13)		x		
• Nombre de cycles de traitement antitumoral systémique (14)		x		
• Type de traitement local antitumoral (chirurgie, irradiation, …) (15)	x	x	x	x

b. *Evaluation de la pertinence du contenu du jeu de données*

EK a coordonné l'évaluation de cette liste dans le cadre de preuves de concepts des quatre cas d'usages d'intérêt, et ce, au sein du groupe de travail CRAB. Les colonnes 2 à 5 du Tableau 1 illustrent la distribution des variables d'intérêt pour chacun des cas d'usages au sein du jeu de données minimales en oncologie et polyvalent, obtenus dans ce contexte et équivalents à un premier niveau de validité de la pertinence des items constitutifs du jeu de données précédemment défini.

c. *Identification de l'absence de structuration parmi les variables d'intérêt des 4 cas d'usages*

Au sein des quatre cas d'usages du travail de thèse, les variables non structurées à extraire par automatique du langage naturel étaient les suivantes et sont représentées en gras et en italique dans le Tableau 1 :

- Stade métastatique ;
- Stade (y)pTNM ;
- Taille (tumorale) ;
- Différenciation (tumorale) ;
- Invasion vasculaire (tumorale) ;
- Invasion périnerveuse (tumorale) ;
- Invasion ganglionnaire (tumorale) ;
- Complétude de la résection microscopique ;
- Statut des récepteurs aux œstrogènes ;
- Statut HER2 ;
- Grade tumoral ;
- Malignité du prélèvement étudié ;
- Type histologique du cancer du sein.

IV. Discussion

1. Synthèse

Le travail de thèse a permis la spécification d'un jeu de données minimales en oncologie polyvalent, et validé à l'heure de la confrontation à quatre cas d'usages très divers dans leurs finalités, la typologie des acteurs impliqués et leurs modalités de réalisation. Cette démarche pragmatique de synthèse autour d'une liste restreinte d'items de la caractérisation d'un patient atteint de cancer s'inscrit d'une part dans une culture d'oncologie clinique quotidienne de synthèse et de caractérisation des cas cliniques, et d'autre part dans l'impulsion de sociétés savantes en vue de l'adoption des dossiers patients informatisés par des praticiens (159).

2. Mise en perspective avec la littérature

Les items retenus dans la première version du jeu de données minimales en oncologie et évaluée à l'AP-HP sont inclus dans les concepts disponibles des jeux de données spécifiques à l'oncologie développés, en France, par le groupe OSIRIS et, aux Etats-Unis, par l'initiative mCODE™, particulièrement axés autour de la médecine de précision et l'expansion consécutive des données -omiques caractérisant les tumeurs dans les critères d'éligibilité des essais cliniques interventionnels (131,183). Ce jalon de travail se rapproche plus du modèle américain que de son équivalent français, en ce que la dynamique de thèse cherche, avant toute chose, à exploiter de façon effective et pragmatique les données de l'EDS francilien, en valorisant le pilotage et la recherche sur données dites de vie réelle. Ainsi, la confrontation du jeu de données OSIRIS à des EDS français avait mis en évidence son manque d'applicabilité, et laissé entrevoir une marge de progression significative pour son utilisation dans un contexte de collecte d'informations de soins de routine, dits de vie réelle, et ce, malgré son souhait initial de 'maintenir le jeu de données minimales aussi réduit que possible'. A cette fin, l'initiative 'OSIRIS – real-world data (RWD)' cherche actuellement à en réduire les dimensions et simplifier le contenu (190).

Le périmètre priorisé dans le contexte des quatre cas d'utilisation secondaire de données abordés correspond aux données nécessaires au calcul d'une première liste d'IQSS (cas d'usage 1, Pilotage), à l'identification de patients à partir de critères d'éligibilité pour l'inclusion dans des essais cliniques (cas d'usage 2, PENELOPE) ou des recherches sur donnés (cas d'usage 3, CovOnco ou 4 CHOTERIA) et à la réalisation d'études épidémiologiques descriptives des modalités de prise en charge de patients (cas d'usage 3, CovOnco). Cette confrontation du contenu du jeu de données minimales aux cas d'usages, et avant toute tentative de standardisation ou de structuration, représente une démarche inédite d'évaluation du jeu de données de façon délibérément précoce par itérations pragmatiques (164). La mise en œuvre des quatre cas d'usage a permis d'évaluer, au sein de l'EDS de l'AP-HP, la disponibilité, la qualité des données correspondant à la version initiale du jeu de données minimales en oncologie ainsi que son utilité dans les différents contextes d'utilisation considérés. Dans cette perspective, plusieurs études ont évalué la disponibilité de tels items constitutifs de ce jeu de données pragmatiques dans les dossiers patients informatisés, révélant des métriques dont les marges de progression semblaient significatives (191). Par ailleurs, une étude de simulation réalisée à l'AP-HP a révélé que les informations clés d'un parcours de soins d'un patient atteint de cancer du sein pouvaient manquer chez jusqu'à la moitié des cas (106).

En conclusion, les contributions de ce chapitre au travail de thèse furent de démontrer qu'il était possible de :

- Spécifier un jeu de données minimales en oncologie dans une perspective pragmatique de réutilisation de données de vie réelle ;
- Conduire une évaluation préliminaire à l'échelle de l'AP-HP de son contenu et de sa pertinence, en rapport avec quatre cas d'usages d'horizons très divers en oncologie, et ce, avant toute velléité de standardisation et de structuration ;
- D'identifier la disponibilité des données d'intérêt de chaque cas d'usage et leur niveau de structuration dans les dossiers patients informatisés ;

- D’identifier la disponibilité sous format textuelle des données d’intérêt dans le même contexte.

S’ensuit, ainsi, la nécessité de :
- Standardiser le jeu de données minimales selon les standards internationaux en optimisant l’expressivité du modèle sélectionné (Partie II de la thèse de modélisation du jeu de données) ;
- Structurer les données textuelles ou de qualité sous-optimale par traitement automatique du langage naturel et en assurer les modalités optimales d’extraction textuelle (Partie III de la thèse d’enrichissement du jeu de données).

Aussi, la deuxième partie de la thèse a cherché à optimiser la standardisation de ce jeu de données minimales selon le modèle de données commun international OMOP, à évaluer l’expressivité de la version 5.3 d’OMOP actuellement déployée sur l’EDS de l’AP-HP, et à estimer la plus-value théorique apportée par son extension oncologie, à savoir la version 5.4 du modèle.

Partie II : Interopérabilité et standardisation du jeu de données minimales polyvalent en oncologie

I. Introduction

Avec l'éclatement nosographique de l'oncologie médicale lié à la génomique, la discipline se développe en autant de sous-populations de patients définies, entre autres caractéristiques, par une anomalie moléculaire tumorale (192). L'interopérabilité des bases de données liées au cancer semble une réponse prometteuse au développement de la recherche en contexte de maladies rares. Sont concernés, ainsi, tant le pilotage comparatif de la qualité et de la sécurité des soins, que la constitution de cohortes de patients partageant des caractéristiques similaires, que la recherche épidémiologique ou que le développement algorithmique appliqué au texte ou à l'image (134). L'utilisation secondaire des données de vie réelle à large échelle que ce soit au niveau régional, national ou international requiert la définition et le respect d'un cadre d'interopérabilité permettant l'exploitation partagée des données d'oncologie dans le cadre du cas d'usage envisagé.

Ce chapitre constitue une contribution au sujet de la standardisation du jeu de données minimales en oncologie, dans le contexte du cas d'usage 2 d'aide au recrutement. Ici, le développement d'un système automatique d'aide au recrutement dans des essais cliniques (SAREC) sur dossiers patients informatisés achoppe sur trois verrous : celui des données, celui des outils de requêtage et celui de l'interopérabilité. Tout d'abord, concernant les données, que la focale se concentre sur les caractéristiques du patient, ou sur les critères d'éligibilité de l'essai clinique, un outil d'aide au recrutement utilisant les dossiers patients informatisés repose sur la mise en correspondance automatique de ces deux types d'informations

médicales. Or, ni les critères d'éligibilité des essais cliniques, ni les données contenues dans les dossiers patients informatisées ne conviennent pour cette tâche d'appariement (193,194).

Certains outils basés sur des algorithmes d'intelligence artificielle prétendent, pour un cas d'usage donné, améliorer la précision et accélérer le processus de préscreening des patients atteints de cancer, incluant l'analyse de données non structurées (195). Ensuite, le développement d'un outil d'aide au recrutement à l'état de l'art nécessite, comme pour toute innovation digitale, de répondre à diverses exigences (prise en compte des besoins des utilisateurs, intégration à l'environnement de travail des utilisateurs, évaluation continue de l'efficacité/efficience). Or, des publications récentes montrent que ces exigences restent souvent non respectées (196). Il existe, ainsi, des méthodologies de développement adaptées à considérer et un cadre d'interopérabilité à prendre en compte dans un contexte où il n'existe pas de représentation standard largement acceptée commune aux critères d'éligibilité et aux données des dossiers patients informatisés, signant le défi d'interopérabilité (197).

A noter, en particulier, que des bases de connaissances conçues pour le partage et la réutilisation des critères d'éligibilité, telles que Trial Bank, Epoch, SysBank ou l'initiative française OSIRIS, peuvent manquer d'adaptabilité et d'évolutivité car elles ne reposent pas sur des normes de données largement mises en œuvre dans les établissements de santé, telles qu'OMOP ou HL7 FHIR (183,198–200). Comme précisé dans la Partie I, des initiatives ou organismes de standardisation ont proposé des modèles d'interopérabilité spécifiques à l'oncologie préconisant le codage des données de santé en utilisant des terminologies le plus souvent internationales. Au niveau français, l'agence du numérique en santé (ANS) définit le cadre d'interopérabilité des systèmes d'information de santé et promeut également l'usage de terminologies de santé internationales, qui devraient être implémentées dans les solutions déployées au niveau des hôpitaux (LOINC, NCBIt, ATC…).

Dans le cadre du projet PENELOPE, le modèle commun de données d'intérêt fut OMOP. Il est ainsi possible d'exécuter, sur n'importe quelle base de données conforme au modèle de données commun OMOP, les requêtes et scripts écrits localement et ainsi réaliser une analyse distribuée. La réalisation d'études de faisabilité distribuées sur ce modèle a été évaluée dans la communauté scientifique (projets EHDEN, EU-Patient Centric Clinical Trial Platforms (EU-PEARL) et European University Hospital Alliance (EUHA)) (201,202). L'extension du modèle OMOP dédiée à l'oncologie a été développée afin de fournir un meilleur cadre de représentation commune de l'évolution naturelle de la maladie tumorale, basée sur la notion de trois *épisodes de la maladie* (DISEASE EPISODES) nouvellement intégrés et sur un ensemble de règles de mise en œuvre pour encoder et aligner les données de soins de routine sources dans les différentes tables du modèle (35,130,187). Par exemple, la maladie y est décrite comme un unique OVERARCHING DISEASE EPISODE parent, qui englobe longitudinalement l'ensemble du parcours de soins d'un patient. En son sein, des EPISODES correspondent à des périodes continues de la maladie, composés de plusieurs *événements* : l'épisode DISEASE EXTENT qui caractérise la dissémination tumorale, et l'épisode DISEASE DYNAMIC qui caractérise l'évolutivité tumorale.

Les objectifs de l'étude préalable OMOP du projet PENELOPE étaient de comparer l'expressivité de deux versions différentes du modèle de données OMOP (v5.3 et 5.4) concernant la représentation du jeu de données minimales développées dans la première

partie de la thèse pour le cas d'usage de l'aide au recrutement en oncologie (cas d'usage 2). Il s'agissait de modéliser les critères d'éligibilité d'essais cliniques utilisés pour le préscreening correspondant au jeu de données minimales précédemment décrit, puis d'évaluer les performances de requêtage correspondant sur l'EDS de l'AP-HP. Notre hypothèse était que la version 5.4 du modèle OMOP, intégrant l'extension oncologie et non encore implémentée à l'AP-HP, augmenterait sa capacité à représenter les données de soin des patients disponibles dans l'entrepôt, et donc leur utilité, lorsqu'elles étaient appliquées à la tâche d'automatisation du préscreening de patients en vue de leur inclusion dans des essais cliniques.

II. Méthodes

1. Identification des essais cliniques cibles et méthodologie globale

Quinze essais cliniques d'urologie allant de la phase I à la phase IV, industriels et académiques, ayant inclus au moins quatre patients, ont été sélectionnés aléatoirement parmi les essais cliniques lancés à l'AP-HP entre 2016 et 2021. Leurs critères d'éligibilité ont été extraits des protocoles correspondants, et en leur sein, les critères de préscreening ont été identifiés manuellement à partir du jeu de données minimales issu de la première partie du travail de thèse. La Figure 2 décrit la méthodologie globale d'évaluation d'un SAREC proposée dans le cadre du projet PENELOPE et qui a été appliquée aux données de l'EDS de l'AP-HP, depuis la sélection des essais cliniques cibles jusqu'à la mesure des performances de l'identification automatique de patients éligibles à une inclusion.

Le calcul de cette performance comporte plusieurs étapes d'évaluation :

- 1/ traduction des critères de préscreening issus des protocoles des essais cliniques en requêtes conformes au modèle OMOP ;
- 2/ taux d'exécution de ces requêtes sur les outils de requête existants au sein de l'EDS de l'AP-HP.

Par exemple, le critère de préscreening « cancer de la prostate métastatique » peut se traduire en concepts standards du vocabulaire de la SNOMED « carcinome de la prostate » (concept_id 4116087), et maladie néoplasique maligne secondaire (concept_id 432851). Puis, cette requête doit être alignée avec la terminologie utilisée à l'AP-HPde la CIM10 « tumeur maligne de la prostate (CIM10 C61), « tumeur maligne secondaire des organes respiratoires et digestifs » (CIM10 C78) et « tumeur maligne secondaire de sièges autres et non précisés » (CIM10 C79). Ensuite, cette requête est appliquée sur l'outil dédié de l'institution francilienne Cohort360.

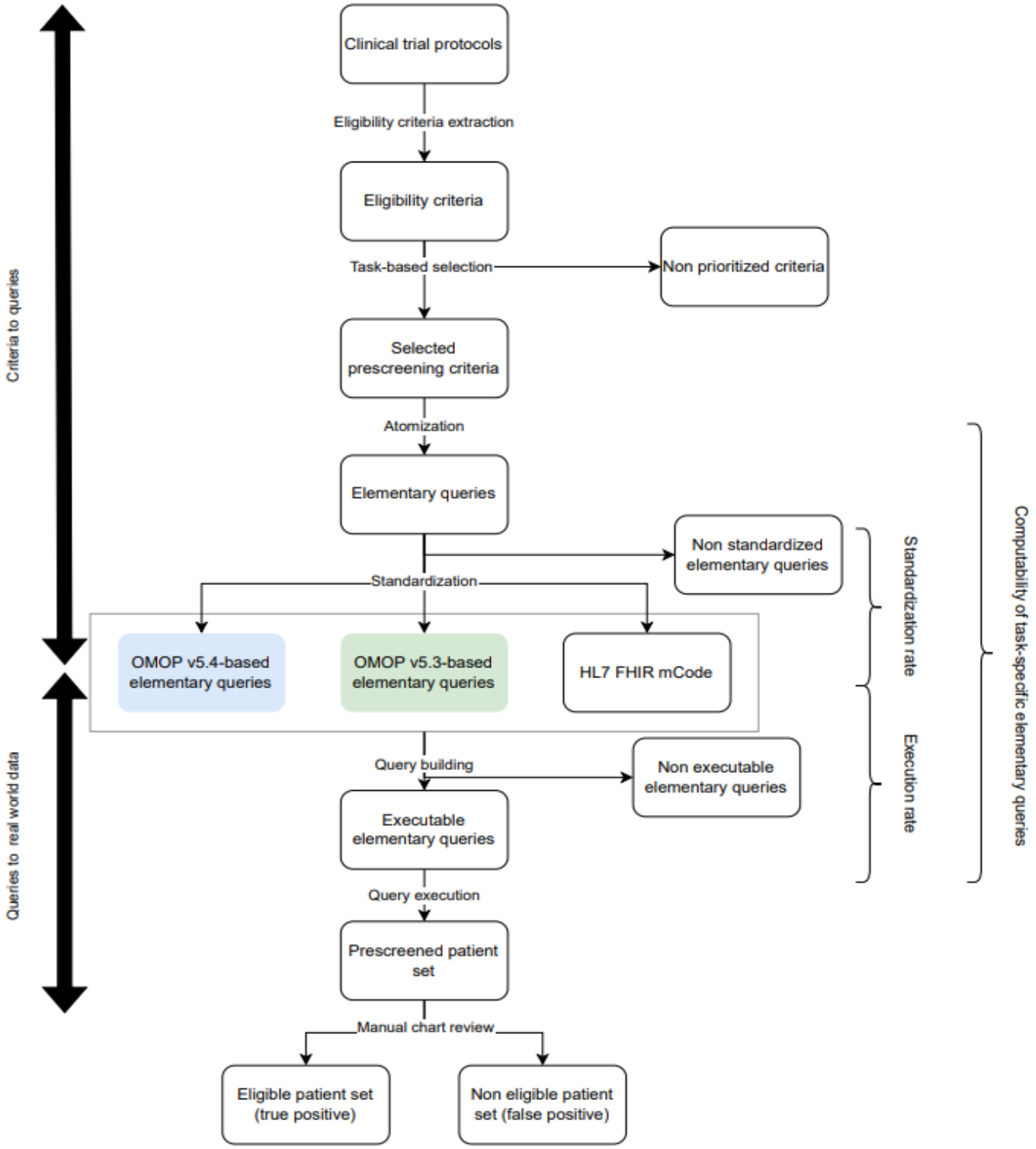


Figure 2. Méthodologie globale du projet PENELOPE pour évaluer l'automatisation du préscreening de patients atteints de cancer éligibles à l'inclusion dans un essai clinique, à partir d'un EDS et selon plusieurs versions du modèle OMOP

Abréviations : CT, essais cliniques; HL7 FHIR, Health Level Seven International Fast Healthcare Interoperability Resources ; mCODE™, minimal Common Oncology Data Elements ; OMOP, Observational Medical Outcomes Partnership.

2. Automatisation de l'identification de patients atteints de cancer éligibles à une inclusion dans un essai clinique

a. Étape 1 : traduction des critères de préscreening en requêtes conformes au modèle OMOP

Deux experts en oncologie médicale et en recherche clinique, d'une part, et en ingénierie des données, d'autre part (EK, Morgan Vaterkowski) ont atomisé les critères de préscreening issus des essais cliniques identifiés pour le cas d'usage, c'est-à-dire qu'ils les ont divisés en un maximum de données élémentaires. Quatre experts en interopérabilité des données, ingénierie des données, oncologie médicale et recherche clinique (Damien Leprovost, Nicolas Griffon, MV, EK) ont aligné ces données élémentaires avec les concepts standards les plus granulaires des tables cliniques OMOP, à l'aide du répertoire de vocabulaires internationaux standards en ligne ATHENA développé par la communauté OHDSI (203). L'alignement a également été réalisé avec les terminologies propres aux entrepôts de données français, telles que la version française de la 10e révision de la classification internationale des maladies (CIM-10) et la classification commune des actes médicaux (CCAM) utilisées dans les EDS, et notamment à l'AP-HP. Par exemple, la donnée élémentaire « tumeur primaire de la prostate » a été alignée avec le concept_id 200962 de la SNOMED CT « Tumeur maligne primaire de la prostate » à partir d'ATHENA, et avec le code CIM-10 C61 « Tumeur maligne de la prostate ». Le taux de données élémentaires et de critères de préscreening connexes qui pouvaient être alignés avec des concepts standards en lien avec les tables OMOP a été calculé.

Ensuite, l'expressivité des deux versions du modèle OMOP a été caractérisée grâce à la représentation d'un parcours de soins d'un patient atteint de cancer. Une première représentation formelle des critères de préscreening basés sur le modèle OMOP sans et avec le module d'extension du cancer (v5.3 et v5.4, respectivement) a été réalisée (130). Concernant l'applications au cas d'usage, les critères de préscreening qui impliquaient que plusieurs données élémentaires soient liées entre elles ont été identifiés. Par exemple, le critère de préscreening suivant « cancer de la prostate métastatique résistant à la castration » nécessite un lien entre les trois données élémentaires suivantes : « cancer de la prostate », « métastatique » et « résistant à la castration ». Pour tous les essais cliniques du cas d'usage de la thèse, l'expressivité du modèle a été évaluée et en particulier la capacité de représenter ces liens entre données élémentaires au sein d'un même critère de préscreening dans les versions 5.3 et 5.4 du modèle OMOP, correspondant au taux de standardisation.

b. Etape 2 : exécution des requêtes OMOP sur l'EDS de l'AP-HP

Cohort360 est un outil d'interrogation en ligne et un générateur de cohortes de patients interfacé à l'EDS de l'AP-HP via une API FHIR. Cohort360 est utilisé pour repérer les patients vivants et éligibles au recrutement dans chaque essai clinique du cas d'usage. Ainsi, pour tous les essais étudiés, ont été générées les cohortes de patients correspondant aux critères de préscreening de tous les essais étudiés. Le taux d'exécution des critères de préscreening et des données élémentaires correspondantes a été évalué pour tous les essais cliniques du cas

d'usage. Trois essais cliniques ont été aléatoirement identifiés en vue d'une revue manuelle de dossiers cliniques. Les taux de vrais et de faux positifs ont ainsi été évalués, après examen minutieux du contenu du dossier électroniques de chaque patient automatiquement identifié comme éligible par la requête de Cohort360. Enfin, les données élémentaires non exécutables sur Cohort360 ont été catégorisées selon la raison incriminée : données manquantes dans l'EDS de l'AP-HP et/ou manque d'expressivité du langage de requête de Cohort360.

III. Résultats

Les résultats du projet PENELOPE ont été présentés sous format d'abstract au symposium annuel 2022 de l'association américaine d'informatique médicale (*American Medical Informatics Association* AMIA), et ont fait l'objet d'une publication sous format d'article original dans le *Journal of Clinical Oncology Clinical Cancer Informatics* (204,205).

1. Caractérisation des essais cliniques du cas d'usage, et identification des données élémentaires relatives à leurs critères de préscreening

Le Tableau Supplémentaire 1 résume les caractéristiques des quinze essais cliniques de notre cas d'usage.

Parmi les 534 critères d'éligibilité des 15 essais cliniques du cas d'usage, 83 ont été identifiés comme relevant de l'étape de préscreening : 60 critères d'inclusion et 23 critères de non-inclusion. Parmi eux, 60 (72%) contenaient plus d'une donnée élémentaire. Les 83 critères de préscreening ont été atomisé en 288 données élémentaires, dont le tratitement de 87 (30%) nécessitait une interprétation médicale. La colonne 2 du Tableau 2 résume la répartition de ces 288 données élémentaires entre les 15 informations minimales de préscreening identifiés lors de la Partie I de la thèse. Parmi les 288 données élémentaires, certaines correspondaient aux sources spécifiques suivantes : observation clinique (n = 17), procédure cytologique ou anatomopathologique (n = 15), imagerie (n = 11) et biologie (n = 2).

Tableau 2. Répartition des données élémentaires au sein des critères de préscreening de 15 protocoles d'essais cliniques de phase I à IV d'urologie sélectionnés aléatoirement parmi ceux ayant cours à l'Assistance Publique – Hôpitaux de Paris entre 2016 et 2021, concernant le projet PENELOPE

	Critères de préscreening (n = 15)	**Données élémentaires* (n = 288)**
Patient		
•	Date de naissance (1)	16
•	Sexe (2)	9
Tumeur primitive		

Critères de préscreening (n = 15)	Données élémentaires* (n = 288)
• Localisation anatomique de la tumeur (topographie) (3)	33
• Type anatomopathologique de la tumeur (morphologie) (4)	32
• Stade tumoral	
○ Stade clinique de la tumeur (cTNM) (5)	46
○ Stade anatomopathologique de la tumeur (pTNM) (6)	6
• Biomarqueurs tumoraux (7)	8
• Nombre de sites métastatiques (8)	2
• Localisation des sites métastatiques (9)	5
• Évolutivité tumorale (10)	14
Traitement antitumoral	
• Type de traitement antitumoral (chirurgie, radiothérapie, chimiothérapie, ...) (11)	71
• Nombre de lignes de traitement antitumoral systémique (1 à n) (12)	47
• Noms et type de médicaments antitumoraux (13)	25
• Nombre de cycles de traitement antitumoral systémique (14)	4
• Type de traitement local antitumoral (chirurgie, irradiation, …) (15)	5

* Une donnée élémentaire peut concerner plusieurs critères de préscreening simultanément (p. ex., « précédemment traité avec du docétaxel » peut informer les critères numérotés 11, 12 et 13).

2. Évaluation du caractère requêtable des critères de préscreening sur l'EDS de l'AP-HP

a. Traduction des critères de préscreening en requêtes OMOP

Parmi 288 données élémentaires correspondant aux critères de préscreening, 262 (88%) d'entre elles ont pu être alignées aux concepts standards OMOP liés aux sept tables suivantes : PERSON, CONDITION_OCCURRENCE, MESURE, OBSERVATION, EXPOSITION, PROCEDURE_OCCURRENCE et SAMPLE. Cet alignement a impliqué que 62 des 83 critères de préscreening pouvaient être représentés de manière exhaustive par des concepts standards OMOP. Les 26 données élémentaires qui n'ont pas pu être alignées avec des concepts standards OMOP concernaient :

- Des contraintes temporelles (p. ex. « s'il est terminé au moins 6 semaines avant le début du traitement à l'étude » (n = 4) ;
- La catégorisation des médicaments (p. ex. « traitement avec au moins un régime contenant du platine » (n = 11) ;
- Une indication clinique thérapeutique (p. ex. « tout traitement antitumoral approuvé pour le carcinome urothélial » (n = 3) ;
- Des biomarqueurs (p. ex. « Le test FGFR montre de faibles niveaux d'expression de FGFR ») (n = 3) ;

- Une stadification tumorale non normalisée (p. ex. « maladie viscérale) (n = 5).

Le Tableau Supplémentaire 2 résume les raisons pour lesquelles 21 critères de préscreening n'ont pas pu être alignés. La Figure 3 illustre la répartition des 262 données élémentaires pouvant être alignées dans les tables OMOP.

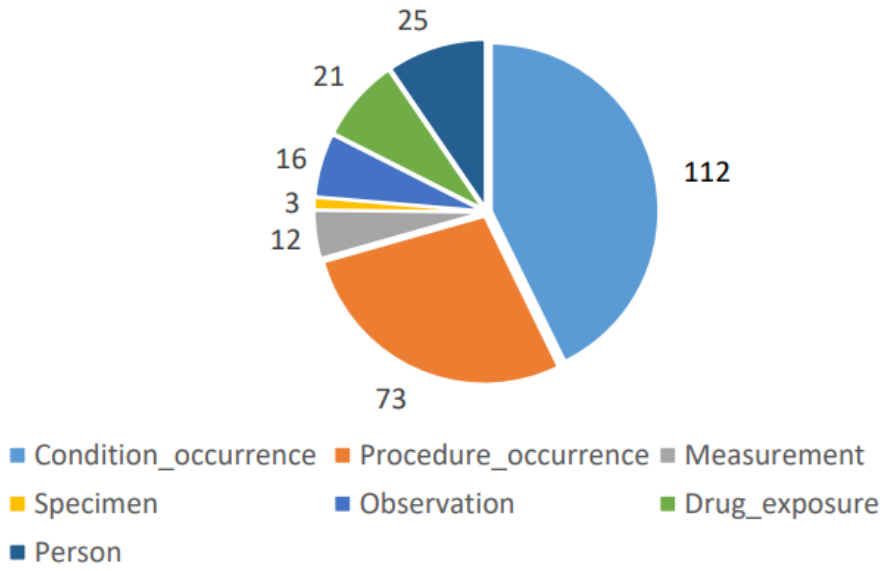


Figure 3. Répartition de 262 donnés élémentaires au sein de 7 tables OMOP, à partir des 288 données élémentaires liées aux 83 critères de préscreening de 15 essais cliniques de phase I-IV, pouvant être alignées avec des concepts standards OMOP

Parmi les 83 critères de préscreening, 60 impliquaient la liaison de plusieurs données élémentaires en leur sein. Par exemple, le critère de préscreening suivant « cancer de la prostate métastatique résistant à la castration » nécessitait un lien entre les trois données élémentaires suivantes : « cancer de la prostate », « métastatique » et « résistant à la castration ».

Dans le modèle OMOP v5.3, ce lien entre données élémentaires était représentable pour 33 des critères de préscreening, principalement parce qu'il impliquait l'âge et le sexe du patient, ce qui correspondait à un taux de standardisation de 40%. La version 5.4 a permis de représenter 29 critères de préscreening supplémentaires : 14 grâce aux tableaux EPISODES uniquement, 3 grâce aux champs polymorphiques uniquement et 12 grâce à la combinaison des deux. Dans cette perspective, le taux de standardisation a atteint 75%. La Figure 4 synthétise ces résultats.

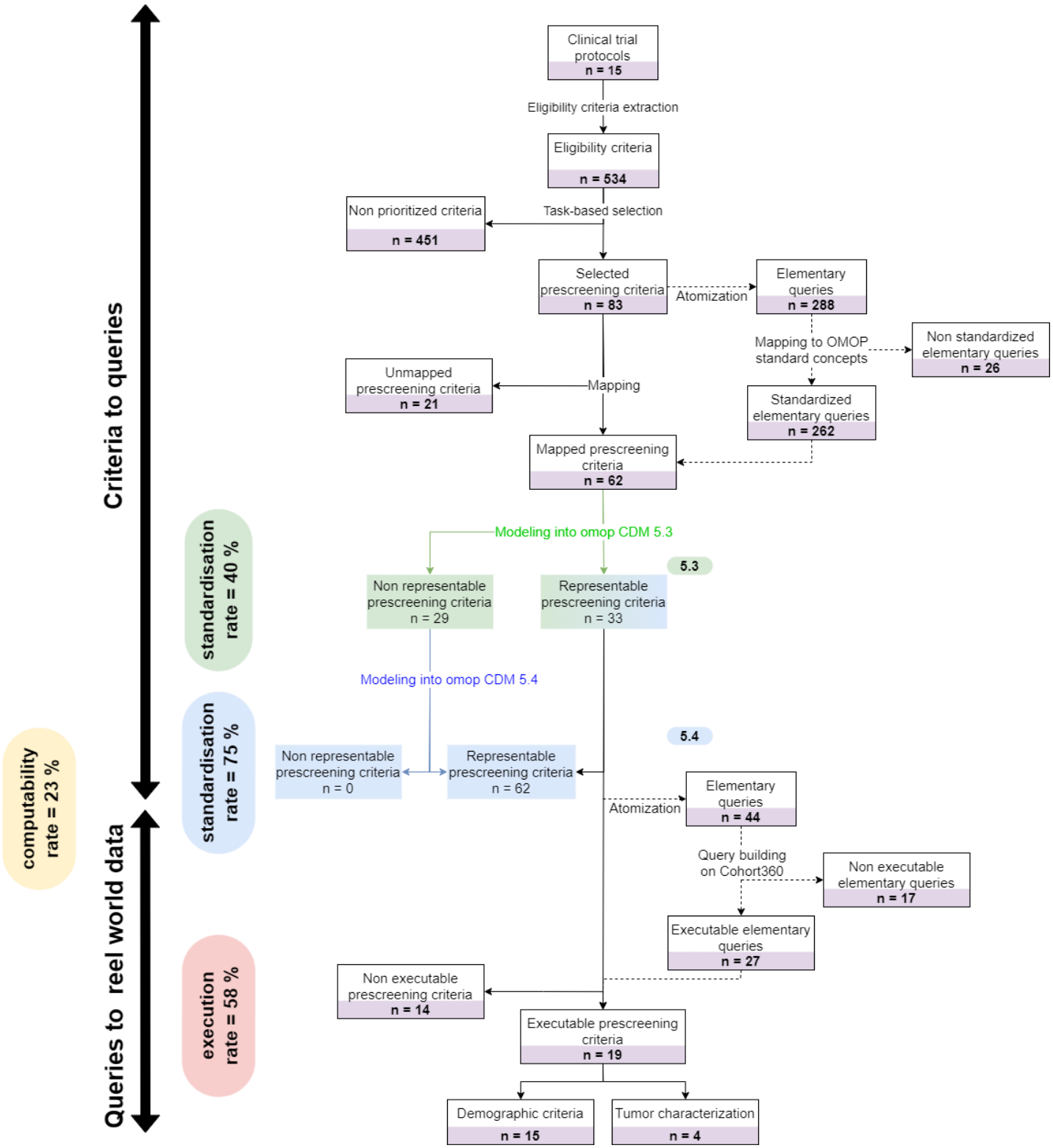


Figure 4. Evaluation du caractère requêtable de 83 critères de préscreening issus de 15 essais cliniques de phase I-IV selon 2 versions du modèle OMOP (v5.3 en vert et v5.4 en bleu) sur l'EDS de l'AP-HP. La standardisation correspond à la représentation exhaustive selon les spécifications du modèle OMOP (criteria to queries). L'exécution correspond au caractère requêtable par Cohort360 sur l'entrepôt de données de l'AP-HP (queries to real world data)

b. Défis liés à la traduction des critères de préscreening en requêtes OMOP v5.3

La traduction des critères de préscreening en requêtes OMOP révéla certaines limites de la version 5.3.

Premièrement, la table OBSERVATION est ambiguë. Malgré quelques cas d'usage clairement définis (antécédents médicaux, antécédents familiaux, ...) la table OBSERVATION est censée capter « toutes les données qui ne peuvent être représentées par aucun autre domaine » (206). Il est donc théoriquement possible de stocker dans cette table toute information caractérisant un patient atteint de cancer. Cependant, il n'existe pas de méthode standardisée pour procéder, d'autant plus qu'il n'existe pas de concept standard pour tous les types d'informations possibles. Dans notre cas, parmi les 262 données élémentaires correspondant à des concepts standards OMOP, 16 pouvaient être représentées uniquement dans la table OBSERVATION, tandis que pour 8 autres données élémentaires (3%), la table OBSERVATION était une option possible.

Deuxièmement, les spécifications d'OHDSI permettent d'aligner une même donnée élémentaire avec plusieurs tables cliniques distinctes. Parmi les 262 données élémentaires de notre cas d'usage, 81 (32%) pouvaient être renseignées dans des tables distinctes en même temps. Par exemple, le statut tumoral métastatique 'M0' pouvait être alternativement représenté dans la table CONDITION_OCCURRENCE (SNOMED) ou bien par MEAS_VALUE (LOINC, NAACCR) dans la table MEASUREMENT. Ces 262 données élémentaires pouvaient être représentées par 328 concept_ids standard distincts disponibles. Par exemple, le cancer de la prostate peut être représenté à l'aide de deux tables différentes, la table CONDITION_OCCURRENCE à l'aide du concept_id 4163261 (SNOMED : tumeur maligne de prostate) et la table MEASUREMENT via le concept_id 45884513 (LOINC : cancer de la prostate). La Figure 5 illustre les implications en termes de manque d'interopérabilité entre deux EDS de cet exemple, en fonction de la multiplicité des tables à disposition dans le modèle 5.3 pour représenter une même donnée élémentaire.

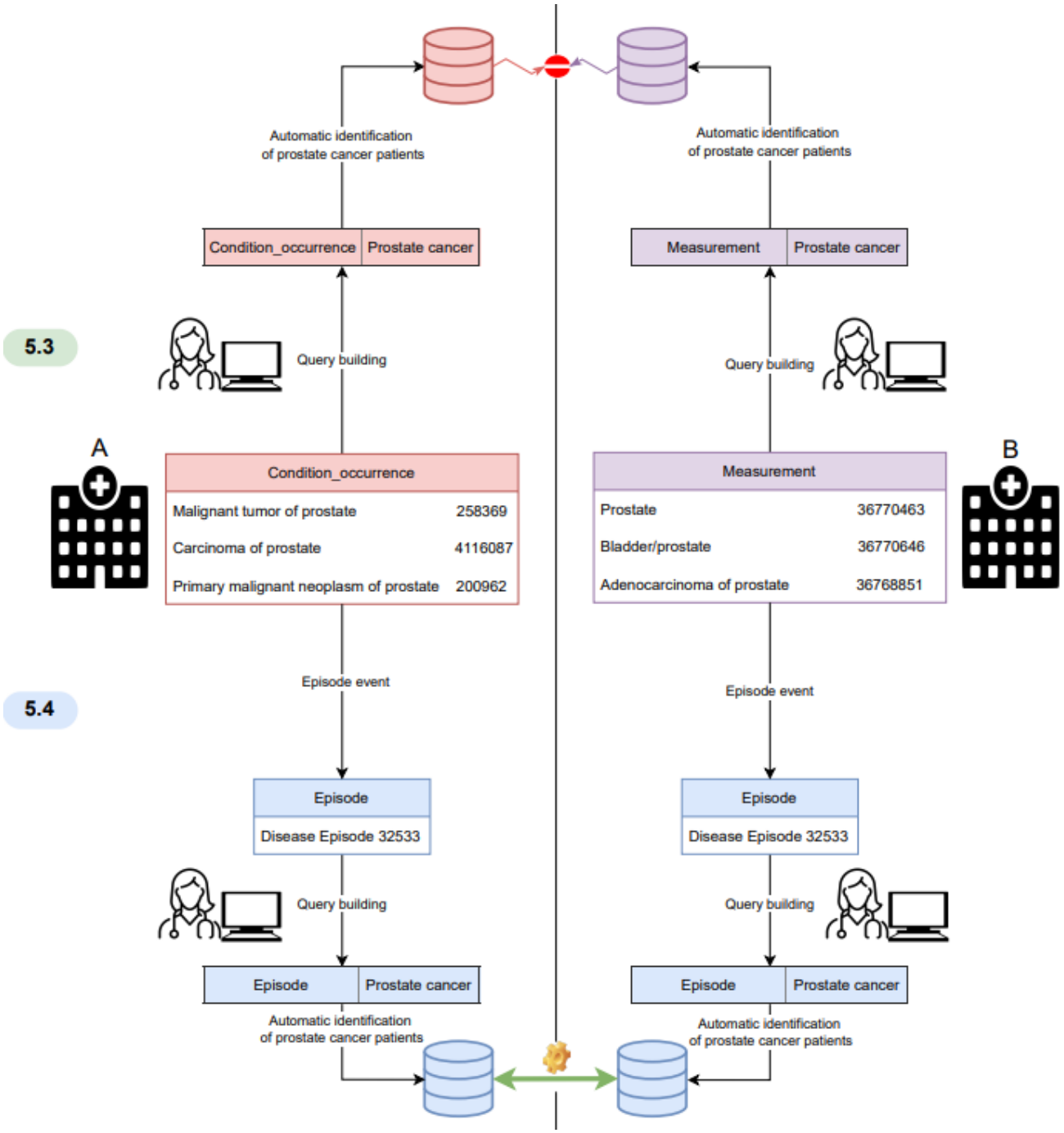


Figure 5. Illustration d'une des limites en interopérabilité liées à la multiplicité des tables pour représenter une même donnée élémentaire, dans les versions 5.3 (vert) et 5.4 (bleu) du modèle OMOP

Troisièmement, le référentiel ATHENA offre plusieurs échelles de représentation d'une même information clinique à l'aide des concepts standards. Par exemple, un cancer urothélial primitif de la vessie peut être représenté dans le tableau CONDITION_OCCURRENCE par, soit, l'ensemble des deux concept_ids distincts SNOMED « carcinome de la vessie » et « carcinome à cellules transitionnelles », soit, le concept_id CIM-O3 « carcinome urothélial, NOS, du col de la vessie ». Dans notre cas d'usage, 54 (21%) des 262 données élémentaires auraient pu être fusionnées en 24 concept_ids précoordonnés distincts - si décision avait été prise de travailler avec une moindre granularité.

Dans la v5.3 du modèle OMOP, la connexion des tables entre elles est centrée autour de la table PERSON, et non sur les événements cliniques correspondants aux différentes étapes de l'évolution de la maladie elle-même. Or, la caractérisation d'un parcours de soins d'un patient atteint de cancer nécessite l'articulation de multiples événements entre eux. La v5.4 du modèle OMOP offre une meilleure expressivité pour traiter cette complexité. La Figure 6 illustre ces limites de représentation du modèle OMOP v5.3 à partir du cas d'un patient de 67 ans dont l'histoire comporte deux cancers : un antécédent de cancer du poumon en 2016 et un antécédent de cancer de la prostate en 2018. La présence de métastases osseuses en 2019, le caractère adénocarcinomateux du cancer en 2020 et le début d'une chimiothérapie sont trois éléments qui ne peuvent être reliés avec aucun cancer primitif spécifique dans la version 5.3 d'OMOP. Même si le concept « cancer de la prostate métastatique » peut être représenté directement dans le modèle OMOP v5.3, il n'est pas possible de relier la topographie de la métastase (ici, osseuse) à un cancer primitif précis. De plus, dans le cas d'une rechute métastatique d'un cancer précédemment traité, le concept précoordonné « cancer de la prostate métastatique » ne peut pas être utilisé car il existe, en amont dans le temps dans le dossier patient informatisé, un code « cancer de la prostate » préexistant ne pouvant être complété ultérieurement que par le code « métastases ». La Figure Supplémentaire 1 illustre la valeur ajoutée théorique de la v5.4 d'OMOP par rapport à la v5.3 dans une situation typique d'oncologie médicale.

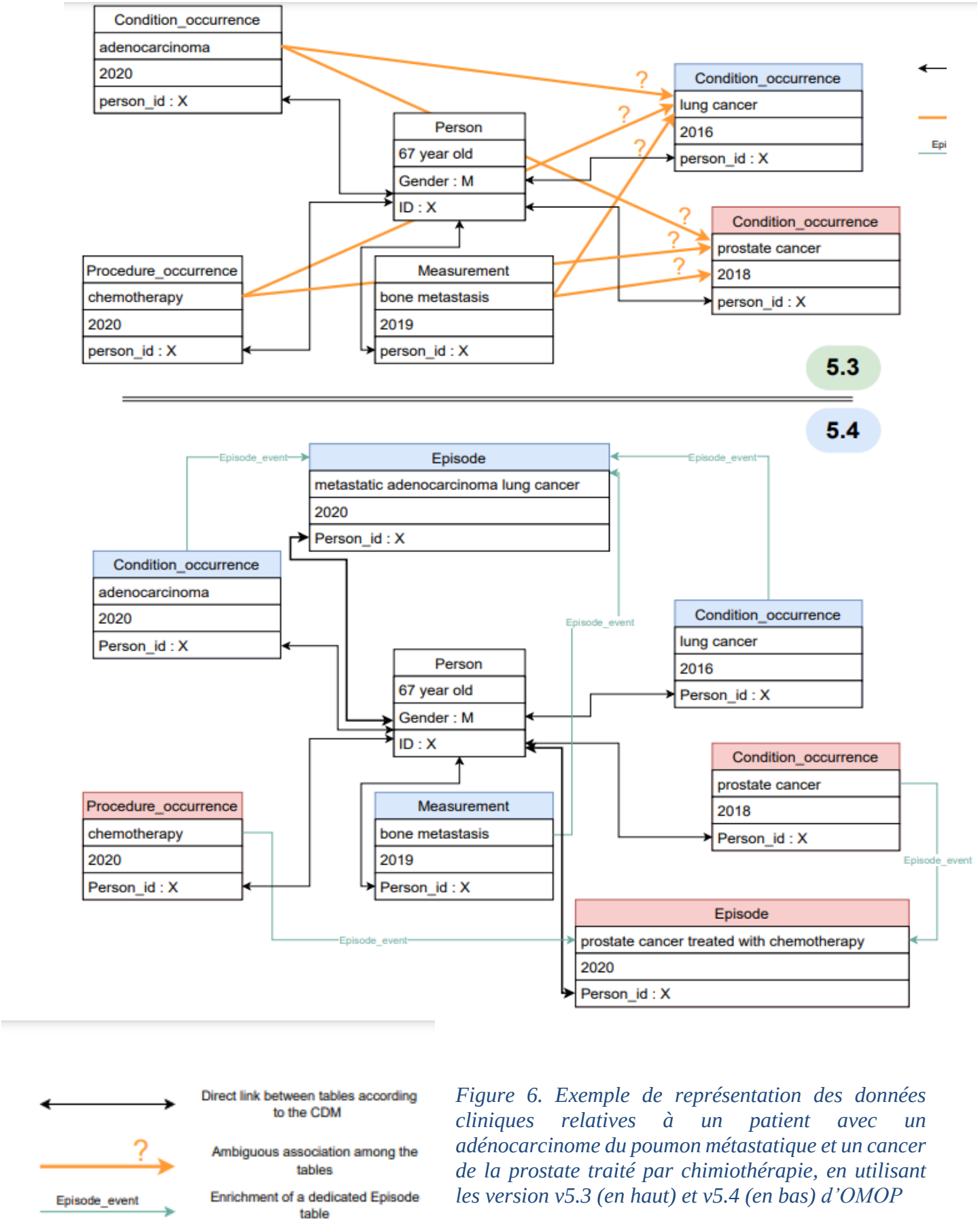


Figure 6. Exemple de représentation des données cliniques relatives à un patient avec un adénocarcinome du poumon métastatique et un cancer de la prostate traité par chimiothérapie, en utilisant les version v5.3 (en haut) et v5.4 (en bas) d'OMOP

c. *Exécution des requêtes OMOP sur l'EDS de l'AP-HP*

Un certain nombre de requêtes élémentaires n'étaient pas exécutables sur l'EDS de l'AP-HP, soit par manque de disponibilité des données cibles (manque de chaînage ou de structuration), soit par manque d'expressivité du langage de requête de Cohort360. Pour les 15 essais cliniques, le taux d'exécution des 83 critères de préscreening initiaux et des 288 données élémentaires correspondantes qui ont pu être calculées a atteint 24% (n = 20) et 44% (n = 126) sur l'EDS AP-HP actuel, respectivement. Parmi les 162 données élémentaires non exécutables, 66 (41%) et 95 (59%) étaient dues à des données structurées mais non intégrées (p. ex., les données de prescriptions et d'administrations de traitement systémique intraveineux antitumoral du logiciel métier CHIMIO), et à des données non structurées (p. ex., la mention du caractère « résistant à la castration » d'un cancer de prostate), respectivement. Sur les 33 critères de préscreening représentables dans la v5.3 du modèle OMOP, 19 ont pu être exécutés sur l'EDS de l'AP-HP, principalement en raison de données démographiques (taux d'exécution de 58%) (Figure 4). En résumé, 19 des 83 critères de préscreening (23%) ont pu être requêtés sur l'EDS de l'AP-HP (Figure 4).

A partir de 3 essais cliniques sélectionnés au hasard au sein du cas d'usage (VESPER, SURF, CASSIOPE), 17, 32 et 63 patients ont été automatiquement identifiés comme potentiellement éligibles à l'inclusion dans ces essais cliniques (Figure 7). Après revue manuelle, la valeur prédictive positive a atteint 53 % (n = 9/17), 41 % (n = 13/32) et 21 % (n = 13/63), respectivement. Les faux positifs étaient principalement dus à une mauvaise classification du type de cancer primitif (0, 15 et 12), de la caractérisation des métastases (4, 1 et 6) et du traitement antitumoral précédemment administré au patient (4, 3 et 22), respectivement.

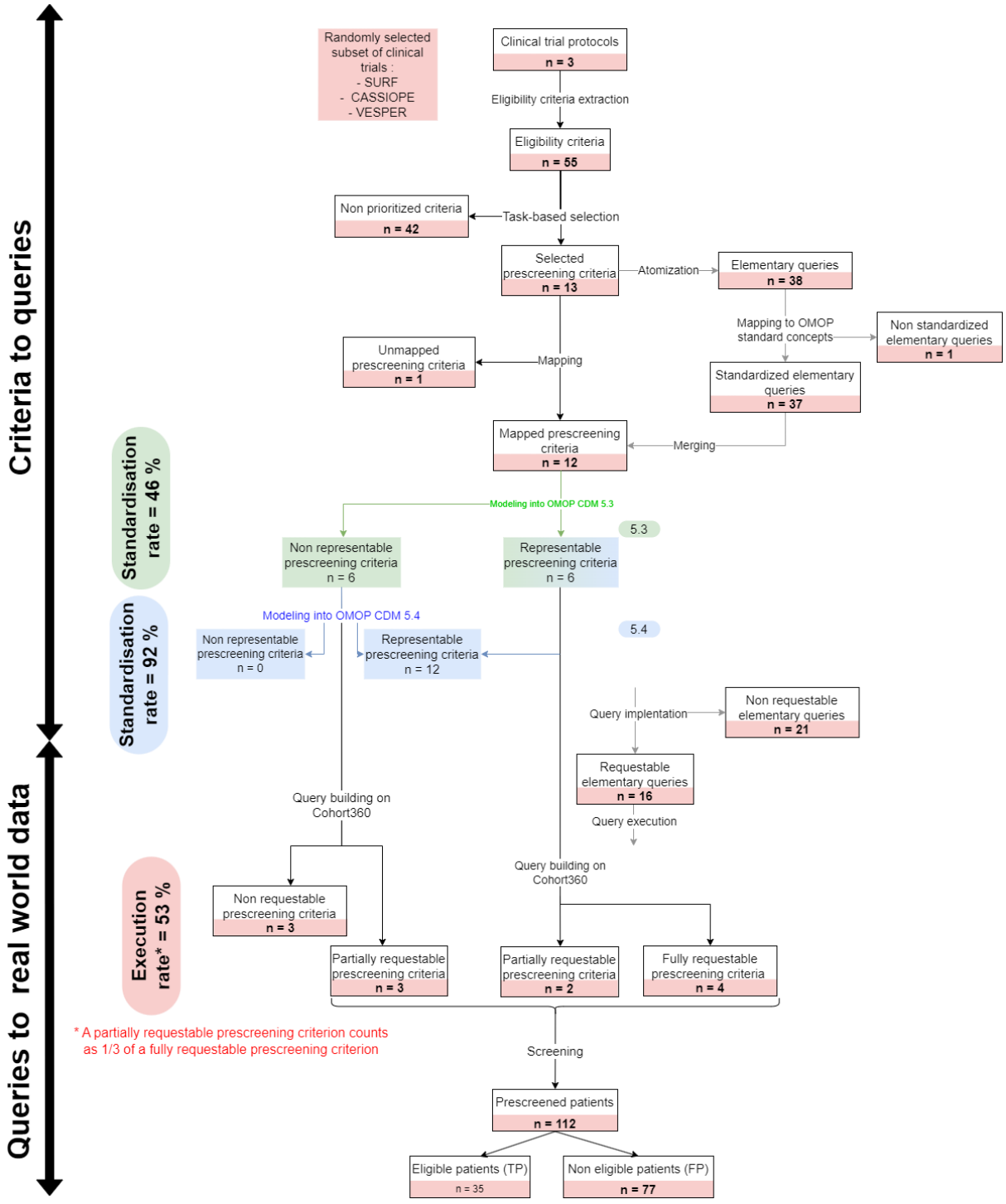


Figure 7. Performance de l'identification automatique des patients éligibles pour l'inclusion dans 3 essais cliniques avec les deux versions du modèle OMOP (v5.3 en vert et v5.4 en bleu) sur l'EDS de l'AP-HP

** Un critère de préscreening partiellement requêtable compte pour 1/3 d'un critère de préscreening complètement requêtable. PF, faux positif; TP, vrai positif.*

Enfin, l'EDS de l'AP-HP s'appuie sur des vocabulaires non standards tels que ceux de la CIM-10 et la CCAM, qui ne chevauchent pas exhaustivement les données élémentaires. Par exemple, la normalisation de la donnée élémentaire « chirurgie du cancer de la vessie autre que RTUTV ou biopsies » nécessite 1/ la modélisation de cette information par le concept_id standard 4270496 « cystectomie », 2/ la cartographie avec plus de 13 concepts distincts correspondants de la CCAM. De plus, les vocabulaires des sources utilisés à l'AP-HP sont parfois moins granulaires que ne l'exigent les critères de préscreening ou les données élémentaires. Dans une telle situation, une légère modification de l'information à modéliser est indispensable pour pouvoir requêter sur l'EDS de l'AP-HP (Figure 8).

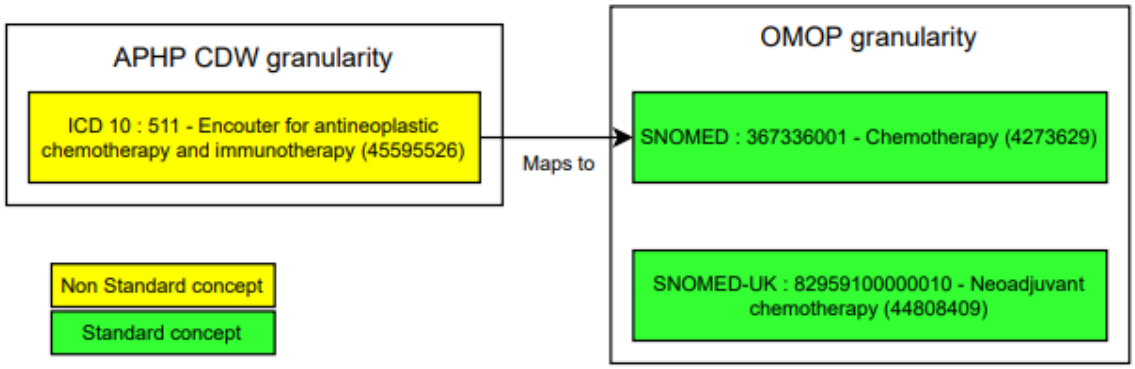


Figure 8. Exemple de perte d'information liée à l'alignement entre le vocabulaire non standard de l'EDS de l'AP-HP et la terminologie standard OMOP, pour la notion de chimiothérapie néoadjuvante

CIM-10, 10e révision de la classification internationale des maladies

IV. Discussion

1. Synthèse

L'initiative PENELOPE promeut l'utilisation d'un modèle de données basé sur des épisodes de maladie, s'appuyant sur un jeu pragmatique de données minimales standardisées cliniques et génomiques, autour de l'optimisation de l'inclusion des patients atteints de cancer dans les essais cliniques à partir des données de vie réelle. La première phase du projet PENELOPE a abordé la question délicate de la représentation formelle des critères de préscreening - dont la définition reste peu consensuelle, principalement basée sur des avis d'experts et non évolutive. Le jeu de données minimales pragmatique développé dans la première partie de la

thèse a permis l'identification des critères de préscreening. A partir de 15 essais cliniques de phase I à IV lancés à l'AP-HP entre 2016 et 2021, 88% des 288 données élémentaires liées à 83 critères de préscreening étaient alignées sur les tables et concepts standards OMOP. Le taux de standardisation de ces 83 critères de préscreening a atteint 75% avec la version 5.4 du modèle OMOP grâce à la mise en œuvre de l'extension cancer, contre 40% atteint avec la version 5.3 du modèle OMOP, soit une augmentation absolue de 35%. Appliqués à l'état actuel de l'EDS de l'AP-HP (v5.3), 58% des 33 critères de préscreening représentables en format OMOP y étaient requêtables, aboutissant à un taux d'exécution de critères de préscreening requêtables de 23%. Pour trois essais cliniques, les valeurs prédictives positives variaient de 21 % à 53 %, principalement en raison d'erreurs de codage.

2. Mise en perspective avec la littérature

a. Approche du préscreening fondée sur les dossiers patients informatisés

Conjointement à l'initiative PENELOPE, la plupart des auteurs s'accordent sur le fait que le requêtage automatique des critères d'éligibilité des essais cliniques sur les dossiers patients informatiques et/ou sur les EDS représente une option pertinente, mais uniquement pour une étape de préscreening qui nécessite une validation manuelle ultérieure (67,71,207). Raghavan *et al.* ont conclu que le traitement automatique des dossiers patients informatisés en oncologie pourrait être une étape de préscreening précieuse « avant un processus de screening manuel des patients » (16). Le projet PENELOPE a utilisé un jeu de données minimales liées au cancer, comme méthode d'identification des critères de préscreening, reposant donc sur l'avis d'experts (AL et EK) : 83 (15,5%) critères de préscreening ont été identifiés comme tels parmi les 534 critères d'éligibilité initiaux selon le jeu de données minimales développé lors de la première partie de la thèse. Différemment, Melzer *et al.* ont élaboré une approche systématique par étapes fondée à la fois sur la disponibilité des informations clés dans les dossiers patients informatisés ainsi que sur leur pouvoir discriminatoire (189). Ils ont classé les 35 critères d'éligibilité de l'essai KATHERINE en 70 données élémentaires compréhensibles, sélectives et simples, dont 76% semblaient disponibles dans l'EDS d'un hôpital universitaire allemand (208,209). Chez 106 patientes atteintes d'un cancer du sein bénéficiant d'une chimiothérapie néoadjuvante, 48 % (n = 3 551) et 27 % (n = 1 995) de l'ensemble des données élémentaires ont pu être extraites de l'examen manuel des dossiers et d'une extraction automatique de l'entrepôt, respectivement. À l'étape suivante, ils ont identifié 17 critères de préscreening parmi les 35 correspondants aux critères d'éligibilité initiaux, et ce, en fonction de leur spécificité par rapport au cancer du sein, de leur disponibilité dans l'EDS ou dans les dossiers patients informatisés, et de leur pouvoir discriminatoire (c.-à-d. avec des données disponibles menant à l'exclusion des patients). L'exécution de ces 17 critères de préscreening sur leur EDS a conduit à une sensibilité, une spécificité et des valeurs prédictives positives et négatives de 100%, 57%, 10% et 100%, respectivement.

Le projet PENELOPE a montré que les requêtes de préscreening exécutables effectuées sur l'EDS AP-HP identifiaient 47% à 79% des cas faussement positifs, principalement en raison d'erreurs de codage des caractéristiques du cancer. Dans la littérature en oncologie, malgré

des améliorations récentes, de telles limitations constituent un obstacle bien connu à une identification automatique et fiable des types primitifs de cancer (210,211). Un outil de préscreening très sensible pourrait être intéressant malgré une faible spécificité pour augmenter le recrutement des patients, mais il impliquerait un effort substantiel de revue manuelle de dossiers patients.

b. Caractérisation d'un outil automatique de préscreening

Fitzer *et al.* ont ajouté certaines conditions aux recommandations élaborées par Schreiweis *et al*. lors de l'élaboration d'un système de recrutement de patients basé sur l'analyse des dossiers patients informatisés (212–214). Conformément à ces recommandations, la démarche de PENELOPE s'inscrit comme une fonctionnalité indispensable au développement du recrutement automatique des patients basé sur les dossiers patients informatisés des EDS (215). La performance d'un outil de préscreening évolutif basé sur les dossiers patients informatisés repose sur sa capacité à aligner les critères de préscreening sur les données des dossiers patients informatisés, c'est-à-dire à les traduire en requêtes standards (étape de traduction) et à exécuter les requêtes sur les données des dossiers patients informatisés (étape d'exécution).

Dans la littérature, l'atomisation des critères d'éligibilité en multiples données élémentaires est une approche validée et une condition préalable pour évaluer l'automatisation de l'identification des patients éligibles aux essais cliniques (216). A partir d'une analyse en aveugle de 292 critères de 20 essais, Wang *et al.* ont suggéré une catégorisation en 4 classes des critères d'éligibilité liés à la facilité d'exécution électronique : de « impossible » à « facile » (217). Roß *et al.* ont suggéré une autre catégorisation selon les caractéristiques suivantes des critères d'éligibilité : sélectivité, compréhensibilité et complexité (avec des concepts multiples, des contraintes temporelles ou des comparaisons complexes, des renseignements supplémentaires ou un jugement clinique requis) (209). Jusqu'à 30% des données élémentaires du cas d'usage de PENELOPE nécessitaient des connaissances médicales supplémentaires, ce qui est cohérent avec la littérature et inhérent au guide d'implémentations d'OMOP (189).

La performance d'un outil de préscreening basé sur les dossiers patients informatisés dépend du taux d'exécution des requêtes traduites sur les données des dossiers patients informatisés, et donc, de leur disponibilité. La disponibilité de 150 données élémentaires couramment requises pour le recrutement des patients à partir des données des dossiers patients informatisés est en cours d'évaluation dans le cadre du projet Electronic Health Records for Clinical Research (EHR4CR) promu par l'union européenne (218). En 2016, Ateya *et al.* ont montré que 74 % des critères d'éligibilité de 228 essais cliniques au Royaume-Uni pouvaient correspondre à des données structurées d'EDS (207). Une étude réalisée dans cinq établissements de santé tertiaires allemands a montré que les données requises pour le recrutement des patients à partir des dossiers patients informatisés étaient complètes dans 35% des cas (71). L'utilisation de ces données élémentaires de soins courants à des fins de recherche clinique semble être possible lorsqu'il s'agit de données démographiques, de diagnostics, de procédures et d'analyses de laboratoire – contrairement aux classifications nosographiques et aux antécédents médicaux (219,220).

Dans le projet PENELOPE, 44% des 288 données élémentaires de préscreening étaient exécutables sur l'EDS de l'AP-HP, ce qui est un taux inférieur à la littérature, car certains flux de données structurées ne sont pas encore intégrés dans l'EDS AP-HP (207). Comme les études réalisées par Gulden *et al.* l'ont montré, la plupart de ces données structurées se réfèrent à des codes de diagnostics et de procédures (221,222). Dans une étude réalisée sur la leucémie lymphoïde chronique et sur le cancer de la prostate, les données non structurées représentaient jusqu'à 59% et 77% des critères d'éligibilité, ce qui est cohérent avec les résultats du cas d'usage de PENELOPE (16). Les performances de l'outil de requêtage déployé dans les hôpitaux sont encore souvent limitées, ce qui renforce la nécessité d'utiliser un jeu de données minimales communes, pertinentes dans le contexte du préscreening, et dérivées des dossiers patients informatisés. Cette convergence permettrait de traiter conjointement les sujets d'interopérabilité et de qualité des données, ainsi que d'obtenir des jeux de données de haute qualité adaptés à la tâche de préscreening en oncologie (223).

c. Modèle commun de données sur le cancer

La traduction des critères d'éligibilité et l'exécution des requêtes correspondantes sur les données des dossiers patients informatisés dépendent profondément du modèle de données sous-jacent à l'outil. L'initiative PENELOPE s'est concentrée sur l'analyse de la pertinence du modèle OMOP dans le contexte du préscreening et, en particulier, de la valeur ajoutée théorique de la v5.4 du modèle OMOP par rapport à la v5.3 (130). En 2020, un modèle dédié aux tumeurs germinales a été développé sur la base du concept similaire aux « épisodes » développés dans le cadre de l'extension oncologie d'OMOP (224). Le présent travail a révélé que la version 5.4 du modèle OMOP pourrait considérablement améliorer la représentation des critères de préscreening des essais cliniques en oncologie. Comme identifié dans la littérature, les limites de structuration reposent majoritairement sur la nécessité d'interprétation médicale experte ainsi que sur les liens structurels entre données élémentaires au sein d'un même critère de préscreening (225). La version 5.4 du modèle OMOP intègre une extension dédiée au cancer et, ce faisant, de nouvelles tables EPISODE et EPISODE_EVENT qui sont dérivées soit de l'information contenue dans les tables CLINIQUES, soit directement à partir des données sources. Ces tables permettent de recentrer les informations autour de la maladie du patient, et non plus autour des caractéristiques démographiques du patient (table PERSON). Les tables EPISODE et EPISODE_EVENT imposent un formalisme autour de l'information clinique d'intérêt, qui devient le point d'entrée centralisant les requêtes ultérieures (Figure 6 et Supp Figure 1). Dans l'extension oncologie d'OMOP, la représentation du diagnostic du cancer combine la caractérisation histologique et topographique, dont les champs « modifiers » permettent le renseignement d'items clés descriptifs tels que le stade, le grade, la latéralité, les biomarqueurs, etc. Par ailleurs, les traitements antitumoraux sont représentés de façon exhaustive, les champs « modifiers » permettant la caractérisation précise des modalités thérapeutiques telles que le nombre de fractions, la dose totale, le rythme d'administration, etc. L'extension oncologie d'OMOP permet également l'alignement avec les dictionnaires de données de l'association nord-américaine des registres centraux du cancer (NAACCR).

Par exemple, dans le format standard OMOP, le « cancer de vessie » peut être représenté alternativement par les deux tables différentes CONDITION_OCCURRENCE ou MEASUREMENT. Dans la version 5.4 du modèle, ces informations informent un champ unique de la table EPISODE_EVENT, ce qui permet l'interopérabilité de plusieurs EDS ayant opté pour des choix initiaux distincts (Figure 5). Par ailleurs, le consortium OHDSI a établi des recommandations relatives aux tables EPISODE et CONDITION_OCCURRENCE, en précisant que la table CONDITION_OCCURRENCE devait être priorisée par rapport à la table MEASUREMENT. Dans la table EPISODE, un domaine REGIMEN a été créé dans le vocabulaire HemOnc pour représenter les protocoles de chimiothérapie. Par exemple, dans la version 5.3 du modèle OMOP, le « cisplatine » doit être séparé de la « gemcitabine » dans la table DRUG_EXPOSURE, tandis que dans la v5.4, le protocole de chimiothérapie concomitant « CISPLATINE GEMCITABINE » peut être représenté directement.

La version 5.4 du modèle OMOP fournit deux champs supplémentaires dans la table MEASUREMENT : measurement_event_id et meas_event_field_concept_id. Ces « champs polymorphiques » permettent de préciser des informations clés, comme le lien entre une métastase et sa tumeur primitive. Comme indiqué précédemment, une information clinique unique peut être soit directement représentée en OMOP par un concept unique, soit par une multiplicité de concepts avec une granularité plus petite – telle que l'option prise dans le présent travail pilote. La v5.4 suggère de faire référence à des concepts d'identification prétraités tels que, p. ex., le concept « carcinome urothélial de haut grade » plutôt que la juxtaposition des concepts « haut grade » et « carcinome à cellules transitionnelles ». Parmi les questions non résolues par le modèle OMOP v5.4, un arbitrage local par des experts demeure nécessaire pour la mise en œuvre de la table EPISODE. Néanmoins, le modèle OMOP est un outil prometteur pour modéliser les données du cancer et promouvoir l'interopérabilité entre les EDS (226–229).

D'autres modèles ont été développés pour accélérer le partage des données en oncologie. Malgré son expressivité, le modèle OSIRIS français pourrait manquer d'évolutivité comme le révèle son adoption encore quelque peu limitée, et le développement actuel de sa version simplifiée et pragmatique OSIRIS_RWD (183,190). La version 3 du modèle HL7 FHIR mCODE™ a été expérimentée à l'échelle nationale et, en raison de son adoption croissante, est une candidate prometteuse (131,132,230–232). Les solutions basées sur FHIR sont, p. ex., des perspectives réalisables concernant les études de faisabilité à partir de critères de préscreening utilisant des concepts standards (230,233–235). La Figure 9 illustre la couverture du modèle de données mCODE™ concernant les 15 éléments du jeu de données minimales correspondant aux critères de préscreening identifiés par l'initiative PENELOPE. Il semble que la version actuelle de mCODE™ pourrait bénéficier d'une plus grande explicitation des règles de modélisation des épisodes, qui fondent la représentation de la maladie cancéreuse, ainsi que de celle des données de biologie moléculaire (236).

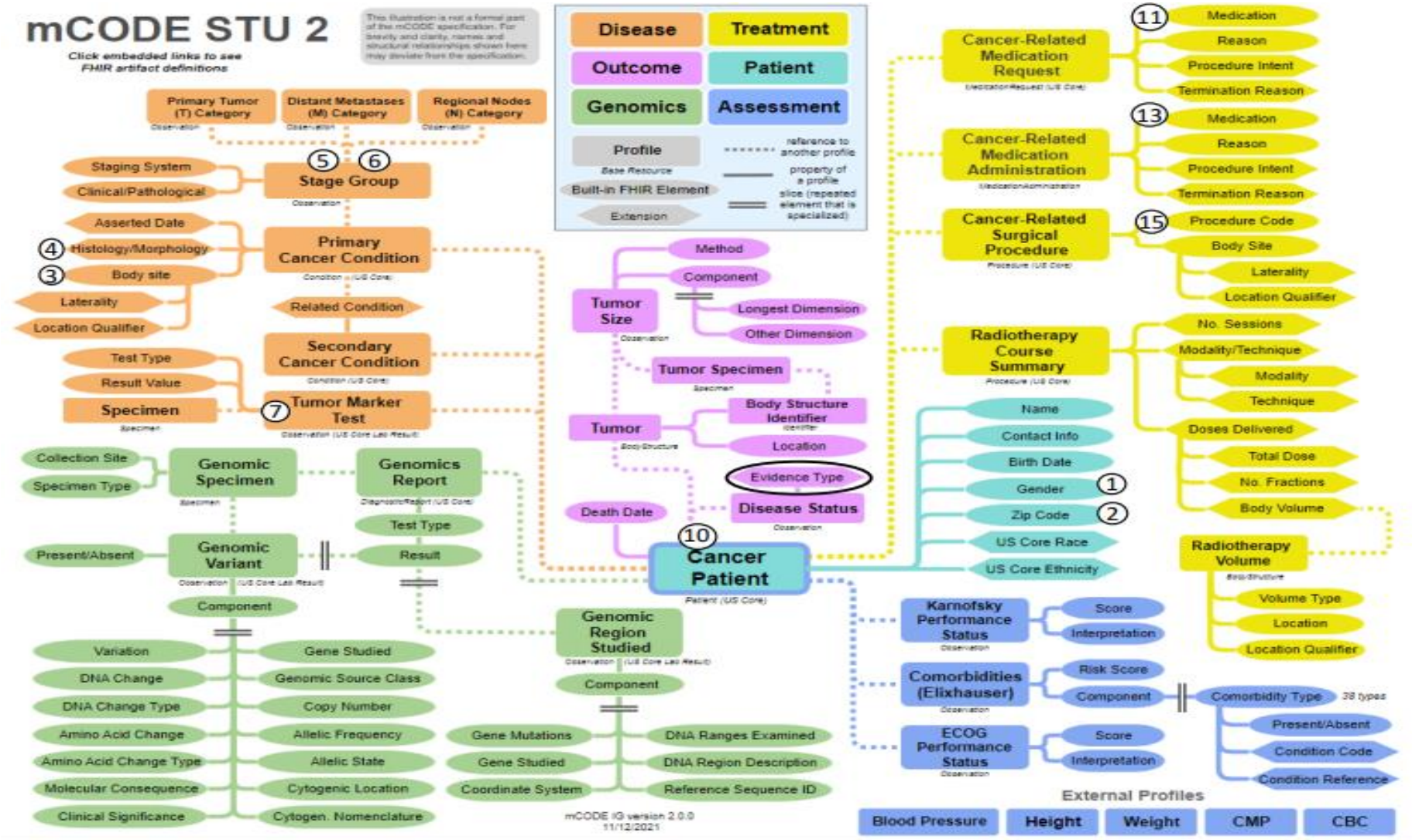


Figure 9. Couverture du modèle de données mCode des 15 éléments du jeu de données minimales correspondant aux critères de préscreening de l'initiative PENELOPE. CBC, numération formule sanguine ; CMP, panel métabolique complet ; ECOG, Eastern Cooperative Oncology Group

Certaines initiatives mettant en œuvre l'automatisation du préscreening ne reposent pas explicitement sur des modèles de données du cancer communs. Dans cette approche, les algorithmes d'extraction textuelle en traitement du langage naturel semblent de prometteuses techniques pour améliorer l'inclusion des patients atteints de cancer dans les essais cliniques (136,229,237,238). Par exemple, certains auteurs ont développé un outil de traitement automatique du langage naturel lié à l'inférence textuelle qui aide les médecins responsables du recrutement des patients en mettant en évidence les zones de textes pertinentes au sein des comptes rendus des dossiers patients informatisés (239). Une limite repose sur la discordance et la variabilité temporelle d'éléments clés tels que le stade du cancer dans les dossiers patients informatisés (240). Les solutions de recherche d'informations, telles que l'outil EMERSE, reposent sur l'implication pragmatique du clinicien en charge du recrutement des patients, sans connaissances préalables en méthodologie de traitement automatique du langage naturel, qui imite le processus de préscreening de la pratique clinique de routine (138,241). Certains auteurs développent des modèles de prédiction basés sur des forêts aléatoires pour évaluer l'éligibilité des patients, sans avoir besoin de traduction ou de représentation des critères (242).

Il est intéressant de noter que certaines initiatives telles que la base de connaissances open source sur les essais cliniques (CTKB) ont tenté de résoudre le problème de l'interopérabilité insuffisante entre les critères d'éligibilité et les données issues des dossiers patients informatisés, grâce à l'utilisation des deux approches : concepts standards promus par le modèle OMOP et adoption large du traitement automatique du langage naturel (243). La CTKB, avec une interface utilisateur basée sur le web, comprend 87 504 concepts standards OMOP liés à 35% de critères d'inclusion et 65% de critères de non-inclusion extraits avec l'outil Criteria2Query parmi 352 110 essais cliniques disponibles sur le répertoire en ligne ClinicalTrials.gov (244). Le traitement automatique du langage naturel semble également prometteur pour pouvoir enrichir des modèles de données communs, puisque dans le projet PENELOPE, 59% des données élémentaires n'étaient pas requêtables sur l'EDS de l'AP-HP en raison de leur manque de structuration (Figure 10).

En conclusion, les contributions de ce chapitre au travail de thèse furent de démontrer qu'il était possible de :

- Mettre en œuvre une méthodologie d'appui à la structuration et la standardisation d'un jeu de données de vie réelle dans le cadre d'un cas d'usage particulier ;
- Définir des critères de préscreening au sein des critères d'inclusion et de non-inclusion des essais cliniques à partir du jeu de données minimales en oncologie défini dans la Partie I de la thèse ;
- Définir et appliquer une méthodologie d'évaluation d'un outil automatique d'identification de patients éligibles à l'inclusion dans des essais cliniques à partir des dossiers patients informatisés ;
- Aligner des critères de préscreening d'essais cliniques à des terminologies respectant les standards internationaux, et promues par le consortium OHDSI ;
- Evaluer la plus-value de l'extension oncologie du modèle OMOP en termes de représentativité des critères de préscreening d'essais cliniques, par rapport à la version 5.3 du modèle commun de données, et en préciser les fondements ;
- Identifier l'étiologie des erreurs de requêtage par un outil automatique sur dossiers patients informatisés, en termes de qualité des données ;

- Articuler le contenu du jeu de données minimales en oncologie avec les spécifications du modèle de données commun mCODE promu par HL7 FHIR.

La partie suivante de la thèse s'attachera à enrichir le jeu de données minimales d'oncologie polyvalent via le développement d'algorithmes dédiés, tout en évaluant la pertinence des différentes méthodes d'extraction textuelle appliquées aux dossiers patients informatiques, et ce, à travers trois cas d'usage distincts : le projet du calcul automatique des indicateurs de qualité et de sécurité des soins EUSOMA concernant le cancer du sein (cas d'usage 1), le projet CovOnco pour les études épidémiologiques (cas d'usage 3) et le projet de vision par ordinateur 'Challenge AI for health' (cas d'usage 4).

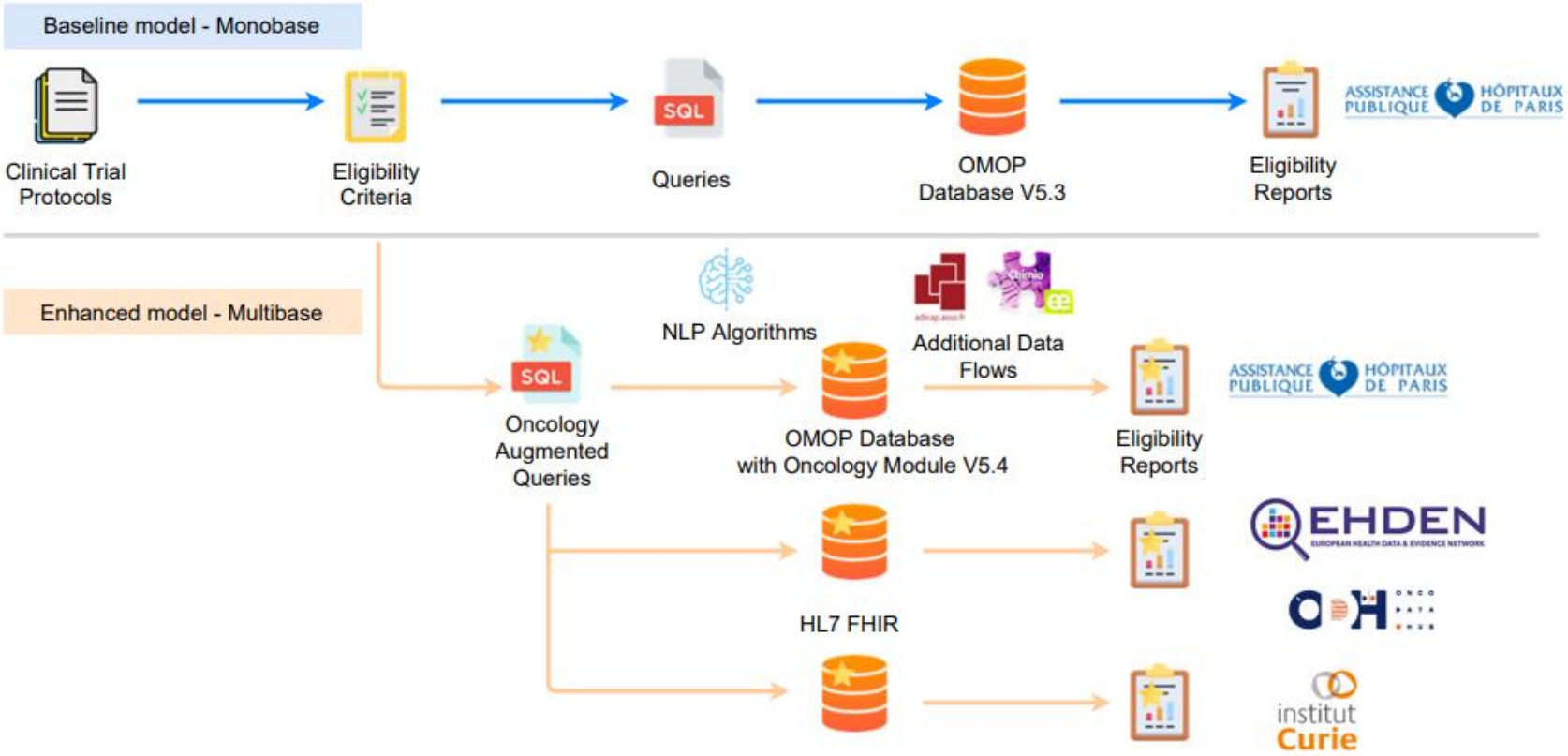


Figure 10. Les modèles de données communs actuels et en perspective de l'initiative PENELOPE

HL7 FHIR, Health Level Seven International Fast Healthcare Interoperability Resources ; TAL, traitement du langage naturel

Partie III : Traitement automatique du langage naturel et structuration du jeu de données minimales polyvalent en oncologie

I. Introduction

Une fois le jeu de données minimales en oncologie défini et sa modélisation optimale identifiée, ses composantes sous format non structuré dans l'EDS de l'AP-HP ont été individualisées. La partie III de la thèse a consisté en la réalisation de travaux de traitement automatique du langage naturel permettant la caractérisation des patients atteints de cancer solide à partir de ce type de supports. Leur structuration a été pensée envisageable via leur extraction semi-automatique à partir de comptes rendus textuels médicaux pertinents. Le lien avec la seconde étape de la thèse a été anticipé via la génération de guides d'annotations et de bibliothèques mises en accès libre sur edsnlp qui propose des outils d'extraction textuelle appliquée à des comptes rendus médicaux français, en vue de l'enrichissement des tables OMOP correspondantes (245). Au-delà de la production d'algorithmes, cette partie de la thèse a permis d'enrichir la réflexion méthodologique concernant la plus-value du traitement automatique du langage naturel par rapport aux données médico-administratives, ainsi qu'au sujet de l'optimisation des tâches manuelles d'annotations, étapes indispensables et aujourd'hui limitantes à l'entraînement de modèles d'apprentissage machine performants.

Les publications produites dans le cadre des projets démonstrateurs ont respecté, concernant les données épidémiologiques, l'extension REporting of studies Conducted using Observational Routinely-collected Data (RECORD) des recommandations STROBE (Strengthening the Reporting of Observational Studies in Epidemiology), alors que les résultats algorithmiques ont suivi la déclaration Minimum information about clinical artificial intelligence modeling (MI-CLAIM) (246,247).

La troisième partie de la thèse a cherché à extraire des items non structurés du jeu de données minimales en oncologie, développé à la Partie I, à savoir les informations nécessaires à la stadification tumorale au diagnostic, et à la pronostication histopathologique des cancers, et ce, en vue de leur structuration. Le travail computationnel s'est déployé en interaction étroite avec une réflexion méthodologique veillant à : 1/ évaluer la plus-value d'initiatives de traitement automatique du langage naturel par rapport aux données structurées médico-administratives existantes, 2/ optimiser les modalités d'annotations manuelles, tâche aujourd'hui limitante du développement d'algorithmes d'apprentissage machine d'extraction textuelle.

1. Extraction textuelle de la stadification tumorale au diagnostic

Ce premier objectif consista en l'extraction automatique de l'information médicale nécessaire à la stadification tumorale au moment du diagnostic d'un cancer solide à partir, d'une part, de l'identification des métastases sur les examens d'imagerie médicale (scanner) et d'imagerie nucléaire (pet-scanner) dédiés au bilan d'extension des cancers récemment diagnostiqués et, d'autre part, à partir de l'évaluation du stade pTNM issus des comptes rendus d'anatomopathologie post-opératoires des cancers opérés. Sur un plan méthodologique, il s'agissait également d'évaluer la plus-value d'algorithmes de traitement automatique du langage naturel par rapport aux données structurées issues du PMSI en termes de codage des lésions métastatiques. Le projet d'épidémiologie CovOnco dans lequel cet objectif s'est intégré a cherché à évaluer l'impact de l'épidémie et des mesures de santé publique associées sur les filières de soin des patients nouvellement atteints de cancer en Ile-de-France (248). Les algorithmes de stadification initiale ont été développés et testés sur une cohorte de patients nouvellement atteints de cancer colorectal. Ils ont ensuite été validés sur une population de patients porteurs d'un nouveau diagnostic de cancer du pancréas, puis de cancer du sein.

2. Extraction textuelle des critères histopronostiques

Ce deuxième objectif consista en l'extraction automatique de l'information médicale nécessaire à la pronostication histologique des tumeurs réséquées (249). Ont été intégrés les critères histopronostiques minimums et communs à tout type de cancer solide réséqué, tels que recommandés par l'INCa et l'ANS (176,177). Ainsi, les variables suivantes ont été extraites des comptes rendus d'anatomopathologie post-opératoire de tumeurs réséqués à l'AP-HP : invasion ganglionnaire (tumorale) (présence / absence), invasion vasculaire (tumorale) (présence / absence), complétude de la résection microscopique (présence / absence),

invasion périnerveuse (tumorale) (présence / absence), score (y)pTNM ((y)p T1-4 N0-3 M0-1), taille (tumorale), différenciation (tumorale) (peu, moyennement, bien).

L'objectif méthodologique sous-jacent à cette tâche d'extraction fut de discuter et évaluer la faisabilité d'une méthode de pré-annotation automatique fondée sur un système de règles, en vue de l'entraînement séquentiel d'un algorithme supervisé d'apprentissage machine. En effet, plusieurs alternatives d'optimisation des tâches d'annotations manuelles sont à disposition des développeurs en traitement automatique du langage naturel, sans qu'elles ne soient réellement hiérarchisées les unes par rapport aux autres dans la littérature. Existent ainsi des systèmes à base de règles fondées sur des expressions régulières qui sont des chaînes de caractères décrivant un ensemble de caractères possibles, et des systèmes statistiques fondés sur des modèles d'apprentissage machine. Initiés à partir du cas d'usage 'Challenge AI for health' dédié au cholangiocarcinome intra-hépatique, cette tâche d'extraction fut élargie à d'autres types primitifs tumoraux fréquents, comme le cancer du sein dans le cas d'usage du calcul automatique des indicateurs qualité internationaux du cancer du sein EUSOMA (17).

II. Identification du stade tumoral au diagnostic

1. Stade tumoral post-opératoire pTNM

a. Méthodes

Les comptes rendus d'anatomopathologie présents dans les dossiers patients informatisés ont été identifiés par les données du logiciel métier DIAMIC présentes dans l'EDS de l'AP-HP. Leur caractère post-opératoire a été identifié à partir de la date structurée de l'édition du compte rendu médical en question, et de celle de l'acte chirurgical de résection tumorale correspondant codé selon la CCAM (Annexe 3c). L'expression régulière d'identification de la date de prélèvement issue du texte était composée de deux sous-parties intégrant toutes les variantes orthographiques :

- la première recherchait la classe d'expressions en lien avec la notion de prélèvement daté :

  "(pr[ée]lev[ée]\sle\s?:?\s{1,3}\d{2}\/\d{2}\/\d{2,4}

- la deuxième recherchait la classe d'expressions en lien avec la date du prélèvement, appliquée à des documents pseudonymisés et dont les dates étaient normalisées sous format JJ/MM/AAAA :

  date\sdu\spr[ée]l[èe]vement\s?:?\s{1,3}\d{2}\/\d{2}\/\d{2,4})"

Ordonnés par ordre chronologique, les premiers comptes rendus d'anatomopathologie trouvés après l'acte de résection tumorale ont été considérés comme post-opératoires.

Le stade anatomopathologique post-opératoire de la tumeur (pTNM) a été répertorié selon la 8e édition de la commission mixte américaine sur le cancer (AJCC) (250). Le score pTNM a été développé dans les années '40 pour classifier le pronostic des cancers, et ce faisant, leur extension géographique (251). Il est composé d'un indice « p » pour « pathology » précise que l'analyse du score est réalisée en anatomopathologie alors que la présence de l'indice « yp » précise que le cancer a été traité avant l'examen anatomopathologique (p. ex., par radiothérapie ou par chimiothérapie). Le « T » pour « tumor » détaille l'extension locale de la tumeur au sein de l'organe primitif concerné, en fonction de la taille du cancer. Le « N » pour « node » décrit la dissémination ganglionnaire loco-régionale du cancer, comme la présence de cellules cancéreuses dans les ganglions de l'aisselle homolatérale en cas de cancer du sein. Enfin, le « M » pour « metastasis » objective une éventuelle extension à distance sous forme de métastases, c'est-à-dire de colonies tumorales dans des organes distincts de celui dont est issu la tumeur primitive. Les stades ypTNM et pTNM ont été distingués en fonction de la présence d'un traitement néoadjuvant à la résection chirurgicale du primitif, et classés selon le risque de rechute tumorale ultérieure : les risques faible et élevé étant définis comme (y)pTxN0 et (y)pTxN1-2, respectivement.

L'expression régulière suivante a été développée en vue de l'extraction automatique du stade pTNM tumoral :

```
([ycpP]{1,2}\s? (T([01234x]|is)[abcdx]?) [,\s] {0,2} [ycp] {0,2}\s? (N[xo01234\+][abcdx]?) *\s?
(M[o01]? [\+x]?)?)|((T([01234x]|is)[abcdx]?) [,\s] {0,2} [ycp] {0,2}\s? (N[xo01234\+][abcdx]?) \s?
(M[o01]? [\+x]?)?)
```

Cette expression régulière permet d'identifier la présence d'un score pTNM au sein d'un compte rendu, et également d'extraire les valeurs associées pour chacune des sous-variables « T », « N », « M » - tâche non assurée par les modèles d'apprentissage machine. En cas de multiples scores pTNM par document, les annotations associées au pire pronostic ont été retenues (Annexe 4). Le jeu de données annotées a été divisé aléatoirement dans le cas d'usage du cancer colorectal en un jeu de développement (50%) et un jeu de test (50%). Les performances de l'expression régulière ont été évaluées sur 100 comptes rendus d'anatomopathologie postopératoires pour le cancer du pancréas, et sur 48 comptes rendus pour le cancer du sein, en guise de jeux de test. Sur les jeux de développement et de test, les paramètres de performance suivants ont été évalués : rappel (sensibilité), précision (valeur positive prédictive), score F1* (moyenne harmonique entre sensibilité et valeur positive prédictive) :

$$^{*}\mathrm{F}_1 = \frac{2\,(\,\text{Sensitivity} * \text{Positive Predictive Value})}{\text{Sensitivity} + \text{Positive Predictive Value}}.$$

b. Résultats

Les méthodes de développement et de validation de ces algorithmes sont référencés dans un article original dédié du projet CovOnco publié en 2021 par l'*International Journal of Cancer*

autour du cancer colorectal (248). Leur validité externe concernant le cancer du pancréas est référencée dans un article original publié en 2023 par l'*International Journal of Cancer* (252). La validité externe concernant le cancer du sein est référencée dans un article original anglais en cours de révision par la *Revue d'Epidémiologie et de Santé Publique*, et ayant fait l'objet d'une présentation orale au congrès de l'EMOIS en 2023 dans le cadre du cas d'usage 1 de la thèse (253).

La Figure 11 illustre la proportion de patients pour lesquels un compte rendu textuel anatomopathologique post-opératoire était disponible, parmi 1 906 patients nouvellement atteints d'un cancer colique et ayant bénéficié d'un séjour en chirurgie pour colectomie, au sein de l'EDS de l'AP-HP, et en fonction de la date de chirurgie. Parmi les 1 099 patients avec un compte rendu anatomopathologique post-opératoire disponible, 802 contenaient un score pTNM mentionné dans ce texte, tel qu'identifié par les expressions régulières correspondantes.

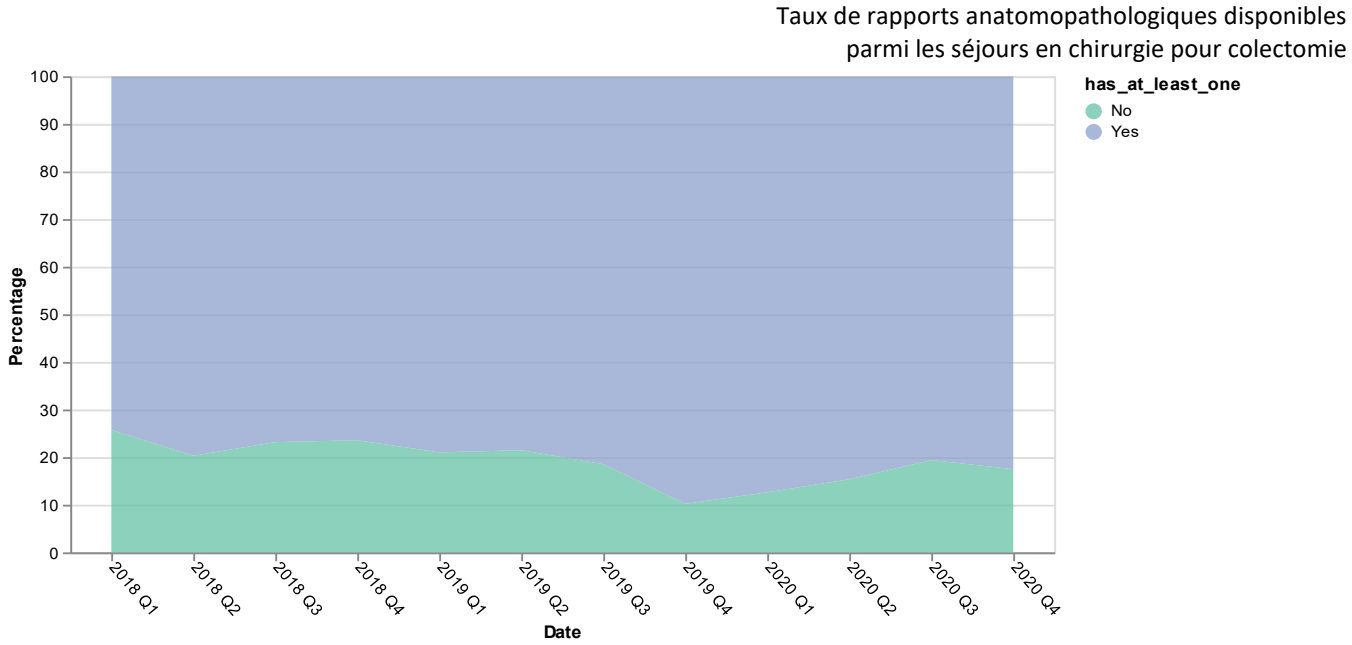


Figure 11. Disponibilité dans l'EDS de l'AP-HP des comptes rendus anatomopathologiques post-opératoires parmi les séjours en chirurgie pour colectomie chez des patients nouvellement atteints de cancer colorectal, en fonction de la date de chirurgie

Concernant le cancer colorectal, 200 comptes rendus d'anatomopathologie postopératoires ont été choisis aléatoirement parmi les 1 099, et annotés. L'algorithme d'identification du score pTNM a été associé à un rappel et une précision de 98% et 96%, identiques sur les jeux de développement et de test. Le score F1 de l'algorithme associé a atteint 0.95 et 0.97 sur le jeu de développement et le jeu de test, respectivement. Concernant le cancer du pancréas, parmi 588 ayant bénéficié d'une exérèse du pancréas pour un cancer invasif dont 573 avaient un compte rendu anatomopathologique postopératoire disponible, 100 comptes rendus d'anatomopathologie postopératoires ont été annotés pour constituer le jeu de test. L'algorithme d'identification pTNM y a été associé à une sensibilité et une valeur prédictive positive de 96% et 92%, respectivement. Concernant le cancer du sein, 94% (n= 3 381) des

3 575 patientes opérées d'un cancer du sein invasif à l'AP-HP disposaient d'un compte rendu d'anatomopathologie post-opératoire, dont 48 ont été annotés comme jeu de test. Les précisions, sensibilité et score F1 y ont atteint des taux de 92%, 93% et 0.92, respectivement.

c. Portabilité

Après la publication princeps de CovOnco objectivant la chute du nombre de diagnostics de nouveaux cancers référés à l'AP-HP pendant les premières vagues de l'épidémie, et ce, sans rattrapage ultérieur de l'activité médicale, deux publications d'épidémiologie ont été réalisées concernant le cancer colorectal : une lettre publiée dans le *Lancet Gastrooncology*, et un article original dans l'*International Journal of Cancer* présentant ces algorithmes d'extraction textuelle (248,254). L'objectif principal de CovOnco était d'évaluer un éventuel impact des retards diagnostics des cancers lors de la pandémie, en termes d'aggravation des stades de dissémination tumorale au diagnostic de nouveau cancer. Ce travail a permis de préciser que, pour les cancers coliques, pancréatiques, broncho-pulmonaires et mammaires, aucune aggravation des stades tumoraux pTNM n'a été objectivée pour les tumeurs opérées. Le Tableau 3 résume le stade tumoral anatomopathologique de 929 cancers du côlon localisés après la résection initiale de la tumeur primaire. Aucune différence statistique n'a été observée entre les deux classes de risque.

Tableau 3. Stade anatomopathologique tumoral post-colectomie, chez 929 cancers du côlon localisés, selon l'année de diagnostic à l'Assistance Publique – Hôpitaux de Paris

Stade tumoral postopératoire	2018	2019	2020
Risque faible, N (%) pT0/T1/T2/T3 et pN0	122 (53%)	132 (49%)	156 (53%)
Risque élevé, N (%) pT4 ou pN1/N2	107 (47%)	137 (51%)	136 (47%)
Valeurs de p *par rapport à la moyenne de 2018-2019*			*0.56*

Des résultats similaires ont été obtenus pour :

- le cancer du pancréas, et référencés dans un article original d'épidémiologie publié en 2023 par l'*International Journal of Cancer* (252) ;
- le cancer broncho-pulmonaire, et référencés dans un article court publié dans le *European Journal of Cancer* en 2022 (255) ;
- le cancer du sein, et référencés dans un article original en cours de révision par le comité éditorial de la revue *Cancer Medicine* avec présentation en poster prévue au congrès de la société européenne d'oncologie médicale (ESMO) en octobre 2023.

Le détail des analyses épidémiologiques issues du développement algorithmique est disponible au sein de l'Annexe 5.

2. Stade métastatique au diagnostic

a. Méthodes

Le statut métastatique des cas lors de la présentation initiale au diagnostic du cancer a été automatiquement extrait des comptes rendus textuels d'imagerie médicale, en deux étapes, dont la méthode a évolué au cours du temps, compte tenu de l'analyse successive de plusieurs types de cancers primitifs différents (côlon, poumon, pancréas, sein).

La première étape a permis d'identifier les comptes rendus des examens d'imagerie d'intérêt disponibles dans l'EDS de l'AP-HP, grâce à leur titre et à la restriction d'une fenêtre temporelle autour de la date de diagnostic de cancer. L'amplitude de cette fenêtre temporelle dépendait du profil évolutif et de l'agressivité du primitif tumoral concerné. Les examens de tomodensitométrie (TDM) thoraco-abdomino-pelvienne (TAP) ont été identifiés comme tels si le titre du compte rendu contenait

- l'une des expressions françaises suivantes :
  SCANNER OU TDM OU TOMODENSITOM[EÉ]TRIE,
- ainsi que la localisation anatomique investiguée par l'examen :
  TAP|TH? ORACO[-|]? ABDO|TH? ORACI[^\s]* *(?:ET)? *ABDO|.

Les examens de tomographies par émissions de positons (TEP) couplées au scanner ont été identifiés comme tels si le titre du compte rendu contenait

- l'expression française se référant à l'utilisation du fluorodésoxyglucose18 :
  \b(((18)?f-fdg)|(grippe(oro)?désoxyglucose)|(fdg-\b18f\b)),
- ou la combinaison des deux expressions françaises concernant le TEP-scanner :
  \b((tep)|(pet)|(tomo(scinti)?graphie par [ée]mission de position(s)?)) (\s|-|\/)
- ou la présence d'une activité nucléaire :
  \b((m[ée]decine nucl[ée]aire)|([0-9] {1,4} ((\s)?) MBQ)|([Aa]ctivité inject[eé](e)?))

La deuxième étape a consisté à identifier la justification médicale de réalisation de l'examen, pour ne se concentrer que sur ceux réalisés dans l'objectif délibéré d'une stadification tumorale initiale. Il s'agissait d'exclure les comptes rendus d'imagerie dans lesquels la description de métastases était fortuite ou non systématique. A cette fin, des modèles de classification ont été développés selon plusieurs méthodes distinctes (forêts aléatoires, réseau de neurones convolutif, etc.) et leurs performances évaluées. L'explicabilité de ces différents modèles de classification a été évaluée par des techniques fondées sur le modèle SHapley Additive exPlanations (SHAP) (256,257). Les imageries d'intérêt ont été annotées comme des examens réalisés délibérément pour évaluer la stadification tumorale initiale du cancer récemment diagnostiqué, caractéristique identifiée par le médecin oncologue annotateur d'après l'analyse des sections « contexte / indication » et « conclusion » du compte rendu d'imagerie. L'annotateur a été invité à classer le compte rendu textuel comme provenant d'une imagerie de stadification tumorale initiale ou non.

La troisième étape fut d'extraire l'éventuelle mention de présence d'une métastase, à l'aide de l'algorithme d'expression régulière composée d'une partie caractérisant la lésion de

- métastatique, disséminée, issue de carcinose, allure secondaire ou lâcher de ballons :

(m[ée]tasta(se|tique)s?)|(diss[ée]min[ée]e?s?)|(carcinose)|(((allure|l[ée]sion|localisation|progression)s?\s)(suspecte?s?)?. {0,30} (secondaire)s?)|(l(a|â)ch(é|e|er)\sde\sballons?)

- lésion cible, associée à une rupture de la corticale, un envahissement des parties molles, ou une ostéolyse :

(l[ée]sions\s(non\s)?cibles)|(rupture. {1,20}corticale)|(envahissement. {0,15}parties\smolles)|((l[i,y]se). {1,20}os)|ost[eé]ol[i,y]|rupture. {1,20}corticale|envahissement. {1,20}parties\smolles

- ostéocondensation secondaire, évolutive, fracturaire ou suspecte :

ost[eé]ocondensa. {1,20} (suspect|secondaire|[ée]volutive)|(l[ée]sion|anomalie|image). {1,20}os. {1,30} (suspect|secondaire|[ée]volutive)|os. {1,30} (l[ée]sion|anomalie|image). {1,20} (suspect|secondaire|[ée]volutive)|(l[ée]sion|anomalie|image). {1,20}L[I,Y]tique|(l[ée]sion|anomalie|image). {1,20}condensant. {1,20} (suspect|secondaire|[ée]volutive)|fracture. {1,30} (suspect|secondaire|[ée]volutive)|((l[l[ée]sion|anomalie|image|nodule). {1,80} (secondaire))|((l[l[ée]sion|anomalie|image|nodule)s.{1,40}suspec?ts?).

L'annotateur a classé les comptes rendus textuels comme rapportant ou non au moins une mention de description de métastase. La qualification des négations et des hypothèses a été prise en charge à l'aide de la bibliothèque edsnlp (v0.7.4)(245). Sur les jeux de développement et de test, les paramètres de performance suivants ont été évalués : rappel (sensibilité), précision (valeur positive prédictive), score F1* (moyenne harmonique entre sensibilité et valeur positive prédictive). L'annotateur était oncologue médical (EK).

Les recouvrements des données structurées concernant l'identification du statut métastatique des cancers fournies par la CIM-10 (codes C78 et C79) furent évaluées chez les patients nouvellement atteints de cancer et chez qui une annotation manuelle par EK retrouvait la présence de métastases sur le compte rendu textuel de TDM TAP et/ou de TEP scanner de stadification initiale. La fenêtre temporelle retenue pour les données du PMSI couvrait la période allant de deux semaines avant la date de réalisation de l'examen jusqu'à trois mois après, compte tenu du caractère rétroactif du codage.

Concernant le cancer colorectal, les TDM TAP effectués entre 90 jours avant et 45 jours après la date de diagnostic du cancer colorectal ont été définis comme comptes rendus d'imagerie d'intérêt. La caractéristique de stadification initiale de l'examen d'imagerie identifié a été évaluée en comparant deux méthodes de classification binaire par apprentissage :

1) un algorithme de forêts aléatoires basé sur la fréquence des mots des sections « contexte / indication » et « conclusion » des comptes rendus textuels,
2) un réseau de neurones convolutif utilisant le modèle de langue Word2Vec pré-entraîné sur l'EDS (258).

La bibliothèque scikit-learn a été utilisée pour implémenter les forêts aléatoires. La bibliothèque pytorch a été utilisée pour développer et former le réseau de neurones. Le jeu

de données annotées a été divisé aléatoirement en un jeu d'entraînement (70 %) et un jeu de test (30 %). Concernant le cancer du pancréas, les TAP effectués entre 90 jours avant et 30 jours après la date de diagnostic du cancer du pancréas ont été définis comme comptes rendus d'imagerie d'intérêt. L'algorithme de forêts aléatoires précédemment développé pour le cas d'usage du cancer colorectal a été utilisé pour identifier l'objectif de réalisation des examens d'imagerie. Concernant le cancer du sein, les TDM TAP et/ou les TEP scanner effectués entre 90 jours avant et 45 jours après la date de diagnostic du cancer du sein ont été définis comme comptes rendus d'imagerie d'intérêt. En plus de l'algorithme de forêt aléatoire et du réseau de neurones convolutif précédemment décrits, un classifieur à renforcement de gradient (*gradient boosting*) basé sur la fréquence des mots des sections « contexte / indication » et « conclusion » des comptes rendus textuels a été évalué.

Le jeu de données annoté a été divisé aléatoirement en un jeu d'apprentissage (65%) et un jeu de test (35%). Une partie du jeu d'apprentissage (35%, soit 55 comptes rendus) a été utilisée comme jeu de développement pour :

- appliquer la méthode d'optimisation Grid Search afin de définir les hyperparamètres optimaux concernant la forêt aléatoire et le classificateur à renforcement de gradient (259) ;
- évaluer la précision et la *loss* du réseau de neurones convolutif à chaque époque d'apprentissage et statuer sur l'arrêt de son entraînement.

b. Résultats

- Identification des imageries de stadification tumorale initiale

Concernant le cas d'usage dédié au cancer colorectal, 436 comptes rendus de TDM TAP ont été annotés manuellement pour évaluer la caractéristique 'stadification initiale' des examens d'imagerie, parmi lesquels 307 ont constitué le jeu de développement. La Figure 13 illustre le taux de patients atteints d'un cancer colorectal nouvellement diagnostiqué et bénéficiant d'un compte rendu de TDM TAP disponible dans l'EDS de l'AP-HP. La complétude de cette donnée ne semblait pas dépendre de la date de diagnostic de cancer du patient.

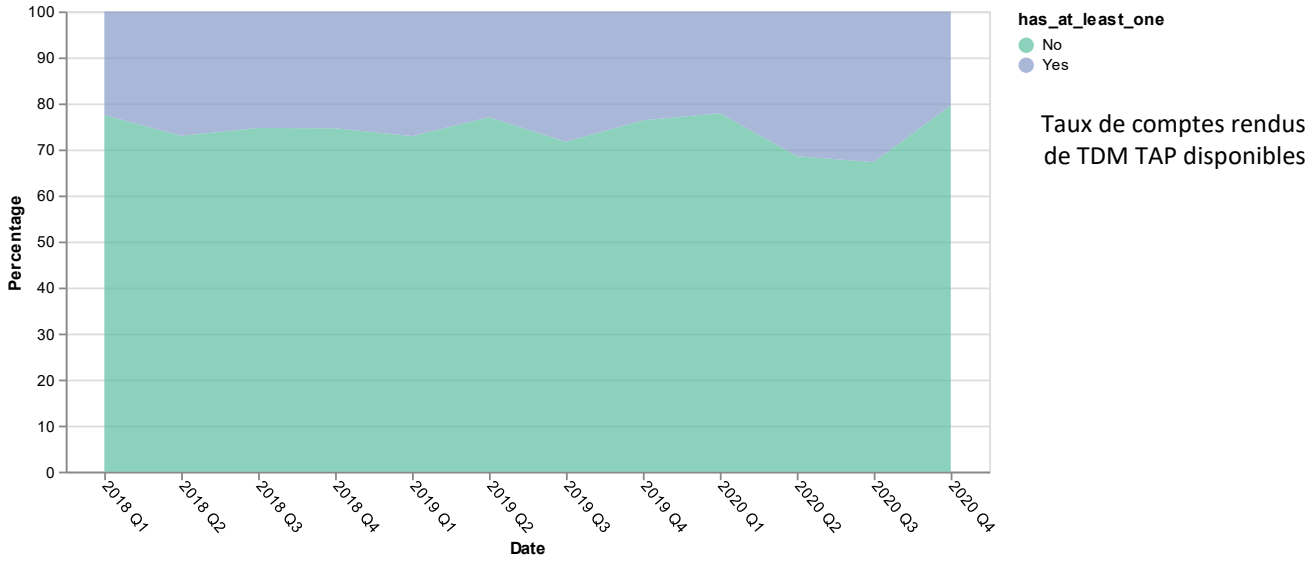


Figure 12. Taux de patients nouvellement diagnostiqués d'un cancer colorectal et ayant au moins un compte rendu de TDM TAP disponible dans l'EDS de l'AP-HP

L'algorithme de forêt aléatoire a été associé à un rappel et une précision de 86% et 78% sur le jeu d'entraînement, et de 73% et 89% sur le jeu de test, respectivement. Le score F1 de la forêt aléatoire a atteint 0,82 et 0,73 sur les jeux d'entraînement et de test, respectivement. Le réseau de neurones a été associé à un rappel de 100% et 68% et une précision de 100% et 83% sur les jeux d'entraînement et de test, respectivement.

Le diagramme récapitulatif SHAP a été utilisé pour évaluer l'importance de chaque mot dans la classification des documents. Les mots ayant le poids le plus élevé pour classer un document comme étant intermédiaire ou non intermédiaire étaient *« bilan », « extension », « recherche », « réévaluation* » et « *masse »* (Figure 13).

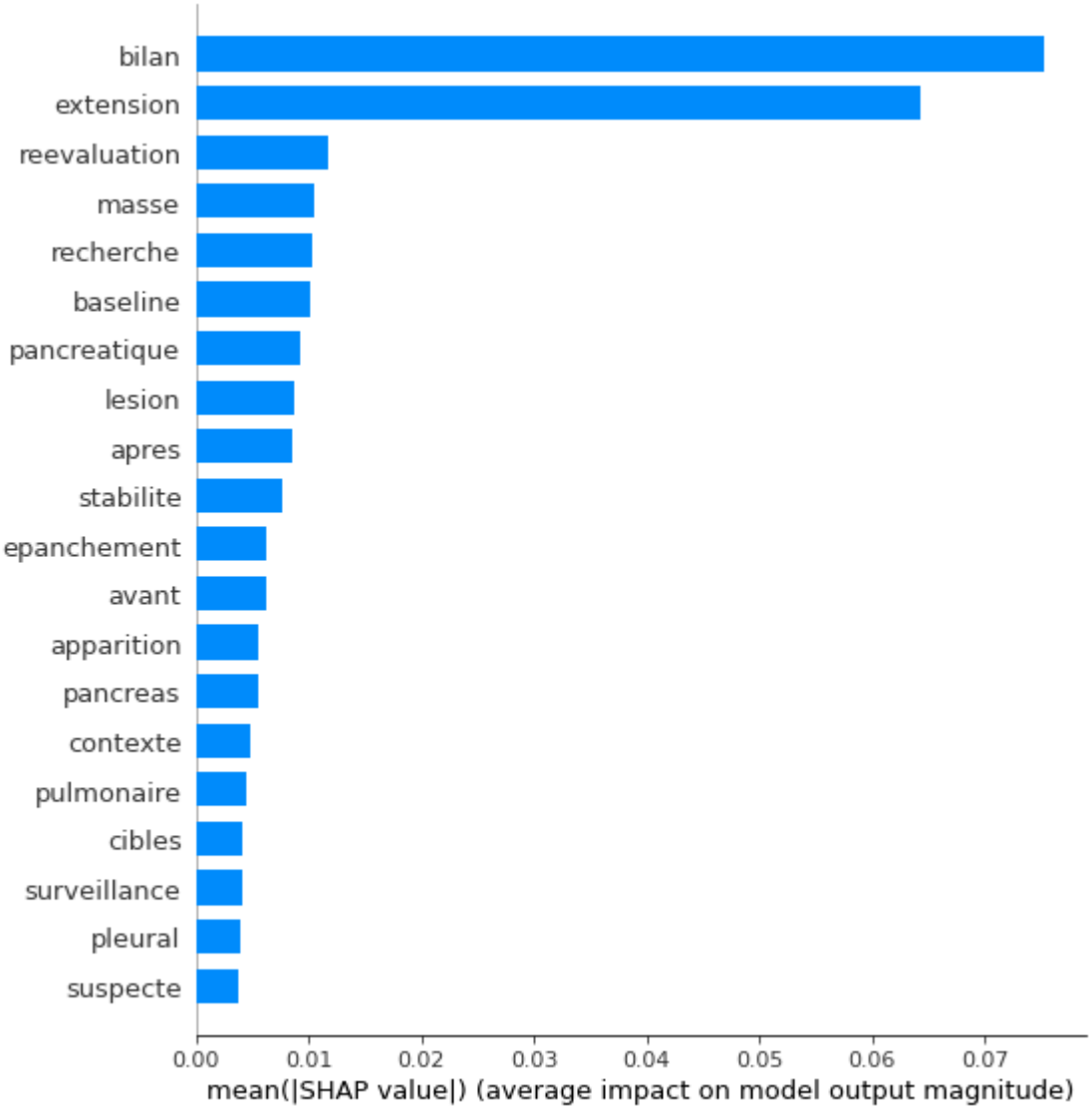


Figure 13. Diagramme récapitulatif SHAP d'évaluation de la pondération croissante des mots utilisés dans l'algorithme de classification des TDM TAP pour le cancer colorectal

Concernant le cancer du pancréas, 100 comptes rendus textuels de TDM TAP ont été annotés manuellement pour constituer le jeu de test. Concernant le cancer du sein, 280 comptes rendus textuels de TEP-scanner ont été annotés manuellement et divisés aléatoirement en un jeu d'apprentissage (65%, soit 182 comptes rendus) et un jeu de test (35%, soit 98 comptes rendus). Les métriques de performances sont synthétisées dans le Tableau 4.

Tableau 4. Performances des modèles de classification de stadification tumorale initiale des imageries sur les différents jeux de tests pour les patients nouvellement référées à l'AP-HP avec un cancer du pancréas et du sein

Modèle	Rappel	Précision	Score F1
Cancer du pancréas (TDM TAP)			
Forêt aléatoire	0.60	0.83	0.70
Cancer du sein (PET-scanner)			
Classifieur à renforcement de gradient	0.89	0.97	0.93
Forêt aléatoire	0.97	0.82	0.89
Réseau de neurones convolutif	0.92	0.86	0.89

- Identification du stade métastatique au sein des comptes rendus d'imagerie de stadification tumorale initiale

A partir de 252 TDM TAP et TEP-scanner de stadification initiale annotés manuellement, dont 70 mentionnaient la présence de métastases, des codes CIM-10 C78 et/ou C79 furent retrouvés chez 53 patients d'entre eux (75%), autour de la date de diagnostic de cancer. Le Tableau 5 résume les métriques de performances concernant l'identification du stade métastatique par expression régulière à partir des comptes rendus de TDM TAP de stadification tumorale initiale, et ce, sur les différents jeux de données.

Tableau 5. Métriques de performance sur les différents jeux de données concernant l'identification du stade métastatique par expression régulière à partir des comptes rendus de TDM TAP de stadification tumorale initiale

Type de cancer	Jeu de développement			Jeu de test		
	Rappel	Précision	Score F1	Rappel	Précision	Score F1
Colorectal	78%	86%	0.82	73%	73%	0.75
Pancréatique		Non évalué		88%	83%	0.85
Mammaire	74%	95%	0.83	87%	77%	0.82

Les performances de concaténation des deux algorithmes (stadification tumorale initiale et mention d'une métastase dans des comptes rendus textuels de TEP-scanner) ont été évaluées et synthétisées dans la Figure 14.

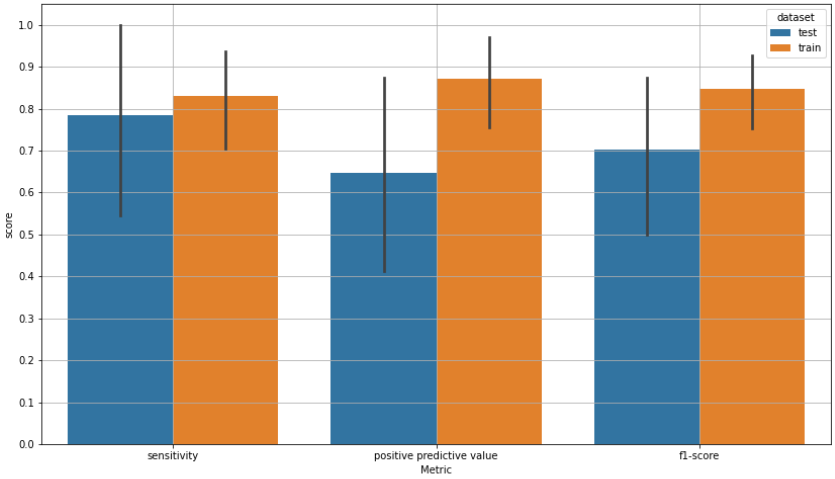


Figure 14. Métriques de performances sur le jeu de test des algorithmes concaténés (algorithme de stadification tumorale initiale et mention de métastase) concernant les comptes rendus de TEP-scanner réalisés autour de la date de diagnostic des cas de cancer du sein nouvellement référés à l'AP-HP

c. Portabilité

Au sein du projet CovOnco, l'utilisation de ces algorithmes a permis d'objectiver l'absence d'aggravation de dissémination tumorale en lien avec les éventuels retards diagnostiques au cours du début de la pandémie de SARS-CoV2. Le Tableau 6 résume le stade tumoral métastatique initial de 782 cas de cancer colorectal, selon l'année du diagnostic de cancer. Aucune différence statistique n'a été observée entre les deux classes de risque.

Tableau 6. Stade tumoral au diagnostic initial de 782 cas de cancer colorectal, selon l'année de diagnostic, à partir des comptes rendus de TDM TAP de stadification tumorale initiale

Année du diagnostic	2018	2019	2020
Cancers du côlon	N=165	N=166	N=145
Stade métastatique N (%)	43 (26%)	62 (37%)	47 (33%)
Valeur de *p* par rapport à la moyenne de 2018-2019			*0.97*
Cancers du rectum	N= 42	N= 40	N=31
Stade métastatique N (%)	11 (26%)	8 (20%)	7 (23%)
Valeur de *p* par rapport à la moyenne de 2018-2019			*0.85*

Des résultats similaires ont été obtenus pour les cancers du pancréas, du poumon et du sein. Le détail des analyses épidémiologiques issues du développement algorithmique est disponible au sein de l'Annexe 5.

III. Identification des critères pronostiques histologiques

Cette partie de thèse a cherché à extraire automatiquement l'information médicale nécessaire à la pronostication histologique des tumeurs réséquées, selon les variables suivantes issues des comptes rendus d'anatomopathologie de l'AP-HP :

- Taille (tumorale) ;
- Différenciation (tumorale) ;
- Invasion vasculaire (tumorale) ;
- Invasion ganglionnaire (tumorale) ;
- Complétude de la résection microscopique ;
- Invasion périnerveuse tumorale ;
- Score (y)pTNM,
- Taille tumorale.

L'objectif méthodologique sous-jacent fut de discuter les modalités d'extraction textuelle, dans un contexte d'optimisation des ressources d'annotations.

1. Méthodes

La population incluse dans le cas d'usage 'Challenge AI for health' a été constituée de patients nouvellement référés à l'AP-HP avec un cholangiocarcinome intra-hépatique opéré (Annexe 4). La population d'intérêt fut identifiée par la présence d'un code CIM-10 C221 (diagnostic principal ou relié), sans avoir été enregistré au cours des deux années précédentes dans l'EDS de l'AP-HP. L'intervention chirurgicale d'exérèse du cholangiocarcinome entre janvier 2015 et décembre 2020 fut identifiée par les codes CCAM listés dans l'Annexe 3c, pour les patients décédés ou revenus à l'AP-HP après juillet 2017. Le compte rendu d'anatomopathologie postopératoire a été identifié manuellement par un oncologue médical (EK) parmi les comptes rendus d'anatomopathologie présents dans le dossier patient électronique jusqu'à deux mois après la date de chirurgie. Les cas de double tumeur primitive ont été exclus, tels que la présence concomitante d'un cholangiocarcinome extra-hépatique, ainsi que les patients opérés après traitement systémique néoadjuvant. L'ensemble des comptes rendus textuels a été divisé en jeux de développement et de test de taille égale.

L'extraction textuelle a ciblé des critères histopronostiques, c'est-à-dire des entités qui sont identifiés par analyse microscopique des pièces opératoires par les anatomopathologistes, et dont la présence grève les chances de survie du patient opéré (249).

Ici, étaient concernées les entités suivantes :

- Invasion ganglionnaire (tumorale) (présence / absence), qui indiquent l'éventuelle présence de cellules tumorales dans les ganglions autour de l'organe porteur d'une tumeur et réséqués dans le même temps opératoire (« adénopathies ») ;
- Invasion vasculaire (tumorale) (présence / absence), qui indique l'éventuelle présence de cellules tumorales dans les microvaisseaux veineux qui vascularisent l'organe porteur de la tumeur ;
- Complétude de la résection microscopique (présence / absence), qui indique dans quelle mesure il existe du tissu sain (exempt de tumeur) entre la marge de résection chirurgicale et le front de développement de la tumeur au sein de l'organe opéré ;
- Invasion périnerveuse (tumorale) (présence / absence), qui indique l'éventuelle présence de cellules tumorales autour des filets nerveux qui innervent l'organe porteur de la tumeur ;
- Score (y)pTNM ((y)p T1-4 N0-3 M0-1), qui synthétise numériquement l'extension géographique de la tumeur selon « T » pour taille de la « tumeur », « N » pour l'éventuelle extension tumorale dans les ganglions « nodes », « M » pour l'éventuelle extension tumorale dans les « métastases » ;
- Taille (tumorale), qui détermine en millimètres ou centimètres le plus grand diamètre de la tumeur réséquée ;
- Différenciation (tumorale) (peu, moyennement, bien), qui indique dans quelle mesure les cellules tumorales ont conservé les caractéristiques phénotypiques de l'organe primitif initial (tumeur bien différenciée) ou non (tumeur peu différenciée) lorsque que l'organe primitif initial n'est pas reconnaissable.

Ces critères histopronostiques sont génériques en oncologie et leur évaluation est commune à tout type de cancer solide réséqué, tel que recommandé par l'INCa et l'ANS (176,177). Les mentions pouvant être multiples dans un compte rendu, un algorithme de post-traitement des entités extraites par expressions régulières a été développé afin d'enrichir un jeu de données de caractérisation à l'échelle du document, de telle sorte que, pour une même entité, l'annotation associée au moindre pronostic soit retenue (Annexe 6). La négation a été prise en charge par des règles de post-traitement dédiées au projet. Sur les jeux de développement et de test, les paramètres de performance suivants ont été évalués : rappel, précision, score F1. L'annotation des 290 comptes rendus divisés à parts égales en jeux de développement et de test a été initialement réalisée par une interne en médecine. Le guide d'annotations est disponible dans l'Annexe 7.

Puis, une oncologue médicale (EK) a corrigé l'annotation de tout le corpus de textes. La validité externe des performances de l'algorithme d'extraction à base de règles fut évaluée sur un jeu de données dédiées au cancer du sein, dans le contexte du cas d'usage consacré à l'automatisation de calcul des indicateurs qualité de sécurité et des soins EUSOMA (17). D'autres expressions régulières spécifiques aux critères du cancer du sein furent développées pour ce projet et ne sont pas présentées en détail dans le présent mémoire de thèse :

- Statut des récepteurs aux œstrogènes ;
- Statut HER2 ;
- Grade tumoral ;
- Malignité du prélèvement étudié ;
- Type histologique du cancer du sein (253).

L’approche d’extraction initialement choisie fut celle d’une pré-annotation automatique des comptes rendus cibles par un système de règles, en vue d’un éventuel complément d’annotation manuelle pour entraîner, dans un second temps, un algorithme supervisé en apprentissage machine responsable de l’extraction textuelle. Cette option méthodologique d’extraction textuelle est illustrée par la Figure 15. Un binôme constitué d’un expert métier et d’un scientifique des données développe des règles permettant la pré-annotation des différentes entités. Cette pré-annotation repose sur la combinaison de différents critères de formalisation :

- Terminologies où sont formalisées les expressions régulières ;
- Schémas où sont formalisés les différents types d'entités à extraire et les éventuels liens hiérarchiques entre ces types ;
- Gestion de la négation ;
- Règles où est formalisée l’imbrication des critères ci-dessus.

Chaque itération de cette pré-annotation permet le croisement entre les annotations réalisées par l’expert sur le jeu de développement, et celles réalisées par les règles. Chaque itération produit ainsi un « cycle d’essais erreurs » au terme duquel l’analyse des faux positifs et faux négatifs de la pré-annotation automatique permet la modification des différents fichiers, en vue d’une nouvelle itération. L’option méthodologique présuppose que le binôme interrompe ces cycles d’itérations quand le développement des règles est estimé optimisé, afin de trancher en faveur de la meilleure modalité d’extraction textuelle : système à base de règles, ou mutation vers un système statistique nécessitant une annotation manuelle pour l’entraînement supervisé correspondant.

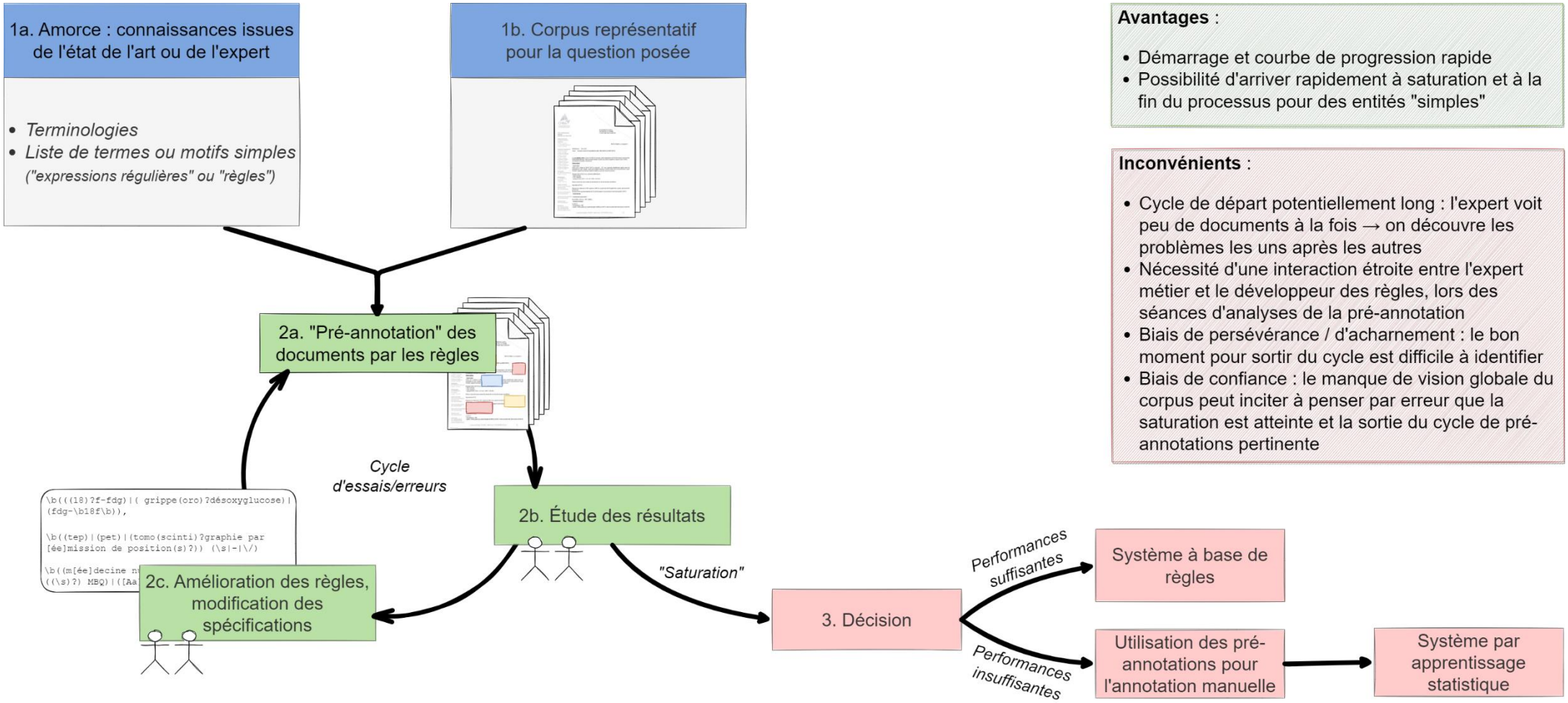


Figure 15. Approche d'extraction textuelle initialement choisie concernant les facteurs histopronostiques des cancers opérés, dans le cadre du cas d'usage du 'Challenge AI for health' pour le cholangiocarcinome : étape initiale de pré-annotation automatique des comptes rendus cibles par un système de règles, en vue d'une éventuelle annotation manuelle pour entraîner un algorithme supervisé en apprentissage machine

2. Résultats

a. Performances d'extraction textuelle obtenues concernant les entités correspondant aux critères histopronostiques

Le diagramme de flux rapportant la constitution du jeu de données du cas d'usage du 'Challenge AI for health' est illustré par la Figure 16.

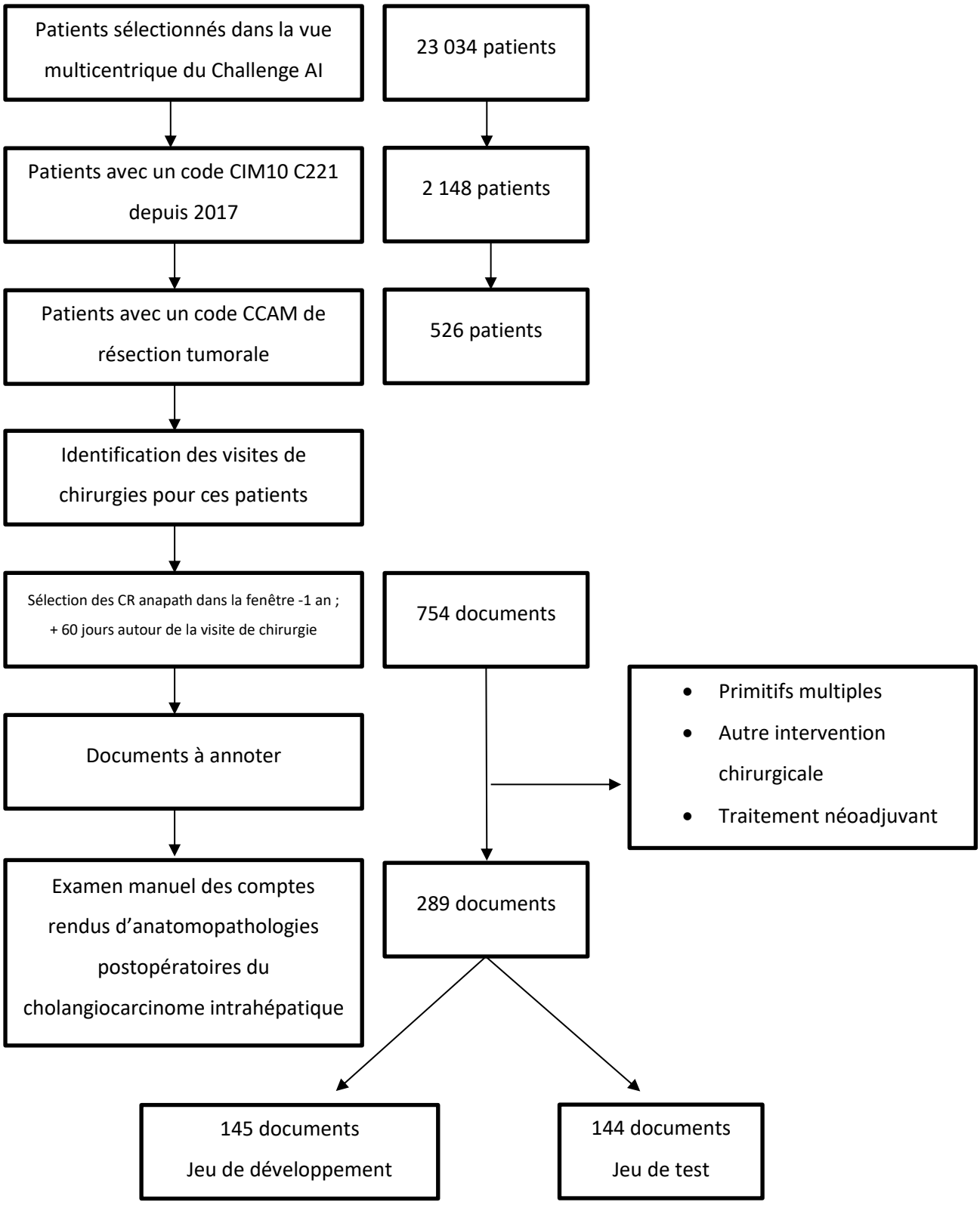


Figure 16. Diagramme de flux de sélection des jeux de données pour le cas d'usage du 'Challenge AI for health'

Les règles ont été élaborées entre avril 2021 et janvier 2022 en collaboration avec une scientifique des données, Sonia Priou, lors de 9 itérations successives impliquant l'analyse de 108 documents visualisés par l'outil d'annotation *brat*, telle qu'illustré par la Figure 17 (260).

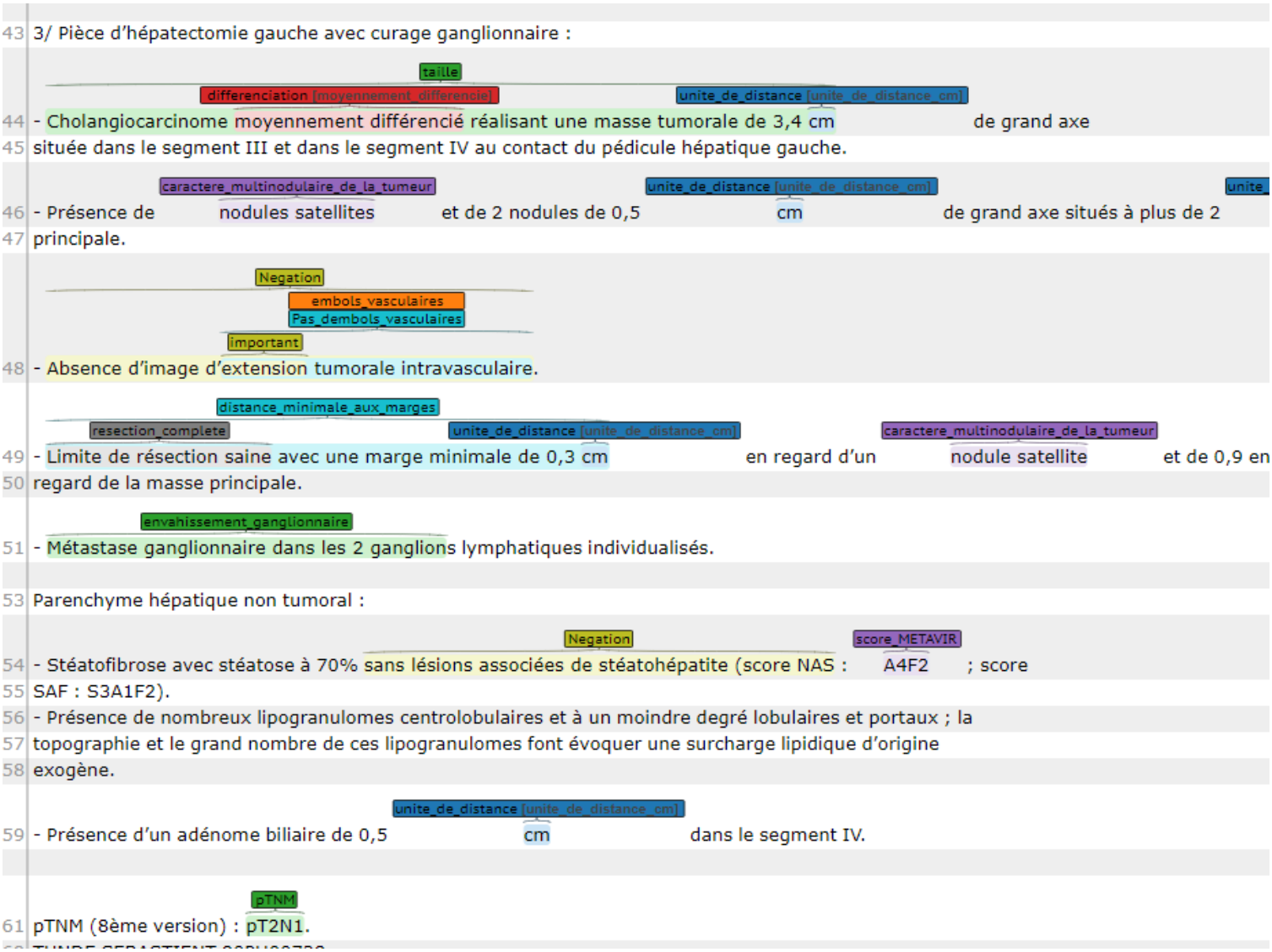


Figure 17. Exemple de visualisation par l'outil BRAT des règles développées en vue de la pré-annotation automatique des critères histopronostiques issus des comptes rendus anatomopathologiques post-opératoires des cholangiocarcinomes opérés à l'AP-HP, dans le cadre du cas d'usage 'Challenge AI for health'

Chaque itération prenait fin lorsque les développeuses pensaient les règles optimisées au regard des annotations et documents analysés lors des séances de travail correspondantes (une-demie à trois journées). La confrontation au jeu de développement annoté permettait à chaque itération d'analyser les annotations résultant en faux positifs et faux négatifs, de manière à optimiser les règles lors de l'itération suivante. Des expressions régulières ont ainsi été élaborées pour chacune des entités d'intérêt, ainsi qu'un ensemble de règles de pré- et post-traitement prenant en charge les négations et priorisant les zones de texte en vue de l'extraction (Annexe 6). La caractérisation de chaque entité a été faite à l'échelle du document,

de telle sorte que, pour une même entité associée à plusieurs annotations au sein d'un même compte rendu, l'annotation associée au moindre pronostic a été retenue (Annexe 6). Par exemple, l'entité 'score pTNM' a été déclinée en quatre sous-entités 'indice' (p, yp), 'T', 'N', 'M' puis synthétisée à l'échelle du document lors de la phase de post-traitement.

Malgré un nombre multiple d'itérations réalisées et d'analyses de documents lors de séances de travail interactif entre SP et EK impliquant 24 versions de la combinaison des quatre fichiers de règles, terminologies, négation et schéma, et ce, sur une période de neuf mois, le niveau des métriques de performances d'extraction n'a jamais été jugé comme atteint pour être en mesure de trancher entre les deux options méthodologiques à suivre : extraction à partir des règles développées, ou nécessité d'un complément manuel pour entraîner un algorithme d'apprentissage machine d'extraction manuelle. L'Annexe 8 retrace les différentes combinaisons de versions des critères de formalisation (schémas, terminologies, négations) du développement des expressions régulières au long des itérations (cycle essais / erreurs de la Figure 15). Ainsi, l'approche initiale d'entraînement d'un modèle d'apprentissage machine à partir d'une pré-annotation automatique fondée sur un modèle de règles puis suivie d'une annotation manuelle fut abandonnée.

Plusieurs difficultés avaient, en effet, été rencontrées. Dès la première itération, la confrontation à trois documents révéla que l'entité 'caractère complet de la résection microscopique' était le lieu d'une hétérogénéité d'explicitation linguistique majeure, ne permettant pas sa formalisation sous forme d'un ensemble de règles homogènes. Par exemple, rares étaient les mentions explicites de « caractère incomplet de la résection tumorale » par le chirurgien. La mention d'une « tranche de résection affleurant/jouxtant avec la marge tumorale » étaient plus fréquentes, tout comme la mention objective du nombre de millimètres séparant la limite de résection de celle de la tumeur (voir guide d'annotations de l'Annexe 7). Décision fut prise de diviser cette entité en quatre entités filles homogènes sur un plan formel, et synonymes sur un plan nosographique : 'complétude de la résection microscopique', 'atteinte tumorale de la marge' et 'distance minimale aux marges', elle-même déclinée en évaluation quantitative (nombre de millimètres) et en évaluation qualitative binaire ('marge de résection inframillimétrique').

Ont également représenté des obstacles majeurs les difficultés de modélisation de l'entité 'taille tumorale' et 'distance minimale aux marges' - elle-même constituant un sous-ensemble de l'entité initialement définie 'caractère complet de la résection microscopique'. A la septième itération, est apparu que le jeu de développement annoté par l'interne en médecine contenait des erreurs en lien avec le flou nosographique existant au sein de cette entité mère, et saillant à l'heure de son explicitation textuelle (261). Les erreurs d'annotations du jeu de développement constitués de 145 comptes rendus ont alors été corrigés à deux reprises par EK, avant que l'option méthodologique initiale ne soit jugée caduque.

Deux questions n'ont pas pu être résolues lors de la campagne de développement des règles pour la pré-annotation :

- Comment savoir, en effet, si la « saturation » des itérations de pré-annotations avait été atteinte, c'est-à-dire que les métriques de performances d'extraction textuelle avaient été optimisées par les règles ?

- Comment, par ailleurs, décider de ne concentrer que sur certaines entités à métriques médiocres l'effort d'une annotation manuelle et un entraînement de modèle statistique, alors que l'ensemble de l'annotation du jeu de test pouvait contenir des erreurs ?

A ce stade, l'apprentissage machine a dorénavant été envisagé comme un système compétitif des règles, plus qu'un système séquentiel d'apprentissage initié par une pré-annotation automatique par règles. Les 290 comptes rendus du corpus de texte (jeux de développement et de test) furent alors entièrement annotés par EK, selon quatre versions successivement améliorées du guide d'annotations disponible dans l'Annexe 7 (262). Cette étape permit la comparaison des performances des systèmes à base de règles et d'apprentissage machine afin d'évaluer la plus-value d'un modèle d'apprentissage supervisé à partir d'une annotation manuelle en termes de performances d'extraction. Les Tableaux 8 et 9 reproduisent les performances atteintes par chacune des deux méthodes d'extraction sur les jeux de développement et de test, respectivement. Sur le jeu de test, pour toutes les entités, les différences de métriques de performances n'ont jamais dépassé 5%, à l'exception de la précision d'extraction de l'entité 'taille de la tumeur'.

En parallèle, la validité externe des règles a été évaluée sur un corpus de textes issus du cas d'usage lié au calcul automatique des indicateurs qualité EUSOMA concernant le cancer du sein (Annexe 1b) (253). Les performances d'extraction en rapport avec les entités communes au 'Challenge AI for health' sont synthétisées dans le Tableau 7, illustrant les médiocres métriques pondérées associées à l'entité 'taille de la tumeur et mauvaises pour l'entité 'distance aux marges'.

Tableau 7. Performance sur un jeu de test des algorithmes d'extraction par expression régulière des entités correspondant à certains critères pronostiques histologiques nécessaires au calcul des indicateurs qualité EUSOMA pour le cancer du sein

Entité	Disponibilité dans les comptes rendus du jeu de test (n=48)	Précision	Rappel	F1-score
Score pTNM	45 (93,8%)	91,9	93,3	92,1
Taille	47 (97,9%)	49,0	50,0	47,0
Invasion vasculaire	42 (87,5%)	86,0	75,0	77,8
Distance minimale à la marge de résection	36 (75,0%)	14,1	27,1	15,1

Tableau 8. Métriques de performance d'extraction relative pour différents critères pronostiques histologiques obtenues par des algorithmes à base de règles et d'apprentissage machine sur le jeu de développement du 'Challenge AI for health'

	Taille	Différenciation	Invasion vasculaire	Invasion périnerveuse	Invasion ganglionnaire	pTNM (indice)	pTNM (T)	pTNM (N)	pTNM (M)	Complétudee de la résection microscopique	Distance minimale aux marges	Atteinte tumorale de la marge
Règles												
Précision %	89,13	95,07	96,3	99,3	96,45	97,9	97,2	98,59	100	100	89,36	77,78
Rappel %	96,09	99,26	94,2	100	98,55	100	100	99,29	100	96,5	98,44	92,92
Apprentissage machine												
Précision %	97,10	95,77	98,52	100	97,90	100	99,30	100	100	98,60	95,80	87,94
Rappel %	96,40	99,27	94,33	99,30	100	100	100	99,30	99,30	100	100	98,41
Différence entre l'apprentissage automatique et les règles												
Précision %	7,97	0,70	2,22	0,70	1,45	2,10	2,10	1,41	0	-1,40	6,44	10,16
Rappel %	0,31	0,01	0,13	-0,70	1,45	0	0	0,01	-0,70	3,50	1,56	5,49

Tableau 9. Métriques de performance d'extraction relative pour différents critères pronostiques histologiques obtenues par des algorithmes à base de règles et d'apprentissage machine sur le jeu de test du 'Challenge AI for health'

	Taille	Différenciation	Invasion vasculaire	Invasion périnerveuse	Invasion ganglionnaire	pTNM (indice)	pTNM (T)	pTNM (N)	pTNM (M)	Complétude de la résection microscopique	Distance minimale aux marges	Atteinte tumorale de la marge
Règles												
Précision %	87,22	98,59	97,76	97,81	97,06	98,6	99,3	99,29	88,11	99,29	80,58	88,72
Rappel %	92,06	99,29	93,57	95,71	94,96	100	99,3	97,89	100	97,89	96,55	92,19
Apprentissage automatique												
Précision %	94,25	99,29	96,29	99,26	96,40	99,30	100	100	89,36	97,08	79,43	86,86
Rappel %	97,03	98,59	94,21	94,37	97,10	99,30	100	98,60	98,44	95,68	98,24	95,20
Différence entre l'apprentissage automatique et les règles												
Précision %	**7,03**	0,70	-1,47	1,45	-0,66	0,70	0,70	0,71	1,25	-2,21	-1,15	-1,86
Rappel %	4,97	-0,70	0,64	-1,34	2,14	-0,70	0,70	0,71	-1,56	-2,21	1,69	3,01

a. Analyse des modalités de développement des règles, permettant l'extraction textuelle des entités correspondants aux critères histopronostiques

Compte tenu d'un niveau relativement similaire des métriques de performances d'extractions des entités obtenues par les deux types de méthodes, s'est posée la question du caractère anticipé et préalable du choix entre ces deux méthodes par type d'entité, plutôt que l'utilisation du système des règles précédant ce choix (Figure 15).

Afin d'informer la pertinence de choix entre l'une et l'autre méthode d'extraction textuelle, les modalités et les ressources nécessaires au développement des règles ont été établies à l'échelle de chaque entité. Ainsi, le nombre de documents, d'entités et d'annotations analysés, ainsi que les efforts nécessaires à l'obtention de ces performances ont été évalués, en fonction du temps, pour chaque entité et pour chaque itération. L'Annexe 8 présente les différentes versions des règles, schémas, terminologies et négations constitutives de chaque itération.

Lors du développement des règles, 108 documents issus du jeu de développement ont été analysés manuellement, avec une hétérogénéité importante du nombre de documents analysés par itération (Tableau 10). Plus de la moitié des entités était analysée lors de la première itération, alors que six documents sur 108 comptes rendus ont été visualisés lors de la première moitié des itérations effectuées (itérations 1 à 5). A chaque itération, deux entités ont été analysées en moyenne à la visualisation de chaque document.

Le développement des expressions régulières a nécessité l'analyse d'un nombre d'annotations (mentions textuelles d'une entité) et de documents très différent en fonction de l'entité considérée, tel que synthétisé dans le Tableau 11. 'Taille tumorale' et 'distance minimale à la marge de résection' ont nécessité l'analyse de 74 et 81 annotations, soit deux à trois fois plus que la moyenne observée pour l'ensemble des entités (37 ; min 1 - max 81), alors qu''invasion de la marge chirurgicale' et 'caractère complet de la résection microscopique' ont nécessité 1 et 5 analyses d'annotations différentes, respectivement. Les proportions étaient similaires pour le nombre total de documents analysés pour chaque entité.

Entre les entités du cas d'usage, il existait une hétérogénéité forte des modalités et ressources nécessaires au développement des expressions régulières correspondantes d'une part, et également une hétérogénéité des métriques de performances associées au jeu de test du cancer du sein (validité externe). Afin de préciser les causes de cette hétérogénéité, l'évolution des métriques de performances d'extraction a été évaluée au fur et à mesure de la modification des expressions régulières, et ce, en fonction de la version de terminologie utilisée lors des itérations effectuées sur le jeu de développement (Figure 18). Cette analyse révéla que pour la majorité des entités à extraire une itération unique permettait l'optimisation significative des performances d'extraction. En revanche, 'taille tumorale' fut associée à une aggravation des performances lors des itérations initiales ainsi qu'à une multitude de tentatives d'optimisation, alors que les résultats précédant montraient que l'apprentissage machine semblait supérieur aux règles pour son extraction à la fois sur les jeux de développement et de test, et que les métriques relatives aux expressions régulières étaient médiocres sur le jeu de validité externe concernant le cancer du sein. L'entité 'distance aux

marges' était associée à un gain absolu de métriques faible relativement aux autres entités, malgré un nombre d'itérations élevées. Cette même entité était associée à des métriques mauvaises lors de son extraction sur le corpus de validation externe du cancer du sein, et ce, malgré un nombre maximal d'analyse d'annotations correspondantes.

Cette analyse fut répétée en fonction du nombre de documents analysés sur le jeu de développement des expressions régulières, et par rapport au jeu de développement final annoté par EK (Tableau 11 et Figure Supplémentaire 2). Fut révélé, ici, que pour les entités ayant posé des difficultés d'extraction telles que 'taille de la tumeur', ou 'distance minimale aux marges', l'analyse d'environ 40% et 90% du jeu de développement n'a été associée à aucune amélioration des métriques de performances.

Tableau 10. Descriptif des entités et documents analysés par itération et au total, lors du développement des règles d'annotation

	Par séance d'itération			En cumulé	
Itération (n=9)	Nombre d'entités analysées (n=9)	Nombre de documents analysés (n=108)	Nombre moyen d'entités analysées par document (min-max)	Nombre d'entités analysées (n=9)	Nombre de documents analysés (n=108)
1	5	3	2 (1-4)	5	3
2	3	3	2 (1-5)	5	3
3	3	1	2 (1-3)	5	4
4	6	5	2 (1-4)	7	5
5	5	3	3 (1-5)	8	6
6	9	11	2 (1-5)	9	17
7	8	51	2 (1-10)	9	55
8	7	93	2 (1-9)	9	108
9	7	80	2 (1-5)	9	108

Tableau 11. Nombre d'analyses nécessaires par entité correspondant aux biomarqueurs pronostiques histologiques sur le jeu de développement des expressions régulières du cas d'usage du 'Challenge AI for health'

Entité	Somme totale du nombre de documents analysés à chaque itération *	Nombre de documents analysés au moins une fois **	Nombre moyen d'analyses d'un même document par entité
Taille	74	40	1.8
Différenciation	33	18	1.8
Invasion vasculaire	30	20	1.5
Invasion périnerveuse	37	20	1.9
Invasion ganglionnaire	45	33	1.4
Score pTNM	29	20	1.5
Complétude de la résection microscopique	5	5	1
Distance minimale aux marges	117	55	1.5
Atteinte tumorale de la marge	1	1	1
Moyenne (min-max)	37 (1-117)	24 (1-55)	1.5 (1-1.9)

** Un même document analysé lors de plusieurs itérations pour la même entité compte pour autant de fois dans la colonne (Figure 18)*
*** Un même document analysé lors de plusieurs itérations pour la même entité n'est compté qu'une fois dans la colonne*

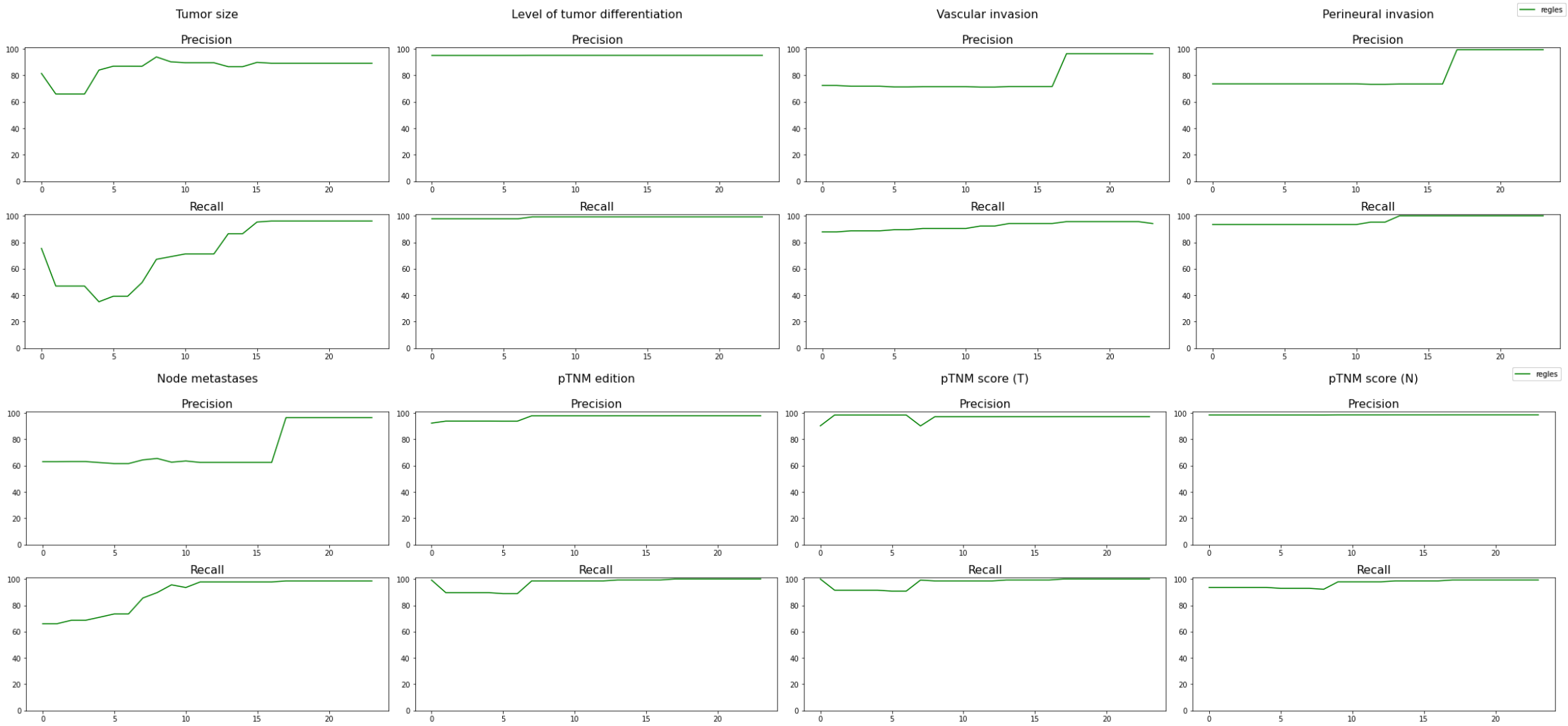


Figure 18. Métriques de performances d'extractions à base de règles en fonction de la version du fichier de terminologie utilisée sur le jeu de développement, par type d'entités correspondant aux critères pronostiques histologiques des cancers opérés du cas d'usage 'Challenge AI for health', et par rapport à l'annotation finale du jeu de développement

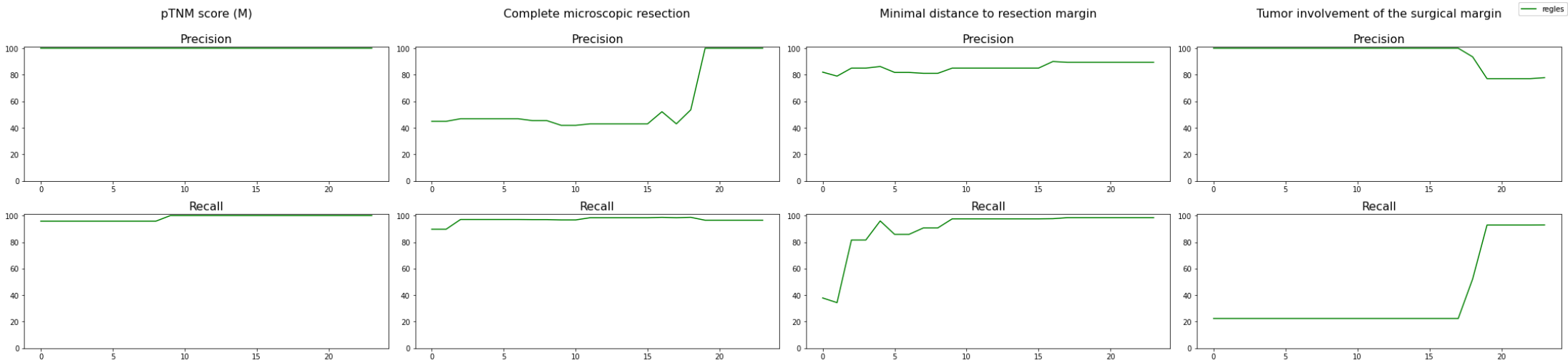


Figure 18 (suite). Métriques de performances d'extractions à base de règles en fonction de la version du fichier de terminologie utilisée sur le jeu de développement, par type d'entités correspondant aux critères pronostiques histologiques des cancers opérés du cas d'usage 'Challenge AI for health', et par rapport à l'annotation finale du jeu de développement

b. *Analyse de l'impact de différentes versions d'annotations sur le développement des règles, permettant l'extraction textuelle des entités correspondants aux critères histopronostiques*

La précédente analyse objectiva que la décision de corriger le jeu de développement survint après l'analyse d'une centaine de comptes rendus concernant l'annotation de l'entité 'distance minimale aux marges'. A titre d'analyses de sensibilité, l'évolution des métriques de performances d'extraction fut évaluée en fonction du type de jeu de développement utilisé comme référence : les trois versions du jeu annoté par l'interne, puis le final annoté par EK. Fut révélé que l'expertise de l'annotateur était associée à une optimisation significative des métriques de performances d'extraction pour les entités filles de 'complétude de la résection chirurgicale'. Ainsi, en substituant les annotations de l'interne par celles d'EK, le rappel associé à cette entité augmenta d'environ 60%, alors que les précisions de 'distance minimale aux marges' et de 'marge de résection envahie' augmentèrent de 30% et 50%, respectivement (Figure Supplémentaire 3).

Portabilité

La tâche d'extraction textuelle précédemment décrite permit la réalisation du 'Challenge AI for health' par neuf candidats dont six ont proposé un algorithme de vision par ordinateur au jury. Les métriques de performance obtenues par les tâches de régression et classification en lien avec les prédictions des différents facteurs pronostiques histologiques sont résumées dans l'Annexe 9, pour chacun des six candidats ayant présenté leurs travaux au jury du 'Challenge AI for health'. Le lauréat devient est ainsi devenue le partenaire du projet « Identification de facteurs pronostiques par outils d'intelligence artificielle appliqués à l'imagerie dans le cholangiocarcinome intra-hépatique (CHOTERIA) » avec l'AP-HP (263). Le projet a été validé par le CSE de l'EDS le 5 septembre 2022 (CSE-22-0016), et l'espace de travail Jupyter dédié a été ouvert le 10 mars 2023. Il est censé se décliner selon les quatre objectifs suivants :

1. Prédiction de critères radiologiques et anatomo-pathologiques
2. Prédiction du pronostic des patients (survie sans rechute, survie globale)
3. Prédiction du profil RNA seq identifié sur les pièces opératoires pour les patients réséqués
4. Prédiction du profil des anomalies moléculaires identifiées sur les pièces opératoires par séquençage ADN pour les patients réséqués.

Cette tâche d'extraction a également permis la réalisation d'une étude de faisabilité de l'automatisation du calcul des IQSS EUSOMA à partir de l'EDS de l'AP-HP a fait l'objet d'une présentation orale au congrès annuel EMOIS en mars 2023 et d'un article original en anglais en cours de révision par la *Revue d'Epidémiologie et de Santé Publique* (253). Globalement, parmi 5 785 patientes diagnostiquées d'un cancer du sein, 5 147 (89 %) avaient des actes liés au cancer enregistrés dans le PMSI, dont 3 732 (73 %) un acte de résection tumorale. Des 34 indicateurs EUSOMA cibles, 9 étaient calculables avec le PMSI seul, et 6 autres le devenaient en utilisant les données présentes dans les comptes rendus d'anatomopathologie. Dix variables élémentaires étaient nécessaires au calcul des 6 indicateurs combinant PMSI et comptes-rendus d'anatomopathologie. Les comptes rendus nécessaires étaient disponibles pour 58,8% à 94,6% des patients, suivant les indicateurs. Une fois les algorithmes d'extraction

textuelle appliqués, les variables nécessaires au calcul des indicateurs étaient possibles à extraire pour 2% à 88% des patients, suivant les indicateurs (Annexe 1b).

IV. Discussion

1. Synthèse

A partir des comptes rendus médicaux contenus dans l'EDS de l'AP-HP, l'extraction textuelle d'items issus du jeu de données minimales en oncologie non structurés ou sous-codés par le PMSI, tels que l'évaluation du stade tumoral au diagnostic et la caractérisation histopronostique des tumeurs, a été associée à des métriques de performance compatibles avec son enrichissement et son utilisation dans des cas d'application variés tels que des études épidémiologiques, de pilotage, ou de vision par ordinateur. La pré-annotation automatique à base de règles initiale en vue de l'éventuel entraînement d'un algorithme d'apprentissage machine d'extraction textuelle représente une option méthodologique qui achoppa sur l'hétérogénéité importante des efforts à mobiliser pour développer les expressions régulières en fonction de la typologie de chaque entité de ce travail. L'utilisation de règles semblait une méthode d'extraction optimale quand certaines conditions linguistiques textuelles de l'évolution du fichier de terminologie utilisé étaient réunies pour une entité donnée, alors que le recours à l'apprentissage machine permettait d'atteindre des métriques de performances d'extraction intéressantes. L'apprentissage supervisé constituait ainsi une option d'extraction textuelle performante, particulièrement lorsque les règles étaient associées à des difficultés d'élaboration inhérentes à l'entité, mais au prix d'une annotation manuelle non nulle. Le rationnel sous-jacent à l'anticipation du choix entre ces deux options méthodologiques d'extraction textuelle demeurait suffisamment peu informé.

2. Mise en perspective avec la littérature

L'hypothèse méthodologique sous-jacente à la tâche d'extraction automatique des critères pronostiques histologiques consistait en une pré-annotation automatique à base de règles, potentiellement complétée par une annotation manuelle pour un apprentissage machine. Le rationnel était de limiter au maximum l'annotation manuelle du corpus de textes cibles, grâce à un système de règles et de terminologies. En effet, les tâches d'annotations manuelles demeurent des obstacles majeurs au développement de modèles d'apprentissage machine, aboutissant à une majorité de corpus de textes annotés de taille moyenne allant de quelques centaines à quelques milliers de documents (154). Plusieurs stratégies ont été développées pour faciliter ces tâches qui semblent fastidieuses. Parmi les nombreux outils d'annotation disponibles pour les utilisateurs, certains semblent se démarquer en termes d'optimisation du

temps et de l'interopérabilité des tâches d'annotations, p. ex. dans le cas des comptes rendus d'anatomopathologie dans le domaine de l'oncologie (264,265).

Pour contrecarrer le défi de la petite taille des corpus annotés manuellement, plusieurs techniques de pré-annotation automatique ont été développées (154). La pré-annotation automatique sémantique est une approche qui a démontré une optimisation du temps et de l'effort consacré à la tâche d'annotation des jeux de données de santé (266). Les techniques utilisant un faible volume de données annotées représentent des alternatives au caractère fastidieux et chronophage de tâches d'annotations sur de larges corpus de textes (267,268). Une méthode fondée sur l'apprentissage actif permet un entraînement interactif en utilisant les corrections réalisées sur la pré-annotation automatique par l'annotateur, en vue de la pré-annotation séquentielle du reste du corpus de textes (269). Similairement, l'usage de codes cliniques pourrait permettre une annotation automatique des corpus dans le contexte d'entraînement supervisé, *modulo* la qualité de l'alignement entre les codes utilisés et les concepts médicaux ainsi que l'inhérent problème du sous-codage (270,271). Enfin, la supervision distante à partir d'une base de connaissances existante semble une option retenue dans la littérature pour annoter automatiquement des corpus de textes de santé (272). L'annotation par la foule semble une alternative aux difficultés liées à la tâche d'annotations, mais, au-delà de l'ambiguïté inhérente au langage naturel, la qualité associée semble encore trop médiocres pour pouvoir servir de corpus d'entraînement ultérieur (273). Quoi qu'il en soit, cette option demeure impropre aux dossiers patients informatisés pour des questions de respect de la confidentialité (274).

Dans le cas d'usage 'Challenge AI for health', les règles ont difficilement pu conduire à une qualité suffisante pour toutes les entités et à l'identification du moment de l'optimisation de leurs performances. La temporalité opportune pour réorienter la méthode et débuter une phase d'annotation manuelle séquentielle en vue de l'entraînement d'un algorithme d'apprentissage machine n'a pas pu être identifiée. Au-delà de la flexibilité cognitive nécessaire au changement de tâches, l'absence de règles pré-établies pour décider d'arrêter le développement des règles et de débuter l'entraînement supervisé a pu expliquer la persistance dans cette tâche, étalée sur une durée de neuf mois de travail (275,276). L'accès à l'explicabilité du système de règles peut expliquer cette persistance, entretenant la croyance que l'itération suivante permettra de finaliser le développement des règles, contrairement aux techniques d'apprentissage machine dont l'explicabilité demeure limitée (277). Alors que les performances du système à base de règles étaient généralement optimisées en un nombre limité d'analyses de documents et d'itérations, certaines entités ont vu se poursuivre la tentative d'optimisation des expressions régulières correspondantes, sans réelle plus-value en termes de métriques de performances d'extraction.

Ensuite, malgré le développement de référentiels, les comptes rendus médicaux souffrent encore beaucoup d'hétérogénéité concernant la transmission de l'information médicale, qui peut être contenue dans le sous-texte. Par exemple, compte tenu d'une part d'un débat scientifique sur la notion, et d'autre part du caractère nécessairement partial et interprétatif des observations médicales, leur explicitation peut devenir une gageure à l'heure du développement d'algorithmes de traitement automatique du langage naturel (261). Ainsi, dans le cas d'usage du 'Challenge AI for health', alors qu'*a priori*, l'entité binaire 'caractère

complet de la résection' ne semblait contenir aucun défi, il a fallu décompenser cette entité en quatre entités filles, dont 'atteinte de la limite de résection'. Cette entité, corollaire transparent sur un plan nosographique de l'entité mère 'caractère complet de la résection microscopique' n'a pourtant été rencontrée qu'une fois dans les 108 documents annotés, et a été assez naturellement associée à des performances meilleures lors du développement de l'algorithme d'apprentissage machine sur le jeu d'entraînement. Or, c'est plus le niveau de complexité sémantique du texte annoté en lien avec la taille de contexte nécessaire à analyser qui conditionne la charge cognitive correspondante, plutôt que sa syntaxe (278). Par ailleurs, les métriques de performances associées aux règles sur le corpus de validité externe du cas d'usage EUSOMA étaient bien trop médiocres pour envisager une généralisabilité des règles correspondantes. Il est ainsi apparu que, en fonction de la caractérisation linguistique des entités, l'approche symbolique pouvait parfaitement suffire, ce qui est concordant avec la littérature. Enfin, leur développement nécessite une interaction étroite entre développeur du modèle et experts du domaine, contrairement aux modèles d'apprentissage machine supervisés pour lesquels l'annotation par l'expert est préalable à l'entraînement.

Dans cet exemple, il semble donc que l'apprentissage machine supervisé puisse tenir toutes ses promesses en termes d'extraction textuelle, *modulo* les limites de la transférabilité de ce type de modèles (139–142,279–282). Néanmoins, les ressources humaines concernant l'annotation manuelle constituent des limites à leur développement et font le lit du développement de techniques d'automatisation de cette tâche chronophage et nécessitant un niveau d'expertise métier substantiel, comme la supervision distante (272). Si l'apprentissage machine est un outil consommateur de ressources en termes de taille de corpus et de volumes d'annotation sur les jeux de données d'entraînement et de test, son empreinte carbone est un défi additionnel qui se révèle de plus en plus pressant (283,284). A l'heure où le gâchis de la recherche clinique en oncologie ne pose plus questions, et où les ressources apparaissent de plus en plus limitées, le concept d'intelligence artificielle durable se développe de plus en plus dans la communauté scientifique (285–288).

Dans ce contexte, des algorithmes d'évaluation de l'empreinte carbone issue des tâches de traitement automatique du langage naturel ont été développés et mis à disposition des scientifiques, certains les intégrant aux métriques d'évaluation et d'autres émettant des recommandations d'optimisation des ressources énergétiques (283,284,289–291). L'entraînement de modèles d'apprentissage machine dits 'à ressource [énergétique] contrainte' semble devenir une voie de recherche à part entière, puisque le recours à l'apprentissage machine pourrait constituer une méthode assez agressive et non durable pour la caractérisation des maladies (292,293). Le modèle d'apprentissage machine utilisé dans le cas d'usage du 'Challenge AI for health' a été *finetuné* à partir de CamemBERT lors de la confrontation de 21 millions de comptes rendus médicaux contenus dans l'EDS de l'AP-HP. Son entraînement a duré deux jours et mobilisé 8 unités de cartes graphiques Tesla V100, soit 10 kgCO2eq de consommation tels qu'évalués par le « Machine Learning Emission calculator » développé par une équipe de l'institut de Montréal pour l'apprentissage d'algorithmes (MILA) (294). La même équipe de l'EDS avait montré que l'entraînement d'un BERT *ex nihilo* était associé à une consommation carbone dix fois plus importante sans plus-value significative en termes d'amélioration des métriques de performances (262). Toutes proportions gardées, le présent travail retrouve également qu'à métriques de performance d'extractions égales, les ressources carbones nécessaires au développement d'algorithmes d'apprentissage machine pourraient être, sinon épargnées, du moins optimisées. En effet,

dans certaines situations médicales, l'usage des règles semble associé à des métriques de performances de pré-annotation textuelle automatique tout à fait satisfaisantes, particulièrement lorsque la terminologie en question est restreinte et homogène formellement comme pour l'identification du stade pTNM et de la présence de métastases (295,296). Les systèmes à base de règles pourraient être associées à des métriques de performances supérieures à des techniques d'apprentissage machine (297).

Pour certaines tâches d'extraction textuelle en santé bien spécifiques, une approche différenciée par typologie d'entités permettrait d'avoir recours à l'apprentissage machine en tant que palliatif performant des règles. Une démarche d'hybridation des modèles à base de règles et d'apprentissage machine pourrait constituer une réponse à ce défi, en vue d'une unique tâche d'extraction textuelle, et ce, différenciée par type d'entités. La Figure 19 illustre la nouvelle option méthodologique proposée par le travail de thèse, qui :

- positionne le choix entre les systèmes à base de règles et à base d'apprentissage machine, préalablement au développement des règles ;
- envisage ce choix de façon différenciée à l'échelle de chaque entité ;
- considère les règles comme une option par défaut, compte tenu d'une moindre empreinte carbone associée.

Pareil recours imbriqué de ces deux types de méthodes a déjà été testé avec succès en oncologie, mais au sein de tâches de traitement automatique du langage naturel distinctes (298). Dans le système proposé, il s'agirait d'identifier, *de façon préalable à la tâche d'annotation*, la complexité et les ressources nécessaires à l'extraction de chaque entité. Une grille d'analyse a été développée en vue de cette caractérisation qui comprend les six dimensions suivantes : discrimination (entre le contenu de ce qui est à annoter de ce qui ne l'est pas), délimitation (des frontières de l'annotation, particulièrement dans le cas d'une entité segmentée en plusieurs mots), expressivité du langage, poids du contexte, ambiguïté, et dimension du label d'annotation (binaire, multiples, etc) (299). Ainsi, un recours différencié à l'échelle de chaque entité aux règles comme système de pré-annotation ou d'annotations, ou à l'apprentissage machine pourrait ainsi être anticipé dès la définition de la tâche d'extraction. L'hypothèse sous-jacente à pareille démarche repose sur la faisabilité d'anticipation et de modélisation initiale de la complexité et des ressources nécessaires au développement de règles, par entité, et préalablement à la tâche d'extraction. Pari est, en effet, tenu que les règles peuvent atteindre des performances d'extraction textuelle similaires aux modèles d'apprentissage machine lorsque certaines conditions textuelles sont remplies, et ce, sans que la plus-value en termes d'empreinte carbone ne soit amoindrie par un « surcoût temporel » lié à la complexité de l'extraction de certaines entités.

Un des écueils au développement des règles du présent cas d'usage fut le caractère très retardé de l'identification de la complexité de certaines entités à modéliser. Si, dès la première itération, l'entité 'caractère complet de la résection chirurgicale' fut divisée en quatre entités filles en raison de sa haute hétérogénéité linguistique, les entités 'taille de la tumeur' ou 'distance minimale aux marges' furent qualifiées de délicates à modéliser par expressions régulières après une étude étirée sur neuf mois, et alors que seulement six documents avaient été analysés lors des itérations 1 à 5 (Tableau 10). L'approche de la méthode hybride d'extraction consiste ainsi à annoter de façon exhaustive en une séance unique de travail un

corpus de textes de taille non nulle (n=40), après avoir testé sa représentativité linguistique par rapport au corpus entier. De cette façon, l'hétérogénéité linguistique des différentes annotations, et de leurs contextes textuels, correspondant à une même entité peut être identifiée dans un temps unique. La faisabilité d'une extraction / pré-annotation automatique par un système à base de règles et d'expressions régulières peut alors être évaluée, grâce à la visualisation de la liste de toutes les annotations réalisées par entité.

Dans un premier temps, il s'agirait d'évaluer le rappel, c'est-à-dire le taux de faux négatifs associés à un système à base de règles. Si le contenu des annotations est jugé relativement homogène, c'est-à-dire relevant de champs lexicaux et de structures syntaxiques proches, l'option des systèmes de règles pourrait être retenue. Ainsi, si dès cette analyse en lien avec le rappel, les annotations sont marquées par une hétérogénéité forte sur un plan lexical ou syntaxique, l'apprentissage machine serait à envisager. Par exemple, la présence de termes de substitution constituerait un signal d'alerte, comme la nécessaire division de l'entité 'caractère complet de la résection microscopique' en quatre entités filles dans le travail présenté ici. Le deuxième temps d'analyse chercherait à évaluer la précision, c'est-à-dire le taux de faux positifs. Les annotations relevées doivent ainsi être spécifiques de l'entité, c'est-à-dire non répandus ailleurs dans le texte. Dans le cas d'usage de la thèse, la présence de termes génériques tels que « nodule » pour l'identification d'un cancer a ainsi augmenté significativement le taux de faux positifs de l'entité 'taille de la tumeur', en raison du caractère non spécifique au cancer de ce mot (« bénin » par rapport à « malin »). Le succès relatif de l'apprentissage machine par rapport aux règles est, ici, très certainement en rapport avec la prise en compte du contexte de l'annotation (300). Ces deux étapes permettent à l'expert de visualiser simultanément la contribution de chaque catégorie d'annotation en termes de participation au rappel et à la précision associés à chaque entité.

Enfin, la correction progressive par un médecin senior du jeu de développement initialement annoté par un interne a été associée, dans le cas d'usage du 'Challenge AI for health', à une amélioration significative des métriques de performances. La reproductibilité des annotations manuelles est une question importante, comme le montre les différences de métriques de performances obtenues lors de la confrontation aux jeux de tests annotés par un médecin junior ou par un senior. En oncologie, la constitution préalable d'un guide d'annotations est indispensable à l'optimisation des tâches d'annotations, mais pourtant insuffisante à optimiser les performances d'extraction d'entités majeures en oncologie comme l'identification des stades tumoraux (pTNM et métastatique) (240,301). Il s'ensuit un nécessaire niveau d'expertise métier de l'annotateur.

En conclusion, les contributions de ce chapitre au travail de thèse furent de démontrer qu'il était possible de :

- Extraire des items du jeu de données minimales en oncologie développé dans la partie I de la thèse à partir de comptes rendus textuels des dossiers patients informatisés de l'AP-HP ;
- Utiliser de façon efficiente les algorithmes de traitement automatique du langage naturel correspondant pour informer des critères de jugement de différents cas d'usage en épidémiologie, pilotage et vision par ordinateur ;
- Analyser les modalités et les ressources nécessaires au développement de règles pour une tâche d'extraction textuelle et identifier ses déterminants et limites ;

- Discuter une méthode de pré-annotation automatique par règles préalable à un éventuel entraînement d'apprentissage machine ;
- Proposer une méthodologie alternative de discrimination *a priori* entre un système à base de règles et un système d'apprentissage machine pour l'extraction textuelle, et ce, à l'échelle de chaque entité.

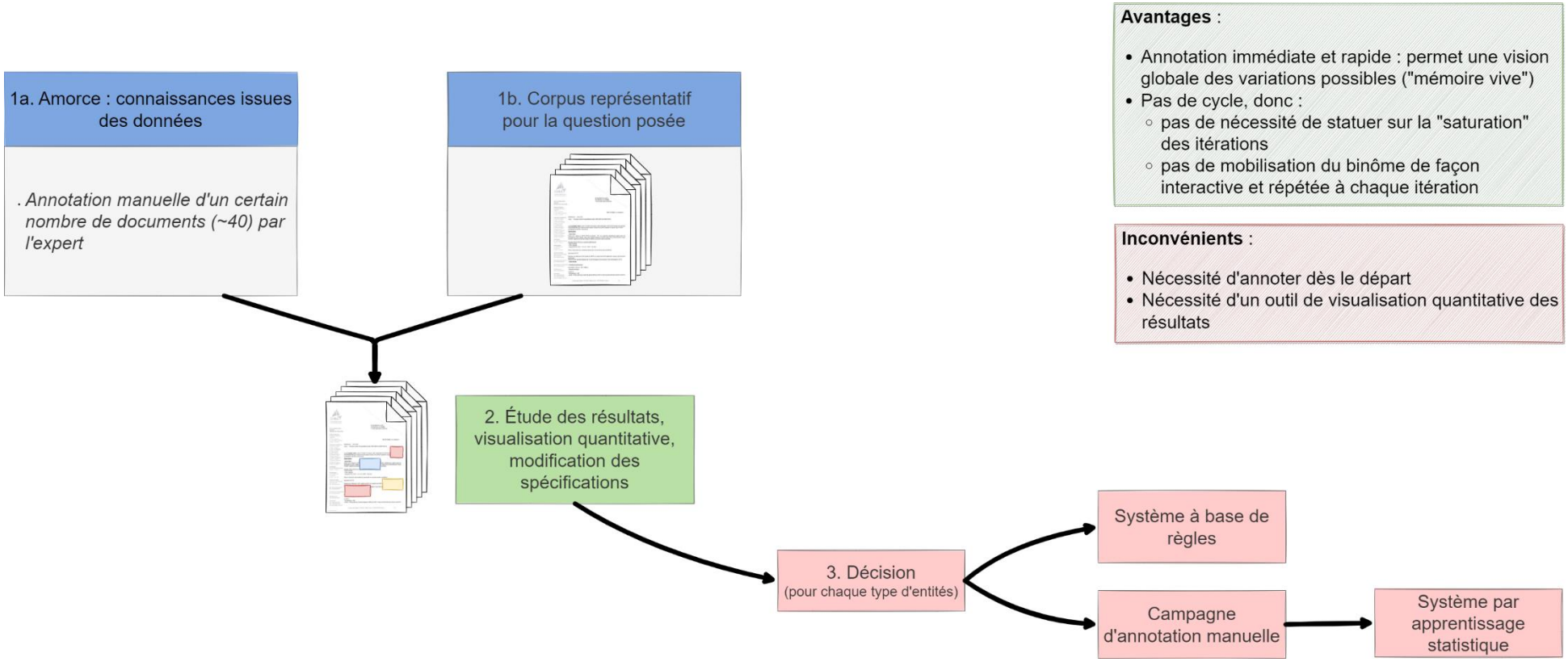


Figure 19. Proposition méthodologie permettant, pour une tâche d'extraction textuelle donnée, de déterminer a priori *entre deux méthodes disponibles : les règles et l'apprentissage machine*

Conclusion et perspectives

I. Synthèse de la thèse

Le présent travail de thèse a permis de démontrer qu'une attention portée à un certain nombre de thématiques des sciences des données permettait l'utilisation efficiente d'un EDS, comme celui de l'AP-HP, et ce, à des fins diverses dans le domaine de l'oncologie. Pilotage, épidémiologie clinique, aide à la décision médicale et développement algorithmique en vision par ordinateur, les applications sont nombreuses et possibles, nonobstant le traitement de points de vigilance transversaux.

En premier lieu, l'identification d'un jeu de données minimales spécifique de l'oncologie constitue une pierre angulaire qui permet de concentrer l'effort de formalisation des items d'intérêt propres à la discipline. Issue de la pratique clinique de synthèse médicale des caractéristiques d'un patient, cet exercice de réduction dimensionnelle des variables contenues dans le dossier médical a pour objectif de modéliser le squelette des données prioritaires, à partir desquelles se déploie l'arborescence de l'évolution de la maladie et du parcours de soins du patient. A partir de 15 items minimaux identifiés dans le présent travail, quatre cas d'usages relevant de perspectives médicales distinctes ont pu être développés avec succès et informer les différentes disciplines d'application : activité de pilotage institutionnel concernant l'automatisation de calculs d'indicateurs de sécurité et de qualité des soins nécessaires à la certification internationale des établissements de santé, épidémiologie clinique concernant l'impact des mesures de santé publique nationales en temps de pandémie sur le retard diagnostic des cancers, aide à la décision concernant l'optimisation du recrutement des patients éligibles à l'inclusion dans des essais cliniques, et développement de réseaux de neurones concernant des modèles de pronostication par vision par ordinateur.

Une deuxième condition nécessaire à l'exploitation sereine et à large échelle des données d'un EDS dans le domaine de l'oncologie, et particulièrement celles correspondant au précédent jeu de données minimales, repose sur leur formalisation optimale et interopérable

entre plusieurs EDS. Dans le cadre de l'initiative française PENELOPE visant à améliorer le recrutement des patients dans des essais cliniques, le présent travail a confirmé et évalué la plus-value de l'extension oncologie du modèle de données commun OMOP développée en 2021. Cette version 5.4 d'OMOP permettait de doubler le taux de formalisation selon le modèle OMOP de critères de préscreening issus d'essais cliniques de phase I à IV. En revanche, seulement 23% de ces critères de préscreening pouvaient être requetés automatiquement sur l'état actuel de l'EDS de l'AP-HP, et ce, *modulo* une valeur prédictive positive ne dépassant pas 30%, liée à une qualité des données structurées et de leur représentation encore sous-optimale. Ce travail propose ainsi une méthodologie globale pour évaluer la performance d'un système d'aide au recrutement dans des essais cliniques : à la fois à partir des métriques habituelles de performance de ce type d'outil telles que sensibilité, spécificité, valeur prédictive positive, valeur prédictive négative, mais aussi à partir d'indicateurs complémentaires caractérisant l'adéquation du modèle pivot choisi et son évolutivité tels que les taux de traduction et d'exécution des requêtes de repérage de patients. Indépendamment du type de modèle de données commun choisi (ici, OMOP), la thèse a démontré que la caractérisation d'un patient atteint de cancer pouvait être réalisée à partir de requêtes exécutables automatiquement sur les dossiers patients informatisés de tout établissement de santé, en vue de leur interopérabilité.

Enfin, le présent travail a permis de montrer dans quelle mesure le traitement automatique du langage naturel pouvait pallier le manque de structuration des données présentes dans un entrepôt, voire celui de la complétude des données structurées disponibles. Informant la stadification tumorale initiale réalisée lors de tout bilan d'extension d'un diagnostic de cancer récent, ainsi que les caractéristiques histopronostiques des tumeurs, le développement de règles et de modèles d'apprentissage machine a permis d'enrichir de façon efficiente le jeu de données minimales en oncologie. La confrontation des métriques de performance d'extraction textuelle et des ressources humaines et techniques nécessaires au développement de systèmes à base de règles et de modèles par apprentissage automatique a permis de valoriser, pour un certain nombre de situations, la première approche et de recommander un niveau d'expertise métier optimal de l'annotateur. L'analyse des modalités d'extraction textuelle a identifié qu'une approche de préannotation automatique à base de règles, avant une phase d'annotation manuelle pour entraînement séquentiel d'un modèle d'apprentissage machine, pourrait être optimisée. Les règles semblent suffire pour les tâches d'extraction textuelle d'une certaine typologie d'entités bien caractérisée sur un plan lexical et sémantique. L'anticipation et la modélisation de cette typologie pourrait être possible en amont de la phase d'extraction textuelle, afin de différencier, en fonction de chaque type d'entité, dans quelle mesure les systèmes à base d'apprentissage machine devraient suppléer aux règles.

II. Perspectives

1. Constitution d'un jeu de données minimales FAIR dans le domaine du cancer

La confrontation de notre posture d'utilisatrice polyvalente à l'état actuel des données de l'EDS de l'AP-HP a permis d'objectiver qu'un certain nombre restreint d'items cliniques, biologiques, radiologiques et thérapeutiques suffisait à caractériser de façon efficiente les patients atteints de cancer solide et de permettre une utilisation secondaire polyvalente des données minimales de ces patients. Quel que soit le cas d'usage envisagé, un jeu de données minimales en oncologie semble constituer une trame transversale et résiliente à une exploitation pragmatique d'un EDS. En vue de la stabilisation et de l'articulation de ce jeu de données au sein de l'entrepôt local, est attendue la constitution d'un dictionnaire de données. Cette approche développée en santé dans les années 1980 permettrait l'identification des métadonnées propres à généraliser l'exploitation du jeu de données minimales de manière collaborative, fiable, reproductible et évolutive (302,303).

Selon les principes de la démarche *findable, accessible, interoperable, reusable* (FAIR) promue à l'heure du partage des données massives en santé, ce jeu de données minimales a pour vocation d'être enrichi et décliné en fonction des spécificités de ses utilisations conjecturelles (96). Certains auteurs ont déjà proposé des modèles et vocabulaires jugés optimaux en vue d'un partage de données d'oncologie respectant ces principes de bonnes pratiques, et ce, particulièrement concernant les données -omics propres à promouvoir la médecine de précision (304,305). Or, il semble que la pratique du partage des données et des codes sous-jacents à la publication d'études en oncologie bénéficierait d'une marge de progression substantielle. Une récente analyse de 306 publications indexées sur PubMed en 2019 en lien avec le champ de l'oncologie observait que si 19% des auteurs déclaraient les données de leur étude disponibles, moins d'1% d'entre eux respectaient les principes FAIR (306). Un constat assez similaire semblerait applicable à la formalisation d'EDS au format OMOP (121). En vue du respect des bonnes pratiques FAIR, la consolidation du jeu de données minimales en oncologie devra s'astreindre à :

- Une description et une caractérisation des données et métadonnées exhaustives (*findable*) ;
- Un stockage et une mise à disposition de ces éléments facilités et spécifiés (*accessible*) ;
- Une perspective d'interopérabilité nationale et internationale (*interoperable*) ;
- Une recherche de transversalité dans les cas d'usages et exploitations par les utilisateurs telles que le pilotage, la recherche observationnelle et (quasi-) interventionnelle, la pédagogie.

A l'échelle locale de l'EDS de l'AP-HP, la consolidation de ce jeu de données minimales en oncologie permettrait de constituer, après validation indispensable par un collège d'experts, une feuille de route pour la standardisation, la structuration, l'enrichissement et la qualification des données cancer d'intérêt de l'entrepôt institutionnel, et ce, malgré l'existence de référentiels internationaux de caractérisation d'un patient atteint de

cancer (307). A l'échelle nationale, ce jeu de données minimales en oncologie consolidé permettrait de guider l'implémentation des modèles de données communs pertinents pour l'interopérabilité des EDS français, au sein de l'initiative PENELOPE et au-delà. La labellisation « sites de recherche intégrée sur le cancer » (SIRIC) de 3 groupes hospitalo-universitaires de l'AP-HP en 2022 par l'INCa en constituerait une éventuelle opportunité (308). Son articulation avec le jeu de données minimales retenu par le projet OSIRIS-RWD permettrait d'en assurer la pertinence, l'applicabilité et, ainsi, l'évolutivité (Figure 20) (190).

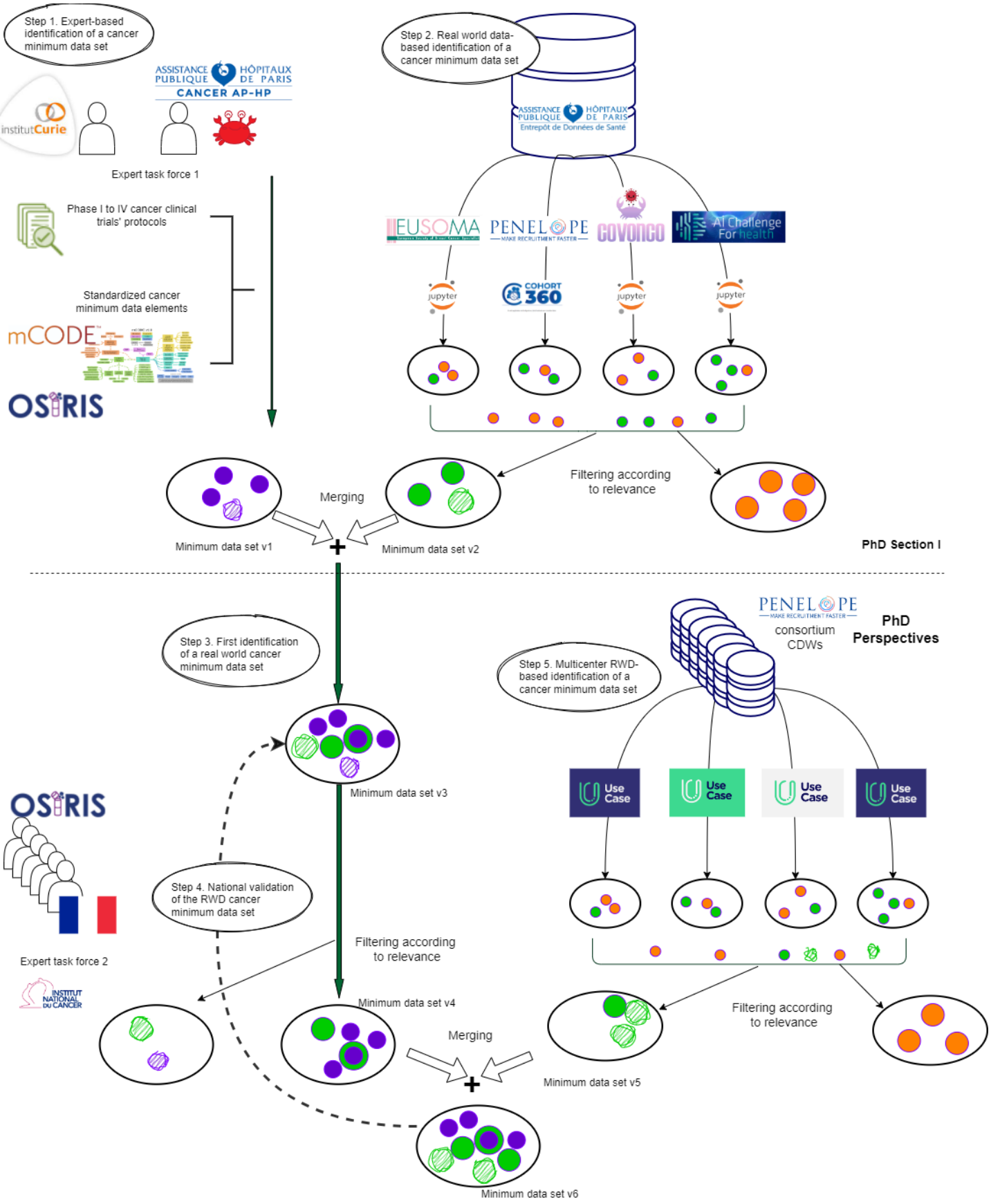


Figure 20. Perspectives des modalités d'amélioration du jeu de données minimales en oncologie développé lors de la thèse à l'échelle nationale

2. Formalisation interopérable de ce jeu de données minimales

Le deuxième verrou identifié par le travail de thèse à l'exploitation pragmatique des données liées au cancer d'un entrepôt comme celui de l'AP-HP réside dans le manque de leur standardisation, faisant échouer toute velléité d'interopérabilité avec d'autres bases de données. Cette approche collaborative est devenue un incontournable de la pratique de pilotage et de recherche en oncologie, compte tenu de l'explosion nosographique continue de la discipline. Avant même la question de l'interopérabilité d'EDS nationale ou internationale, le choix des modalités de formalisation de leur contenu conditionne leur exploitation, et ce, particulièrement dans le domaine de l'oncologie. Cette discipline, comme d'autres, a trait à une pathologie le plus souvent chronique dont l'évolution se réalise souvent par étapes longitudinales caractérisées par des lignes thérapeutiques, et dont les caractéristiques sont interdépendantes. Ainsi, la représentation du parcours de soins d'un patient atteint de cancer nécessite des relations entre différentes données caractéristiques de ces épisodes séquentiels. Si la direction de l'EDS de l'AP-HP a décidé, conjointement au HDH, d'opter pour l'adoption du modèle de données commun proposé par OHDSI, son implémentation actuelle est associée à une marge d'amélioration importante en ce qui concerne le cancer. En effet, la version 5.3 n'intègre pas l'extension oncologie développée en 2021, alors que son recours permet, comme le présent travail l'a démontré, d'améliorer substantiellement la représentation des données d'un patient atteint de cancer, et notamment en ce qui concerne la modélisation de son parcours de soins longitudinal qui se déploie à partir de l'histoire naturelle de sa pathologie.

L'initiative PENELOPE pourrait être la première tentative nationale d'interopérer les EDS autour du partage de données liées au cancer afin d'améliorer l'identification automatique des patients atteints de cancer éligibles à une inclusion dans les essais cliniques, et ce, selon des référentiels internationaux. Le premier jalon, déployé dans la thèse, a utilisé le jeu de données minimales développé dans la première partie pour identifier les critères de préscreening parmi ceux d'inclusion et de non-inclusion d'essais cliniques en oncologie médicale. Ce faisant, nous avons défini une méthodologie globale pour évaluer la qualité de l'ensemble du processus, depuis le traitement des critères d'éligibilité à l'identification automatique d'une cohorte de patients partageant ces caractéristiques. Parmi les perspectives du projet PENELOPE, l'extension de ce travail pilote aux autres établissements de santé partenaires du consortium permettra d'évaluer la validité externe de pareille démarche. Le programme 'ACCES AP-HP', financé à hauteur de dix millions d'euros par BPI France, permettra d'encadrer ce nouveau jalon puisque l'initiative PENELOPE en constitue un des cinq cas d'usage, en vue « de constituer et consolider, au sein des établissements participants [consortium HUGO à Nantes, Angers, Rennes et Tours ; EDEN4Health à Reims ; laboratoires INSERM LIMICS de Sorbonne, LTSI de Rennes 1 et CRC de Paris Cité], un jeu de données minimal permettant de développer et valider des algorithmes de repérage de patients atteints de cancer afin de faciliter leur inclusion dans des essais cliniques (préscreening). En lien avec l'ANS, l'INCa et le HDH, ce projet repose sur la conception et la validation d'un modèle de données dédié à la maladie cancéreuse basé sur les standards internationaux d'interopérabilité » (309). La validité externe, ici évaluée, pourra être étayée par les métriques standards de valeurs prédictive positive et négative de l'identification automatique des cohortes de patients éligibles. Elle pourra également être enrichie des

métriques présentées dans le présent travail : taux de calculabilité résultant de la combinaison entre le taux de standardisation des critères d'éligibilité en requêtes standards (*Criteria To Queries*), et celui d'exécution des requêtes sur les dossiers patients informatisés (*Queries To Real-World data*). De ce fait, seront évaluées à la fois l'interopérabilité des données sources, et également leur qualité.

Certains points d'attention, parmi les limites objectivées lors du présent travail de thèse, seront mis en exergue lors de la poursuite du projet PENELOPE. Une étape inhérente à sa méthodologie réside dans le reformatage des critères de préscreening en données élémentaires (standardisation), ici effectuée manuellement à l'aide du dictionnaire ATHENA (203). Le recours à l'outil ATLAS proposé par OHDSI permettrait, dans la prochaine phase du projet PENELOPE, l'accès à la liste exhaustive et récursive des concepts standards fils connexes (310). Par ailleurs, ATLAS, ou tout autre outil de reformatage automatique de critères textuels en données élémentaires standardisées comme CTKB, permettra d'accélérer, fiabiliser et sécuriser l'évolutivité et la cohérence de cette transformation chronophage, subjective et partiale des critères en requêtes OMOP standards (243). La formalisation des critères de préscreening devra également constituer un point de vigilance du déploiement ultérieur du projet PENELOPE. En effet, l'hétérogénéité de leur formulation a représenté un sujet d'achoppement du présent travail. Or, un consortium multipartite comprenant l'institut nation du cancer américain (NCI), des représentants des industries pharmaceutiques, la Food and Drug Administration (FDA) et des investigateurs d'essais a lancé une initiative visant à simplifier et à harmoniser les critères d'éligibilité des essais cliniques en oncologie (311).

Est ainsi attendu que les prochaines recommandations émises par la FDA amélioreront la qualité de ces éléments clés du processus de recrutement des patients, et feront la promotion de guides d'implémentation d'outils de repérage de patients ("cohort builders") s'appuyant sur les initiatives d'OHDSI ou d'HL7 FHIR Vulcan "données de vie réelle pour la recherche" (intégrant les spécifications d'mCODE en ce qui concerne l'oncologie. Ces deux grandes propositions américaines de modélisation des données d'oncologie résultent de perspectives quelque peu distinctes. Le modèle d'OHDSI a pour objectif la conduite d'études observationnelles à partir de bases de données interopérées. La démarche OMOP repose sur l'adoption d'un format et d'une représentation (vocabulaires) communes, associée à des outils d'analyses dédiés. L'histoire de la maladie y est représentée en épisodes (diagnostic, réponse, progression, etc.) selon une perspective globale et longitudinale symbolisée par la table OVERARCHING_DISEASE. Ainsi, grande est la place laissée aux jalons représentés par les lignes thérapeutiques qui scandent l'histoire naturelle de la pathologie. Le modèle mCODE™, quant à lui, formalise ces épisodes séquentiels comme des événements tumoraux à part entière, bien que possiblement reliés entre eux, et autour desquels se déploient le reste des données. L'alignement réalisé entre les propositions formelles identifie quelques différences d'approches comme, p. ex., le caractère obligatoire de certains champs et éléments en OMOP qui sont optionnels ou sans valeur pertinente correspondante en FHIR, sur la présence d'extensions utilisées pour ajouter des données en FHIR et la table FACT_RELATIONSHIPS d'OMOP qui relie les diagnostics primaires et secondaires, ou encore, p. ex., la cristallisation des données morphologiques et topographiques du diagnostic tumoral en OMOP, maintenues séparées en FHIR (312,313). Toujours est-il qu'un atelier dédié à « créer des opportunités pour faire progresser les systèmes de santé apprenants en oncologie en favorisant l'échange de données cliniques et de recherche entre mCODE™ et l'extension oncologie d'OMOP » eut lieu

lors du symposium OHDSI en octobre 2022 (314). L'horizon semble ainsi la mise en place de systèmes d'informations de santé apprenants et durables basés sur un standard FHIR d'échanges de données cliniques et sur un modèle de données optimisé pour l'analyse et la conduite d'études observationnelles en santé, à savoir OMOP. Dans ce contexte, plusieurs initiatives d'alignement entre les deux formalisations de données liées au cancer ont été développées par FHIR, et notamment autour d'items clés tels que la description histologique et topographique d'un cancer primitif, sa stadification, les traitements anti-tumoraux et les biomarqueurs (315).

Dans ce contexte, les perspectives du projet PENELOPE pourraient consister en :

- L'automatisation de la transformation des critères de préscreening en données élémentaires standards et l'évaluation des outils développés à cette fin ;
- La mise en œuvre à l'AP-HP et chez ses partenaires d'une version OMOP intégrant l'extension oncologie, impliquant la nécessaire génération de règles d'implémentation fondées sur des avis consensuels d'experts, ainsi que l'installation de l'outil ATLAS ;
- L'implémentation de l'API HL7 FHIR en vue de l'adoption de mCODE™, ainsi que des fonctionnalités d'alignement avec OMOP ;
- L'évaluation de la plus-value de ces efforts de standardisation des données des entrepôts en termes de performances obtenues par les outils existants de générations de cohortes de patients (impact sur les taux de traduction et d'exécution et sur la performance globale des outils de préscreening automatique) ;
- L'optimisation de la qualité et de l'enrichissement des données clés, et particulièrement via les techniques de traitement automatique du langage naturel.

Dans la veine pragmatique de la dynamique de la thèse, la menée à bien de ces perspectives pourra se réaliser à l'occasion de l'implication de l'AP-HP dans des programmes de structuration et d'interopérabilité de son EDS. Au niveau national, le volet santé du programme « France 2030 » a permis au gouvernement français de lancer en 2023, via l'agence de l'innovation en santé, le Paris Saclay Cancer Cluster (PSCC) soutenu à hauteur de 100 millions d'euros pendant 10 ans par l'Etat et de 50 millions par Sanofi (316). Un des trois plans du PSCC cherche justement à soutenir la structuration des données autour de l'initiative OSIRIS-RWD de 4 premiers établissements de santé, dont l'AP-HP qui recevra à cette fin une enveloppe d'environ 3 millions d'euros. Au niveau européen, l'institution de santé francilienne est impliquée dans le projet Breast cancer benchmark qui vise à tester l'interopérabilité des EDS de 9 CHU européens autour du modèle OMOP dans le cadre de l'initiative EUHA (202). L'objectif du projet est de faire interopérer les données de 150 patients atteints de cancer du sein en vue d'informer neuf indicateurs de qualité et de sécurité des soins spécifiquement identifiés pour ce projet à partir des référentiels EUSOMA et du consortium international pour la mesure des résultats en santé (ICHOM) (317). D'autres projets internationaux auxquels participent actuellement des dizaines d'établissements de santé tels qu'IDEA4RC autour de la caractérisation des cancers rares, ou EUCAIM concernant « l'analyse fédérée et distribuée des données d'imagerie du cancer » impliquent des médecins chercheurs de l'AP-HP comme investigateurs principaux, et représentent, de fait, des projets structurants pour les données d'oncologie de l'institution (318,319).

En vue de leur accompagnement et de leur déploiement, le présent travail de thèse a permis l'ouverture, début 2023, d'un espace de travail purement axé autour de la recherche et du développement dédié à la consolidation des données cancer de l'entrepôt de l'AP-HP (Projet CSE-22-24_OncOMOP-Terabase). La responsabilité scientifique en est, entre autres, partagée avec les co-porteuses de la mission *Intelligence artificielle et Cancer* au niveau institutionnel (320). Cet espace permet, notamment, les travaux en traitement automatique du langage naturel en vue de la structuration des données clés d'oncologie.

3. Recours au traitement automatique du langage naturel en vue de l'enrichissement du jeu de données minimales en oncologie, suivant une méthodologie d'extraction textuelle durable

Le manque de structuration et/ou de qualité de certaines informations d'intérêt d'oncologie disponibles au sein de l'EDS de l'AP-HP a permis d'entériner la pertinence d'une approche d'enrichissement de ce jeu de données minimales par des techniques de traitement automatique du langage naturel. Après le développement et la validation, à l'occasion du présent travail de thèse, d'algorithmes permettant l'extraction d'une dizaine d'entités clés en vue de la caractérisation d'un patient atteint de cancer solide, l'effort devrait se poursuivre pour finaliser l'enrichissement du jeu de données minimales. L'identification du dictionnaire de données correspondant permettrait de servir de trame de guide d'annotations pour le programme d'extraction textuelle par traitement automatique du langage naturel des items concernés, que l'option méthodologique retenue soit celle d'une approche symbolique (règles) ou statistique (apprentissage machine). Ce référentiel commun des différents chercheurs impliqués dans le développement du traitement automatique du langage naturel appliqué au contenu oncologique de l'EDS de l'AP-HP permettrait de mutualiser l'effort de structuration des données liées au cancer, et limiter la redondance des efforts mobilisés tout en assurant la pérennité et la qualité des algorithmes mis à libre disposition de la communauté scientifique et des utilisateurs au sein de la bibliothèque edsnlp (245,321,322).

Ce développement en traitement automatique du langage naturel permettrait, ainsi, d'informer certains des champs névralgiques des tables OMOP ou des *ressources* FHIR concernant les informations clés en oncologie et, ce faisant, améliorer l'interopérabilité des données cancer de l'entrepôt de l'AP-HP (323). En effet, dans la première étape du projet PENELOPE déployée lors de la thèse, 33% des 288 données élémentaires correspondant aux 83 critères de préscreening du cas d'usage n'étaient pas requêtables en raison d'une simple absence de structuration, et ce, indépendamment de toute problématique de standardisation. Au-delà d'une question qualitative de l'absence de structuration des données d'intérêt, le traitement automatique du langage naturel représente également une opportunité précieuse pour pallier les données sous-codées du PMSI. Ainsi, le développement et la validation d'une expression régulière concernant la stadification métastatique initiale a suppléé avec succès aux codes CIM-10 correspondants et dont la fiabilité put être remise en question au sein des données de codage de l'AP-HP (271).

Parmi des perspectives méthodologiques, l'évaluation de la validité externe de ces algorithmes doit être envisagée concernant les types de cancers solides non traités ici, ainsi que les autres étapes du parcours de soins dépassant le bilan d'extension initial. En effet, une fragilité inhérente au traitement automatique du langage naturel réside dans la dépendance de ses performances algorithmiques au type de support de référence concerné (151). Ainsi, dans les jeux de développement correspondants, l'explicitation du mot 'métastase' variait tellement entre les comptes rendus d'imagerie initiale par rapport à ceux produits lors de l'évaluation de l'efficacité d'un traitement antitumoral, que décision fut prise, pour la thèse, de développer des classifieurs de documents en fonction de l'intentionnalité médicale sous-jacente à la réalisation de l'imagerie. Par exemple, lors de métastases connues, le sous-texte peut être amené à réduire à l'expression de 'lésions' – terme médical on ne peut plus générique – la référence à une métastase. Or, la structuration et la standardisation de cette information est fondamentale en raison de son sous-codage PMSI d'une part, et de son absence de relation directe à un type de cancer primitif donné – comme objectivé par la thèse dans le projet PENELOPE. Cette validation constitue un prérequis chronophage à la généralisabilité des algorithmes développés et validés sur les différents cas d'usages du présent travail de thèse. Quoi qu'il en soit, les prochains développements algorithmiques envisagés devront être associés, à l'heure de leur dissémination scientifique au sein de publications dédiées, à une traçabilité optimale en termes de méthodes et résultats (324).

La thèse a, finalement, contribué à enrichir la réflexion méthodologique concernant les modalités d'extraction textuelle automatique d'une dizaine d'entités issues du jeu de données minimales en oncologie. Les suites du travail consisteront à évaluer, en prospectif et pour un autre cas d'usage, la nouvelle proposition méthodologique à ressources contraintes concernant l'extraction textuelle des informations clés du cancer contenues dans les dossiers patients informatisés. Dans cette perspective, les systèmes à base de règles correspondent à l'approche par défaut par rapport aux modèles d'apprentissage machine, compte tenu de leur plus faible empreinte carbone. Le recours aux modèles statistiques est, ainsi, mobilisé uniquement lorsqu'un type d'entité semble *a priori* complexe à modéliser par règles. A partir de la caractérisation initiale syntaxique et sémantique des annotations réalisées en une séance de travail unique par un annotateur expert du domaine médical, l'hypothèse consiste à penser que certains critères linguistiques permettent d'anticiper la facilité et le succès avec lesquels un système à base de règles pourra être développé par typologie d'entités (Figure 19).

Cette méthodologie consiste en plusieurs étapes, illustrées ici à l'aide d'un exemple fondé sur l'extraction de l'entité « invasion tumorale vasculaire » :
1) annoter par un expert en une séance de travail unique un jeu de comptes rendus pour tous les types d'entités ;
2) visualiser pour chaque entité la distribution des annotations correspondant à chaque entité (Tableau 12) ;
3) réaliser un regroupement de ces annotations en catégories définies par un plus petit commun dénominateur sémantique (Tableau 12) ;
4) évaluer la distribution de ces catégories parmi les annotations correspondantes (Figure 21) (rappel) ;

5) évaluer la distribution de ces catégories en dehors des annotations dans le reste du texte libre (précision) ;
6) conclure en déterminant si la modélisation par règles est faisable, c'est-à-dire que les catégories sont peu nombreuses et homogènes sur un plan linguistique, et spécifiques des annotations.

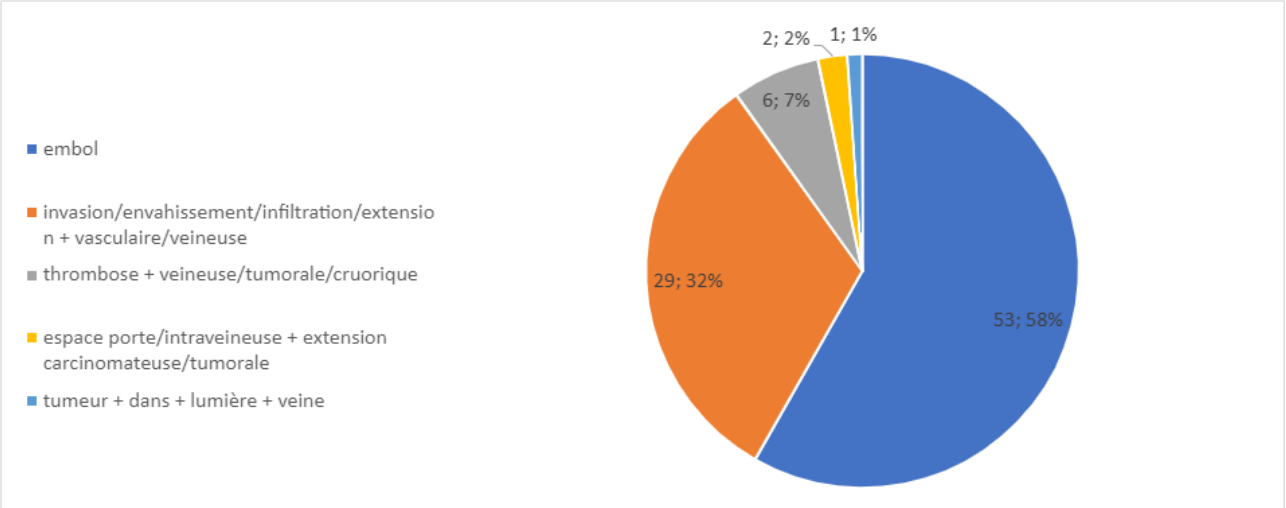


Figure 21. Distribution des catégories linguistiques de l'ensemble des annotations correspondant à l'entité « invasion tumorale vasculaire », à partir de 40 comptes rendus d'anatomopathologie post-opératoires aléatoirement identifiés du jeu de développement, pour le cas d'usage 'Challenge AI for health'

Tableau 12. Distribution des premières annotations correspondant à l'entité « invasion tumorale vasculaire » au sein de 40 comptes rendus du jeu de développement, et définition de 5 catégories en fonction du plus petit dénominateur sémantique

Contenu des annotations	Nombre d'occurrences	Catégorie correspondante définie
emboles	26	1 = embol
Absence d'invasion vasculaire	5	2 = invasion/envahissement/infiltration/extension + vasculaire/veineuse
Absence d'embole	4	1 = embol
embole	4	1 = embol
Absence d'embol	2	1 = embol
embols	2	1 = embol
emboles	2	1 = embol
thrombose veineuse tumorale	2	3 = thrombose + veineuse/tumorale/cruorique
et d'emboles	2	1 = embol
invasion vasculaire	2	2 = invasion/envahissement/infiltration/extension + vasculaire/veineuse
envahissements vasculaires	2	2 = invasion/envahissement/infiltration/extension + vasculaire/veineuse
sans invasion vasculaire	2	2 = invasion/envahissement/infiltration/extension + vasculaire/veineuse
Emboles	1	1 = embol
veine porte gauche est perméable,sans infiltration	1	2 = invasion/envahissement/infiltration/extension + vasculaire/veineuse
espaces portes les plus proches de la tumeur sont pour certains le siège d'une extension carcinomateuse	1	4 = espace porte/intraveineuse + extension carcinomateuse/tumorale

La suite du travail de thèse consiste donc, d'une part, à évaluer la pertinence de cette méthode sur un nouveau cas d'usage, et d'autre part, à développer un outil interactif de visualisation de ces critères par l'annotateur amené à trancher entre les deux options méthodologiques d'extraction (325,326). A l'occasion de la poursuite de l'enrichissement du jeu de données minimales, l'angle méthodologique initié pendant la thèse concernant le traitement automatique du langage naturel pourra ainsi se préciser.

L'objectif de cette étude consiste à valider automatiquement le score de pronostication français PRONOPALL dédié à l'oncologie médicale, et ce, à partir de quatre items dont deux ne sont pas structurés dans l'EDS de l'AP-HP semble une opportunité pertinente (327). Cette initiative permettrait d'augmenter la taille de la population de validité externe du score par rapport aux études réalisées et fondées sur un recueil manuel des informations d'intérêt (328). Ce faisant, le cas d'usage étendrait à la démarche de pronostication les travaux de structuration des données cancer de l'entrepôt, tout en s'ouvrant à une nouvelle typologie de comptes rendus médicaux et en s'inscrivant dans une thématique médicale importante et déjà étayée par les techniques de traitement automatique du langage naturel (329,330). Ainsi, le calcul du nombre de sites métastatiques requis par le calcul du score PRONOPALL nécessite, abstraction faite des codes PMSI correspondants et sous-codés dans l'entrepôt de l'AP-HP, leur identification à partir de comptes rendus médicaux textuels, comme ceux d'imagerie, mais aussi de réunions de concertations pluridisciplinaires, de consultations et d'hospitalisations de jour et conventionnelles. La topographie anatomique des sites métastatiques à caractériser laisse anticiper un formalisme linguistique étroit et homogène pour certaines localisations secondaires comme le foie (ex : métastases 'dans le foie', 'hépatiques', 'du foie'), et particulièrement hétérogène et large pour d'autres comme l'os (ex : métastases 'de la troisième vertèbre lombaire', 'humérales', 'de la voûte crânienne', 'costales', etc.). Pourrait ainsi se confirmer qu'un système à base de règles permettrait la modélisation rapide et efficace de la première typologie d'entités, alors qu'un modèle d'apprentissage machine demeurerait nécessaire pour l'extraction de la seconde. Si nécessaire, d'autres cas d'usages envisagés par l'équipe Cancer Research Application on Big data (CRAB), qui a accompagné étroitement les travaux de thèse, pourraient représenter autant d'opportunités intéressantes de contenus et de méthodes pour l'approfondissement et l'exploitation des techniques de traitement automatique du langage naturel autour des données cancer de l'EDS de l'AP-HP.

Bibliographie

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin [Internet]. 2021 May 1 [cited 2021 Jul 5];71(3):209–49. Available from: https://onlinelibrary.wiley.com/doi/full/10.3322/caac.21660
2. Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM. Projecting cancer incidence and deaths to 2030: The unexpected burden of thyroid, liver, and pancreas cancers in the united states. Vol. 74, Cancer Research. American Association for Cancer Research Inc.; 2014. p. 2913–21.
3. Malone ER, Oliva M, Sabatini PJB, Stockley TL, Siu LL. Molecular profiling for precision cancer therapies. Vol. 12, Genome medicine. NLM (Medline); 2020. p. 8.
4. Park JJH, Siden E, Zoratti MJ, Dron L, Harari O, Singer J, et al. Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. Trials [Internet]. 2019 Sep 18 [cited 2021 Nov 19];20(1). Available from: https://pubmed.ncbi.nlm.nih.gov/31533793/
5. Laronga C, Gray JE, Siegel EM, Lee JH, Fulp WJ, Fletcher M, et al. Florida initiative for quality cancer care: Improvements in breast cancer quality indicators during a 3-year interval. J Am Coll Surg. 2014;219(4):638-645.e1.
6. Ministère des Solidarités et de la Santé. Stratégie nationale de santé 2018-2022. 2022;1–53.
7. Andreano A, Anghinoni E, Autelitano M, Bellini A, Bersani M, Bizzoco S, et al. Indicators based on registers and administrative data for breast cancer: routine evaluation of oncologic care pathway can be implemented. J Eval Clin Pract [Internet]. 2016 Feb 1 [cited 2022 Sep 9];22(1):62–70. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/jep.12436
8. Guarneri V, Pronzato P, Bertetto O, Roila F, Amunni G, Bortolami A, et al. Use of electronic administrative databases to measure quality indicators of breast cancer care: Experience of five regional oncology networks in Italy. J Oncol Pract. 2020 Feb 1;16(2):81.
9. Houzard S, Courtois E, Le Bihan Benjamin C, Erbault M, Arnould L, Barranger E, et al. Monitoring breast cancer care quality at national and local level using the French National Cancer Cohort. Clin Breast Cancer. 2022 May 21;
10. Bevan G, Hood C. What's measured is what matters: targets and gaminf in the English public health care system. Public Adm [Internet]. 2006;84(3):517–38. Available from: https://doi.org/10.1111/j.1467-9299.2006.00600.x
11. Deriu PL, Basso S, Mastrilli F, Orecchia R. OECI accreditation of the European Institute of Oncology of Milan: strengths and weaknesses. Tumori. 2015;101 Suppl:S21-4.
12. Ringborg U, Pierotti M, Storme G, Tursz T. Managing cancer in the EU: the Organisation of European Cancer Institutes (OECI). Eur J Cancer. 2008 Apr;44(6):772–3.
13. Amster A, Jentzsch J, Pasupuleti H, Subramanian KG. Completeness, accuracy, and computability of National Quality Forum-specified eMeasures. J Am Med Informatics Assoc. 2015;22(2):409–16.
14. Ahmad FS, Rasmussen L V., Persell SD, Richardson JE, Liss DT, Kenly P, et al. Challenges to electronic clinical quality measurement using third-party platforms in primary care practices: The healthy hearts in the heartland experience. JAMIA Open. 2019;2(4):423–8.
15. Schorer AE, Moldwin R, Koskimaki J, Bernstam E V, Venepalli NK, Miller RS, et al. Chasm Between Cancer Quality Measures and Electronic Health Record Data Quality. JCO Clin Cancer Informatics [Internet]. 2022;(6):e2100128. Available from: https://doi.org/10.1200/CCI.21.00128
16. Raghavan P, Chen JL, Fosler-Lussier E, Lai AM. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci. 2014;2014:218–23.
17. Biganzoli L, Marotti L, Hart CD, Cataliotti L, Cutuli B, Kühn T, et al. Quality indicators in breast cancer care: An update from the EUSOMA working group. Eur J Cancer. 2017 Nov;86:59–81.
18. van Dam PA, Tomatis M, Marotti L, Heil J, Wilson R, Rosselli Del Turco M, et al. The effect of EUSOMA certification on quality of breast cancer care. Eur J Surg Oncol J Eur Soc Surg Oncol Br Assoc Surg Oncol. 2015 Oct;41(10):1423–9.
19. Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Nikšić M, et al. Global surveillance of trends in cancer survival: analysis of individual records for 37,513,025 patients diagnosed with one of 18 cancers during 2000–2014 from 322 population-based registries in 71 countries (CONCORD-3). Lancet (London, England) [Internet]. 2018 Mar 3 [cited 2022 Sep 29];391(10125):1023. Available from: /pmc/articles/PMC5879496/
20. Kempf E, Lemoine N, Tergemina-clain G, Turpin A, Postel-Vinay S, Lanoy E, et al. A Case-Control Study Brings to Light the Causes of Screen Failures in Phase 1 Cancer Clinical Trials. PLoS One. 2016;11(5):3–9.
21. Fouad MN, Lee JY, Catalano PJ, Vogt TM, Zafar SY, West DW, et al. Enrollment of patients with lung and colorectal cancers onto clinical trials. J Oncol Pract. 2013;9(2).
22. Bennette CS, Ramsey SD, McDermott CL, Carlson JJ, Basu A, Veenstra DL. Predicting Low Accrual in the National Cancer Institute's Cooperative Group Clinical Trials. J Natl Cancer Inst. 2016;108(2):1–7.
23. Stensland KD, McBride RB, Latif A, Wisnivesky J, Hendricks R, Roper N, et al. Adult cancer clinical trials that fail to complete: an epidemic? J Natl Cancer Inst. 2014 Sep;106(9).
24. Benson AB 3rd, Pregler JP, Bean JA, Rademaker AW, Eshler B, Anderson K. Oncologists' reluctance to accrue patients onto clinical trials: an Illinois Cancer Center study. J Clin Oncol Off J Am Soc Clin Oncol. 1991 Nov;9(11):2067–75.
25. Comis RL, Miller JD, Aldigé CR, Krebs L, Stoval E. Public attitudes toward participation in cancer clinical trials. J Clin

Oncol Off J Am Soc Clin Oncol. 2003 Mar;21(5):830–5.
26. Asher N, Raphael A, Wolf I, Pelles S, Geva R. Oncologic patients' misconceptions may impede enrollment into clinical trials: a cross-sectional study. BMC Med Res Methodol. 2022 Jan;22(1):5.
27. Jenkins V, Farewell D, Batt L, Maughan T, Branston L, Langridge C, et al. The attitudes of 1066 patients with cancer towards participation in randomised clinical trials. Br J Cancer. 2010 Dec;103(12):1801–7.
28. Ulrich CM, Ratcliffe SJ, Zhou Q, Huang L, Hochheimer C, Gordon T, et al. Association of Perceived Benefit or Burden of Research Participation With Participants' Withdrawal From Cancer Clinical Trials. JAMA Netw open. 2022 Nov;5(11):e2244412.
29. Ulrich CM, Knafl K, Foxwell AM, Zhou Q, Paidipati C, Tiller D, et al. Experiences of Patients After Withdrawal From Cancer Clinical Trials. JAMA Netw open. 2021 Aug;4(8):e2120052.
30. Hillman SL, Jatoi A, Strand CA, Perlmutter J, George S, Mandrekar SJ. Rates of and Factors Associated With Patient Withdrawal of Consent in Cancer Clinical Trials. JAMA Oncol. 2023 Jun;
31. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. Contemp Clin trials Commun. 2018 Sep;11:156–64.
32. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. Contemp Clin Trials Commun [Internet]. 2018 Sep 1 [cited 2022 Sep 12];11:156. Available from: /pmc/articles/PMC6092479/
33. Mills EJ, Seely D, Rachlis B, Griffith L, Wu P, Wilson K, et al. Barriers to participation in clinical trials of cancer: a meta-analysis and systematic review of patient-reported factors. Lancet Oncol [Internet]. 2006 Feb [cited 2022 Sep 29];7(2):141–8. Available from: https://pubmed.ncbi.nlm.nih.gov/16455478/
34. OHDSI. OMOP Common Data Model Oncology Extension Documentation [Internet]. Available from: https://ohdsi.github.io/CommonDataModel/oncology.html
35. Kostos L, Hong W, Lee B, Tran B, Lok SW, Anton A, et al. Cancer clinical trial vs real-world outcomes for standard of care first-line treatment in the advanced disease setting. Int J Cancer. 2021;149(2):409–19.
36. Sa R, Xu Y, Pan X, Wang Y, Lin Z, Zhang X, et al. A bibliometric analysis of research progress on pharmacovigilance and cancer from 2002 to 2021. Front Oncol. 2023;13:1078254.
37. Yu CT, Farhat Z, Livinski AA, Loftfield E, Zanetti KA. Characteristics of Cancer Epidemiology Studies that Employ Metabolomics: A Scoping Review. Cancer Epidemiol biomarkers Prev a Publ Am Assoc Cancer Res cosponsored by Am Soc Prev Oncol. 2023 Jul;
38. Cigarini F, Daolio J, Caviola G, Pellegri C, Cavuto S, Guberti M, et al. Impact of COVID-19 on cancer care pathways in a comprehensive cancer center in northern Italy. Front public Heal. 2023;11:1187912.
39. Elanko A, Khan J, Hamady ZZ, Malik H. Cancer surgery sustainability in the light of COVID-19 pandemic. Eur J Surg Oncol. 2020 Jun 1;46(6):1174–5.
40. Nicolas Revel directeur général de la Caisse nationale de l'assurance maladie (CNAM). Hearing before the Commission des Affaires Sociales, Sénat(15 April, 2020). 2020.
41. Jazieh AR, Akbulut H, Curigliano G, Rogado A, Alsharm AA, Razis ED, et al. Impact of the COVID-19 Pandemic on Cancer Care: A Global Collaborative Study. JCO Glob Oncol. 2020 Nov 28;(6):1428–38.
42. Nab M, Vehmendahl R van, Somers I, Schoon Y, Hesselink G. Delayed emergency healthcare seeking behaviour by Dutch emergency department visitors during the first COVID-19 wave: a mixed methods retrospective observational study. BMC Emerg Med 2021 211 [Internet]. 2021 May 1 [cited 2021 Jul 16];21(1):1–9. Available from: https://bmcemergmed.biomedcentral.com/articles/10.1186/s12873-021-00449-9
43. Adelhoefer S, Berning P, Solomon SB, Maybody M, Whelton SP, Blaha MJ, et al. Decreased public pursuit of cancer-related information during the COVID-19 pandemic in the United States. Cancer Causes Control [Internet]. 2021 Jun 1 [cited 2021 Jul 16];32(6):577–85. Available from: https://pubmed.ncbi.nlm.nih.gov/33683506/
44. Blay JY, Boucher S, Vu B Le, Cropet C, Chabaud S, Perol D, et al. Delayed care for patients with newly diagnosed cancer due to COVID-19 and estimated impact on cancer mortality in France. ESMO Open [Internet]. 2021 Jun 1 [cited 2021 Jul 16];6(3):100134. Available from: http://www.esmoopen.com/article/S2059702921000934/fulltext
45. Curigliano G, Banerjee S, Cervantes A, Garassino MC, Garrido P, Girard N, et al. Managing cancer patients during the COVID-19 pandemic: an ESMO multidisciplinary expert consensus. Ann Oncol [Internet]. 2020 Oct 1 [cited 2021 Jul 16];31(10):1320–35. Available from: http://www.annalsofoncology.org/article/S0923753420399488/fulltext
46. Jazieh AR, Chan SL, Curigliano G, Dickson N, Eaton V, Garcia-Foncillas J, et al. Delivering Cancer Care During the COVID-19 Pandemic: Recommendations and Lessons Learned From ASCO Global Webinars. https://doi.org/101200/GO2000423. 2020 Sep 30;(6):1461–71.
47. American Society of Clinical Oncology. Cancer screening, diagnosis, staging, and surveillance. Updated June 22, 2020. Accessed July 13, 2020. [Internet]. Available from: https://www.asco.org/asco-coronavirus-resources/care-individuals-cancer-during- covid-19/cancer-screening-diagnosis-staging 2.
48. Communiqué de presse INCa. Organisation de la reprise des traitements et des dépistages des cancers : assurer la qualité et la sécurité des parcours de soins pour les patients en tenant compte des situations territoriales. 2020.
49. Ministère des Solidarités et de la Santé. Continuité des activités des Centres régionaux de coordination des dépistages des cancers (CRCDC). Inst Natl du cancer. 2020;
50. Weill A, Drouin J, Desplas D, Cuenot F, Dray-Spira R, Zureik M. Usage des médicaments de ville en France durant l ' épidémie de la Covid-19 – point de situation jusqu'au 13 septembre 2020. EPI-PHARE GIS ANSM - CNAM [Internet]. 2020; Available from: https://www.ansm.sante.fr/S-informer/Points-d-information-Points-d-information/Usage-

des-medicaments-de-ville-en-France-durant-l-epidemie-de-Covid-19-point-de-situation-a-la-fin-du-confinement-Point-d-Information
51. Hanna TP, King WD, Thibodeau S, Jalink M, Paulin GA, Harvey-Jones E, et al. Mortality due to cancer treatment delay: systematic review and meta-analysis. BMJ. 2020 Nov;371:m4087.
52. Hartman HE, Sun Y, Devasia TP, Chase EC, Jairath NK, Dess RT, et al. Integrated Survival Estimates for Cancer Treatment Delay Among Adults With Cancer During the COVID-19 Pandemic. JAMA Oncol [Internet]. 2020;48109:1–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/33119036
53. Maringe C, Spicer J, Morris M, Purushotham A, Nolte E, Sullivan R, et al. The impact of the COVID-19 pandemic on cancer deaths due to delays in diagnosis in England, UK: a national, population-based, modelling study. Lancet Oncol. 2020 Aug 1;21(8):1023–34.
54. Schork NJ. Artificial Intelligence and Personalized Medicine. Cancer Treat Res. 2019;178:265–83.
55. Aerts HJWL. The Potential of Radiomic-Based Phenotyping in Precision Medicine: A Review. JAMA Oncol. 2016 Dec;2(12):1636–42.
56. Région Ile de France. Un plan pour faire de l'Île-de-France la 1re région européenne en Intelligence artificielle [Internet]. 2018. Available from: https://www.iledefrance.fr/un-plan-pour-faire-de-lile-de-france-la-1re-region-europeenne-en-intelligence-artificielle
57. Région Ile de France. Stratégie Smart Santé Paris Région 2020-2022. 2020.
58. Assistance Publique - Hôpitaux de Paris. La Région Île-de-France lance le « AI for Health Challenge 2020 » : deuxième Challenge international sur l'oncologie avec l'AP-HP et l'Institut Curie [Internet]. Dossier de presse. 2020. Available from: https://www.aphp.fr/contenu/la-region-ile-de-france-lance-le-ai-health-challenge-2020-deuxieme-challenge-international
59. Banales JM, Marin JJG, Lamarca A, Rodrigues PM, Khan SA, Roberts LR, et al. Cholangiocarcinoma 2020: the next horizon in mechanisms and management. Nat Rev Gastroenterol Hepatol. 2020 Sep;17(9):557–88.
60. Kirstein MM, Vogel A. Epidemiology and Risk Factors of Cholangiocarcinoma. Visc Med. 2016 Dec;32(6):395–400.
61. Pastorino R, Vito C De, Migliara G, Glocker K, Binenbaum I, Ricciardi W, et al. Benefits and challenges of Big Data in healthcare : an overview of the European initiatives. 2019;29:23–7.
62. Dougoud-Chauvin V, Lee JJ, Santos E, Williams VL, Battisti NML, Ghia K, et al. Using Big Data in oncology to prospectively impact clinical patient care: A proof of concept study. J Geriatr Oncol. 2018 Nov;9(6):665–72.
63. Tsai CJ, Riaz N, Gomez SL. Big Data in Cancer Research: Real-World Resources for Precision Oncology to Improve Cancer Care Delivery. Semin Radiat Oncol. 2019 Oct;29(4):306–10.
64. Barker AD, Lee JSH. Translating "Big Data" in Oncology for Clinical Benefit: Progress or Paralysis. Cancer Res. 2022 Jun;82(11):2072–5.
65. Hong N, Zhang N, Wu H, Lu S, Yu Y, Hou L, et al. Preliminary exploration of survival analysis using the OHDSI common data model: a case study of intrahepatic cholangiocarcinoma. BMC Med Inform Decis Mak [Internet]. 2018 Dec 7 [cited 2022 May 23];18(Suppl 5). Available from: https://pubmed.ncbi.nlm.nih.gov/30526572/
66. Gao P, Shen X, Zhang X, Jiang C, Zhang S, Zhou X, et al. Precision environmental health monitoring by longitudinal exposome and multi-omics profiling. Genome Res. 2022 Jun;32(6):1199–214.
67. Cuggia M, Besana P, Glasspool D. Comparing semi-automatic systems for recruitment of patients to clinical trials. Int J Med Inform [Internet]. 2011 [cited 2021 Nov 23];80(6):371–88. Available from: https://pubmed.ncbi.nlm.nih.gov/21459664/
68. Dugas M, Lange M, Müller-Tidow C, Kirchhof P, Prokosch HU. Routine data from hospital information systems can support patient recruitment for clinical studies. Clin Trials [Internet]. 2010 Apr [cited 2021 Nov 23];7(2):183–9. Available from: https://pubmed.ncbi.nlm.nih.gov/20338903/
69. Griffon N, Pereira H, Djadi-Prat J, García MT, Testoni S, Cariou M, et al. Performances of a Solution to Semi-Automatically Fill eCRF with Data from the Electronic Health Record: Protocol for a Prospective Individual Participant Data Meta-Analysis. Stud Health Technol Inform [Internet]. 2020 Jun 16 [cited 2021 Nov 23];270:367–71. Available from: https://pubmed.ncbi.nlm.nih.gov/32570408/
70. Nordo AH, Levaux HP, Becnel LB, Galvez J, Rao P, Stem K, et al. Use of EHRs data for clinical research: Historical progress and current applications. Learn Heal Syst [Internet]. 2019 Jan 1 [cited 2021 Nov 23];3(1). Available from: /pmc/articles/PMC6508843/
71. Köpcke F, Trinczek B, Majeed RW, Schreiweis B, Wenk J, Leusch T, et al. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. BMC Med Inform Decis Mak [Internet]. 2013 [cited 2021 Nov 23];13(1). Available from: https://pubmed.ncbi.nlm.nih.gov/23514203/
72. Schreiweis B, Trinczek B, Köpcke F, Leusch T, Majeed RW, Wenk J, et al. Comparison of Electronic Health Record System Functionalities to support the patient recruitment process in clinical trials. Int J Med Inform [Internet]. 2014 Nov 1 [cited 2021 Nov 23];83(11):860–8. Available from: https://pubmed.ncbi.nlm.nih.gov/25189709/
73. Claerhout B, Kalra D, Mueller C, Singh G, Ammour N, Meloni L, et al. Federated electronic health records research technology to support clinical trial protocol optimization: Evidence from EHR4CR and the InSite platform. J Biomed Inform [Internet]. 2019 Feb 1 [cited 2021 Nov 23];90. Available from: https://pubmed.ncbi.nlm.nih.gov/30611012/
74. Kush R, Alschuler L, Ruggeri R, Cassells S, Gupta N, Bain L, et al. Implementing Single Source: The STARBRITE Proof-of-Concept Study. J Am Med Inform Assoc [Internet]. 2007 Sep [cited 2021 Nov 23];14(5):662. Available from: /pmc/articles/PMC1975790/

75. Murphy EC, Ferris FL, O'Donnell WR. An Electronic Medical Records System for Clinical Research and the EMR–EDC Interface. Invest Ophthalmol Vis Sci [Internet]. 2007 Oct [cited 2021 Nov 23];48(10):4383. Available from: /pmc/articles/PMC2361387/
76. El Fadly AN, Rance B, Lucas N, Mead C, Chatellier G, Lastic PY, et al. Integrating clinical research with the Healthcare Enterprise: from the RE-USE project to the EHR4CR platform. J Biomed Inform [Internet]. 2011 Dec [cited 2021 Nov 23];44 Suppl 1(SUPPL. 1). Available from: https://pubmed.ncbi.nlm.nih.gov/21888989/
77. Ethier JF, Curcin V, McGilchrist MM, Choi Keung SNL, Zhao L, Andreasson A, et al. eSource for clinical trials: Implementation and evaluation of a standards-based approach in a real world trial. Int J Med Inform [Internet]. 2017 Oct 1 [cited 2021 Nov 23];106:17–24. Available from: https://pubmed.ncbi.nlm.nih.gov/28870379/
78. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, et al. Using electronic health records for clinical research: the case of the EHR4CR project. J Biomed Inform [Internet]. 2015 Feb 1 [cited 2021 Nov 23];53:162–73. Available from: https://pubmed.ncbi.nlm.nih.gov/25463966/
79. Chow CJ, Habermann EB, Abraham A, Zhu Y, Vickers SM, Rothenberger DA, et al. Does Enrollment in Cancer Trials Improve Survival? J Am Coll Surg. 2008;216(4):774–81.
80. Dugas M, Lange M, Berdel WE, Müller-Tidow C. Workflow to improve patient recruitment for clinical trials within hospital information systems - A case-study. Trials. 2008 Jan 11;9.
81. Köpcke F, Kraus S, Scholler A, Nau C, Schüttler J, Prokosch HU, et al. Secondary use of routinely collected patient data in a clinical trial: An evaluation of the effects on patient recruitment and data acquisition. Int J Med Inform [Internet]. 2013 Mar [cited 2022 Oct 3];82(3):185–92. Available from: https://pubmed.ncbi.nlm.nih.gov/23266063/
82. Chubak J, Ziebell R, Greenlee RT, Honda S, Hornbrook MC, Epstein M, et al. The Cancer Research Network: a platform for epidemiologic and health services research on cancer prevention, care, and outcomes in large, stable populations. Cancer Causes Control. 2016 Nov;27(11):1315–23.
83. Lacey JVJ, Chung NT, Hughes P, Benbow JL, Duffy C, Savage KE, et al. Insights from Adopting a Data Commons Approach for Large-scale Observational Cohort Studies: The California Teachers Study. Cancer Epidemiol biomarkers Prev a Publ Am Assoc Cancer Res cosponsored by Am Soc Prev Oncol. 2020 Apr;29(4):777–86.
84. Seneviratne MG, Seto T, Blayney DW, Brooks JD, Hernandez-Boussard T. Architecture and Implementation of a Clinical Research Data Warehouse for Prostate Cancer. EGEMS (Washington, DC). 2018 Jun;6(1):13.
85. Woo C, Cioffi GN, Bej TA, Wilson B, Briggs JM, Markt SC, et al. Data Matching to Support Analysis of Cancer Epidemiology Among Veterans Compared With Non-Veteran Populations-An Exemplar in Brain Tumors. JCO Clin cancer informatics. 2021 Sep;5:985–94.
86. Eschrich SA, Teer JK, Reisman P, Siegel E, Challa C, Lewis P, et al. Enabling Precision Medicine in Cancer Care Through a Molecular Data Warehouse: The Moffitt Experience. JCO Clin cancer informatics. 2021 May;5:561–9.
87. Adlung L, Cohen Y, Mor U, Elinav E. Machine learning in clinical decision making. Med (New York, NY). 2021 Jun;2(6):642–65.
88. He T, Puppala M, Ezeana CF, Huang Y-S, Chou P-H, Yu X, et al. A Deep Learning-Based Decision Support Tool for Precision Risk Assessment of Breast Cancer. JCO Clin cancer informatics. 2019 May;3:1–12.
89. Kempf E, Azria E, Kempf A. Computer-based risk prediction models: Ethical issues of Adjuvant! Online use in early-stage breast cancerOutils informatiques de prédiction de risque: Les enjeux éthiques d'Adjuvant ! Online dans le cancer du sein localis. Gynecol Obstet Fertil. 2016;44(2).
90. Doutreligne M, Degremont A, Jachiet P-A, Lamer A, Tannier X. Good practices for clinical data warehouse implementation: A case study in France. PLOS Digit Heal. 2023 Jul;2(7):e0000298.
91. Sweeney SM, Hamadeh HK, Abrams N, Adam SJ, Brenner S, Connors DE, et al. Challenges to Using Big Data in Cancer Box 1 : Patient Perspective by Access to Data : The Role of Privacy New Research Model : :1175–82.
92. Price WN 2nd, Cohen IG. Privacy in the age of medical big data. Nat Med. 2019 Jan;25(1):37–43.
93. Berman JJ. Confidentiality issues for medical data miners. Artif Intell Med. 2002;26(1–2):25–36.
94. Avis commun du CCNE et du CNPEN, Avis 143 du CCNE, Avis 5 du CNPEN. Plateformes de données de santé : enjeux d'éthique. Com Consult Natl d'Ethique pour les Sci la vie la santé. 2023;1–73.
95. Hecht group C. How Much Does A Clinical Data Warehouse Cost? [Internet]. 2022. Available from: https://www.hechtgroup.com/how-much-does-a-clinical-data-warehouse-cost/#google_vignette
96. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data [Internet]. 2016;3(1):160018. Available from: https://doi.org/10.1038/sdata.2016.18
97. Roland M, Torgerson DJ. What are pragmatic trials ? Br Med J (Clin Res Ed). 1998;316(7127):285.
98. Bocquet F, Campone M, Cuggia M. The Challenges of Implementing Comprehensive Clinical Data Warehouses in Hospitals. Vol. 19, International journal of environmental research and public health. Switzerland; 2022.
99. Bocquet F, Raimbourg J, Bigot F, Simmet V, Campone M, Frenel J-S. Opportunities and Obstacles to the Development of Health Data Warehouses in Hospitals in France: The Recent Experience of Comprehensive Cancer Centers. Int J Environ Res Public Health. 2023 Jan;20(2).
100. Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N. Initializing a hospital-wide data quality program. The AP-HP experience. Comput Methods Programs Biomed [Internet]. 2019 Nov 1 [cited 2021 Nov 23];181. Available from: https://pubmed.ncbi.nlm.nih.gov/30497872/
101. Grimberg F, Asprion PM, Schneider B, Miho E, Babrak L, Habbabeh A. The Real-World Data Challenges Radar: A Review on the Challenges and Risks regarding the Use of Real-World Data. Digit biomarkers. 2021;5(2):148–57.

102. Syed R, Eden R, Makasi T, Chukwudi I, Mamudu A, Kamalpour M, et al. Digital Health Data Quality Issues: Systematic Review. J Med Internet Res. 2023 Mar;25:e42615.
103. Bian J, Lyu T, Loiacono A, Viramontes TM, Lipori G, Guo Y, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. J Am Med Inform Assoc. 2020 Dec;27(12):1999–2010.
104. Cooper GS, Yuan Z, Stange KC, Amini SB, Dennis LK, Rimm AA. The utility of Medicare claims data for measuring cancer stage. Med Care. 1999 Jul;37(7):706–11.
105. Nordstrom BL, Whyte JL, Stolar M, Mercaldi C, Kallich JD. Identification of metastatic cancer in claims data. Pharmacoepidemiol Drug Saf. 2012 May;21 Suppl 2:21–8.
106. Priou S, Lame G, Jankovic M, Chatellier G, Bey R, Tournigand C, et al. Why Are Data Missing in Clinical Data Warehouses? A Simulation Study of How Data Are Processed (and Can Be Lost). Stud Health Technol Inform. 2023 May;302:202–6.
107. Yang DX, Khera R, Miccio JA, Jairam V, Chang E, Yu JB, et al. Prevalence of Missing Data in the National Cancer Database and Association With Overall Survival. JAMA Netw open. 2021 Mar;4(3):e211793.
108. Diaz-Garelli F, Strowd R, Lawson VL, Mayorga ME, Wells BJ, Lycan TWJ, et al. Workflow Differences Affect Data Accuracy in Oncologic EHRs: A First Step Toward Detangling the Diagnosis Data Babel. JCO Clin cancer informatics. 2020 Jun;4:529–38.
109. Aiello Bowles EJ, Tuzzio L, Ritzwoller DP, Williams AE, Ross T, Wagner EH, et al. Accuracy and complexities of using automated clinical data for capturing chemotherapy administrations: implications for future research. Med Care. 2009 Oct;47(10):1091–7.
110. Carroll NM, Burniece KM, Holzman J, McQuillan DB, Plata A, Ritzwoller DP. Algorithm to Identify Systemic Cancer Therapy Treatment Using Structured Electronic Data. JCO Clin cancer informatics. 2017 Nov;1:1–9.
111. Lauzanne O, Frenel J-S, Baziz M, Campone M, Raimbourg J, Bocquet F. Optimizing the Retrieval of the Vital Status of Cancer Patients for Health Data Warehouses by Using Open Government Data in France. Int J Environ Res Public Health. 2022 Apr;19(7).
112. Joubert M, Dufour J-C, Falco L, Aymard S, Fieschi M. Towards interoperability of heterogeneous health databases: application to a tumor samples bank. Stud Health Technol Inform. 2004;107(Pt 2):1251–5.
113. Gagalova KK, Leon Elizalde MA, Portales-Casamar E, Görges M. What You Need to Know Before Implementing a Clinical Research Data Warehouse: Comparative Review of Integrated Data Repositories in Health Care Institutions. JMIR Form Res. 2020 Aug;4(8):e17687.
114. Alonso-Calvo R, Perez-Rey D, Paraiso-Medina S, Claerhout B, Hennebert P, Bucur A. Enabling semantic interoperability in multi-centric clinical trials on breast cancer. Comput Methods Programs Biomed. 2015 Mar;118(3):322–9.
115. London JW, Fazio-Eynullayeva E, Palchuk MB, Sankey P, McNair C. Effects of the COVID-19 Pandemic on Cancer-Related Patient Encounters. JCO Clin Cancer Informatics. 2020 Jul 27;(4):657–65.
116. Commission européenne. Règlement du parlement européen et du conseil relatif à l'espace européen des données de santé [Internet]. 2022. Available from: https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:52022PC0197&from=EN
117. Combes S, Bacry E, Fontbonne C. [Health Data Hub in France, use cases in oncology and radiation oncology]. Cancer Radiother J la Soc Fr Radiother Oncol. 2020 Oct;24(6–7):762–7.
118. Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Vol. 57, Methods of information in medicine. Germany; 2018. p. e50–6.
119. Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009;16(5):624–30.
120. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby J V, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc. 2014;21(4):578–82.
121. Puttmann D, de Groot R, de Keizer N, Cornet R, Elbers PWG, Dongelmans D, et al. Assessing the FAIRness of databases on the EHDEN portal: A case study on two Dutch ICU databases. Int J Med Inform. 2023 Aug;176:105104.
122. Observational Health Data Sciences and Informatics (OHDSI) [Internet]. Available from: https://ohdsi.org
123. SHRINE (USA) [Internet]. Available from: https://www.i2b2.org/work/shrine.html
124. PCORnet (USA) [Internet]. Available from: https://pcornet.org/
125. https://www.health-data-hub.fr/ [Internet]. Available from: https://www.health-data-hub.fr/
126. https://www.medizininformatik-initiative.de [Internet]. Available from: https://www.medizininformatik-initiative.de
127. https://www.ehden.eu/ [Internet]. Available from: https://www.ehden.eu/
128. Lemordant P, Bouzille G, Mathieu R, Thenault R, Gibaud B, Garde C, et al. How to Optimize Connection Between PACS and Clinical Data Warehouse: A Web Service Approach Based on Full Metadata Integration. Stud Health Technol Inform. 2022 Jun;290:27–31.
129. Shin SJ, You SC, Park YR, Roh J, Kim J-H, Haam S, et al. Genomic Common Data Model for Seamless Interoperation of Biomedical Data in Clinical Practice: Retrospective Study. J Med Internet Res. 2019 Mar;21(3):e13249.
130. Belenkaya R, Gurley MJ, Golozar A, Dymshyts D, Miller RT, Williams AE, et al. Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. JCO Clin Cancer Informatics

[Internet]. 2021 Jan [cited 2022 May 9];5(5):12–20. Available from: https://pubmed.ncbi.nlm.nih.gov/33411620/
131. Osterman TJ, Terry M, Miller RS. Improving Cancer Data Interoperability : The Promise of the Minimal Common Oncology Data Elements (mCODE) Initiative. JCO Clin Cancer Inf. 2020;993–1001.
132. The Initiative to Create a Core Cancer Model and Foundational EHR Data Elements. mCODE™: Minimal Common Oncology Data Elements [Internet]. Available from: https://mcodeinitiative.org/
133. mCODE dataset [Internet]. Available from: https://mcodeinitiative.org/
134. Foran DJ, Chen W, Chu H, Sadimin E, Loh D, Riedlinger G, et al. Roadmap to a Comprehensive Clinical Data Warehouse for Precision Medicine Applications in Oncology. Cancer Inform. 2017;16:1176935117694349.
135. Jung HA, Jeong O, Chang DK, Park S, Sun J-M, Lee S-H, et al. Real-time autOmatically updated data warehOuse in healThcare (ROOT): an innovative and automated data collection system. Transl lung cancer Res. 2021 Oct;10(10):3865–74.
136. Savova GK, Danciu I, Alamudun F, Miller T, Lin C, Bitterman DS, et al. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records [Internet]. Vol. 79, Cancer Research. American Association for Cancer Research Inc.; 2019 [cited 2020 Oct 7]. p. 5463–70. Available from: https://cancerres.aacrjournals.org/content/79/21/5463
137. Saha A, Burns L, Kulkarni AM. A scoping review of natural language processing of radiology reports in breast cancer. Vol. 13, Frontiers in oncology. Switzerland; 2023. p. 1160167.
138. Hanauer DA, Barnholtz-Sloan JS, Beno MF, Del Fiol G, Durbin EB, Gologorskaya O, et al. Electronic Medical Record Search Engine (EMERSE): An Information Retrieval Tool for Supporting Cancer Research. JCO Clin Cancer Informatics [Internet]. 2020 Sep [cited 2020 Oct 7];4(4):454–63. Available from: https://pubmed.ncbi.nlm.nih.gov/32412846/
139. Baclic O, Tunis M, Young K, Doan C, Swerdfeger H, Schonfeld J. Challenges and opportunities for public health made possible by advances in natural language processing. Can Commun Dis Rep. 2020 Jun;46(6):161–8.
140. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ. 2015;350:1–6.
141. Petch J, Kempainnen J, Pettengell C, Aviv S, Butler B, Pond G, et al. Developing a Data and Analytics Platform to Enable a Breast Cancer Learning Health System at a Regional Cancer Center. JCO Clin cancer informatics. 2023 Mar;7:e2200182.
142. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Heal Informatics [Internet]. 2018 Sep 1 [cited 2020 Feb 19];22(5):1589–604. Available from: https://pubmed.ncbi.nlm.nih.gov/29989977/
143. Vincent M, Douillet M, Lerner I, Neuraz A, Burgun A, Garcelon N. Using Deep Learning to Improve Phenotyping from Clinical Reports. Stud Health Technol Inform [Internet]. 2022 Jun 6 [cited 2022 Dec 29];290:282–6. Available from: https://pubmed.ncbi.nlm.nih.gov/35673018/
144. Kehl KL, Xu W, Gusev A, Bakouny Z, Choueiri TK, Riaz I Bin, et al. Artificial intelligence-aided clinical annotation of a large multi-cancer genomic dataset. Nat Commun. 2021 Dec;12(1):7304.
145. Sarwar T, Seifollahi S, Chan J, Zhang X, Aksakalli V, Hudson I, et al. The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges. ACM Comput Surv. 2023;55(2).
146. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl. 2023;82(3):3713–44.
147. Rezaeenour J, Ahmadi M, Jelodar H, Shahrooei R. Systematic review of content analysis algorithms based on deep neural networks. Multimed Tools Appl. 2023;82(12):17879–903.
148. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: A methodical review [Internet]. Vol. 27, Journal of the American Medical Informatics Association. Oxford University Press; 2020 [cited 2020 Oct 7]. p. 457–70. Available from: https://academic.oup.com/jamia/article/27/3/457/5651084
149. Yuan C, Xie Q, Ananiadou S. Zero-shot Temporal Relation Extraction with ChatGPT. In: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks [Internet]. Toronto, Canada: Association for Computational Linguistics; 2023. p. 92–102. Available from: https://aclanthology.org/2023.bionlp-1.7
150. Jahan I, Laskar MTR, Peng C, Huang J. Evaluation of ChatGPT on Biomedical Tasks: A Zero-Shot Comparison with Fine-Tuned Generative Transformers. In: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks [Internet]. Toronto, Canada: Association for Computational Linguistics; 2023. p. 326–36. Available from: https://aclanthology.org/2023.bionlp-1.30
151. Carrell DS, Schoen RE, Leffler DA, Morris M, Rose S, Baer A, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. J Am Med Inform Assoc. 2017 Sep;24(5):986–91.
152. Roelofs E, Persoon L, Nijsten S, Wiessler W, Dekker A, Lambin P. Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. Radiother Oncol. 2013 Jul;108(1):174–9.
153. Hossain E, Rana R, Higgins N, Soar J, Barua PD, Pisani AR, et al. Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review. Comput Biol Med. 2023 Mar;155:106649.
154. Spasic I, Nenadic G. Clinical Text Data in Machine Learning: Systematic Review. JMIR Med informatics. 2020 Mar;8(3):e17984.
155. Tan HX, Teo CHD, Ang PS, Loke WPC, Tham MY, Tan SH, et al. Combining Machine Learning with a Rule-Based Algorithm to Detect and Identify Related Entities of Documented Adverse Drug Reactions on Hospital Discharge Summaries. Drug Saf. 2022 Aug;45(8):853–62.

156. Chang YC, Dai HJ, Wu JCY, Chen JM, Tsai RTH, Hsu WL. TEMPTING system: A hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries. J Biomed Inform. 2013;46(SUPPL.).
157. Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, et al. An end-to-end hybrid algorithm for automated medication discrepancy detection. BMC Med Inform Decis Mak. 2015 May;15:37.
158. Ford I, Ph D, Norrie J, Sc M. Pragmatic Trials. N Engl J Med. 2016;375(5):454–63.
159. ASCO, NCI. Clinical Oncology Requirements for the EHR [Internet]. American Society of Clinical Oncology. 2009. p. 1–26. Available from: papers3://publication/uuid/7DC6400E-D889-47F3-A2BC-09936022A0D8
160. Rollet Q, Bouvier V, Moutel G, Launay L, Bignon A-L, Bouhier-Leporrier K, et al. Multidisciplinary team meetings: are all patients presented and does it impact quality of care and survival – a registry-based study. BMC Health Serv Res [Internet]. 2021;21(1):1032. Available from: https://doi.org/10.1186/s12913-021-07022-x
161. Lin T, Pham J, Paul E, Conron M, Wright G, Ball D, et al. Impacts of Multidisciplinary Meeting Presentation: Drivers and Outcomes from a Population Registry Retrospective Cohort study. J Thorac Oncol [Internet]. 2021;16(3, Supplement):S146–7. Available from: https://www.sciencedirect.com/science/article/pii/S1556086421002720
162. Gérvas J, Pérez Fernández M. Minimum basic data set in general practice: definitions and coding. Fam Pract. 1992 Sep;9(3):349–52.
163. Fernández-Navarro P, López-Abente G, Salido-Campos C, Sanz-Anquela JM. The Minimum Basic Data Set (MBDS) as a tool for cancer epidemiological surveillance. Eur J Intern Med. 2016 Oct;34:94–7.
164. Choquet R, Maaroufi M, De Carrara A, Messiaen C, Luigi E, Landais P. A methodology for a minimum data set for rare diseases to support national centers of excellence for healthcare and research. J Am Med Informatics Assoc. 2015;22(1):76–85.
165. Daniel C. 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. Rev Med Interne [Internet]. 2020 May 1 [cited 2020 Oct 20];41(5):303–7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7164890/
166. Daniel C, Salamanca E. Les bases de données des hôpitaux. L'Entrepôt de données de l'AP–HP. In: Villani C;, Nordlinglinger B, editors. Santé et intelligence artificielle. CNRS. 1996. p. CNRS.
167. Greater Paris University Hospitals. Cohort360 - A web application to find patients, build cohorts and visualize health records [Internet]. Available from: https://github.com/aphp/Cohort360
168. Méthodologie nationale de référence MR-004 régissant les traitements de données à caractère personnel à des fins de recherche d'intérêt public n'impliquant pas la personne humaine [Internet]. Available from: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037187498
169. mCODE [Internet]. Available from: http://hl7.org/fhir/us/mcode/
170. Bhatt N, Cotter MM. Breast cancer pathology reporting: an audit to compare standards with minimum datasets in a district general hospital. Vol. 14, The breast journal. United States; 2008. p. 118–9.
171. Maughan NJ, Morris E, Forman D, Quirke P. The validity of the Royal College of Pathologists' colorectal cancer minimum dataset within a population. Br J Cancer. 2007 Nov;97(10):1393–8.
172. Delahunt B, Srigley JR, Judge M, Amin M, Billis A, Camparo P, et al. Dataset for the reporting of renal biopsy for tumour: recommendations from the International Collaboration on Cancer Reporting (ICCR). J Clin Pathol. 2019 Sep;72(9):573–8.
173. Loughrey MB, Webster F, Arends MJ, Brown I, Burgart LJ, Cunningham C, et al. Dataset for Pathology Reporting of Colorectal Cancer: Recommendations From the International Collaboration on Cancer Reporting (ICCR). Ann Surg. 2022 Mar;275(3):e549–61.
174. Kaur MR, Colloby PS, Martin-Clavijo A, Marsden JR. Melanoma histopathology reporting: are we complying with the National Minimum Dataset? J Clin Pathol. 2007 Oct;60(10):1121–3.
175. Lewis JSJ, Adelstein DJ, Agaimy A, Carlson DL, Faquin WC, Helliwell T, et al. Data Set for the Reporting of Carcinomas of the Nasopharynx and Oropharynx: Explanations and Recommendations of the Guidelines From the International Collaboration on Cancer Reporting. Arch Pathol Lab Med. 2019 Apr;143(4):447–51.
176. Institut National du Cancer. Mise à jour 2011 des comptes rendus d'anatomopathologie : données minimales à renseigner pour une tumeur primitive. 2011;
177. ASIP Santé. Volet Compte Rendu Structuré d'Anatomie et de Cytologie Pathologiques (CR-ACP) Spécifications fonctionnelles. 2018;1–36.
178. Stone E, Rankin N, Phillips J, Fong K, Currow DC, Miller A, et al. Consensus minimum data set for lung cancer multidisciplinary teams: Results of a Delphi process. Respirology. 2018 Oct;23(10):927–34.
179. Sigurdardottir KR, Kaasa S, Rosland JH, Bausewein C, Radbruch L, Haugen DF. The European Association for Palliative Care basic dataset to describe a palliative care cancer population: Results from an international Delphi process. Palliat Med. 2014 Jun;28(6):463–73.
180. Pheby DF, Etherington DJ. Improving the comparability of cancer registry treatment data and proposals for a new national minimum dataset. J Public Health Med. 1994 Sep;16(3):331–40.
181. Cheong CM, Golder AM, Horgan PG, McMillan DC, Roxburgh CSD. Evaluation of clinical prognostic variables on short-term outcome for colorectal cancer surgery: An overview and minimum dataset. Cancer Treat Res Commun. 2022;31:100544.
182. Milani A, Mauri S, Gandini S, Magon G. Oncology Nursing Minimum Data Set (ONMDS): can we hypothesize a set of prevalent Nursing Sensitive Outcomes (NSO) in cancer patients? Ecancermedicalscience. 2013;7:345.
183. Guérin J, Laizet Y, Le Texier V, Chanas L, Rance B, Koeppel F, et al. OSIRIS: A Minimum Data Set for Data Sharing and

Interoperability in Oncology. JCO Clin Cancer Informatics. 2021;(5):256–65.
184. HL7 FHIR. The Vulcan Real World Data project [Internet]. 2023. Available from: http://hl7.org/fhir/uv/vulcan-rwd/#overview
185. CodeX [Internet]. Available from: https://codex.hl7.org/
186. Benda L, Charles C, Rimaud G, Vlaar T, Dia N, Menu-Branthomme A, et al. Health Data Hub et interopérabilité du Système national des données de santé. Rev Epidemiol Sante Publique [Internet]. 2023;71:101457. Available from: https://www.sciencedirect.com/science/article/pii/S0398762023000354
187. Belenkaya R, Gurley M, Dymshyts D, Araujo S, Williams A, Chen RJ, et al. Standardized Observational Cancer Research Using the OMOP CDM Oncology Module. Stud Health Technol Inform [Internet]. 2019 Aug 21 [cited 2022 Mar 3];264:1831–2. Available from: https://pubmed.ncbi.nlm.nih.gov/31438365/
188. Heudel P, Livartowski A, Arveux P, Willm E, Jamain C. [The ConSoRe project supports the implementation of big data in oncology]. Bull Cancer [Internet]. 2016 Nov 1 [cited 2022 Mar 7];103(11):949–50. Available from: https://pubmed.ncbi.nlm.nih.gov/27816168/
189. Melzer G, Maiwald T, Prokosch HU, Ganslandt T. Leveraging Real-World Data for the Selection of Relevant Eligibility Criteria for the Implementation of Electronic Recruitment Support in Clinical Trials. Appl Clin Inform [Internet]. 2021 Jan 1 [cited 2022 Oct 3];12(1):17–26. Available from: https://pubmed.ncbi.nlm.nih.gov/33440429/
190. Institut National du Cancer. OSIRIS - REAL WORLD DATA [Internet]. 2022. Available from: https://www.e-cancer.fr/Professionnels-de-la-recherche/Recherche-translationnelle/OSIRIS-projet-national-sur-le-partage-des-donnees
191. Li EC, D'Amato SL, Barr TR, Weisberg T. The "Big 6 Spotlight": Baseline assessment and implementation of a process to systematically collect critical oncology data elements for quality improvement and research. J Clin Oncol [Internet]. 2012 Dec 1;30(34_suppl):312. Available from: https://doi.org/10.1200/jco.2012.30.34_suppl.312
192. Fountzilas E, Tsimberidou AM. Overview of precision oncology trials: challenges and opportunities. Expert Rev Clin Pharmacol. 2018 Aug;11(8):797–804.
193. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: A literature review. J Biomed Inform [Internet]. 2010;43(3):451–67. Available from: http://dx.doi.org/10.1016/j.jbi.2009.12.004
194. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med. 2009;48(1):38–44.
195. Calaprice-Whitty D, Galil K, Salloum W, Zariv A, Jimenez B. Improving Clinical Trial Participant Prescreening With Artificial Intelligence (AI): A Comparison of the Results of AI-Assisted vs Standard Methods in 3 Oncology Trials. Ther Innov Regul Sci [Internet]. 2020 Jan 1 [cited 2022 Mar 9];54(1):69–74. Available from: https://pubmed.ncbi.nlm.nih.gov/32008227/
196. Kukhareva P V., Weir C, Del Fiol G, Aarons GA, Taft TY, Schlechter CR, et al. Evaluation in Life Cycle of Information Technology (ELICIT) framework: Supporting the innovation life cycle from business case assessment to summative evaluation. J Biomed Inform [Internet]. 2022;127(November 2021):104014. Available from: https://doi.org/10.1016/j.jbi.2022.104014
197. Weng C, Shah NH, Hripcsak G. Deep phenotyping: Embracing complexity and temporality-Towards scalability, portability, and interoperability. Vol. 105, Journal of biomedical informatics. United States; 2020. p. 103433.
198. Sim I, Olasov B, Carini S. The Trial Bank system: capturing randomized trials for evidence-based medicine. AMIA Annu Symp Proc. 2003;7(2):1076.
199. Shankar RD, Martins SB, O'Connor MJ, Parrish DB, Das AK. Epoch: An ontological framework to support clinical trials management. Proc HIKM 2006 Int Work Healthc Inf Knowl Manag. 2006;25–32.
200. Carini S, Sim I. SysBank: A Knowledge Base for Systematic Reviews of Randomized Clinical Trials. [cited 2022 Sep 12]; Available from: http://www.ai.sri.com/~gkb/
201. Commission européenne. EU Patient- cEntric clinicAl tRial pLatform [Internet]. 2019. Available from: https://cordis.europa.eu/project/id/853966
202. European University Hospital Alliance [Internet]. Available from: https://www.euhalliance.eu/
203. Observational Health Data Sciences and Informatics. ATHENA [Internet]. 2015. Available from: https://athena.ohdsi.org/search-terms/start
204. Kempf E, Vaterkowski M, Griffon N, Leprovost D, Breant S. How to improve cancer Patients ENrollment within clinical trials from rEal Life databases using the OMOP oncology Extension : the French PENELOPE initiative. AMIA 2022 Annu Symp. 2022;abstract N.
205. Kempf E, Vaterkowski M, Leprovost D, Griffon N, Ouagne D, Breant S, et al. How to Improve Cancer Patients ENrollment in Clinical Trials From rEal-Life Databases Using the Observational Medical Outcomes Partnership Oncology Extension: Results of the PENELOPE Initiative in Urologic Cancers. JCO Clin Cancer Informatics [Internet]. 2023 May 11;(7):e2200179. Available from: https://doi.org/10.1200/CCI.22.00179
206. OHDSI. OBSERVATION Table, OMOP common data model [Internet]. Available from: https://ohdsi.github.io/CommonDataModel/cdm53.html#OBSERVATION
207. Ateya MB, Delaney BC, Speedie SM. The value of structured data elements from electronic health records for identifying subjects for primary care clinical trials Healthcare Information Systems. BMC Med Inform Decis Mak. 2016 Jan 11;16(1).
208. von Minckwitz G, Huang C-S, Mano MS, Loibl S, Mamounas EP, Untch M, et al. Trastuzumab Emtansine for Residual Invasive HER2-Positive Breast Cancer. N Engl J Med [Internet]. 2019 Feb 14 [cited 2022 Oct 20];380(7):617–28.

Available from: https://www.nejm.org/doi/full/10.1056/nejmoa1814017
209. Ross J, Tu S, Carini S, Sim I. Analysis of Eligibility Criteria Complexity in Clinical Trials. Summit on Translat Bioinforma [Internet]. 2010 Mar 1 [cited 2022 Oct 3];2010:46. Available from: /pmc/articles/PMC3041539/
210. Lyu HG, Stein LA, Saadat L V, Phicil SN, Haider A, Raut CP. Assessment of the Accuracy of Disease Coding Among Patients Diagnosed With Sarcoma. JAMA Oncol. 2018 Sep;4(9):1293–5.
211. Abraha I, Montedori A, Serraino Di, Orso M, Giovannini G, Scotti V, et al. Accuracy of administrative databases in detecting primary breast cancer diagnoses: A systematic review. BMJ Open. 2018;8(7):1–18.
212. Fitzer K, Haeuslschmid R, Blasini R, Altun FB, Hampf C, Freiesleben S, et al. Patient Recruitment System for Clinical Trials: Mixed Methods Study About Requirements at Ten University Hospitals. JMIR Med informatics [Internet]. 2022 Apr 20 [cited 2022 Oct 17];10(4):e28696. Available from: http://www.ncbi.nlm.nih.gov/pubmed/35442203
213. Schreiweis B, Brandner A, Bergh B. Applicability of Different Electronic Record Types for Use in Patient Recruitment Support Systems: Comparative Analysis. JMIR Form Res 2021;5(9)e13790 https//formative.jmir.org/2021/9/e13790 [Internet]. 2021 Sep 21 [cited 2022 Oct 17];5(9):e13790. Available from: https://formative.jmir.org/2021/9/e13790
214. Schreiweis B, Bergh B. Requirements for a patient recruitment system [Internet]. Vol. 210, Studies in health technology and informatics. Stud Health Technol Inform; 2015 [cited 2022 Sep 26]. p. 521–5. Available from: https://pubmed.ncbi.nlm.nih.gov/25991202/
215. Trinczek B, Köpcke F, Leusch T, Majeed RW, Schreiweis B, Wenk J, et al. Design and Multicentric Implementation of a Generic Software Architecture for Patient Recruitment Systems Re-Using Existing HIS Tools and Routine Patient Data. Appl Clin Inform [Internet]. 2014 [cited 2022 Oct 3];5(1):264. Available from: /pmc/articles/PMC3974260/
216. Girardeau Y, Doods J, Zapletal E, Chatellier G, Daniel C, Burgun A, et al. Leveraging the EHR4CR platform to support patient inclusion in academic studies: challenges and lessons learned. BMC Med Res Methodol [Internet]. 2017 Feb 28 [cited 2021 Nov 23];17(1):36. Available from: https://pubmed.ncbi.nlm.nih.gov/28241798/
217. Wang AY, Lancaster WJ, Wyatt MC, Rasmussen L V., Fort DG, Cimino JJ. Classifying Clinical Trial Eligibility Criteria to Facilitate Phased Cohort Identification Using Clinical Data Repositories. AMIA . Annu Symp proceedings AMIA Symp. 2017;2017:1754–63.
218. Doods J, Lafitte C, Ulliac-Sagnes N, Proeve J, Botteri F, Walls R, et al. A European inventory of data elements for patient recruitment. Stud Health Technol Inform. 2015;210:506–10.
219. Bruland P, McGilchrist M, Zapletal E, Acosta D, Proeve J, Askin S, et al. Common data elements for secondary use of electronic health record data for clinical trial execution and serious adverse event reporting. BMC Med Res Methodol [Internet]. 2016 Nov 22 [cited 2022 Oct 3];16(1):1–10. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0259-3
220. Doods J, Botteri F, Dugas M, Fritz F. A European inventory of common electronic health record data elements for clinical trial feasibility. Trials [Internet]. 2014 Jan 10 [cited 2022 Oct 3];15(1). Available from: https://pubmed.ncbi.nlm.nih.gov/24410735/
221. Gulden C, Landerer I, Nassirian A, Altun FB, Johanna A. Extraction and Prevalence of Structured Data Elements in Free-Text Clinical Trial Eligibility Criteria. Stud Heal Technol Inf [Internet]. 2019 [cited 2022 Oct 17];258:226–30. Available from: https://pubmed.ncbi.nlm.nih.gov/30942751/
222. Vass A, Reinecke I, Boeker M, Prokosch HU, Gulden C. Availability of Structured Data Elements in Electronic Health Records for Supporting Patient Recruitment in Clinical Trials. Stud Health Technol Inform. 2022 Jun 6;290:130–4.
223. Cohen, Marlene Z ; Bagley Thompson, Cheryl ; Yates, Bernice ; Zimmerman, Lani ; Pullen CH. Implementing Common Data Elements Across Studies to Advance Research. Nurs Outlook 2015. 2015;63(2):181–8.
224. Ci B, Yang DM, Krailo M, Xia C, Yao B, Luo D, et al. Development of a Data Model and Data Commons for Germ Cell Tumors. JCO Clin Cancer Inform. 2020;4.
225. Dieter J, Dominick F, Knurr A, Ahlbrandt J, Ückert F. Analysis of Not Structurable Oncological Study Eligibility Criteria for Improved Patient-Trial Matching. Methods Inf Med. 2021 May 1;60(1–2):9–20.
226. Ahmadi N, Peng Y, Wolfien M, Zoch M, Sedlmayr M. OMOP CDM Can Facilitate Data-Driven Studies for Cancer Prediction: A Systematic Review. Int J Mol Sci [Internet]. 2022 Oct 5 [cited 2022 Oct 20];23(19):11834. Available from: https://pubmed.ncbi.nlm.nih.gov/36233137/
227. Yoo S, Yoon E, Boo D, Kim B, Kim S, Paeng JC, et al. Transforming Thyroid Cancer Diagnosis and Staging Information from Unstructured Reports to the Observational Medical Outcome Partnership Common Data Model. Appl Clin Inform [Internet]. 2022 May 1 [cited 2022 Oct 20];13(3):521. Available from: /pmc/articles/PMC9200482/
228. Jeon H, You SC, Kang SY, Seo SI, Warner JL, Belenkaya R, et al. Characterizing the Anticancer Treatment Trajectory and Pattern in Patients Receiving Chemotherapy for Cancer Using Harmonized Observational Databases: Retrospective Study. JMIR Med Inf [Internet]. 2021 Apr 1 [cited 2022 Oct 20];9(4):e25035–e25035. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8058693
229. Ryu B, Yoon E, Kim S, Lee S, Baek H, Yi S, et al. Transformation of Pathology Reports Into the Common Data Model With Oncology Module: Use Case for Colon Cancer. J Med Internet Res [Internet]. 2020 Dec 1 [cited 2022 May 23];22(12):e18526. Available from: https://pubmed.ncbi.nlm.nih.gov/33295294/
230. Ott, Simon ; Rinner, Christoph ; Duftschmid G. Expressing Patient Selection Criteria Based on HL7 V3 Templates Within the Open-Source Tool ART-DECOR. Stud Heal Technol Inf [Internet]. 2019 [cited 2022 Sep 26];260:226–33. Available from: https://pubmed.ncbi.nlm.nih.gov/31118342/
231. Augustinov G, Duftschmid G. Can the Austrian Nation-Wide EHR System Support the Recruitment of Trial Patients? Stud Heal Technol Inf. 2019;259:87–90.

232. Gulden C, Mate S, Prokosch HU KS, Gulden C, Mate S, Prokosch HU, Kraus S, Gulden C, Mate S, Prokosch HU KS. Investigating the Capabilities of FHIR Search for Clinical Trial Phenotyping. Stud Heal Technol Inf. 2018;253:3–7.
233. Gruendner J, Deppenwiese N, Folz M, Köhler T, Kroll B, Prokosch H, et al. Architecture for a privacy preserving feasibility query portal for distributed COVID-19 Fast Healthcare Interoperability Resources (FHIR) patient data repositories : Design and Implementation Study Table of Contents. 2022;
234. Rafee A, Riepenhausen S, Neuhaus P, Meidt A, Dugas M, Varghese J. ELaPro, a LOINC-mapped core dataset for top laboratory procedures of eligibility screening for clinical trials. BMC Med Res Methodol. 2022 Dec 1;22(1).
235. Chang E, Mostafa J. The use of SNOMED CT, 2013-2020: a literature review. J Am Med Inform Assoc. 2021 Aug;28(9):2017–26.
236. Botsis T, Murray JC, Ghanem P, Balan A, Kernagis A, Hardart K, et al. Precision Oncology Core Data Model to Support Clinical Genomics Decision Making. JCO Clin cancer informatics. 2023 Apr;7:e2200108.
237. Delorme J, Charvet V, Wartelle M, Lion F, Thuillier B, Mercier S, et al. Natural Language Processing for Patient Selection in Phase I or II Oncology Clinical Trials. JCO Clin Cancer Informatics [Internet]. 2021;(5):709–18. Available from: https://doi.org/10.1200/CCI.21.00003
238. Löbe M, Stäubert S, Goldberg C, Haffner I, Winter A. Towards Phenotyping of Clinical Trial Eligibility Criteria. Stud Health Technol Inform. 2018;248:293–9.
239. Shivade C, Hebert C, Lopetegui M, de Marneffe MC, Fosler-Lussier E, Lai AM. Textual inference for eligibility criteria resolution in clinical trials. J Biomed Inform. 2015 Dec 1;58:S211–8.
240. Warner JL, Levy MA, Neuss MN, Warner JL, Levy MA, Neuss MN. Feasibility and accuracy of extracting cancer stage information from narrative electronic health record data. J Oncol Pract. 2016 Feb 1;12(2):e169–79.
241. Kamal J, Pasuparthi K, Rogers P, Buskirk J, Mekhjian H. Using an information warehouse to screen patients for clinical trials: a prototype. AMIA Annu Symp Proc. 2005;5(6):1004.
242. Köpcke F, Lubgan D, Fietkau R, Scholler A, Nau C, Stürzl M, et al. Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data. BMC Med Inform Decis Mak. 2013 Dec 9;13(1).
243. Liu H, Chi Y, Butler A, Sun Y, Weng C. A knowledge base of clinical trial eligibility criteria. J Biomed Inform [Internet]. 2021;117(November 2020):103771. Available from: https://doi.org/10.1016/j.jbi.2021.103771
244. National Library of Medicine. ClinicalTrials.gov [Internet]. 2000. Available from: https://clinicaltrials.gov/
245. Dura B, Wajsburt P, Petit-Jean T, Cohen A, Jean C, Bey R. EDS-NLP: efficient information extraction from French clinical notes (v0.7.4). Zenodo [Internet]. 2022. Available from: https://doi.org/10.5281/zenodo.7428752
246. Elm E von, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ Br Med J [Internet]. 2007 Oct 20 [cited 2021 Sep 15];335(7624):806. Available from: /pmc/articles/PMC2034723/
247. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020 269 [Internet]. 2020 Sep 9 [cited 2021 Sep 6];26(9):1320–4. Available from: https://www.nature.com/articles/s41591-020-1041-y
248. Kempf E, Priou S, Lamé G, Daniel C, Bellamine A, Sommacale D, et al. 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. Int J Cancer [Internet]. 2021 Jan 17 [cited 2022 Jan 20];accepted(September 2021):1–10. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/ijc.33928
249. Vogel A, Bridgewater J, Edeline J, Kelley RK, Klümpen HJ, Malka D, et al. Biliary tract cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up ☆. Ann Oncol [Internet]. 2023;34(2):127–40. Available from: http://dx.doi.org/10.1016/j.annonc.2022.10.506
250. Amin M, Greene F, Edge S, Compton C, Gershenwald J, Brookland R, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. [Internet]. CA: a cancer journal for clinicians United States: CA Cancer J Clin; Mar, 2017 p. 93–9. Available from: https://pubmed.ncbi.nlm.nih.gov/28094848/
251. Brierley J. The evolving TNM cancer staging system: an essential component of cancer care. C Can Med Assoc J = J l'Association medicale Can. 2006 Jan;174(2):155–6.
252. Kempf E, Priou S, Lamé G, Laurent A, Guével E, Tzedakis S, et al. No changes in clinical presentation, treatment strategies and survival of pancreatic cancer cases during the Sars-Cov-2 outbreak: a retrospective multicenter cohort study on real-world data. Int J Cancer. 2023;
253. Guével E;, Priou S;, Lamé G;, Flicoteaux R;, Chatellier G;, Tournigand C;, et al. Indicateurs de Qualité de Soins et de Santé : Utilisation du Traitement Automatique des Langues pour le calcul et suivi dans le cas du cancer du sein. In: EMOIS congress.
254. Priou S, Lamé G, Chatellier G, Tournigand C, Kempf E. Effect of the COVID-19 pandemic on colorectal cancer care in France [Internet]. Vol. 6, The Lancet Gastroenterology and Hepatology. Elsevier Ltd; 2021 [cited 2021 Jul 5]. p. 342–3. Available from: https://pubmed.ncbi.nlm.nih.gov/33857441/
255. Priou S, Lamé G, Zalcman G, Wislez M, Bey R, Chatellier G, et al. Influence of the SARS-CoV-2 outbreak on management and prognosis of new lung cancer cases, a retrospective multicentre real-life cohort study. Eur J Cancer. 2022;173:33–40.
256. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;2017-Decem(Section 2):4766–75.

257. Li R, Shinde A, Liu A, Glaser S, Lyou Y, Yuh B, et al. Machine Learning-Based Interpretation and Visualization of Nonlinear Interactions in Prostate Cancer Survival. JCO Clin cancer informatics. 2020 Jul;4:637–46.
258. Kim Y. Convolutional Neural Networks for Sentence Classification [Internet]. p. 1746–51. Available from: http://nlp.stanford.edu/sentiment/
259. Radzi SFM, Karim MKA, Saripan MI, Rahman MAA, Isa INC, Ibahim MJ. Hyperparameter Tuning and Pipeline Optimization via Grid Search Method and Tree-Based AutoML in Breast Cancer Prediction. J Pers Med. 2021 Sep;11(10).
260. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. brat: a Web-based Tool for {NLP}-Assisted Text Annotation. In: Proceedings of the Demonstrations at the 13th Conference of the {E}uropean Chapter of the Association for Computational Linguistics [Internet]. Avignon, France: Association for Computational Linguistics; 2012. p. 102–7. Available from: https://aclanthology.org/E12-2021
261. Rami-Porta R. The Evolving Concept of Complete Resection in Lung Cancer Surgery. Cancers (Basel). 2021 May;13(11).
262. Dura B, Jean C, Tannier X, Calliger A, Bey R, Neuraz A, et al. Learning structures of the French clinical language: development and validation of word embedding models using 21 million clinical reports from electronic health records. ArXiv [Internet]. 2022 Jul 26 [cited 2023 Jan 26];abs/2207.1. Available from: https://arxiv.org/abs/2207.12940v1
263. Région Ile de France. Guerbet et Magic LEMP lauréats du 2e « AI for Health » Challenge [Internet]. 2021. Available from: https://pharmaceutiques.com/actualites/innovations/guerbet-et-magic-lemp-laureats-du-2e-ai-for-health-challenge/
264. Neves M, Ševa J. An extensive review of tools for manual annotation of documents. Brief Bioinform. 2021 Jan;22(1):146–63.
265. Giachelle F, Irrera O, Silvello G. MedTAG: a portable and customizable annotation tool for biomedical documents. BMC Med Inform Decis Mak. 2021 Dec;21(1):352.
266. Névéol A, Islamaj Doğan R, Lu Z. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. J Biomed Inform. 2011 Apr;44(2):310–8.
267. Kholghi M, Sitbon L, Zuccon G, Nguyen A. Active learning reduces annotation time for clinical concept extraction. Int J Med Inform. 2017 Oct;106:25–31.
268. Zhang E, Thurier Q, Boyle L. Improving Clinical Named-Entity Recognition with Transfer Learning. Stud Health Technol Inform. 2018;252:182–7.
269. Gobbel GT, Garvin J, Reeves R, Cronin RM, Heavirland J, Williams J, et al. Assisted annotation of medical free text using RapTAT. J Am Med Inform Assoc. 2014;21(5):833–41.
270. Kim B-H. Medical Codes Prediction from Clinical Notes: From Human Coders to Machines. 2022;(BayLearn):2–4. Available from: http://arxiv.org/abs/2210.16850
271. Jin Y, Xiong Y, Shi D, Lin Y, He L, Zhang Y, et al. Learning from undercoded clinical records for automated International Classification of Diseases (ICD) coding. J Am Med Informatics Assoc [Internet]. 2023 Mar 1;30(3):438–46. Available from: https://doi.org/10.1093/jamia/ocac230
272. Su P, Li G, Wu C, Vijay-Shanker K. Using distant supervision to augment manually annotated data for relation extraction. PLoS One. 2019;14(7):e0216913.
273. Leonardelli E, Menini S, Palmero Aprosio A, Guerini M, Tonelli S. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators{'} Disagreement. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing [Internet]. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. p. 10528–39. Available from: https://aclanthology.org/2021.emnlp-main.822
274. Li D, Rastegar Mojarad M, Li Y, Sohn S, Mehrabi S, Komandur Elayavilli R, et al. A Frequency-based Strategy of Obtaining Sentences from Clinical Data Repository for Crowdsourcing. Stud Health Technol Inform. 2015;216:1033–4.
275. Kupis L, Goodman ZT, Kornfeld S, Hoang S, Romero C, Dirks B, et al. Brain Dynamics Underlying Cognitive Flexibility Across the Lifespan. Cereb Cortex. 2021 Oct;31(11):5263–74.
276. Uddin LQ. Cognitive and behavioural flexibility: neural mechanisms and clinical considerations. Nat Rev Neurosci. 2021 Mar;22(3):167–79.
277. Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Trans neural networks Learn Syst. 2021 Nov;32(11):4793–813.
278. Tomanek K, Hahn U, Lohmann S, Ziegler J. A cognitive cost model of annotations based on eye-tracking data. Proc Annu Meet Assoc Comput Linguist. 2010;2010-July(July):1158–67.
279. Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. Nat Biomed Eng [Internet]. 2022;6(12):1330–45. Available from: https://doi.org/10.1038/s41551-022-00898-y
280. Kehl KLK, Xu W, Lepisto E, Al. E, Elmarakeby H, Hassett MJ, et al. Natural Language Processing to Ascertain Cancer Outcomes From Medical Oncologist Notes. JCO Clin Cancer Inf [Internet]. 2020 Sep 5 [cited 2020 Oct 7];4(4):680–90. Available from: https://pubmed.ncbi.nlm.nih.gov/32755459/
281. Mitchell JR, Szepietowski P, Howard R, Reisman P, Jones JD, Lewis P, et al. A Question-and-Answer System to Extract Data From Free-Text Oncological Pathology Reports (CancerBERT Network): Development Study. J Med Internet Res 2022;24(3)e27210 https//www.jmir.org/2022/3/e27210 [Internet]. 2022 Mar 23 [cited 2022 May 23];24(3):e27210.

Available from: https://www.jmir.org/2022/3/e27210
282. Sorin V, Barash Y, Konen E, Klang E. Deep-learning natural language processing for oncological applications [Internet]. Vol. 21, The Lancet Oncology. Lancet Publishing Group; 2020 [cited 2021 Feb 4]. p. 1553–6. Available from: http://www.thelancet.com/article/S147020452030615X/fulltext
283. Lannelongue L, Grealey J, Inouye M. Green Algorithms: Quantifying the Carbon Footprint of Computation. Adv Sci [Internet]. 2021 Jun 1 [cited 2022 Dec 29];8(12). Available from: /pmc/articles/PMC8224424/
284. Grealey J, Lannelongue L, Saw W-YY, Marten J, McRossed D Sign©ric G, Ruiz-Carmona S, et al. The Carbon Footprint of Bioinformatics. Mol Biol Evol [Internet]. 2022 Mar 1 [cited 2022 Dec 29];39(3). Available from: /pmc/articles/PMC8892942/
285. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet (London, England). 2009 Jul;374(9683):86–9.
286. Richie C. Environmentally sustainable development and use of artificial intelligence in health care. Bioethics. 2022 Jun;36(5):547–55.
287. Ducret M, Mörch C-M, Karteva T, Fisher J, Schwendicke F. Artificial intelligence for sustainable oral healthcare. J Dent. 2022 Dec;127:104344.
288. Alami H, Rivard L, Lehoux P, Hoffman SJ, Cadeddu SBM, Savoldelli M, et al. Artificial intelligence in health care: laying the Foundation for Responsible, sustainable, and inclusive innovation in low- and middle-income countries. Global Health. 2020 Jun;16(1):52.
289. Savci P, Das B. Comparison of pre-trained language models in terms of carbon emissions, time and accuracy in multi-label text classification using AutoML. Heliyon. 2023 May;9(5):e15670.
290. Bloomfield PS, Clutton-Brock P, Pencheon E, Magnusson J, Karpathakis K. Artificial Intelligence in the NHS: Climate and Emissions☆,☆☆. J Clim Chang Heal [Internet]. 2021;4:100056. Available from: https://www.sciencedirect.com/science/article/pii/S2667278221000535
291. Bannour N, Ghannay S, Névéol A, Ligozat AL. Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. Sustain 2021 - 2nd Work Simple Effic Nat Lang Process Proc Sustain. 2021;11–21.
292. Hahn U, Oleynik M. Medical Information Extraction in the Age of Deep Learning. Yearb Med Inform [Internet]. 2020 Aug 1 [cited 2021 Feb 3];29(1):208–20. Available from: /pmc/articles/PMC7442512/?report=abstract
293. Huang T, Luo T, Yan M, Zhou JT, Goh R. RCT: Resource Constrained Training for Edge AI. IEEE Trans neural networks Learn Syst. 2022 Aug;PP.
294. Lacoste A, Luccioni AS, Schmidt V, Dandres T. Quantifying the Carbon Emissions of Machine Learning. ArXiv [Internet]. 2019;abs/1910.0. Available from: https://api.semanticscholar.org/CorpusID:204823751
295. Foufi V, Lanteri S, Gaudet-Blavignac C, Remy P, Montet X, Lovis C. Automatic Annotation Tool to Support Supervised Machine Learning for Scaphoid Fracture Detection. Stud Health Technol Inform. 2018;255:210–4.
296. Nguyen AN, Lawley MJ, Hansen DP, Bowman R V., Clarke BE, Duhig EE, et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Informatics Assoc. 2010 Jul;17(4):440–5.
297. Malmasi S, Ge W, Hosomura N, Turchin A. Comparison of Natural Language Processing Techniques in Analysis of Sparse Clinical Data: Insulin Decline by Patients. AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci. 2019;2019:610–9.
298. Santus E, Schuster T, Tahmasebi AM, Li C, Yala A, Lanahan CR, et al. Exploiting Rules to Enhance Machine Learning in Extracting Information From Multi-Institutional Prostate Pathology Reports. JCO Clin Cancer Informatics [Internet]. 2020 Oct 2 [cited 2020 Oct 7];4(4):865–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/33006906
299. Fort K, Nazarenko A, Rosset S. Modeling the complexity of manual annotation tasks: A grid of analysis. 24th Int Conf Comput Linguist - Proc COLING 2012 Tech Pap. 2012;(December):895–910.
300. Yang B, Wang L, Wong DF, Shi S, Tu Z. Context-aware Self-Attention Networks for Natural Language Processing. Neurocomputing [Internet]. 2021;458:157–69. Available from: https://www.sciencedirect.com/science/article/pii/S0925231221009048
301. Keegan NM, Vasselman SE, Barnett ES, Nweji B, Carbone EA, Blum A, et al. Clinical annotations for prostate cancer research: Defining data elements, creating a reproducible analytical pipeline, and assessing data quality. Prostate. 2022 Aug;82(11):1107–16.
302. Tao S, Zeng N, Hands I, Hurt-Mueller J, Durbin EB, Cui L, et al. Web-based interactive mapping from data dictionaries to ontologies, with an application to cancer registry. BMC Med Inform Decis Mak. 2020 Dec;20(Suppl 10):271.
303. Linnarsson R, Wigertz O. The data dictionary--a controlled vocabulary for integrating clinical databases and medical knowledge bases. Methods Inf Med. 1989 Apr;28(2):78–85.
304. Vesteghem C, Brøndum RF, Sønderkær M, Sommer M, Schmitz A, Bødker JS, et al. Implementing the FAIR Data Principles in precision oncology: review of supporting initiatives. Brief Bioinform. 2020 May;21(3):936–45.
305. van der Velde KJ, Singh G, Kaliyaperumal R, Liao X, de Ridder S, Rebers S, et al. FAIR Genomes metadata schema promoting Next Generation Sequencing data reuse in Dutch healthcare and research. Sci data. 2022 Apr;9(1):169.
306. Hamilton DG, Page MJ, Finch S, Everitt S, Fidler F. How often do cancer researchers make their data and code available and what factors are associated with sharing? BMC Med. 2022 Nov;20(1):438.
307. Li Y, Luo Y-H, Wampfler JA, Rubinstein SM, Tiryaki F, Ashok K, et al. Efficient and Accurate Extracting of Unstructured EHRs on Cancer Therapy Responses for the Development of RECIST Natural Language Processing Tools: Part I, the Corpus. JCO Clin Cancer Informatics [Internet]. 2020 Sep 4 [cited 2020 Oct 7];2000(4):383–91. Available from: https://pubmed.ncbi.nlm.nih.gov/32364754/

308. INCa. Labellisation de sites de recherche intégrée sur le cancer (SIRIC) [Internet]. 2022. Available from: https://www.e-cancer.fr/Institut-national-du-cancer/Appels-a-projets/Appels-a-projets-resultats/SIRIC2022
309. Communiqué de presse APHP. Le projet ACCES AP-HP est lauréat de l'appel à projets « Entrepôts de données de santé hospitaliers » [Internet]. 2023. Available from: Le projet ACCES AP-HP est lauréat de l'appel à projets « Entrepôts de données de santé hospitaliers »
310. Reinecke I, Gruhl M, Pinnau M, Altun FB, Folz M, Zoch M, et al. An OHDSI ATLAS Extension to Support Feasibility Requests in a Research Network. Stud Health Technol Inform. 2022 Jun;295:515–6.
311. Gerber DE, Singh H, Larkins E, Ferris A, Forde PM, Selig W, et al. A New Approach to Simplifying and Harmonizing Cancer Clinical Trials—Standardizing Eligibility Criteria. JAMA Oncol [Internet]. 2022 Aug 4 [cited 2022 Aug 5]; Available from: https://jamanetwork.com/journals/jamaoncology/fullarticle/2794872
312. FHIR. OMOP 5.2 to FHIR R4 Mappings [Internet]. 2019. Available from: http://build.fhir.org/ig/HL7/cdmh/profiles.html
313. FHIR. mCODE and Research Standards : OMOP [Internet]. 2020. Available from: https://confluence.hl7.org/display/COD/OMOP
314. OHDSI. 2022 OHDSI Symposium [Internet]. 2022. Available from: https://www.ohdsi.org/ohdsi2022symposium/
315. FHIR. Connectathon 30 FHIR-OMOP Oncology [Internet]. 2022. Available from: https://confluence.hl7.org/display/FHIR/2022-05+FHIR-OMOP+Oncology
316. Agence nationale de la recherche (ANR). FRANCE 2030 : « PARIS SACLAY CANCER CLUSTER », PREMIER LAURÉAT DE L ' AMI FRANCE 2030 : 1 MD € POUR RELANCER ET CONSOLIDER LA POLITIQUE DE SITE DE. 2022.
317. Ong WL, Schouwenburg MG, van Bommel ACM, Stowell C, Allison KH, Benn KE, et al. A Standard Set of Value-Based Patient-Centered Outcomes for Breast Cancer: The International Consortium for Health Outcomes Measurement (ICHOM) Initiative. JAMA Oncol. 2017 May;3(5):677–85.
318. IDEA4RC. Intelligent ecosystem to improve the governance, the sharing, and the re-use of health data for rare cancers [Internet]. 2022. Available from: https://www.idea4rc.eu/
319. European Commission. European Cancer Imaging Initiative, EUCAIM project [Internet]. 2022. Available from: https://digital-strategy.ec.europa.eu/en/policies/cancer-imaging
320. Kempf E, Fournier L, Guettier C, Mission IA et cancer de l'Assistance Publique - Hôpitaux de Paris. Intelligence artificielle et cancer. L'intenat Paris / Forum des spécialités - Oncol. 2023;24–7.
321. Datta S, Bernstam E V., Roberts K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes [Internet]. Vol. 100, Journal of Biomedical Informatics. Academic Press Inc.; 2019 [cited 2020 Oct 7]. p. 103301. Available from: https://doi.org/10.1016/j.jbi.2019.103301
322. Feng J, Phillips R V, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. npj Digit Med [Internet]. 2022;5(1):66. Available from: https://doi.org/10.1038/s41746-022-00611-y
323. Zong N, Wen A, Stone DJ, Sharma DK, Wang C, Yu Y, et al. Developing an FHIR-Based Computational Pipeline for Automatic Population of Case Report Forms for Colorectal Cancer Clinical Trials Using Electronic Health Records. JCO Clin Cancer Informatics [Internet]. 2020 Sep 5 [cited 2021 Feb 1];(4):201–9. Available from: https://ascopubs.org/doi/10.1200/CCI.19.00116
324. Fu S, Wang L, Moon S, Zong N, He H, Pejaver V, et al. Recommended practices and ethical considerations for natural language processing-assisted observational research: A scoping review. Clin Transl Sci. 2023 Mar;16(3):398–411.
325. Trivedi G, Pham P, Chapman WW, Hwa R, Wiebe J, Hochheiser H. NLPReViz: an interactive tool for natural language processing on clinical text. J Am Med Inform Assoc. 2018 Jan;25(1):81–7.
326. Trivedi G, Dadashzadeh ER, Handzel RM, Chapman WW, Visweswaran S, Hochheiser H. Interactive NLP in Clinical Care: Identifying Incidental Findings in Radiology Reports. Appl Clin Inform. 2019 Aug;10(4):655–69.
327. Barbot A-C, Mussault P, Ingrand P, Tourani J-M. Assessing 2-month clinical prognosis in hospitalized patients with advanced solid tumors. J Clin Oncol Off J Am Soc Clin Oncol. 2008 May;26(15):2538–43.
328. Duval L, Lam Y-H, Pons-Tostivint E, Bennouna J, Matysiak-Budnik T, Lepeintre A, et al. Réévaluation du score Pronopall : une étude rétrospective multicentrique. Bull Cancer [Internet]. 2022;109(4):457–64. Available from: https://www.sciencedirect.com/science/article/pii/S0007455122000169
329. Lindvall C, Lilley EJ, Zupanc SN, Chien I, Udelsman B V., Walling A, et al. Natural Language Processing to Assess End-of-Life Quality Indicators in Cancer Patients Receiving Palliative Surgery. J Palliat Med [Internet]. 2019 Feb 1 [cited 2020 Oct 7];22(2):183–7. Available from: https://pubmed.ncbi.nlm.nih.gov/30328764/
330. Poort H, Zupanc SN, Leiter RE, Wright AA, Lindvall C. Documentation of Palliative and End-of-Life Care Process Measures among Young Adults Who Died of Cancer: A Natural Language Processing Approach. J Adolesc Young Adult Oncol [Internet]. 2020 Feb 1 [cited 2020 Oct 7];9(1):100–4. Available from: https://pubmed.ncbi.nlm.nih.gov/31411524/
331. Kempf E, Lamé G, Layese R, Priou S, Chatellier G, Chaieb H, et al. 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. Eur J Cancer [Internet]. 2021 Feb [cited 2021 Apr 12];150(0):260–7. Available from: www.sciencedirect.com
332. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin. 2023 Jan;73(1):17–48.
333. Daly AA, Rolph R, Cutress RI, Copson ER. A Review of Modifiable Risk Factors in Young Women for the Prevention of Breast Cancer. Breast cancer (Dove Med Press. 2021;13:241–57.
334. Giaquinto AN, Sung H, Miller KD, Kramer JL, Newman LA, Minihan A, et al. Breast Cancer Statistics, 2022. CA Cancer

J Clin. 2022;72(6):524–41.
335. Peintinger F. National Breast Screening Programs across Europe. Breast Care (Basel). 2019 Dec;14(6):354–8.
336. Wojtyla C, Bertuccio P, Wojtyla A, La Vecchia C. European trends in breast cancer mortality, 1980–2017 and predictions to 2025. Eur J Cancer [Internet]. 2021;152:4–17. Available from: https://doi.org/10.1016/j.ejca.2021.04.026
337. Bosch G, Posso M, Louro J, Roman M, Porta M, Castells X, et al. Impact of the COVID-19 pandemic on breast cancer screening indicators in a Spanish population-based program: a cohort study. Elife. 2022 Jun;11.
338. Mayo M, Potugari B, Bzeih R, Scheidel C, Carrera C, Shellenberger RA. Cancer Screening During the COVID-19 Pandemic: A Systematic Review and Meta-analysis. Mayo Clin proceedings Innov Qual outcomes. 2021 Dec;5(6):1109–17.
339. Li T, Nickel B, Ngo P, McFadden K, Brennan M, Marinovich ML, et al. A systematic review of the impact of the COVID-19 pandemic on breast cancer screening and diagnosis. The Breast [Internet]. 2023;67:78–88. Available from: https://www.sciencedirect.com/science/article/pii/S0960977623000012
340. Ng JS, Hamilton DG. Assessing the impact of the COVID-19 pandemic on breast cancer screening and diagnosis rates: A rapid review and meta-analysis. J Med Screen [Internet]. 2022 May 20;29(4):209–18. Available from: https://doi.org/10.1177/09691413221101807
341. Gathani T, Dodwell D, Horgan K. The impact of the first 2 years of the COVID-19 pandemic on breast cancer diagnoses: a population-based study in England. Br J Cancer [Internet]. 2023 Nov 12 [cited 2022 Nov 21];128(3):481–3. Available from: https://www.nature.com/articles/s41416-022-02054-4
342. Grimm LJ, Lee C, Rosenberg RD, Burleson J, Simanowith M, Jr TF, et al. Impact of the COVID-19 Pandemic on Breast Imaging : An Analysis of the National Mammography Database. J Am Coll Radiol [Internet]. 2022;19(8):919–34. Available from: https://doi.org/10.1016/j.jacr.2022.04.008
343. Alagoz O, Lowry KP, Kurian AW, Mandelblatt JS, Ergun MA, Huang H, et al. Impact of the COVID-19 Pandemic on Breast Cancer Mortality in the US: Estimates From Collaborative Simulation Modeling. J Natl Cancer Inst. 2021 Nov;113(11):1484–94.
344. Adachi K, Kimura F, Takahashi H, Kaise H, Yamada K, Ueno E, et al. Delayed Diagnosis and Prognostic Impact of Breast Cancer During the COVID-19 Pandemic. Clin Breast Cancer. 2023 Apr;23(3):265–71.
345. Cairns A, Jones VM, Cronin K, Yocobozzi M, Howard C, Lesko N, et al. Impact of the COVID-19 Pandemic on Breast Cancer Screening and Operative Treatment. Am Surg. 2022 Jun;88(6):1051–3.
346. İlgün AS, Özmen V. The Impact of the COVID-19 Pandemic on Breast Cancer Patients. Eur J breast Heal. 2022 Jan;18(1):85–90.
347. Yadav PB, Rana D, Bharti D, Divya P, Ms D, Gupta A. Impact of COVID-19 pandemic on breast cancer care : report from a regional cancer centre. Lancet Oncol [Internet]. 23:S17. Available from: http://dx.doi.org/10.1016/S1470-2045(22)00416-8
348. Resende CAA, Fernandes Cruz HM, Costa e Silva M, Paes RD, Dienstmann R, Barrios CHE, et al. Impact of the COVID-19 Pandemic on Cancer Staging: An Analysis of Patients With Breast Cancer From a Community Practice in Brazil. JCO Glob Oncol [Internet]. 2022 Nov 9;(8):e2200289. Available from: https://doi.org/10.1200/GO.22.00289
349. Fancellu A, Sanna V, Piredda C, Ariu L, Piana GQ, Deiana G, et al. The COVID-19 outbreak may be associated to a reduced level of care for breast cancer. A comparative study with the pre-COVID era in an Italian Breast Unit. Eur J Cancer [Internet]. 2020 Oct 1;138:S16–7. Available from: https://doi.org/10.1016/S0959-8049(20)30562-1
350. Ferlay J, Partensky C, Bray F. More deaths from pancreatic cancer than breast cancer in the EU by 2017. Acta Oncol [Internet]. 2016 Oct 2 [cited 2022 Jul 23];55(9–10):1158–60. Available from: https://pubmed.ncbi.nlm.nih.gov/27551890/
351. Saad AM, Turk T, Al-Husseini MJ, Abdel-Rahman O. Trends in pancreatic adenocarcinoma incidence and mortality in the United States in the last four decades; A SEER-based study. BMC Cancer [Internet]. 2018 Jun 25 [cited 2022 Jul 23];18(1):1–11. Available from: https://bmccancer.biomedcentral.com/articles/10.1186/s12885-018-4610-4
352. Lukács G, Kovács Á, Csanádi M, Moizs M, Repa I, Kaló Z, et al. Benefits of timely care in pancreatic cancer: A systematic review to navigate through the contradictory evidence [Internet]. Vol. 11, Cancer Management and Research. Dove Medical Press Ltd; 2019 [cited 2021 Apr 12]. p. 9849–61. Available from: https://pubmed.ncbi.nlm.nih.gov/31819622/
353. Gastrointestinal cancers: Pancreatic cancer in the COVID-19 era | ESMO [Internet]. [cited 2022 Jul 23]. Available from: https://www.esmo.org/guidelines/cancer-patient-management-during-the-covid-19-pandemic/gastrointestinal-cancers-pancreatic-cancer-in-the-covid-19-era
354. Malagón T, Yong JHE, Tope P, Miller WH, Franco EL, Ali R, et al. Predicted long-term impact of COVID-19 pandemic-related care delays on cancer mortality in Canada. Int J Cancer [Internet]. 2022 Apr 15 [cited 2022 Jul 23];150(8):1244–54. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/ijc.33884
355. Parmar A, Eskander A, Sander B, Naimark D, Irish JC, Chan KKW. Impact of cancer surgery slowdowns on patient survival during the COVID-19 pandemic: a microsimulation modelling study. CMAJ [Internet]. 2022 Mar 21 [cited 2022 Mar 24];194(11):E408–14. Available from: https://www.cmaj.ca/content/194/11/E408
356. Kędzierska-Kapuza K. Pancreatic Cancer Surgery During COVID-19 Pandemic– A High-Volume Polish Centre Experience. Clin Surg. 2020;5(12):1–10.
357. Siegel RL, Wagle NS, Cercek A, Smith RA, Jemal A. Colorectal cancer statistics, 2023. CA Cancer J Clin. 2023;73(3):233–54.

358. Faivre J, Lepage C, Dancourt V. Le dépistage organisé du cancer colorectal en France et en Europe : historique et état des lieux. Bull Epidémiologique Hebd. 2009;2–3:17–9.
359. Meyer A, Drouin J, Zureik M, Weill A, Dray-Spira R. Colonoscopy in France during the COVID-19 pandemic. Int J Colorectal Dis [Internet]. 2021 May 1 [cited 2021 Jul 16];36(5):1073–5. Available from: https://pubmed.ncbi.nlm.nih.gov/33409566/
360. Challine A, Lazzati A, Dousset B, Voron T, Parc Y, Lefevre JH. Colorectal screening: We have not caught up. A surge of colorectal cancer after the coronavirus disease 2019 (COVID-19) pandemic? Surgery [Internet]. 2021 Apr 1 [cited 2021 Jul 16];169(4):991–3. Available from: http://www.surgjournal.com/article/S0039606020308564/fulltext
361. Morris EJAA, Goldacre R, Spata E, Mafham M, Finan PJ, Shelton J, et al. Impact of the COVID-19 pandemic on the detection and management of colorectal cancer in England: a population-based study. Lancet Gastroenterol Hepatol [Internet]. 2021 Mar 1 [cited 2021 Feb 9];6(3):199–208. Available from: https://pubmed.ncbi.nlm.nih.gov/33453763/
362. Thierry AR, Pastor B, Pisareva E, Ghiringhelli F, Bouché O, Fouchardière CD La, et al. Association of COVID-19 Lockdown With the Tumor Burden in Patients With Newly Diagnosed Metastatic Colorectal Cancer. JAMA Netw Open [Internet]. 2021 Sep 1 [cited 2021 Sep 10];4(9):e2124483–e2124483. Available from: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2784019
363. Radulovic RS, Cuk V V., Juloski JT, Arbutina DD, Krdzic ID, Milic L V., et al. Is Colorectal Cancer Stage Affected by COVID-19 Pandemic? Chirurgia (Bucur) [Internet]. 2021 [cited 2021 Jul 5];116(3):331. Available from: https://www.revistachirurgia.ro/pdfs/2021-3-331.pdf
364. Aguiar S, Riechelmann RP, de Mello CAL, da Silva JCF, Diogenes IDC, Andrade MS, et al. Impact of COVID-19 on colorectal cancer presentation [Internet]. Vol. 108, The British journal of surgery. NLM (Medline); 2021 [cited 2021 Jul 5]. p. e81–2. Available from: https://academic.oup.com/bjs/article/108/2/e81/6065715
365. Savu E, Vasile L, Serbanescu M-S, Alexandru DO, Gheonea IA, Pirici D, et al. Clinicopathological Analysis of Complicated Colorectal Cancer: A Five-Year Retrospective Study from a Single Surgery Unit. Diagnostics (Basel, Switzerland). 2023 Jun;13(12).
366. Jonge L de, Worthington J, Wifferen F van, Iragorri N, Peterse EFP, Lew J-B, et al. Impact of the COVID-19 pandemic on faecal immunochemical test-based colorectal cancer screening programmes in Australia, Canada, and the Netherlands: a comparative modelling study. Lancet Gastroenterol Hepatol [Internet]. 2021 Apr 1 [cited 2021 Jul 16];6(4):304–14. Available from: http://www.thelancet.com/article/S2468125321000030/fulltext
367. Osterlund P, Salminen T, Soveri L-M, Kallio R, Kellokumpu I, Lamminmäki A, et al. Repeated centralized multidisciplinary team assessment of resectability, clinical behavior, and outcomes in 1086 Finnish metastatic colorectal cancer patients (RAXO): A nationwide prospective intervention study. Lancet Reg Heal – Eur [Internet]. 2021 Apr 1 [cited 2021 Jul 14];3:100049. Available from: http://www.thelancet.com/article/S2666776221000260/fulltext
368. Ho KMA, Banerjee A, Lawler M, Rutter MD, Lovat LB. Predicting endoscopic activity recovery in England after COVID-19: a national analysis. Lancet Gastroenterol Hepatol [Internet]. 2021 [cited 2021 Mar 22];381–90. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2468125321000583
369. Mazzone PJ, Gould MK, Arenberg DA, Chen AC, Choi HK, Detterbeck FC, et al. Management of Lung Nodules and Lung Cancer Screening During the COVID-19 Pandemic: CHEST Expert Panel Report. Chest [Internet]. 2020 Jul 1 [cited 2021 Sep 17];158(1):406–15. Available from: https://pubmed.ncbi.nlm.nih.gov/32335067/
370. Valluri S, Lakshmi HN, Sunkavalli C. Incidental Findings in CT Scans on Screening for COVID-19. Indian J Surg Oncol. 2023 Jun;14(2):318–23.
371. Chang SH, Zervos M, Kent A, Chachoua A, Bizekis C, Pass H, et al. Safety of patients and providers in lung cancer surgery during the COVID-19 pandemic. Eur J Cardio-thoracic Surg [Internet]. 2020 Dec 1 [cited 2021 Sep 17];58(6):1222–7. Available from: https://pubmed.ncbi.nlm.nih.gov/33150417/
372. Leclère JB, Fournel L, Etienne H, Al Zreibi C, Onorati I, Roussel A, et al. Maintaining Surgical Treatment of Non-Small Cell Lung Cancer During the COVID-19 Pandemic in Paris. Ann Thorac Surg [Internet]. 2021 May 1 [cited 2021 Sep 18];111(5):1682–8. Available from: https://pubmed.ncbi.nlm.nih.gov/33038341/
373. Challine A, Dousset B, de'Angelis N, Lefèvre JH, Parc Y, Katsahian S, et al. Impact of coronavirus disease 2019 (COVID-19) lockdown on in-hospital mortality and surgical activity in elective digestive resections: A nationwide cohort analysis. Surgery. 2021 Dec 1;170(6):1644–9.
374. Pages PB, Cottenet J, Bonniaud P, Tubert-Bitter P, Piroth L, Cadranel J, et al. Impact of the SARS-CoV-2 Epidemic on Lung Cancer Surgery in France: A Nationwide Study. Cancers (Basel) [Internet]. 2021 Dec 1 [cited 2022 Feb 21];13(24). Available from: https://pubmed.ncbi.nlm.nih.gov/34944896/
375. Lièvre A, Turpin A, Ray-Coquard I, Malicot K Le, Thariat J, Ahle G, et al. Risk factors for Coronavirus Disease 2019 (COVID-19) severity and mortality among solid cancer patients and impact of the disease on anticancer treatment: A French nationwide cohort study (GCO-002 CACOVID-19). Eur J Cancer [Internet]. 2020 Dec 1 [cited 2021 Oct 19];141:62–81. Available from: http://www.ejcancer.com/article/S0959804920310431/fulltext
376. Benderra MA, Aparicio A, Leblanc J, Wassermann D, Kempf E, Galula G, et al. Clinical Characteristics, Care Trajectories and Mortality Rate of SARS-CoV-2 Infected Cancer Patients: A Multicenter Cohort Study. Cancers 2021, Vol 13, Page 4749 [Internet]. 2021 Sep 23 [cited 2022 Feb 21];13(19):4749. Available from: https://www.mdpi.com/2072-6694/13/19/4749/htm
377. Elkrief A, Kazandjian S, Bouganim N. Changes in Lung Cancer Treatment as a Result of the Coronavirus Disease 2019

Pandemic. JAMA Oncol [Internet]. 2020 Nov 1 [cited 2021 Sep 17];6(11):1805–6. Available from: https://jamanetwork.com/journals/jamaoncology/fullarticle/2770258
378. Glasbey JC, Ademuyiwa AO, Adisa A, AlAmeer E, Arnaud AAPAAP, Ayasra F, et al. Effect of COVID-19 pandemic lockdowns on planned cancer surgery for 15 tumour types in 61 countries: an international, prospective, cohort study. Lancet Oncol [Internet]. 2021 Nov 1 [cited 2022 May 3];0(0):1–11. Available from: http://www.thelancet.com/article/S1470204521004939/fulltext%0Ahttp://www.thelancet.com/article/S1470204521004939/abstract%0Ahttps://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(21)00493-9/abstract
379. Ouvrage collectif édité par L'INCa. Algorithme de sélection des hospitalisations liées à la prise en charge du cancer dans les bases nationales d'activité hospitalière de court séjour «algorithme cancer». 2013;1–192.
380. Dinmohamed AG, Visser O, Verhoeven RHA, Louwman MWJ, van Nederveen FH, Willems SM, et al. Fewer cancer diagnoses during the COVID-19 epidemic in the Netherlands [Internet]. Vol. 21, The Lancet Oncology. Lancet Publishing Group; 2020 [cited 2020 Nov 13]. p. 750–1. Available from: https://pubmed.ncbi.nlm.nih.gov/32359403/
381. Kaufman HW, Chen Z, Niles J, Fesko Y. Changes in the Number of US Patients With Newly Identified Cancer Before and During the Coronavirus Disease 2019 (COVID-19) Pandemic. JAMA Netw open [Internet]. 2020 Aug 3 [cited 2020 Oct 20];3(8):e2017267. Available from: https://jamanetwork.com/
382. Team TNUCI of S (NCIS) W. A segregated-team model to maintain cancer care during the COVID-19 outbreak at an academic center in Singapore [Internet]. Vol. 31, Annals of Oncology. Elsevier Ltd; 2020 [cited 2020 Nov 13]. p. 840–3. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7174823/
383. Lai AG, Pasea L, Banerjee A, Denaxas S, Katsoulis M, Chang WH, et al. Estimating excess mortality in people with cancer and multimorbidity in the COVID-19 emergency. medRxiv [Internet]. 2020;15(11):2020.05.27.20083287. Available from: https://www.medrxiv.org/content/10.1101/2020.05.27.20083287v1
384. Arduino PG, Conrotto D, Broccoletti R. The outbreak of Novel Coronavirus disease (COVID-19) caused a worrying delay in the diagnosis of oral cancer in north-west Italy: the Turin Metropolitan Area experience [Internet]. Oral Diseases. Blackwell Publishing Ltd; 2020 [cited 2020 Nov 13]. Available from: https://pubmed.ncbi.nlm.nih.gov/32306459/
385. Panzuto F, Maccauro M, Campana D, Faggiano A, Massironi S, Pusceddu S, et al. Impact of the SARS-CoV2 pandemic dissemination on the management of neuroendocrine neoplasia in Italy: a report from the Italian Association for Neuroendocrine Tumors (Itanet). J Endocrinol Invest [Internet]. 2020 Aug 16 [cited 2020 Oct 20]; Available from: http://www.ncbi.nlm.nih.gov/pubmed/32803662
386. Cortiula F, Pettke A, Bartoletti M, Puglisi F, Helleday T. Managing COVID-19 in the oncology clinic and avoiding the distraction effect. Vol. 31, Annals of Oncology. Elsevier Ltd; 2020. p. 553–5.
387. Communiqué de presse. Covid-19 : Doctolib alerte sur la chute de fréquentation des cabinets et s'engage pour permettre aux patients de retourner consulter [Internet]. 2020 [cited 2020 Nov 13]. Available from: www.community.doctolib.com
388. Sud A, Torr B, Jones ME, Broggio J, Scott S, Loveday C, et al. 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. 2020 Aug 1 [cited 2020 Dec 11];21(8). Available from: www.gov.uk/guidance/national-
389. Sud A, Jones ME, Broggio J, Loveday C, Torr B, Garrett A, et al. Collateral damage: the impact on outcomes from cancer surgery of the COVID-19 pandemic. Ann Oncol. 2020 Aug 1;31(8):1065–74.
390. Lai AG, Pasea L, Banerjee A, Hall G, Denaxas S, Chang WH, et al. 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 [Internet]. 2020 [cited 2020 Dec 11];10(11):1–9. Available from: http://bmjopen.bmj.com/
391. Tempero M. COVID-19 and Cancer: Unintended Consequences. Vol. 18, Journal of the National Comprehensive Cancer Network : JNCCN. NLM (Medline); 2020. p. 1147.
392. van de Haar J, Hoes LR, Coles CE, Seamon K, Fröhling S, Jäger D, et al. Caring for patients with cancer in the COVID-19 era [Internet]. Vol. 26, Nature Medicine. Nature Research; 2020 [cited 2020 Nov 13]. p. 665–71. Available from: https://pubmed.ncbi.nlm.nih.gov/32405058/
393. Qadan M, Hong TS, Tanabe KK, Ryan DP, Lillemoe KD. A Multidisciplinary Team Approach for Triage of Elective Cancer Surgery at the Massachusetts General Hospital During the Novel Coronavirus COVID-19 Outbreak. Ann Surg [Internet]. 2020 Jul 1 [cited 2020 Nov 13];272(1):e20–1. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7188033/
394. Ueda M, Martins R, Hendrie PC, McDonnell T, Crews JR, Wong TL, et al. Managing Cancer Care during the COVID-19 Pandemic: Agility and Collaboration Toward a Common Goal. JNCCN J Natl Compr Cancer Netw [Internet]. 2020 Apr 1 [cited 2020 Nov 13];18(4):366–9. Available from: https://pubmed.ncbi.nlm.nih.gov/32197238/
395. Jones D, Neal RD, Duffy SRG, Scott SE, Whitaker KL, Brain K. Impact of the COVID-19 pandemic on the symptomatic diagnosis of cancer: the view from primary care [Internet]. Vol. 21, The Lancet Oncology. Lancet Publishing Group; 2020 [cited 2020 Nov 13]. p. 748–50. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7251992/
396. Dubois RN. CANCER PREVENTION RESEARCH | COMMENTARY. 2020;
397. Dockter AG, Angelos GC. Molecular-based alternatives for colorectal cancer screening during the COVID-19 pandemic. Surg Technol Int. 2020 May 1;36:1–5.
398. American Cancer Society Cancer Action Network (ACS. COVID-19 Pandemic Impact on Cancer Patients and Survivors [Internet]. Available from: https://www.fightcancer.org/sites/default/files/National Documents/COVID19-Ongoing-

Impact-Polling-Memo.pdf
399. Yang SC, Wang J Der, Wang SY. Considering lead-time bias in evaluating the effectiveness of lung cancer screening with real-world data. Sci Rep [Internet]. 2021 Dec 1 [cited 2022 Jan 7];11(1). Available from: https://pubmed.ncbi.nlm.nih.gov/34108586/
400. Prada-Ramallal G, Takkouche B, Figueiras A. Bias in pharmacoepidemiologic studies using secondary health care databases: A scoping review [Internet]. Vol. 19, BMC Medical Research Methodology. BioMed Central Ltd.; 2019 [cited 2020 Dec 11]. p. 1–14. Available from: https://doi.org/10.1186/s12874-019-0695-y
401. Khan SA, Tavolari S, Brandi G. Cholangiocarcinoma: Epidemiology and risk factors. Liver Int Off J Int Assoc Study Liver. 2019 May;39 Suppl 1:19–31.
402. Kendall T, Verheij J, Gaudio E, Evert M, Guido M, Goeppert B, et al. Anatomical, histomorphological and molecular classification of cholangiocarcinoma. Liver Int Off J Int Assoc Study Liver. 2019 May;39 Suppl 1:7–18.
403. Zhu AX, Macarulla T, Javle MM, Kelley RK, Lubner SJ, Adeva J, et al. Final Overall Survival Efficacy Results of Ivosidenib for Patients With Advanced Cholangiocarcinoma With IDH1 Mutation: The Phase 3 Randomized Clinical ClarIDHy Trial. JAMA Oncol. 2021 Nov;7(11):1669–77.
404. Casak SJ, Pradhan S, Fashoyin-Aje LA, Ren Y, Shen Y-L, Xu Y, et al. FDA Approval Summary: Ivosidenib for the Treatment of Patients with Advanced Unresectable or Metastatic, Chemotherapy Refractory Cholangiocarcinoma with an IDH1 Mutation. Clin cancer Res an Off J Am Assoc Cancer Res. 2022 Jul;28(13):2733–7.
405. Job S, Rapoud D, Dos Santos A, Gonzalez P, Desterke C, Pascal G, et al. Identification of Four Immune Subtypes Characterized by Distinct Composition and Functions of Tumor Microenvironment in Intrahepatic Cholangiocarcinoma. Hepatology. 2020 Sep;72(3):965–81.
406. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. Nat Rev Clin Oncol [Internet]. 2019 Nov 1 [cited 2020 Feb 27];16(11):703–15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31399699

Figures et tables supplémentaires

Tableau Supplémentaire 1. Caractérisation des 15 essais cliniques de phase I – IV sélectionnés au hasard et ouverts entre 2016 et 2021 à l'AP-HP pour les patients atteints d'un cancer urologique, utilisés dans le projet PENELOPE

Numéro	Phase	Type de cancer primitif	Promotion	Bras expérimental	Bras contrôle
NCT01812369 VESPER	3	Cancer localisé de vessie	Académique	Gemcitabine J1 et J8 Cisplatine J1, toutes les 3 semaines, 4 cycles	ddMVAC toutes les 2 semaines, 6 cycles
NCT02961257 CABASTY	3	Cancer de prostate métastatique	Académique	Cabazitaxel 25 mg/m² J1 toutes les 3 semaines	Cabazitaxel 16 mg/m² J1 et J15, toutes les 4 semaines
NCT02495974 PRÉMISSE	4	Cancer de prostate métastatique résistant à la castration	Industrielle	Enzalutamide une fois par jour	NA
NCT03314324 ODENZA	2	Cancer de prostate métastatique résistant à la castration	Académique	Darolutamide deux fois par jour	Enzalutamide une fois par jour
NCT04191096 MK-3475-991	3	Cancer de prostate métastatique hormono-sensible	Industrielle	Pembrolizumab + Enzalutamide + déprivation androgénique	Placebo + Enzalutamide + déprivation androgénique
NCT02960906 BIONIKK	2	Carcinome rénal métastatique à cellules claires	Académique	Nivolumab ou Nivolumab–Ipilimumab, ou VEGFR-TKI	NA
NCT02603432 B9991001	3	Cancer urothélial avancé	Industrielle	Avelumab plus soins de conforts	Soins de confort exclusifs
NCT02689167 SURFER	2	Cancer avancé à cellules rénales	Académique	Sunitinib 2/3 : 2 semaines « avec » en alternance avec 1 semaine « sans »	Sunitinib 4/6 : 4 semaines « avec » en alternance avec 2 semaines « sans »

Numéro	Phase	Type de cancer primitif	Promotion	Bras expérimental	Bras contrôle
NCT03419572 CASSIOPE	4	Carcinome rénal métastatique	Industrielle	Cabozantinib	NA
NCT03013335 NIVOREN	2	Carcinome rénal métastatique	Académique	Nivolumab	NA
NCT03549715 NEMIO	1	Carcinome urothélial infiltrant de la vessie	Académique	Durvalumab + tremelimumab + ddMVAC	Durvalumab + ddMVAC
NCT02302807 GO29294	3	Cancer avancé de la vessie	Industrielle	Atézolizumab J1 tous les 21 jours	Chimiothérapie (vinflunine, paclitaxel ou docétaxel)
NCT03661320 CA017078	3	Cancer de la vessie	Industrielle	Nivolumab + gemcitabine/cisplatine	Gemcitabine/cisplatine
NCT03799835 ALBAN	3	Cancer de la vessie	Académique	BCG + atézolizumab	BCG uniquement
NCT01976741	1	Tous les types de tumeurs solides	Industrielle	Augmentation et expansion de la dose de rogaratinib	NA

Abréviations : BCG, bacille de Calmette-Guérin ; ddMVAC, méthotrexate J1 Vinblastine J2 doxorubicine J2 cisplatine J2 chaque cycle toutes les 2 semaines ; NA, non applicable ; VEGFR-TKI, Inhibiteur de la tyrosine kinase du récepteur du facteur de croissance de l'endothélium vasculaire

Tableau Supplémentaire 2. Résumé des 21 critères de préscreening n'ayant pas pu être alignés à des concepts standards OMOP

Nom de l'étude	Critères de préscreening non alignés à des concepts standards OMOP	Données élémentaires rendant l'alignement impossible
CABASTY	Patient âgé ≥ 65 ans avec mCRPC précédemment traité par docétaxel	Résistant à la castration
PREMISE	mCRPC	Résistant à la castration
MK-3475-991	Présente une maladie métastatique + ≥2 lésions osseuses à la scintigraphie osseuse et/ou une maladie viscérale	Maladie viscérale
MK-3475-991	A reçu une pharmacothérapie antérieure, une radiothérapie traitement ou chirurgie du cancer de la prostate métastatique avec les exceptions suivantes : * Jusqu'à 3 mois d'ADT avec agonistes ou antagonistes de la LHRH ou orchidectomie avec ou sans antiandrogènes concomitants de première génération avant la randomisation, sans preuve radiographique de progression de la maladie ou d'augmentation de l'APS avant la randomisation si le participant n'a pas été traité par docétaxel pour un cancer de la prostate métastatique. * Peut avoir 1 cycle de radiothérapie palliative ou de traitement chirurgical pour traiter les symptômes résultant d'une maladie métastatique si elle a été administrée au moins 4 semaines avant la randomisation * Pour les participants atteints d'une maladie métastatique à faible volume (définie comme <4 lésions osseuses), peut avoir 1 cycle de radiothérapie définitive (c.-à-d. EBRT) à la prostate si elle a été administrée au moins 4 semaines avant la randomisation. *Jusqu'à 6 cycles de docétaxel avec administration finale du traitement dans les 2 mois suivant la randomisation et aucune preuve de progression de la maladie pendant ou après la fin du traitement par docétaxel. Chez ces participants, jusqu'à 6 mois d'ADT avec agonistes ou antagonistes de la LHRH ou d'orchidectomie avec ou sans antiandrogènes concomitants de première génération sont autorisés.	Pour traiter les symptômes résultant d'une maladie métastatique s'il a été administré au moins 4 semaines avant la randomisation À la prostate s'il a été administré au moins 4 semaines avant la randomisation.
B9991001	Traitement systémique adjuvant ou néoadjuvant antérieur dans les 12 mois suivant la randomisation	Dans les 12 mois suivant la randomisation
BIONIKK	Traitement systémique antérieur par le VEGF ou le traitement ciblé par les récepteurs du VEGF (y compris, mais sans s'y limiter, le sunitinib, le pazopanib, l'axitinib, le tivozanib et le bevacizumab). Les sujets traités dans un cadre adjuvant et avec un intervalle libre de plus de 1 an après la fin du traitement pourraient être inclus.	Traitement systémique antérieur par le VEGF ciblé (y compris, mais sans s'y limiter, le sunitinib, le pazopanib, l'axitinib, le tivozanib et le bevacizumab).
BIONIKK	Traitement antérieur avec un anticorps anti--1, anti--L1, anti--L2, anti-CD137 ou antiCTLA-4, ou tout autre anticorps ou médicament ciblant spécifiquement la costimulation des lymphocytes T ou les voies de point de contrôle.	Traitement antérieur avec un anticorps anti--L1, anti--L2, anti-CD137, ou tout autre anticorps ou médicament ciblant spécifiquement la costimulation des lymphocytes T ou les voies de point de contrôle.
CASSIOPE	A déjà reçu au moins un traitement ciblant le VEGF	Traitement ciblé par le VEGF

Nom de l'étude	Critères de préscreening non alignés à des concepts standards OMOP	Données élémentaires rendant l'alignement impossible
NIVOREN	Avoir reçu au moins un traitement antiangiogénique systémique antérieur, y compris, mais sans s'y limiter : sunitinib, sorafénib, pazopanib, axitinib et bevacizumab, dans un contexte avancé ou métastatique	Traitement antiangiogénique systémique antérieur, y compris, mais sans s'y limiter : sunitinib, sorafénib, pazopanib, axitinib et bevacizumab
NIVOREN	Les patients ayant reçu un traitement antérieur avec des anticorps anti--1, anti--L1, anti--L2, anti-CD137 ou anti-CTLA-4 (ou tout autre anticorps ou médicament ciblant spécifiquement les voies de contrôle ou la costimulation des lymphocytes T).	Traitement antérieur avec des anticorps anti--1, anti--L1, anti--L2, anti-CD137 ou anti-CTLA-4 (ou tout autre anticorps ou médicament ciblant spécifiquement les voies de contrôle ou la costimulation des lymphocytes T).
NEMIO	Absence de métastases, confirmée par une tomodensitométrie ou une IRM de base négative du bassin, de l'abdomen et de la poitrine pas plus de 4 semaines avant la randomisation. Les patients atteints d'une maladie N1 de stade clinique sont éligibles si le ganglion lymphatique unique mesure < 15 mm dans l'axe le plus court	Pas plus de 4 semaines avant la randomisation
NEMIO	Tout traitement antitumoral approuvé pour le carcinome urothélial, y compris la chimiothérapie, ou l'immunothérapie avant le début du traitement à l'étude. Il convient de noter que les injections intravésicales antérieures de CANCER DU SEING sont autorisées si elles sont administrées pour un carcinome urothélial non invasif sur le plan musculaire.	Tout traitement antitumoral approuvé pour le carcinome urothélial, y compris la chimiothérapie, ou l'immunothérapie avant le début du traitement à l'étude. S'il est administré pour des traitements non invasifs sur le plan musculaire
GO29294	Progression de la maladie pendant ou après le traitement avec au moins un régime contenant du platine (p. ex. GC, MVAC, CarboGem, etc.) pour la récidive inopérable, localement avancée ou métastatique du cancer urothélial de vessie Un régime est défini comme des patients recevant au moins deux cycles d'un régime contenant du platine. Les patients ayant reçu une chimiothérapie adjuvante/néoadjuvante antérieure et ayant progressé dans les 12 mois suivant le traitement avec un régime adjuvant/néoadjuvant contenant du platine seront considérés comme des patients de deuxième ligne. Les patients peuvent ne pas avoir reçu plus de deux schémas thérapeutiques antérieurs (y compris le régime à base de platine requis) pour leur cancer urothélial de vessie. Les patients doivent avoir démontré une progression de la maladie pendant ou après tous les schémas thérapeutiques antérieurs.	Traitement avec au moins un régime contenant du platine (p. ex. GC, MVAC, CarboGem, etc.)

Nom de l'étude	Critères de préscreening non alignés à des concepts standards OMOP	Données élémentaires rendant l'alignement impossible
CA017078	Traitement antérieur du cancer de la vessie (y compris cancer urothélial de vessie envahissant la musculeuse) par chimiothérapie systémique, immunothérapie, radiothérapie ou chirurgie pour le cancer de la vessie autre que la RTUTV ou les biopsies. Le BCG intravésical ou la chimiothérapie préalable est autorisé s'il est terminé au moins 6 semaines avant le début du traitement à l'étude.	S'il est terminé au moins 6 semaines avant le début du traitement à l'étude.
CA017078	Traitement antérieur avec un anticorps anti--1, anti--L1, anti--L2 ou anti-CTLA-4, ou tout autre anticorps ou médicament ciblant spécifiquement la costimulation des lymphocytes T ou les voies de point de contrôle.	Anticorps anti--L1, anti--L2 ou tout autre anticorps ou médicament ciblant spécifiquement la costimulation des lymphocytes T ou les voies de point de contrôle.
ALBAN	Tout carcinome urothélial non invasif sur le plan musculaire à haut risque confirmé histologiquement (tumeurs à histologie mixte autorisées si l'histologie du carcinome urothélial est prédominante) défini sur la RTUTV comme l'un des éléments suivants : tumeur T1 et/ou haut grade (OMS 2004) et/ou grade 3 (OMS 1973) et/ou carcinome in situ (CIS)	Tumeurs histologiques mixtes autorisées si l'histologie du carcinome urothélial est prédominante Non invasif sur le plan musculaire
ALBAN	Au moins une (deuxième) résection supplémentaire de la tumeur primaire a été réalisée dans l'un des cas suivants (directives EAU, 2017)] : tumeurs T1 à la discrétion du médecin, RTUTV initiale incomplète, pas de muscle dans l'échantillon	Aucun muscle dans l'échantillon (peut être omis si des tumeurs TaLG / G1 ou CIS primaire seulement ont été trouvés).
16443	Les sujets inscrits dans les cohortes d'expansion MTD de la partie 1, de la partie 2 et de la partie 3 de l'étude doivent présenter des niveaux élevés d'expression de FGFR basés sur l'analyse d'échantillons de biopsie tumorale archivée ou fraîche. Les sujets atteints d'un cancer de la vessie présentant de faibles niveaux globaux d'expression de FGFR peuvent être inclus si l'activation des mutations FGFR3 est confirmée. (Seulement pour les cohortes d'expansion MTD de l'étude Partie 1 (tous les arrivants), Partie 2 (sqNSCLC + LAC +cancer du sein + SCCHN) et Partie 3 (sqNSCLC + LAC + CANCER DU SEIN)).	FGFR
16443	Le test FGFR montre de faibles niveaux d'expression de FGFR (uniquement pour les cohortes d'expansion MTD de l'étude Partie 1 (tous les arrivants), Partie 2 (sqNSCLC + LAC + SCCHN) et Partie 3 (sqNSCLC + LAC))	FGFR
16443	Le test d'expression FGFR / mutation FGFR montre de faibles niveaux d'expression FGFR et l'absence de mutation activatrice dans le gène FGFR3 (uniquement pour la partie 2 et la partie 3 (CANCER DU SEIN))	FGFR

Abréviations : mCRPC, cancer de la prostate métastatique résistant à la castration ; EAU, Association Européenne d'Urologie ; DMT, dose maximale tolérée ; SCCHN, carcinome épidermoïde de la tête et du cou; sqNSCLC, cancer du poumon épidermoïde non à petites cellules; RTUTV, résection transurétrale de la tumeur de la vessie; OMS, Organisation mondiale de la Santé

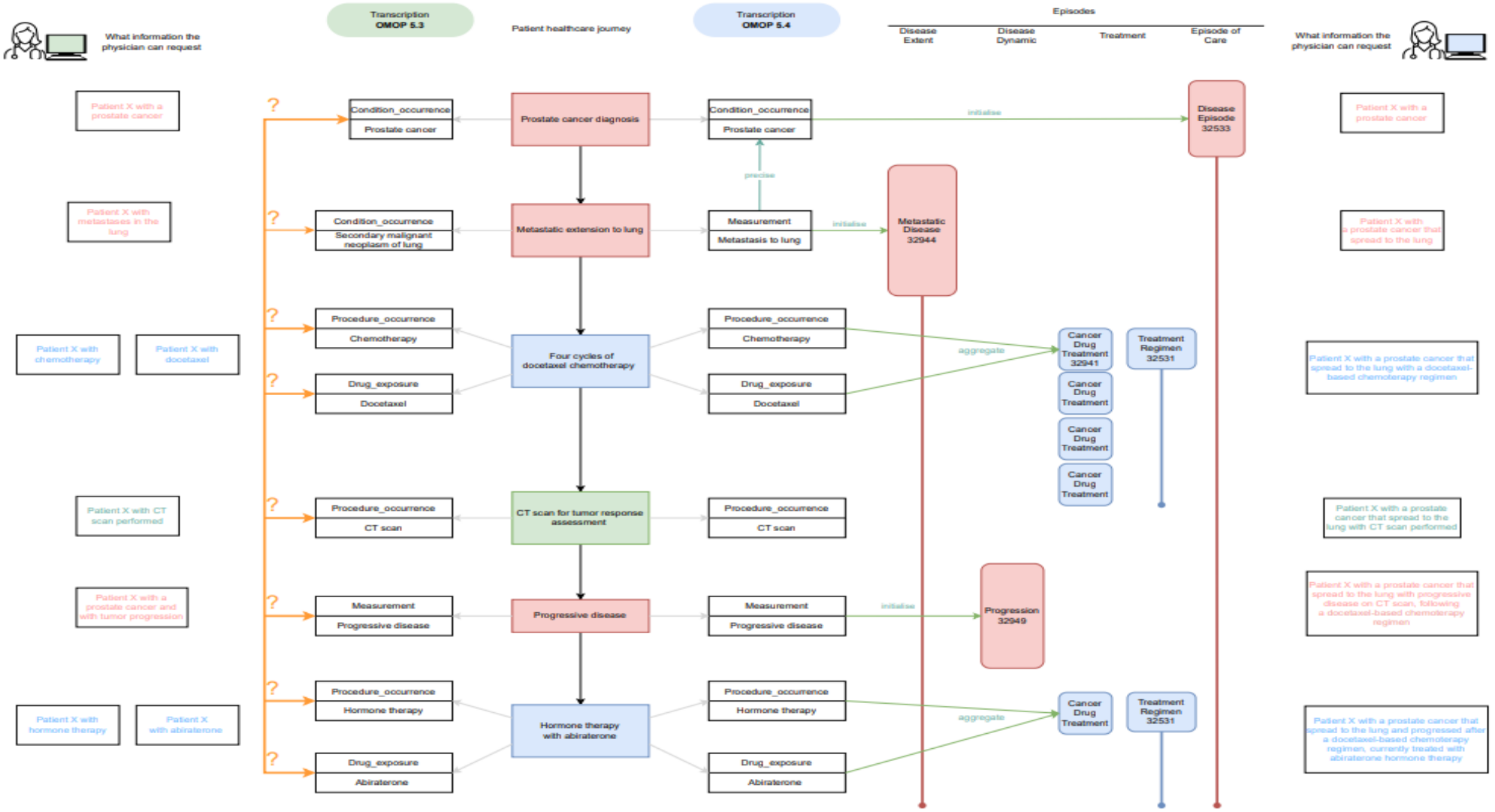


Figure Supplémentaire 1. Exemple de la valeur ajoutée de la version 5.4 du modèle OMOP v5.4 (à droite) par rapport à la version v5.3 (à gauche) en termes d'informations opérables du parcours d'un patient atteint de cancer.

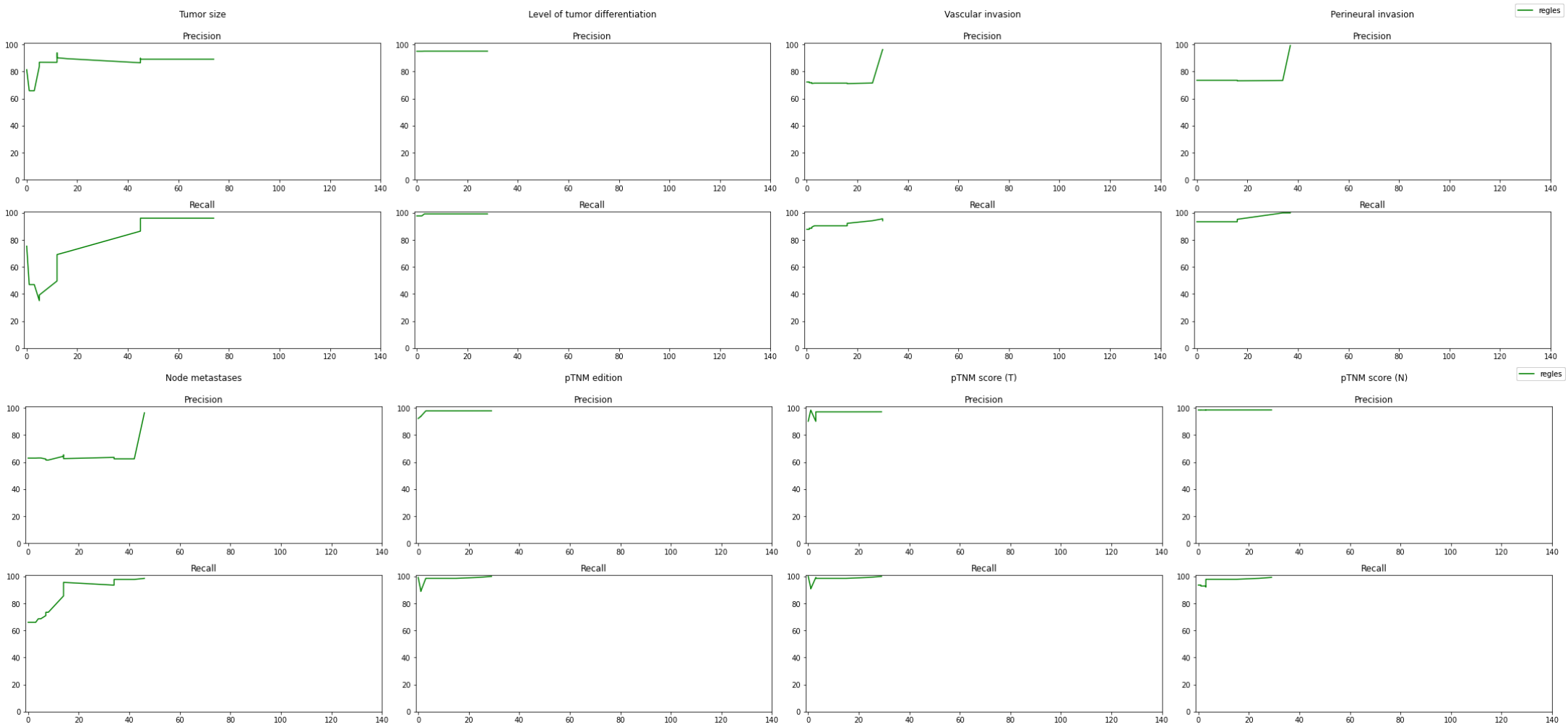


Figure Supplémentaire 2. Métriques de performances d'extractions à base de règles en fonction du nombre de documents analysés issus du jeu de développement, par type d'entités correspondant aux critères pronostiques histologiques des cancers opérés du cas d'usage 'Challenge AI for health', et par rapport à l'annotation finale du jeu de développement

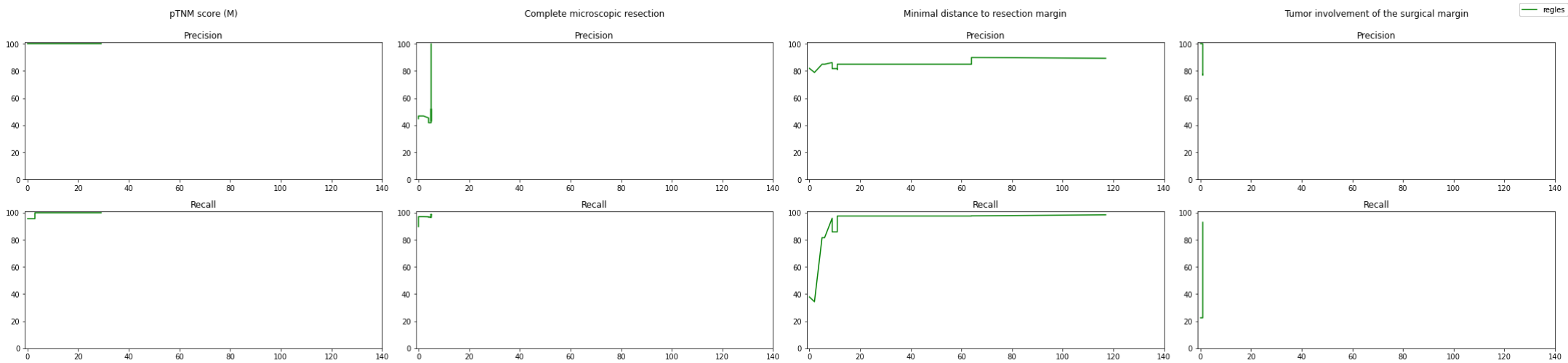


Figure Supplémentaire 2 (suite). Métriques de performances d'extractions à base de règles en fonction du nombre de documents analysés issus du jeu de développement, par type d'entités correspondant aux critères pronostiques histologiques des cancers opérés du cas d'usage 'Challenge AI for health', et par rapport à l'annotation finale du jeu de développement

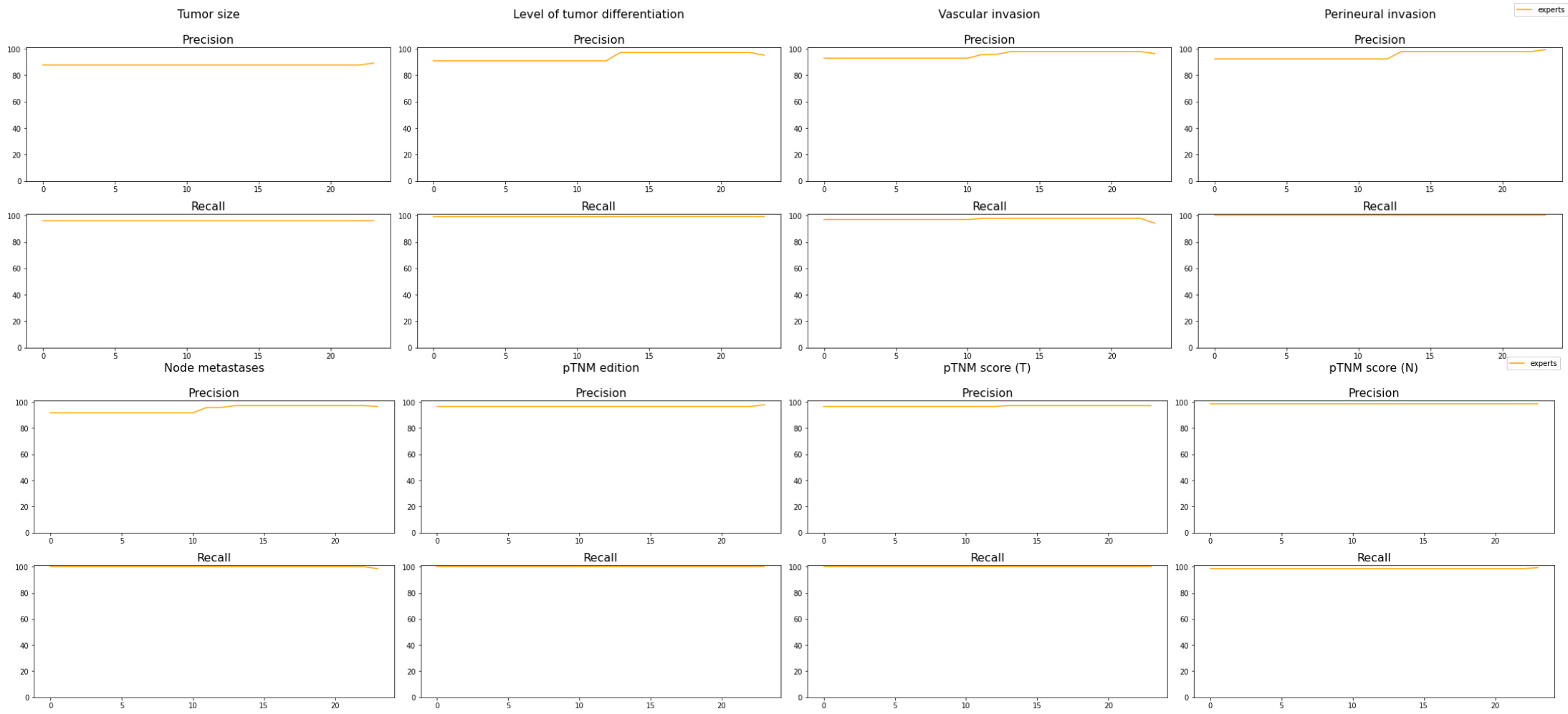


Figure Supplémentaire 3. Métriques de performances d'extractions à base de règles sur le jeu de développement, en fonction de la version du jeu de développement considérée, par type d'entités correspondant aux critères pronostiques histologiques des cancers opérés du cas d'usage 'Challenge AI for health'

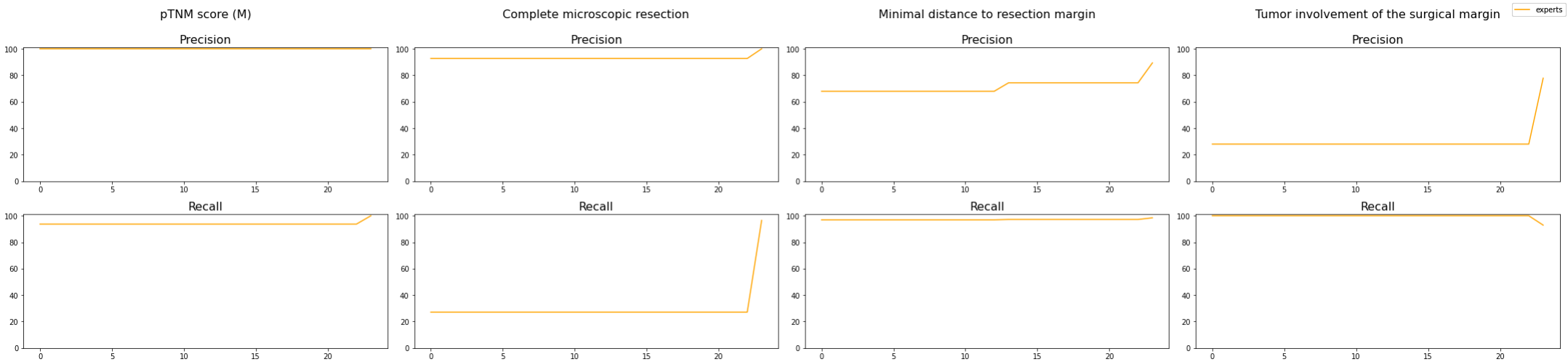


Figure Supplémentaire 3 (suite). Métriques de performances d'extractions à base de règles sur le jeu de développement, en fonction de la version du jeu de développement considérée, par type d'entités correspondant aux critères pronostiques histologiques des cancers opérés du cas d'usage 'Challenge AI for health

Annexes

Annexe 1a. Méthodologie du cas d'usage concernant l'automatisation du calcul des indicateurs de qualité et de sécurité des soins EUSOMA du cancer du sein

Le cas d'usage d'application des algorithmes de structuration des items du jeu de données minimales liées au cancer cherchait à étudier la faisabilité du calcul automatisé d'IQSS pour le cancer du sein, à partir de données des dossiers patients informatisés. Les IQSS concernés choisis étaient ceux proposés par l'European Society of Breast Cancer Specialists (EUSOMA), et mesurés à partir du PMSI et du contenu textuel des comptes rendus d'anatomopathologie contenus dans l'EDS de l'AP-HP (17,18). Ont été incluses les patientes majeures de sexe féminin nouvellement référées et opérées de leur cancer à l'AP-HP pour un cancer du sein entre le 1er janvier 2019 et le 30 juin 2021. Les patientes ayant plusieurs cancers ont été exclues. Les hospitalisations liées à un cancer du sein ont été identifiées via la présence d'un code CIM10 C50 (cancer invasif) ou D05 (cancer *in situ*) en diagnostic principal ou relié dans le PMSI. L'identification des traitements a été rendue possible par les codes CIM10 du PMSI pour la chimiothérapie (Z511), la radiothérapie (Z510), les soins palliatifs (Z515) et les codes CCAM pour la chirurgie (Annexe 3c). Les intitulés des IQSS ont été traduits en français (Annexes 1a et 1c). Chacun des 34 IQSS proposé par l'EUSOMA a été décomposé en variables élémentaires nécessaires à son calcul, et disponibles dans les données du PMSI, dans un compte-rendu textuel d'anatomopathologie, ou dans une autre source de données. Les IQSS ont été classés en 3 groupes selon la source des données nécessaires à leur calcul : PMSI uniquement ; PMSI et compte-rendu textuel d'anatomopathologie ; autres. Le présent cas d'usage s'est restreint aux indicateurs évaluables par le PMSI et les comptes rendus textuels d'anatomopathologie.

Trois types de comptes rendus d'anatomopathologie ont été identifiés :

- diagnostics : datés avant l'administration du premier traitement antitumoral ;
- pré-opératoires : datés d'au moins trois jours avant la date de la chirurgie ;
- post-opératoires : datés dans une fenêtre de trois jours avant ou après la date de la première chirurgie.

Les variables élémentaires ont été annotées soit par un ingénieur (Etienne Guével) et une oncologue médicale (EK) sur le jeu de développement constitué de 259 comptes rendus échantillonnés aléatoirement parmi ceux édités avant juin 2021, soit uniquement par EK sur le jeu de test constitué de 48 comptes rendus échantillonnés aléatoirement parmi ceux édités après juin 2021. Des expressions régulières ont été développées pour extraire les variables élémentaires à partir des 259 comptes rendus du jeu de développement, à l'aide de la bibliothèque logicielle edsnlp v0.7.4 (12). Enfin, la calculabilité des IQSS a été établie à partir de la combinaison de la disponibilité des comptes rendus au sein de l'EDS de l'AP-HP, de la présence des variables élémentaires dans les comptes rendus, et des performances des algorithmes d'extraction des variables élémentaires.

Annexe 1b. Indicateurs de qualité et de sécurité des soins EUSOMA considérés pour le cancer du sein, et variables élémentaires nécessaires à leur calcul

Intitulé de l'IQSS	Cohorte d'intérêt	Population d'intérêt	Variables élémentaires nécessaires au calcul de l'IQSS	Sources de données nécessaires au calcul des variables élémentaires	Nombre (%) de patientes pour lesquelles toutes les sources de données nécessaires sont disponibles	Nombre (%) de patientes pour lesquelles toutes les variables élémentaires sont disponibles dans les sources, et pour lesquelles l'IQSS est calculable
2. Ratio de diagnostics bénins/diagnostics malin	Patientes opérées	3575	date de la chirurgie du primitif malignité du prélèvement étudié	PMSI CRA opératoire	3381 (94,6 %)	3147 (88,0 %)
3.b Taux de diagnostic ACP préopératoire	Patientes opérées	3575	date de la chirurgie du primitif malignité du prélèvement étudié type de technique anatomopathologique utilisée	PMSI CRA pré opératoire CRA pré opératoire	2101 (58,8 %)	1781 (49,8 %)
4.a Taux d'évaluation de facteurs pronostiques de la composante invasive (12 facteurs pronostiques)*	Cancer invasif pour les patientes opérées	3532	date de chimiothérapie	PMSI	2709 (76,7 %)	622 (17,6 %)
			type de cancer date de chirurgie du primitif grade de la tumeur statut des récepteurs aux œstrogènes	PMSI PMSI CRA diagnostique CRA diagnostique		

Intitulé de l'IQSS	Cohorte d'intérêt	Population d'intérêt	Variables élémentaires nécessaires au calcul de l'IQSS	Sources de données nécessaires au calcul des variables élémentaires	Nombre (%) de patientes pour lesquelles toutes les sources de données nécessaires sont disponibles		Nombre (%) de patientes pour lesquelles toutes les variables élémentaires sont disponibles dans les sources, et pour lesquelles l'IQSS est calculable	
			statut HER2	CRA diagnostique				
			type histologique	CRA diagnostique				
			score pTNM	CRA opératoire				
			embolies vasculaires	CRA opératoire				
			distance aux marges	CRA opératoire				
				CRA opératoire				
			taille de la tumeur	CRA opératoire				
4.b Taux d'évaluation de facteurs pronostiques de la composante non invasive (5 facteurs pronostiques)	Cancer *in situ* pour les patientes opérées	43	date de chirurgie du primitif	PMSI	39	(90,7 %)	1	(2,3 %)
			type de cancer	PMSI				
			grade de la tumeur	CRA diagnostique				
			type histologique	CRA diagnostique				
			statut des récepteurs aux œstrogènes	CRA diagnostique				
			distance aux marges	CRA opératoire				
			taille de la tumeur	CRA opératoire				
		43	type de cancer	PMSI	39	(90,7 %)	20	(46,5 %)

Intitulé de l'IQSS	Cohorte d'intérêt	Population d'intérêt	Variables élémentaires nécessaires au calcul de l'IQSS	Sources de données nécessaires au calcul des variables élémentaires	Nombre (%) de patientes pour lesquelles toutes les sources de données nécessaires sont disponibles	Nombre (%) de patientes pour lesquelles toutes les variables élémentaires sont disponibles dans les sources, et pour lesquelles l'IQSS est calculable
11.d Taux de chirurgie conservatrice pour les patientes avec un cancer in situ de petite taille	Cancer *in situ* pour les patientes opérées		date de chirurgie du primitif	PMSI		
			type de chirurgie du primitif	PMSI		
			taille de la tumeur	CRA opératoire		
13.a Taux de chimiothérapie adjuvante chez les patientes ayant un cancer du sein invasif, T > 1 cm ou N + et RE+	Cancer invasif pour les patientes opérées	3532	date de chirurgie du primitif	PMSI	3342 (94,6 %)	1588 (44,9 %)
			date de chimiothérapie	PMSI		
			statut des récepteurs aux œstrogènes	CRA opératoire		
			taille de la tumeur	CRA opératoire		
			score pTNM	CRA opératoire		

Abréviations : CRA, compte rendu anatomopathologique ; IQSS, indicateurs de qualité et de sécurité des soins ; PMSI, programme de médicalisation du système d'information

Annexe 2. Glossaire du cas d'usage PENELOPE

« **Screening** » ou **sélection de patients potentiellement éligibles** : dans le contexte de l'aide au recrutement, le processus de « screening » désigne l'identification et la sélection de patients potentiellement éligibles à une étude clinique. Le résultat de ce processus est une liste de patients potentiellement éligibles (**statut de patient : screené**), qui seront contactés et invités à une visite d'inclusion permettant de vérifier l'éligibilité du patient et son acceptation formelle (consentement) de participation à l'étude clinique. Le screening est un processus fastidieux et long et le plus souvent manuel.

« **Inclusion** » : le processus d'inclusion désigne la phase consistant à s'assurer que le patient remplit l'ensemble des critères d'éligibilité du protocole de l'étude ET qu'il est volontaire pour participer à l'étude clinique. Le respect de ces critères permet de recruter une population homogène de participants dont la sécurité par rapport aux traitements proposés dans l'essai est garantie ainsi que leur appartenance à la population cible. Le résultat de ce processus est l'inclusion du patient dans l'étude clinique (**statut du patient : inclus**).

« **Préscreening** » ou « **eScreening** » dans le contexte de l'aide au recrutement, le processus de « préscreening » ou eScreening » est une phase préalable au screening utilisant des services numériques pour **automatiser le plus possible la sélection de patients potentiellement éligibles** (**statut du patient : préscreené**) et ainsi faciliter/accélérer le « screening ». Les solutions de préscreening peuvent notamment utiliser les données de santé de patients collectées lors des soins et stockées dans les dossiers patients informatisés, les EDS, les bases PMSI ou registres hospitaliers de spécialités, etc. Les patients préscreenés font l'objet d'une vérification manuelle de leur statut [Thadani09].

« **Système d'aide au recrutement dans des essais cliniques (SAREC) » (Clinical Trial Recruitement Support System (CTRSS)** : un système d'aide à la décision permettant d'automatiser en partie le processus de recrutement de patients dans des essais cliniques (67)[Cuggia11]. Les CTRSS font partie des interventions digitales en appui au recrutement. Parmi les CTRSS on distingue en particulier les services numériques qui utilisent les données de santé de patients collectées lors des soins pour identifier et sélectionner automatiquement des patients potentiellement éligibles (EHR-based CTRSS).

« Etude de faisabilité (d'une étude clinique) » : ensemble d'analyses permettant de s'assurer du bon déroulement de l'étude clinique. Un des objectifs d'une étude de faisabilité est d'identifier des sites ayant la capacité de conduire l'étude selon les exigences du promoteur de la recherche. Le questionnaire de faisabilité adressé aux sites potentiels permet au promoteur de s'assurer des capacités des sites. Il est de plus en plus attendu par les promoteurs que les sites fournissent des éléments probants de leur capacité à inclure des patients.

« **Potentiel d'inclusion** » : l'estimation du potentiel d'inclusion d'un site est un des objectifs du questionnaire de **faisabilité** qu'un promoteur de recherche adresse à un site afin de s'assurer que ce site a la capacité de conduire l'étude cliniques selon ses exigences. Le potentiel d'inclusion d'un site s'exprime par un nombre de patients que le site sera en capacité d'inclure pendant la durée de l'étude clinique, éventuellement associé à la cinétique de

recrutement. L'estimation de cette capacité à inclure repose actuellement le plus souvent sur des déclarations d'expert locaux (« leaders d'opinion » (PU-PH), investigateurs potentiels identifiés par le promoteur de la recherche). Les données de santé de patients collectées lors des soins peuvent être utilisées pour estimer automatiquement le potentiel d'inclusion. On peut faire l'hypothèse que l'effectif de patients respectant les critères d'éligibilité/d'inclusion d'un étude cliniques les années passées (file active de patients) est corrélé au nombre de patients qui seront effectivement inclus dans l'étude clinique et qu'il représente à ce titre un estimateur du nombre de patients potentiellement éligibles par le site. Les files actives de patients au niveau d'un site constituent ainsi un indicateur intéressant pour la sélection des sites.

« Algorithme de sélection de patients » : désigne un processus (suite d'étapes permettant d'obtenir un résultat à partir d'éléments fournis en entrée) qui peut être partiellement ou totalement automatisé et qui permet d'identifier au sein d'une source de données relatives à des patients (base de données médico-administrative, dossiers patients informatisés, registre de pathologies, registre de patients, etc.) des patients selon un ensemble de caractéristiques préalablement définies. Les algorithme de sélection de patients sont développés dans différents contextes : identification de patients éligibles à une prise en charge (p. ex. « screening » de patients éligibles à l'entrée d'une filière de prise en charge de l'ostéoporose), identification de patients potentiellement éligibles à l'inclusion dans un essai clinique ou dans une cohorte* épidémiologique, identification de patients dont les données sont susceptibles d'être utilisées dans le cadre d'une étude d'association phénome-génome. Les **algorithmes de « phénotypage »**[1] sont un sous-ensemble des algorithmes de sélection de patients ne prenant en compte que des caractéristiques phénotypiques (signes, symptômes, diagnostics : « phénome ») à l'exclusion de caractéristiques génétiques (« génome ») ou liées à l'exposition à des facteurs externes (« exposome »).

[1] L'utilisation d'algorithme de « phénotypage » n'est pas appropriée dans le cadre de la sélection de patients potentiellement éligibles à des études cliniques dont les critères d'éligibilité ne sont pas exclusivement liés à des caractéristiques de « phénome ».

ATHENA est un répertoire de vocabulaires internationaux standards en ligne développé par la communauté OHDSI, attribuant des **concept_ids** aux informations médicales d'intérêt correspondant à des « **concepts** », en lien avec leur alignement au sein de terminologies standards et non standards existantes.

Annexe 3a. Rationnel, synthèse et discussion du projet CovOnco

La première question épidémiologique d'intérêt fut d'évaluer l'évolution du nombre de nouveaux cas de cancer référés à l'AP-HP à partir du premier confinement. Comparée aux données de 2018 et 2019 pour la même période temporelle, la diminution de nouveaux cas de cancer a atteint 40%, quels que soient la classe d'âge et le type de cancer. Ces travaux ont fait l'objet d'une publication dans le *European Journal of Cancer* (331). Ce projet a permis d'évaluer l'évolution du nombre et la caractérisation en termes de stadification tumorale et de parcours de soins, des patients nouvellement référés à l'AP-HPavec un diagnostic de cancer solide lors de l'épidémie de Sars-Cov2. Un des enjeux était d'informer le pilotage des soins, en mettant ces analyses à dispositions des soignants et des différentes instances décideuses de l'institution. Le projet a bénéficié de mises à jour successives de la cohorte de patients, entre son ouverture le 12 juin 2020 et le 20 juin 2023.

Justification des types de cancers d'intérêt retenus dans le projet CovOnco

Après avoir mis en évidence une diminution du nombre de nouveaux cas référés à l'AP-HP entre mars et mai 2020 par rapport aux années précédentes, le projet CovOnco s'est intéressé aux cancers aux enjeux de santé publique les plus importants soit en termes de pronostic péjoratif (cancer du pancréas), soit en termes d'incidence et prévalence, cibles d'un programme national de dépistage organisé (cancer du côlon et cancer du sein), soit des deux (cancer du poumon) (331,332).

- Le cancer du sein

Le cancer du sein constitue un enjeu de santé publique majeur qui représente plus de 30 % des cancers féminins et dont la prévalence et les taux d'incidence continuent d'augmenter dans les pays développés malgré l'identification de facteurs de risque modifiables de ce type de cancer (332,333). Aux États-Unis, le risque de mortalité de le cancer du sein a certes diminué de 43% depuis son pic à la fin des années 80, principalement en raison de la précocité des diagnostics de cancer réalisés à un stade tumoral localisé (334). Malgré une certaine hétérogénéité de mise en œuvre en Europe depuis la fin des années 90, les programmes nationaux de dépistage de le cancer du sein ont en effet été associés à une réduction du risque de mortalité de 25% à 30% pour les femmes âgées de 50 à 74 ans (335). Par conséquent, les études de modélisation épidémiologique font état d'un chiffre de près d'un demi-million de décès par cancer du sein évités en Europe entre 1994 et 2025 (336).
Dans le cas spécifique de l'épidémie de SARS-CoV-2, de nombreux programmes nationaux de dépistage ont été interrompus pendant plusieurs mois en 2020 (337). Une méta-analyse a montré que le nombre de patients diagnostiqués d'un cancer du sein dans les suites d'une procédure de dépistage a chuté de 40 % au début de la pandémie par rapport à la période qui la précédait, un résultat confirmé par d'autres revues systématiques de littérature (338–340). Des études nationales ont montré que certains programmes nationaux de dépistage dédiés au cancer du sein ont pu être rétablis et recouvrer le niveau d'activité d'avant l'épidémie, comme au Royaume-Uni p. ex., mais pas encore aux États-Unis (341,342). De plus, les modélisateurs ont anticipé une augmentation de la mortalité par cancer du sein d'ici 2030 en raison des retards de diagnostic et de traitement en lien avec la pandémie de SARS-CoV-2, dans le monde entier (343,344). Cependant, les évaluations en situation réelle de l'impact de la pandémie sur les trajectoires de soins et les devenirs des patients récemment diagnostiqués

d’un cancer du sein dans la littérature incluent un faible nombre de patients et/ou manquent de suivi longitudinal clinique à moyen et long terme des patients concernés (344–349).

- Le cancer du pancréas

Le cancer du pancréas est une maladie très agressive, dont l’incidence ne cesse de croître et dont le pronostic reste sombre, malgré les innovations récentes en matière de procédure diagnostique et de traitement (1,350,351). Une temporalité resserrée des délais diagnostiques a été associée significativement à une amélioration de la survie des patients concernés, de sorte que le traitement antitumoral doit être initié le plus rapidement possible (352). Même si les recommandations internationales des sociétés savantes d’oncologie ont classé certains cas de cancer du pancréas nouvellement diagnostiqués comme « hautement prioritaires » de soins pendant la pandémie de SARS-CoV-2, d'autres catégories de patients, tels que les patients âgés nouvellement porteurs d’un cancer du pancréas, ont été considérées comme des « priorités moyennes » (353). En raison des retards potentiels de diagnostic et d’initiation de traitement spécifique, les modélisateurs ont anticipé une augmentation de la mortalité spécifique au cancer du pancréas dans les années à venir (53,354,355). Néanmoins, les évaluations empiriques de l'impact de la pandémie sur les trajectoires de soins et les résultats des nouveaux cas de cancer du pancréas manquent d’exhaustivité (356).

- Le cancer du côlon

Le cancer du côlon représente la deuxième cause de mortalité par cancer aux Etats-Unis, et fait la cible d’un dépistage organisé en France depuis 2003, qui sera porté effectivement à l’échelle nationale en 2008 (357,358). Au début de la pandémie de SARS-CoV-2, les décideurs politiques ont interrompu les programmes nationaux de dépistage du cancer, y compris celui dédié au cancer colorectal (47,49). Plusieurs études ont mis en évidence une diminution significative du nombre de cas de cancers colorectaux nouvellement référés aux hôpitaux tertiaires et des procédures de colectomies connexes effectuées pendant les périodes de confinements nationaux, et ce, sans rattrapage séquentiel (359–361). Des données préliminaires ont montré que les patients atteints d'un nouveau cancer colorectal diagnostiqué après le début de la période d'épidémie de SARS-CoV-2 auraient pu être plus susceptibles de souffrir de tumeurs plus avancées lors de la présentation clinique initiale et de survivre moins longtemps, comparativement à la période précédant l’épidémie (51,52,362–364). Par ailleurs, le développement local du cancer colorectal peut provoquer des complications potentiellement mortelles telles que l'occlusion intestinale, la perforation intestinale ou l'hémorragie digestive nécessitant des interventions chirurgicales à risque non nul et effectuées sans délai (365). Les retards diagnostiques du cancer colorectal peuvent également augmenter le taux de tumeurs non résécables, donc sans objectif de rémission complète à long terme (366). Cette situation clinique est associée à une diminution des taux de survie globale des patients concernés, car la résection complète des sites tumoraux, y compris les métastases, est la référence thérapeutique de la prise en charge du cancer du côlon (367). Une étude de modélisation anglaise prévoyait, en mars 2021, un recouvrement du taux d’activité standard atteint plus de deux ans après le début de l’épidémie, avec un rattrapage de 162 735 coloscopies à réaliser dans cet intervalle de temps et correspondant à une augmentation transitoire de procédures de +130% (368).

- Le cancer du poumon

Aux Etats-Unis, les prédictions épidémiologiques anticipent qu'en 2023, le cancer du poumon sera diagnostiqué chez environ 240 000 individus, et entraînera environ 130 000 décès, avec un taux de médiane de survie globale à 5 ans atteignant 25% (332). L'interruption transitoire des procédures de diagnostic du cancer du poumon, telles que le suivi de nodules pulmonaires suspects, au cours de la première vague de l'épidémie pourrait expliquer la réduction du nombre de patients nouvellement diagnostiqués, et ce, dès 2020 (369). Certains nouveaux diagnostics de cancer du poumon pourraient avoir été faits de façon fortuite à un stade tumoral précoce, à l'occasion de tomodensitométries thoraciques réalisées en urgence chez les suspicions cliniques d'infection par le SARS-CoV-2, mais cette hypothèse est faiblement étayée par la littérature disponible (370). De nombreuses études mirent en évidence que les procédures curatives de résection carcinologique pulmonaire semblaient faisables et non grevées d'une surmortalité pendant les pics épidémiques comparées aux années antérieures, ce qui rend la diminution d'accès aux soins tertiaires une réelle perte de chance pour les patients (371–374). Néanmoins, la vigilance reste de mise car, en 2020, la survenue d'une infection par SARS-CoV-2 chez des patients en cours de traitement antitumoral systémique (chimiothérapie, thérapies ciblées, immunothérapie) semblait représenter un facteur pronostic péjoratif significatif, correspondant probablement au surrisque de 30% à 43% étayé par la littérature d'oncologie générale (375,376). Au total, il semblait que les modifications de parcours de soins des patients diagnostiqués d'un cancer du poumon pendant l'épidémie aient particulièrement concerné les situations palliatives exclusives alors que les indications curatives ont pu être maintenues de façon optimale (377,378).

Méthodes

Le projet CovOnco s'est principalement structuré autour des données administratives de l'AP-HP utilisées pour le système national de pilotage hospitalier, à savoir le Programme de Médicalisation des Systèmes d'Information (PMSI). Un dossier PMSI contient, pour le séjour hospitalier de chaque patient, les diagnostics principaux, reliés et associés codés selon la CIM-10, les procédures diagnostiques et thérapeutiques codées selon la CCAM, l'âge, le sexe et le lieu de résidence du patient. Son séjour à l'hôpital est décrit par les variables suivantes : groupe lié au diagnostic, dates et modalités d'entrée et de sortie du séjour (domicile, autre structure hospitalière, décès), et informations supplémentaires sur les services spécifiques (p. ex. unité de soins intensifs) fréquentés. S'inspirant de l'algorithme développé par l'INCa en 2013, les patients atteints de cancer ont été sélectionnés à l'aide d'une liste des codes CIM-10 dédiés au cancer C00 à D48, à l'exclusion des tumeurs bénignes (D10 – D36) (Annexe 3b) (379). La population d'intérêt a compris les patients référés entre le 1er janvier 2019 et le 31 décembre 2020 vers l'un des 28 hôpitaux universitaires de l'AP-HP où était déployé ORBIS, le système d'information de soins. Afin de bénéficier d'au moins deux ans d'historique médical dans les dossiers patients informatisés, seuls les hôpitaux ayant déployé le logiciel métier ORBIS avant janvier 2016 ont été pris en compte. Ont été inclus les patients 1/ avec un code CIM-10 correspondant au cancer primitif d'intérêt (diagnostic principal ou relié), non enregistré précédemment au moins au cours des dix-huit à vingt-quatre mois précédents (en fonction du type de cancer primitif d'intérêt), et 2/ sans aucun autre code de cancer primitif afin d'éliminer les cas de cancers multiples. Les patients âgés ont été définis comme ceux de plus de 70 ans. Pour des raisons réglementaires, les analyses se sont restreintes aux patients qui étaient soit décédés au moment de l'analyse, soit passés à l'AP-HP depuis juillet 2017.

Nombre de nouveaux cas référés avec un cancer pendant les pics épidémiques
Nous avons comparé le nombre mensuel de cancers enregistrés en 2020 avec ceux enregistrés en 2018 et en 2019 et avec le nombre moyen enregistré entre 2018 et 2019. Nous avons effectué des comparaisons agrégées entre les périodes de 3 mois correspondant au premier confinement français (1er mars au 31 mai 2020) et au post-confinement (1er juin au 31 septembre 2020). Nous avons comparé ces indicateurs entre les périodes temporelles précédant l'épidémie et la suivant, avec un intérêt particulier pour les deux périodes de confinement en France (17 mars-11 mai 2020, 30 octobre-15 décembre 2020).

Délais de prise en charge initiale
Nous avons évalué le temps écoulé entre la première réunion de concertation pluridisciplinaire (RCP) d'un patient et le début de son traitement antitumoral. La date de la RCP était disponible dans les informations structurées liées à chaque rapport RCP, et nous avons identifié la date de traitement à l'aide des données des réclamations. Deux analyses distinctes ont été effectuées en fonction de l'événement qui s'est produit en premier. Pour les patients ayant reçu un diagnostic de cancer réalisé dans les hôpitaux de l'AP-HP, nous avons également calculé le délai médian entre le rapport de pathologie avant le traitement du cancer et la survenue du premier traitement.

Identification des parcours de soins et devenirs cliniques
Pour analyser les stratégies de traitement, nous avons classé les patients qui ont reçu des soins anti-tumoraux en au moins trois catégories mutuellement exclusives : chirurgie de la tumeur primitive correspondant aux stratégies curatives (indépendamment de tout traitement périopératoire du cancer) selon les codes de la CCAM (Annexe 3c), traitement antitumoral systémique parentéral exclusif (chimiothérapie CIM-10 Z511, indépendamment de la radiothérapie) correspondant aux stratégies palliatives actives, meilleurs soins de soutien exclusifs (CIM-10 Z515 sans traitement antitumoral actif du cancer). Nous avons calculé la moyenne mobile sur 3 mois (moyenne du mois précédent, du mois associé et du mois suivant) de la distribution du traitement, dans les 18 mois suivant le premier code de diagnostic du cancer.

Devenir clinique des patients
L'infection par le SARS-CoV-2 a été définie comme un résultat positif à la PCR, un test sérologique positif ou la présence de l'un des codes U071 de la CIM-10 l'année suivant la date du diagnostic de cancer. Nous avons analysé la survie globale (SG) des patients cancer du sein en utilisant le lien systématiquement établi entre le CDW de l'AP-HP et le registre national des décès (RND) supervisé par l 'Institut national de la statistique et des études économiques (INSEE). La SG d'un patient a été définie comme le temps écoulé entre la date de la première occurrence d'un code de la CIM-10 av. J.-C. et la date du décès du patient. Les patients vivants ont été censurés à la date de la dernière mise à jour de la NDR dans le CDW de l'AP-HP (juin 2022).

Analyse statistique
Les courbes de survie ont été tracées à l'aide de la méthode de Kaplan-Meier. Nous avons comparé la SG des patients référés en 2019 à ceux référés en 2020 (globalement, par classe d'âge et par catégorie de traitement antitumoral). Pour ce faire, nous avons estimé les hazard

ratios (HR) et leurs intervalles de confiance (IC) à 95 % à l'aide d'un modèle de risque proportionnel de Cox variant dans le temps, avec l'âge comme covariable constante et l'infection par le SARS-CoV-2 comme covariable variant dans le temps. D'autres variables catégorielles ont été comparées à un test χ^2 et des variables quantitatives à un test U de Mann-Whitney. L'extraction finale des données a été effectuée le 5 décembre 2022. La signification statistique a été fixée à $p<0,05$. Tous les calculs ont été effectués à l'aide de Python version 3.7 (www.python.org).

Résultats

Entre janvier et septembre 2020, 28 348, 27 272 et 23 734 patients atteints d'un nouveau cancer ont été référés à l'AP-HP Grand Paris en 2018, 2019 et 2020, respectivement. La diminution des nouveaux cas de cancer a été constante dans tous les types de tumeurs et s'est poursuivie après le confinement : -30% et -9% pour le côlon, -27% et -6% pour le poumon, -29% et -14% pour le sein, -33% et -12% pour les cancers de la prostate, respectivement, de mars à mai et de juin à septembre 2020 par rapport à la moyenne 2018-2019, respectivement. Des schémas de diminution similaires ont été observés pour les tumeurs de mauvais pronostic : -34% et -16% pour le pancréas, -30% et -17% pour la vessie, -32% et -7% pour le système nerveux central et -29% et -9% pour les cancers du foie, respectivement.
Globalement et particulièrement pour les types de cancers primitifs d'intérêt, aucune augmentation de délais de prise en charge initiale, et aucune aggravation des stades tumoraux au diagnostic a été identifié (Annexe 5). La distribution des catégories des traitements antitumoraux n'a pas été modifiée entre la période pré-pandémique et après 2020. Les éventuels surrisque de mortalité à un an identifiés ont été associées à des infections par le SARS-CoV-2 lui-même.

Discussion

Les politiques de santé publique appliquées à la pandémie de SARS-CoV-2 ont entraîné une baisse substantielle des nouveaux cas de cancer en région parisienne. Les nouveaux cas ont encore été inférieurs aux attentes en septembre, ce qui montre que la situation n'était pas revenue à la normale même après la levée du confinement. Ces résultats suggèrent un retard dans les nouveaux diagnostics et le traitement des cas de cancer.

Nos résultats sont conformes à la littérature publiée. Aux Pays-Bas, le nombre de diagnostics de cancer a diminué de 25 % pour les cancers autres que les cancers de la peau et de 60 % pour les cancers de la peau (à l'exclusion des carcinomes basocellulaires) après la mise en œuvre des politiques de « distanciation sociale » (380). Une étude transversale américaine a montré que l'incidence hebdomadaire de six principaux types de cancer a chuté de 46% pendant la crise pandémique par rapport au niveau de référence (381). Cette diminution est conforme à d'autres études et a atteint 52 % et 49 % pour les cancers du sein et colorectal, respectivement (382). Une étude anglaise a montré une diminution de 76% des références urgentes pour un diagnostic précoce dans 8 centres de soins tertiaires, pendant la crise (383). Une étude italienne avait rapporté une diminution du dépistage du cancer buccal pendant le 1er confinement (384). De même, une étude observationnelle multicentrique américaine a conclu que l'incidence des nouveaux cancers avait été divisée par 2 entre avril 2019 et avril 2020 sur une population de 28 millions de patients (115). Dans cette étude, la baisse de l'incidence du cancer était plus élevée pour les cancers du sein, de la prostate et du mélanome.

Le dépistage du cancer du sein et du cancer colorectal avait diminué de 89 % et 85 %, respectivement. Une étude italienne a conclu que l'incidence des tumeurs neuroendocriniennes avait diminué de 77% pendant le confinement national (385).

Les résultats de notre étude confirment que l' *effet de distraction* induite par l'épidémie de SARS-CoV-2 sur la prise en charge des patients cancéreux (386). La pandémie peut entraîner des conséquences imprévues subtiles telles que le stress psychologique et l'isolement social, la diminution de la prestation de soins pour des conditions potentiellement mortelles, des perturbations économiques et logistiques dans l'administration des médicaments. Dans une étude française réalisée sur 125 000 cabinets de professionnels de santé, les patients ont déclaré que la peur de la contamination (38%), la réticence à déranger indûment un médecin en pleine crise (28%) et les cabinets médicaux fermés (17%) étaient les principales raisons de ne pas consulter un médecin (387). Dans un contexte de réduction des ressources en soins de santé, les efforts doivent être équilibrés entre la menace aiguë de la propagation du virus et les problèmes cliniques à plus long terme liés aux maladies graves, telles que le cancer. Retarder le diagnostic de la tumeur pourrait empêcher les tumeurs curables d'être traitées efficacement et nuire considérablement aux résultats cliniques pour les patients (388,389).

Grâce à une étude de modélisation, Lai *et al.* ont estimé que cette prise en charge insuffisante pourrait entraîner un excès de 6 270 et 33 890 décès à un an chez les patients anglais et américains atteints d'un cancer, respectivement (390). Une autre étude modèle anglaise a évalué que les décès supplémentaires liés au cancer sur 5 ans dus à la pandémie atteindraient 3 500 pour 4 principaux types de cancer (53). Les sociétés internationales de lutte contre le cancer ont exhorté les patients atteints de cancer à revenir dans les cliniques, tandis que le SARS-CoV-2 continuait de se propager lors des vagues suivantes (391). Anticipant une perte d'opportunités pour les patients atteints de cancer, de nombreuses communautés scientifiques ont publié des recommandations relatives aux soins contre le cancer pendant la pandémie (392–394). Les médecins généralistes ont plaidé en faveur d'un diagnostic rapide du cancer symptomatique (395). Les programmes nationaux de dépistage du cancer devraient se poursuivre tout en revendiquant la sécurité des patients en ce qui concerne le risque de contamination par le virus afin de réduire l'arriéré dans les programmes habituels de dépistage du cancer (396). Des procédures de dépistage innovantes, non invasives et résilientes face à des ressources limitées devraient être élaborées et déployées (397). Dans une enquête américaine menée auprès de 3 055 patients et survivants du cancer, la moitié d'entre eux ont signalé un certain impact de la pandémie sur leurs soins de santé et 27% des patients traités activement ont exprimé un certain retard dans leur traitement (398).

Un biais de délai pourrait expliquer la survie plus faible chez les patients diagnostiqués après l'épidémie de SARS-CoV-2 (399). Nos résultats sont basés sur des bases de données administratives et doivent donc être interprétés avec prudence, car ce type d'études épidémiologiques comporte un risque de biais internes (400). Cependant, comme le codage du dossier médical est obligatoire pour des raisons de financement, la probabilité de biais sur le nombre de cas enregistrés est faible.

Annexe 3b. Catégorisation topographique des cancers primitifs en fonction de la nomenclature de la classification internationale des maladies, 10e révision (CIM-10)

Topographie du cancer primitif	Code issu de la classification internationale des maladies, 10e révision (CIM-10)
Anus	C21
Voies biliaires	C23 C24 D01.5 D37.6
Vessie	C66 C67 C68 D09.0 D09.1 D41.2 D41.3 D41.4 D41.7 D41.9
Intestin grêle	C17 D01.4 D37.2
Sein	C50 D05 D48.6
Col utérin	C53 D06
Système nerveux central	C70 C71 C72.0 C72.2 C72.3 C72.8 C72.9 D42 D43.0 D43.1 D43.2 D43.4 D43.7 D43.9
Colon	C18 C19 D01.0 D01.1 D37.3 D37.4
Carcinome de primitif indéterminé	C76 C80 C97 D09.7 D09.9 D48.7 D48.9 D48.3
Autre digestif	C26 C48 D01.7 D01.9 D37.7 D37.9 D48.4
Endomètre	C54 C55 D07.0 D39.0
Œil	C69 D09.2
Estomac	C16 D00.2 D37.1
Autre gynécologique	C51 C52 C57 C58 D07.1 D07.3 D39.2 D39.7 D39.9
Tête et cou	C0 C10 C11 C12 C13 C14 C30 C31 C32 D00.0 D02.0 D37.0 D38.0
Lymphome de Hodgkin	C81
Rein	C6.4 C6.5 D41.0 D41.1
Leucémie	C91 C92 C93 C94.0 C94.1 C94.2 C94.3 C94.4 C94.5 C94.7 C95
Foie	C22
Lung	C33 C34 D02.1 D02.2 D38.1
Maladie maligne immunoprolifération	C88 C94.6 D45 D46 D47

Topographie du cancer primitif	Code issu de la classification internationale des maladies, 10e révision (CIM-10)
Mélanome	C43 D03
Mésothéliome	C45.0 C45.1 C45.2 C45.7 C45.9
Myélome	C90
Lymphome non-Hodgkin	C82 C83 C84 C85 C86
Œsophage	C15 D00.1
Ostéosarcome	C40 C41 D48.0
Autre endocrinien	C74 C75 D09.3 D44.1 D44.2 D44.3 D44.4 D44.5 D44.6 D44.7 D44.8 D44.9 D444.0 D444.8
Autre tumeur maligne	C96
Ovaire	C56 D39.1
Pancréas	C25
Autre pulmonaire	C37 C38 C39 D02.3 D02.4 D38.2 D38.3 D38.4 D38.5 D38.6
Système nerveux périphérique	C47 C72.1 C72.4 C72.5 D43.3 D48.2
Prostate	C61 D07.5 D40.0
Rectum	C20 D01.2 D37.5
Autre cutané	C44 D04 D48.5
Tissus mous	C46 C49 D48.1
Testicule	C62
Thyroïde	C73 D44.0
Urothélial	C60 C63 D07.4 D07.6 D40.7 D40.9

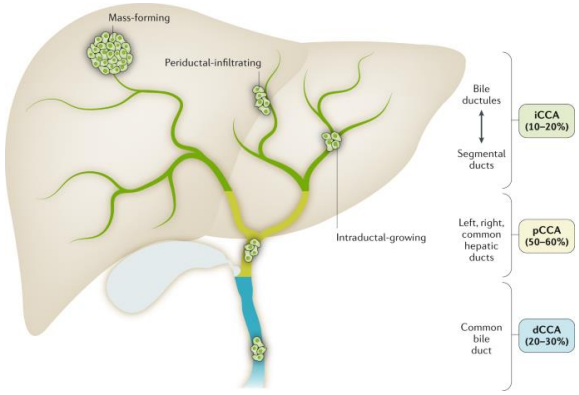
Annexe 3c. Liste des codes reliés aux résections chirurgicales des cancers colorectaux, pancréatiques, pulmonaire, mammaires invasifs et de cholangiocarcinomes selon la Classification Commune des Actes médicaux (CCAM)

Type de cancer et de chirurgie	Code CCAM
Cancer colorectal	
Colectomie	HJFA004, HJFA001, HJFA002, HJFA006, HJFA007, HJFA011, HJFA012, HJFA014, HJFA017, HJFA019, HJFC023, HJFC031, HJFA008, HHFA018, HHFA009, HHFA026, HHFA023, HHFA006, HHFA022, HHFA008, HHFA021, HHFA017, HHFA010, HHFA014, HHFA005, HHFA024, HHFA004, HHFA002, HJFC023, HJFA012HHFC296, HHFC040, HHFA030, HHFA031, HJFA006, HJFA007, HJFA019, HJFA005, HJFA003, HJFA018, HJFD002, HJFA004, HJFA002, HJFA001, HJFA015, HJFA016, HHFA029, HJFA017, HHFA028
Cancer du sein	
Tumorectomie	
• avec curage	QEFA001, QEFA008
• sans curage	QEFA004, QEFA016, QEFA017, QEFA018
Mastectomie	
• avec curage	QEFA003, QEFA005, QEFA010, QEFA020
• sans curage	QEFA007, QEFA012, QEFA013, QEFA015
Cancer du pancréas	HGFA014, HNFA001, HNFA002, HNFA004, HNFA005, HNFA006, HNFA007, HNFA008, HNFA010, HNFA011, HNFA013, HNFC001, HNFC002, HNFC028
Cancer du poumon	GFFA001-002, GFFA004, GFFA006-013, GFFA015-016, GFFA018-019, GFFA021-031, GFFA033-034, GFFC002
Cholangiocarcinome	HLFA003-07, HLFA009-12, HLFA0014, HLFA0017-20, HLFC002-04, HLFC027, HLFC032, HLFC037, HLFC801

Annexe 4. Rationnel et méthodologie du projet 'Challenge AI for health' dans sa version développée à l'Assistance Publique – Hôpitaux de Paris

Rationnel scientifique

Les cholangiocarcinomes constituent un groupe de maladies cancéreuses hétérogènes. Cette affliction est relativement rare, représentant environ 3% des tumeurs digestives et environ 10% des tumeurs primaires du foie. L'incidence du cholangiocarcinome varie fortement d'un continent à l'autre, s'approchant de 100 par 100 000 dans certaines régions d'Asie et étant inférieure à 1 par 100 000 en Amérique du Nord. Son incidence est en revanche croissante au niveau mondial sur les dernières décennies (59). Son pronostic est mauvais avec un taux de survie à 5 ans ne dépassant pas 10 %. La résection chirurgicale, unique option curative, n'est réalisée que dans un tiers des cas, en raison d'un diagnostic généralement posé à un stade avancé. Le développement tumoral du cholangiocarcinome demeure en effet silencieux et ce, jusqu'à un stade évolué. Cette pathologie hétérogène se classe selon la localisation anatomique du cancer : extra-hépatique, périhilaire, et intrahépatique. Cette dernière catégorie tumorale a vu son incidence augmenter récemment dans les pays développés alors que son taux de mortalité a triplé. Les traitements systémiques de chimiothérapie peuvent être discutés en péri-opératoire et sont indiqués aux stades avancés inopérables. Les options thérapeutiques systémiques sont limitées et leur efficacité reste moyenne en termes de survie globale (60).



1.

Figure 22. Classification anatomique des cholangiocarcinomes - image reproduite de (59)

Il serait utile de disposer d'éléments permettant de prédire l'agressivité de ces tumeurs et l'efficacité des traitements locaux et généraux. Des classifications moléculaires ont mis en évidence des profils évolutifs et potentielles cibles thérapeutiques distincts au sein du cholangiocarcinome intra-hépatique : les tumeurs développées aux dépens des petits canaux biliaires seraient associées à des mutations de *KRAS* et *TP53*, alors que les cancers des plus gros canaux arboreraient plus fréquemment des anomalies du *FGFR-2* et d'*IDH-1 /-2* (401). De récentes classifications moléculaires de la tumeur et de son microenvironnement ouvrent le chemin pour l'identification et le déploiement de thérapies ciblées de nouvelle génération (402). L'ivosidenib, un inhibiteur d'IDH semble doubler la survie globale en 3ème ligne des cholangiocarcinomes avancés mutés *IDH1*, et a, de ce fait, reçu une autorisation de mise sur le marché aux Etats-Unis et en Europe, alors qu'il est, dans l'attente de son prix de remboursement, la cible d'un accès précoce en France (403,404). En parallèle, la prise en

compte du micro-environnement tumoral dans la caractérisation moléculaire du cholangiocarcinome intra-hépatique a révélé des profils pronostiques spécifiques ainsi que des formes « chaudes » potentiellement sensibles aux traitements par immunothérapie pour 10 % des cas. D'autres facteurs pronostiques, tels que les caractéristiques de la fibrose tumorale, ont été plus récemment mis en évidence. Ce type d'analyse génique demeure pour le moment réservé au domaine de la recherche translationnelle, et nécessite un accès, encore trop mal réparti sur le territoire français, aux séquenceurs (405).

La question de la détermination de ces sous-groupes moléculaires à partir des colorations standards d'anatomopathologie (HES) ou des images de scanners reste entière et permettrait de simplifier et de rendre accessible la pronostication moléculaire des tumeurs des patients, via la réalisation de biopsies virtuelles. Les outils d'intelligence artificielle rassemblés sous le vocable « radiomics » modélisent les données qualitatives d'imagerie en les transformant en informations quantitatives, permettant ainsi une interprétation objective, reproductible et automatisée. Les modèles d'apprentissage profond qui s'entraînent sur ces jeux de données sont capables d'« apprendre » à discerner la typologie des lésions et leurs caractéristiques, afin d'en préciser les spécificités en termes de pronostic. La « radiomics » cherche également à évaluer le profil moléculaire tumoral en fonction des caractéristiques d'imagerie correspondante, suivant la voie de la pronostication ouverte par la bio-informatique. L'anatomopathologie est un horizon prometteur pour l'utilisation des outils d'intelligence artificielle en raison de la numérisation possible des images des prélèvements de tissus, et de la grande variabilité d'interprétation intra- et inter-observateurs. La caractérisation de ces images permet un raffinement dans la classification des tumeurs, à des fins d'évaluation pronostique (406). L'imagerie biomédicale a le potentiel d'offrir de manière efficace, efficiente et non-invasive des informations critiques sur le sous-type de la tumeur, notamment sur la caractérisation des anomalies moléculaires tumorales et de guider le choix de la stratégie thérapeutique la plus adaptée au patient. Ce potentiel est encore à dévoiler pour les cholangiocarcinomes, car l'absence de bases de données de qualité et de grande échelle a, jusqu'ici, empêché le déploiement de ce type d'études.

Méthodologie générale

Les données mises à disposition appartenaient à 140 patients opérés d'un cholangiocarcinome intra-hépatique. L'objectif du projet consistait à prédire, à partir des images de 600 coupes de scanners abdomino-pelviens standards pré-opératoires avec injections de produits de contraste iodés, les caractéristiques histopronostiques des tumeurs opérées suivantes : taille de la tumeur, niveau de différenciation, nombre de nodules, exhaustivité de la résection chirurgicale, présence d'invasion néoplasique périnerveuse, présence d'embole tumoral, stade tumoral pTNM, présence d'invasion tumorale ganglionnaire. Les tâches de classification étaient binaires ou multi-classes, et nécessitaient l'extraction textuelle des informations correspondantes à partir des comptes rendus d'anatomopathologie postopératoires des tumeurs réséquées. Ainsi, une étape indispensable et préalable au développement de tels modèles de prédiction consistait à structurer ces informations textuelles afin de pouvoir les intégrer dans les modèles de vision par ordinateur précédemment décrits.

Annexe 5. Résultats de CovOnco concernant les stades pTNM et métastatiques des nouvelles tumeurs du pancréas, poumon et mammaires référées à l'Assistance Publique – Hôpitaux de Paris

Cancer du pancréas

Parmi 588 patients opérés d'un cancer du pancréas, 464 ont bénéficié de cette chirurgie comme première étape des traitements antitumoraux et 456 (98%) avaient un compte rendu anatomopathologique post-opératoire disponible (385 (83%) patients avaient un score pTNM identifiable dans le compte rendu). La répartition des groupes de risque de pTNM n'a pas varié d'une période à l'autre : 50 % contre 52 % pour la catégorie à faible risque, 50 % contre 48 % pour la catégorie à risque élevé en 2019 et 2020-2021, respectivement (p = 0,80). Parmi les 124 patients opérés après traitement néoadjuvant, 117 (94%) avaient un compte rendu anatomopathologique post-opératoire disponible (et 111 (89%) patients avaient un score ypTNM identifiable dans le compte rendu). La répartition des groupes de risque ypTNM n'a pas varié d'une période à l'autre : 59 % contre 70 % pour la catégorie à faible risque, 41 % contre 30 % pour la catégorie à risque élevé en 2019 et 2020-2021, respectivement (p = 0,30). Parmi 855 patients nouvellement référés à l'AP-HP avec un cancer invasif du pancréas et pour lesquels un compte rendu de TDM TAP de stadification tumorale initiale était disponible, la proportion de cancers métastatiques ne différait pas entre 2019 (36 %) et 2020-21 (33 %) (p = 0,39).

Cancer du poumon

Concernant le cancer du poumon, parmi les 2 213 comptes rendus d'anatomopathologie disponibles dans les suites d'une résection de nouveau cancer bronchique dans l'EDS de l'AP-HP, 1 893 présentaient une évaluation du score pTNM. La répartition des groupes à risque de pTNM n'a pas changé au fil du temps, soit 62 % contre 64 % pour les groupes à faible risque et 38 % contre 36 % pour les catégories à risque élevé en 2018-2019 par rapport à 2020, respectivement (p = 0,40).

Parmi les 2 602 patients présentant un compte rendu de TDM TAP de stadification tumorale initiale dans l'EDS de l'AP-HP, les taux de cancer métastatique n'ont pas changé au fil du temps, étant de 29% et 28% en 2018-2019 et 2020, respectivement (p = 0,78).

Cancer du sein

Concernant le cancer du sein, parmi 2 924 patients du projet CovOnco qui ont bénéficié d'une résection de cancer du sein, 2 762 (94 %) avaient un compte rendu anatomopathologique post-opératoire disponible, dont 1 628 (59 %) avaient un score (y)pTNM mentionné dans le rapport. Parmi les 1 332 patients ayant bénéficié d'une résection initiale du cancer du sein et de données disponibles dans l'EDS de l'Assistance Publique – Hôpitaux de Paris, la répartition entre les groupes à risque pTNM n'a pas changé au fil du temps : 168 (25%) et 148 (23%) pour la catégorie à haut risque en 2019 et 2020, respectivement (p = 0,37). Parmi les 296 patients opérés après traitement néoadjuvant, la répartition entre les groupes à risque ypTNM ne différait pas au fil du temps : 55 (37%) et 67 (45%) pour la catégorie à haut risque en 2019 et 2020, respectivement (p = 0,31). Un autre article anglais étudiant la faisabilité du calcul des indicateurs qualité EUSOMA est en cours de révision par le comité éditorial de la *Revue*

d'Epidémiologie et de Santé Publique, après avoir été présenté à l'oral lors du congrès EMOIS de 2023. Les résultats en termes de portabilité du développement de l'expression régulière relative au score pTNM sont présentés dans la partie relative à l'extraction textuelle des critères histopronostiques.
Parmi les 725 patients pour lesquels un TEP-scanner et/ou un TDM TAP de stadification tumorale initiale étaient disponibles (18 % de la population totale), la proportion de cancers métastatiques ne différait pas au fil du temps : 15 % et 15 % en 2019 et 2020, respectivement.

Annexe 6. Algorithme de post-traitement en vue d'une caractérisation à l'échelle du document, pour les entités extraites correspondant aux critères histologiques pronostiques des tumeurs réséquées

- La section Conclusion est la section d'extraction prioritaire, sur le reste du document
- Si plusieurs tailles de tumeur sont extraites, la plus grande doit être conservée (la section Conclusion étant prioritaire)
- Si plusieurs niveaux de différenciation tumorale sont extraits, le mauvais pronostic doit être maintenu (« médiocre » > « léger » > « bien »)
- Si l'exhaustivité de la résection tumorale multiple est extraite, le mauvais pronostic doit être conservé (« résection incomplète » > « résection complète »)
- Si des atteintes tumorales ganglionnaires multiples sont extraites, le mauvais pronostic doit être conservé (« atteinte tumorale ganglionnaire » > « atteinte tumorale sans ganglion »)
- Si plusieurs niveaux d'invasion tumorale périnerveuse sont extraits, le mauvais pronostic doit être maintenu (« invasion tumorale périnerveuse » > « pas d'invasion tumorale périnerveuse »)
- Il est extrait de plusieurs niveaux de marge de résection avec atteinte tumorale, le mauvais pronostic doit être maintenu (« marge de résection avec atteinte tumorale » > « marge de résection sans atteinte tumorale »)
- Si plusieurs distances entre le front tumoral et la marge de résection sont extraites, les plus petites doivent être conservées (description numérique et linguistique)
- Si plusieurs scores pTNM sont extraits, celui lié au pronostic le plus sombre doit être conservé (« M1 » > « M0 », puis « N2 » > « N1 » > « N0 », puis « T4 » >"T3">"T2"> « T1 »

Gold standard (annotation)	Sortie de l'algorithme	Classification
Pas d'annotation	Pas d'annotation	VP
Pas d'annotation	Envahissement ganglionnaire	FP
Pas d'annotation	Pas d'envahissement ganglionnaire	FP
Envahissement ganglionnaire	Pas d'annotation	FN
Envahissement ganglionnaire	Envahissement ganglionnaire	VP
Envahissement ganglionnaire	Pas d'envahissement ganglionnaire	FP
Pas d'envahissement ganglionnaire	Pas d'annotation	FN
Pas d'envahissement ganglionnaire	Envahissement ganglionnaire	FP
Pas d'envahissement ganglionnaire	Pas d'envahissement ganglionnaire	VP

Annexe 7. Guide d'annotations

Remarques générales

Ne sont pas annotés :

- les articles définis, chiffres et indéfinis,
- les zones de textes correspondant aux « Renseignements / contexte cliniques »,
- les intitulés de sections décrivant les différents éléments organiques analysés (« ganglion inter-aortico-cave », etc)
    - à distinguer des descriptions suivant « : » qui sont annotés de part et d'autre des « : »

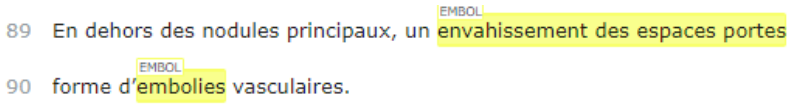
Ne sont pas exclus :

- les comptes rendus d'anatomopathologie en format de tableur.

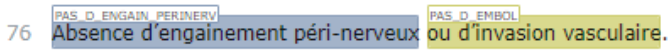
Une annotation ne peut pas englober plus d'une phrase.

Le respect de la grammaire n'est pas une limite à l'annotation.

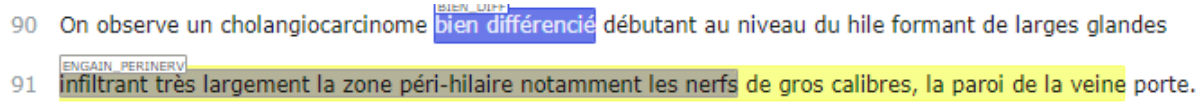
Au sein d'une même phrase, les redondances sémantiques sont annotées séparément :



Les marques de négation sont intégrées dans les annotations « pas de ». En cas de liste, le « ni de » / « ou de » est annoté. Ex CR 87084 :



Il peut y avoir des superpositions d'annotations. CR 1516519 :



Liste des entités

1. Taille tumorale
2. Différenciation
    a. Bien différencié
    b. Moyennement différencié
    c. Peu différencié
3. Complétude de la résection microscopique
    a. Résection complète
    b. Résection incomplète
4. Distance minimale entre la tumeur et les marges de résection

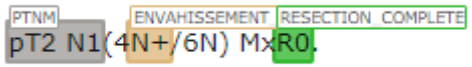
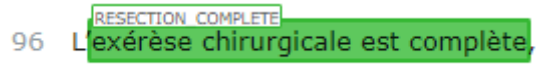
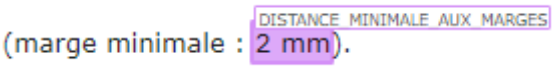
    Seuil millimétrique de la distance minimale entre la tumeur et les marges de résection
    a. Marge infra-millimétrique
    b. Marge supra-millimétrique
5. Envahissement d'une limite de résection
    a. Limite envahie
    b. Limite non envahie
6. Présence d'engainement périnerveux
    a. Engainement périnerveux
    b. Pas d'engainement périnerveux
7. Invasion tumorale vasculaire
    a. Embols vasculaires
    b. Pas d'embol vasculaire
    c. Envahissement intra-vasculaire par la tumeur
    d. Pas d'envahisement intra-vasculaire par la tumeur
8. Présence d'envahissement ganglionnaire
    a. Envahissement ganglionnaire
    b. Pas d'envahissement ganglionnaire
9. Classification TNM de la tumeur

    Type de classification employée

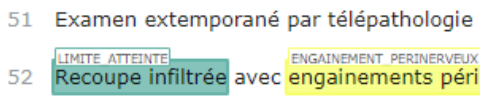
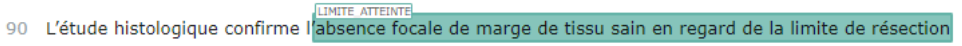
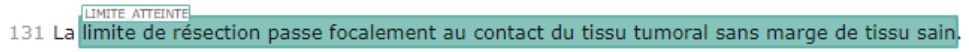
Exemples d'annotations par entité

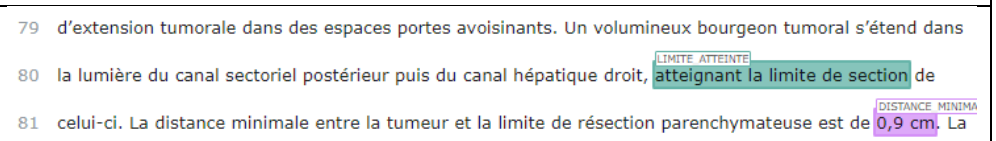
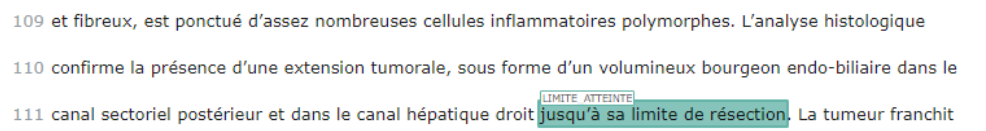
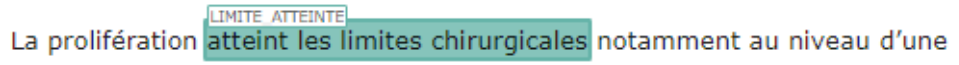
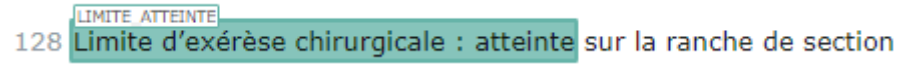
Label	Section textuelle d'intérêt	Définition	Exemples
Taille	Description macroscopique et description microscopique	Il s'agit de la taille du nodule tumoral : - Une dimension (plus grand axe) - En deux ou trois dimensions, dont le grand axe sera tiré en post-processing En cas de présence de plusieurs nodules tumoraux, l'ensemble des lésions est pris en compte dans une même annotation. La taille est décrite en « unité de distance (^{2}) » suivie d'une dimension qui peut être écrite en abrégé : centimètre (cm), millimètre (mm). En cas de format sous tableur des résultats pour de multiples nodules, l'ensemble des tailles est annoté en un temps (cf exemple 2).	Taille: TAILLE 16 x 16 x 8 cm 2 Localisations hépatiques mesurant TAILLE 52 et 12mm d'un adénocarcinome Nodule (n°) 1 2 Localisation Foie droit Foie droit Taille TAILLE (mm) 52 12 Diagnostic Cholangiocarcinome Cholangiocarcinome Capsule (oui/non/franchie) Non Non bien limitée, de TAILLE 2 cm x 1,2 cm, homogène, CR 909147

Label	Section textuelle d'intérêt	Définition	Exemples
		En pratique, une section d'annotation implique la présence minimale de l'unité de mesure. La taille des lésions non tumorales n'est pas annotée.	nodules satellites ou à distance mesurant de TAILLE 0,5 à 2,2 cm de diamètre. CR 4909147 106 La pièce pèse 1,1 g et mesure 1,5 x 1,3 x 0,9 cm. A la coupe, elle comporte un nodule l 107 capsulaire de TAILLE 0,3 cm de grand axe x TAILLE 0,2 cm d'épaisseur.
Différenciation	Description microscopique exclusivement	La différenciation de la tumeur est catégorisée en trois niveaux par ordre pronostic péjoratif croissant : - Bien différencié - Moyennement différencié - Peu (ou moins bien) différencié Lorsque co-existent plusieurs contingents distincts par la différenciation au sein du même fragment de phrase, celui associé au pire pronostic est annoté. En cas de format non-structuré, seul le niveau de différentiation d'intérêt est annoté (cf exemple). Si un contingent est décrit comme « mieux différencié que », la catégorie	CR 9592449 83 - Cholangiocarcinome périphérique bien à MOYEN_DIFF moyennement différencié. CR 11891 70 tumorale qu'au niveau des nodules satellites, il correspond à un adénocarcinome MOYEN_DIFF mieux différencié

Label	Section textuelle d'intérêt	Définition	Exemples
		annotée correspond à l'échelle progressive peu / moyen / bien	
Résection complète	Description microscopique exclusivement	Interprétation par l'anatomopathologiste selon lequel le chirurgien a réséqué l'ensemble de la tumeur. L'annotation inclut l'item « exérèse » et le caractère « complet ». Cette information peut être contenue dans une abréviation dédiée.	CR 698063  CR 3015295 
Résection incomplète	Description microscopique exclusivement	Interprétation par l'anatomopathologiste selon lequel le chirurgien n'a pas réséqué l'ensemble de la tumeur, c'est-à-dire qu'une limite de résection est atteinte. Cette information peut être contenue dans une abréviation dédiée.	pR1
Distance aux marges	Description macroscopique	Il s'agit de la distance minimale entre la limite de la résection chirurgicale et la lésion tumorale.	Ex 1 

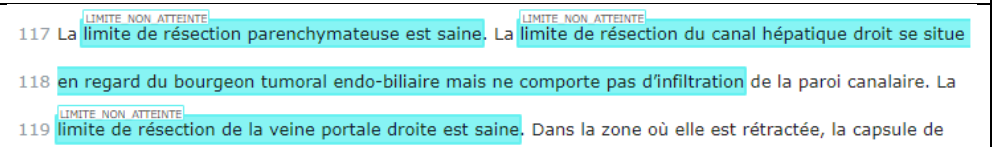
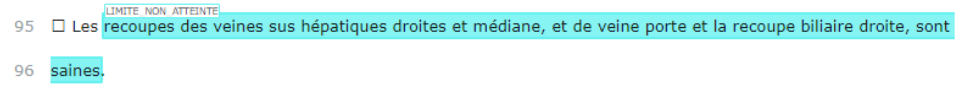
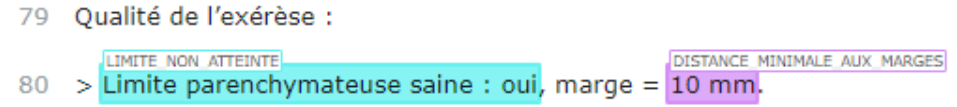
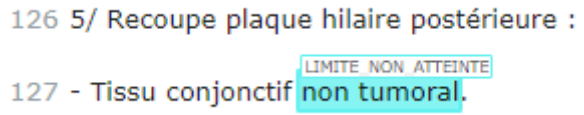
Label	Section textuelle d'intérêt	Définition	Exemples
	et description microscopique	Elle est décrite en « unité de distance » en une dimension, suivie d'une dimension qui peut être écrite en abrégé : centimètre (cm), millimètre (mm). En cas de format sous tableur des résultats pour de multiples nodules, l'ensemble des distances aux marges est annoté en un temps dans la mesure où une seule unité de mesure est précisée (cf exemple 2). La distance peut être exprimée en référence à un seuil donné (cf exemple 3).	Ex 2 PAS_D_EMBOLS_VASCULAIRES Emboles vasculaires Non Non Distance/ limite de résection DISTANCE_MINIMALE_AUX_MARGES 12 >15mm Nécrose (%) 0 0 Congélation (oui/non) Oui Non CR 46086 54 > Marge parenchymateuse macroscopique (distance tumeur ou nodule satellite / tranche de section) 55 : < 2 cm (DISTANCE_MINIMALE_AUX_MARGES 11 mm). CR 296508 : 50 inflammatoire. La tumeur arrive à DISTANCE_MINIMALE_AUX_MARGES moins de 0,1 mm des limites encrées LIMITE_INFRAMILLIMETRIQUE . La tumeur est située à moins de 1 mm de la tranche de section.
Marge infra-millimétrique	Description microscopique exclusivement	Mention exclusivement textuelle du caractère infra-millimétrique de la distance minimale aux marges. Est annoté l'item « inframillimétrique » exclusivement.	inframillimétrique. CR 2221374 120 La LIMITE_NON_ATTEINTE limite de résection parenchymateuse passe en tissu sain avec une marge LIMITE_INFRA_MILL inframillimétrique.

Label	Section textuelle d'intérêt	Définition	Exemples
Marge supra-millimétrique	Description microscopique exclusivement	Mention exclusivement textuelle du caractère supra-millimétrique de la distance minimale aux marges. Est annoté l'item « supramillimétrique ».	
Limite atteinte	Description macroscopique et microscopique	Mention du caractère envahi par la tumeur de la limite de résection. La mention de la limite (recoupe, tranche, limite de résection) est toujours contenue dans l'annotation. Les expressions « tranche » et « limite » sont toujours caractérisées par le qualificatif exhaustif de type « de section », « d'exérèse », « chirurgicale », « d'exérèse chirurgicale », « de section chirurgicale ».	CR 2640824  CR 4909147 :  CR 0882817 (contre-exemple)  CR 894954 :

Label	Section textuelle d'intérêt	Définition	Exemples
		S'y ajoute la mention de l'infiltration / « au contact » / atteinte, quand elle est disponible au sein de la même phrase. Il peut également s'agir de la négation du module ci-dessous, qui à ce moment est intégré à l'annotation. L'item « tumeur » n'est pas annoté (car peut être « loin » de la description de l'atteinte de la limite). La négation est intégrée à l'annotation si elle est nécessaire sémantiquement (CR 4909147). En cas de doute médical, l'annotation n'est pas réalisée. Les lésions de CIS sont négligées.	 CR 894954 :  CR 729589 :  **Ex : « marge envahie par la tumeur » : 'marge envahie' suffit** CR 415797 :  CR 8918091

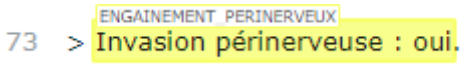
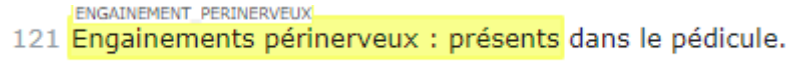
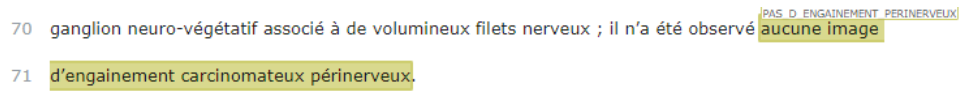
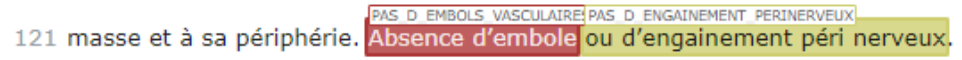
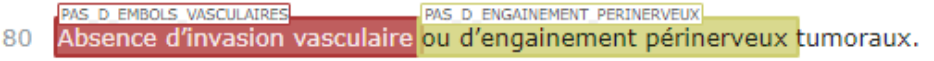
Label	Section textuelle d'intérêt	Définition	Exemples
			49 4-Recoupe de la voie biliaire principale 50 Un fragment de 1 cm de grand axe est reçu. 51 Examen extemporané (Dr Benchene) : « très suspect ». 52 Après fixation formolée, inclusion en paraffine (bloc 2), on observe la paroi d'un canal dont l'épithélium est 53 largement abrasé et électro coagulé rendant l'interprétation difficile. Cette paroi contient quelques récessus 54 glandulaires normaux. Focalement, il existe quelques amas cellulaires avec des noyaux modérément atypiques 55 et hyperchromatiques ainsi que des glandes tortueuses et dilatées avec des atypies cytonucléaires légères. 147 LIMITE_ATTEINTE Recoupe de la voie biliaire principale : carcinomateux CR 199433 : 76 - LIMITE_ATTEINTE Recoupe biliaire : tumorale. CR 11891 : 86 L'envahissement au niveau du parenchyme hépatique arrive sur les prélèvements effectués au 87 niveau du hile (5A) LIMITE_ATTEINTE au contact des limites d'exérèse modifiées par l'électrocoagulation, par contre 88 les LIMITE_NON_ATTEINTE limites des vaisseaux du hile (5B) sont saines. CR 8934451 161 3/ LIMITE_ATTEINTE Recoupe biliaire basse : 162 - Microfoyers d'infiltration carcinomateuse.

Label	Section textuelle d'intérêt	Définition	Exemples
Limite non atteinte	Description macroscopique et microscopique	Mention du caractère non envahi par la tumeur de la limite de résection. La mention de la limite (recoupe, tranche, limite de résection) est toujours contenue dans l'annotation. S'y ajoutent : - Soit une négation du module ci-dessus : mention de l'infiltration / atteinte - Soit par caractérisation d'une limite saine (quand aucune négation n'est explicitée) - Soit par caractérisation de l'absence de tumeur (quand la localisation de la recoupe est implicite dans la phrase, cf contexte). Les expressions « tranche » et « limite » sont toujours caractérisées par leur qualificatif exhaustif de type « de section chirurgicale », « d'exérèse chirurgicale ».	Prélèvements après encrage de la LIMITE_NON_ATTEINTE tranche de section : foie non tumoral 88 LIMITE_NON_ATTEINTE Recoupe saine. CR 415797 129 Grand épiploon, patch de diaphragme, LIMITE_NON_ATTEINTE recoupe voie biliaire du hile hépatique, recoupe des veines sus 130 hépatiques droite et médiane, recoupe de veine porte: saines. CR 29054 ? à « à distance des limites d'exérèse » 141 La lésion tumorale le tissu adipeux adjacent. 142 Présence de nombreux ENGAINEMENT_PERINERVEUX engainements péri-neveux. Absence d'PAS_D_EMBOLS_VASCULAIRES embole vasculaire tumoral. 143 La lésion est située LIMITE_NON_ATTEINTE à distance des limites d'exérèse latérales encrées (marge minimale : CR 758907 53 Les prélèvements au niveau de la LIMITE_NON_ATTEINTE recoupe veineuse portale, artérielle et canalaire sont sains. CR 894954

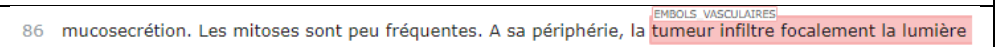
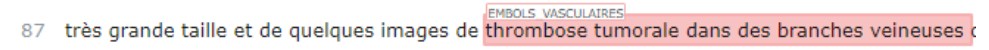
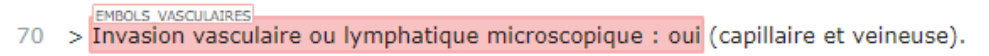
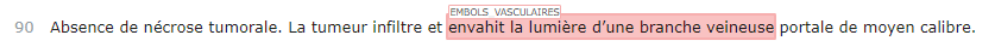
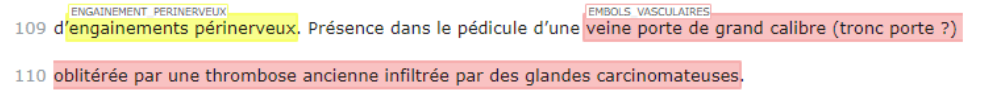
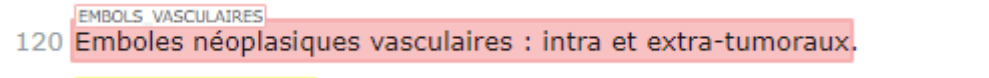
Label	Section textuelle d'intérêt	Définition	Exemples
		Les sections peuvent être annotées quand elles sont indispensables à la qualification topographique de l'envahissement tumoral (cf CR 8934451) et qu'elles précèdent immédiatement cette qualification. (Si elles sont redondantes avec le contenu de la description, elles ne sont pas annotées.) Autrement, l'annotation de la section n'est pas réalisée, mais seulement le caractère (non) envahi sans topographie.	 CR 15797  CR 546086 :  CR 556983  CR 68723 :

Label	Section textuelle d'intérêt	Définition	Exemples
			167 6/ RECOUPE BILIAIRE DISTALE 168 Examen macroscopique : 169 Il a été transmis un segment biliaire de 1 cm de longueur et 0,3 cm de diamètre. 170 Examen histologique : 171 La paroi biliaire comporte un foyer de glandes péribiliaires hyperplasiques. Elle est par ailleurs normale 172 LIMITE_NON_ATTEINTE sans modification dysplasique de l'épithélium de surface. CR 8934451 161 3/ LIMITE_ATTEINTE Recoupe biliaire basse : 162 - Microfoyers d'infiltration carcinomateuse. 163 4/ Recoupe biliaire haute (canal hépatique gauche et canal du IV) : 164 - Foyer d'hyperplasie épithéliale avec BilIN de bas et de haut grade. 165 - LIMITE_NON_ATTEINTE Absence de lésion carcinomateuse infiltrante. CR 551525

Label	Section textuelle d'intérêt	Définition	Exemples
			38 1) RECOUPE BASSE DE LA VOIE BILIAIRE : 39 Examen extemporané par télépathologie (A. Bay et C. Guettier) : 40 Modifications hypertrophiques et dysplasiques de l'épithélium probablement en rapport avec la prothèse 41 biliaire. Hyperplasie des glandes péribiliaires. LIMITE_NON_ATTEINTE Absence de lésion tumorale infiltrante. 42 Examen histologique : 43 L'étude histologique définitive confirme les constatations de l'examen extemporané. La muqueuse biliaire 44 est focalement ulcérée et tapissée par un enduit fibrinoleucocytaire. L'épithélium est irrégulièrement 45 hyperplasique avec quelques ébauches papillaires en surface. Les cellules épithéliales sont basophiles 46 avec des noyaux discrètement irréguliers sans atypie majeure cependant. Le chorion héberge un infiltrat 47 inflammatoire polymorphe d'intensité modérée. Les glandes péribiliaires sont irrégulièrement 48 hyperplasiques. LIMITE_NON_ATTEINTE Absence de lésion tumorale infiltrante. CR 1631539 210 LIMITES CHIRURGICALES : 211 - LIMITE_NON_ATTEINTE Pancréatique : saine 212 - LIMITE_ATTEINTE Biliaire (si DPC) : envahie 213 - LIMITE_NON_ATTEINTE Rétropéritonéale : saine
Engainement périnerveux	Description microscopique exclusivement	Présence d'un engainement périnerveux par les cellules tumorales : il s'agit d'une catégorisation binaire.	CR 2640824 58 infiltration carcinomateuse s'étend vers la paroi de la voie biliaire de plus petit calibre. Les filets nerveux du 59 tissu péricanalaire sont le siège de nombreux ENGAINEMENT_PERINERVEUX engainements carcinomateux périnerveux.

Label	Section textuelle d'intérêt	Définition	Exemples
		Le qualificatif tumoral n'est annoté que s'il est compris entre l'invasion et le « périnerveux ».	CR 46086 :  CR 968723 
Pas d'engainement périnerveux	Description microscopique exclusivement	Les mentions relatives à la négation sont intégrées à l'annotation. En cas de liste de négation, le module de négation n'est pas annoté de façon répétée s'il s'agit du 2ème item de la liste. le "ou" est intégré à l'annotation. Il sera intégré en tant que contexte par le NER, au risque d'intégrer dans l'annotation un autre item d'intérêt. L'absence d'engainement peut être annotée dans la description de filets nerveux normaux.	CR 894954  CR 15797  CR 824334  CR 343678 

Label	Section textuelle d'intérêt	Définition	Exemples
Invasion tumorale vasculaire	Description microscopique exclusivement	Présence d'embol(s) tumoral(ux) vasculaire(s) : il s'agit d'une catégorisation binaire. Sont annotés les items : embol(ie)s (car impliquant la présence de tumeur dans un vaisseau). Les qualificatifs « tumoraux +/- endovasculaires » ne sont pas nécessaires à annoter. Sont inclus les thromboses veineuses tumorales, et « l'envahissement / l'invasion par la tumeur des structures veineuses et artérielles » Quand l'item « embol » est absent du texte, l'annotation comporte les items suivants : « veine » + « dedans » + « ~~tumeur »~~ si et seulement si présents au sein de la même phrase. Le qualificatif « tumoral » est intégré ssi il est pris entre les notions d'« infiltration » et de « veine ». L'annotation du qualificatif (« portale ») de la veine n'est pas réalisée (Ex 4 et 3),	CR 11891 : 57 périganglionnaire, où l'on observe des EMBOL embolies néoplasiques intravasculaires. CR … 69201701 infiltrer la graisse périhilaire. La EMBOLS VASCULAIRES veine sus-hépatique gauche et des branches de la veine porte gauche sont thrombosées et envahies par la prolifération tumorale. La branche gauche de l'artère hépatique est saine. On CR 230105 Au niveau du hile hépatique, on observe un EMBOLS VASCULAIRES embole néoplasique détruisant et comblant complètement la veine portale mais également une infiltration pariétale de l'artère hépatique. D'autre part, il existe de très CR 32385 45 du nodule. On note des EMBOLS VASCULAIRES cellules tumorales dans la lumière d'une veine portale CR 698063 81 HEPATECTOMIE DROITE : la tumeur répond à un cholangiocarcinome bien à DIFFERENCIATION peu différencié. L'architecture est 82 faite tantôt de glandes de petite taille, tantôt de massifs cribriformes ou de travées. Les cellules tumorales sont 83 d'assez petites taille au cytoplasme éosinophile et aux noyaux arrondis. Elles présentent des atypies 84 cytonucléaires modérées. Le stroma est fibreux et abondant. Elle EMBOLS VASCULAIRES infiltre la paroi de la veine sus-hépatique CR 9698063 118 EMBOLS VASCULAIRES Invasion de la paroi de la veine sus-hépatique moyenne. CR 2063806

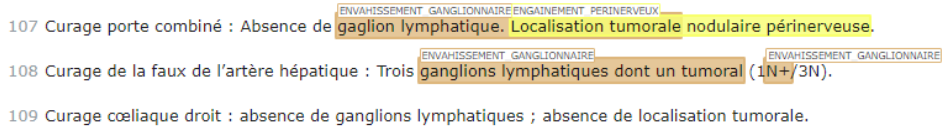
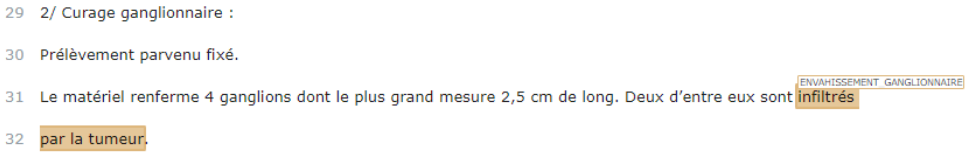
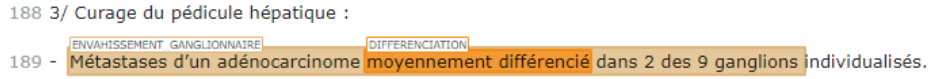
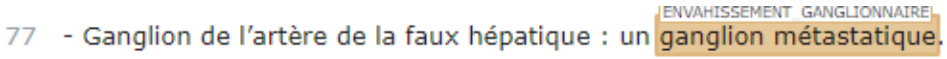
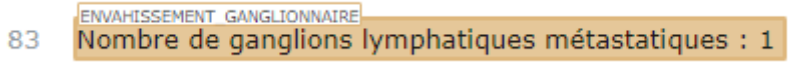
Label	Section textuelle d'intérêt	Définition	Exemples
		sauf quand elle est « prise » dans le texte d'intérêt (cf Ex 1) Une veine peut être décrite comme un « sinusoïde ». Les caractéristiques, dont topographiques, des veines ne sont pas annotées. En cas de format « : », l'ensemble est annoté, dans la mesure où ce qui suit les « : » précise ou non l'atteinte de la structure anatomique analysée. En cas de « thrombose tumorale veineuse », les deux premiers items sont annotés (car il peut y avoir des thromboses veineuses d'origine non-tumorale). Un « enserrement » n'est pas une « invasion » et n'est pas annoté. Quand le caractère « veineux » de la structure envahie est sous-entendu, il ne peut pas être annoté (cf CR 111468726)	 CR 1757374 :  CR 46086  CR 546086 :  CR 758973 :  CR 968723 :  CR 968723 :  CR 968723 :

Label	Section textuelle d'intérêt	Définition	Exemples
		L'invasion unique de la paroi vasculaire ne constitue pas une « invasion vasculaire » (elle n'est pas non plus annotée comme une « absence d'invasion vasculaire »). L'invasion de la lumière vasculaire constitue une « invasion vasculaire tumorale ».	194 - Extension tumorale dans le pédicule portal avec emboles vasculaires veineux et lymphatiques et 195 engainements périnerveux. 196 - Thrombose tumorale d'une veine portale de grand calibre. CR 11891 89 En dehors des nodules principaux, un envahissement des espaces portes est parfois observé sous 90 forme d'embolies vasculaires. CR 2221374 99 centrale de la masse. Il s'y associe une fibrose abondante ponctuée de quelques calcifications au sein de 100 laquelle sont identifiées des structures portales et une image de veine collectrice oblitérée par une 101 thrombose ancienne qui renferme quelques amas de cellules carcinomateuses. CR 111468726 82 fibreuses. Certaines sont le siège d'un embol néoplasique. La section veineuse portale est comprimée par 83 la tumeur et sa paroi est envahie. Sur un plan de coupe, sa lumière est comblée par un thrombus fibreux et 84 tumoral. Plusieurs sections canalaires sont revêtues par un épithélium tubulo-papillaire végétant en CR 2063806 86 mucosecrétion. Les mitoses sont peu fréquentes. A sa périphérie, la tumeur infiltre focalement la lumière 87 des sinusoïdes. Absence de réaction inflammatoire péri-tumorale. Il est observé plusieurs nodules

Label	Section textuelle d'intérêt	Définition	Exemples
Pas d'invasion vasculaire tumorale	Description microscopique exclusivement	Les mentions relatives à la négation sont annotées. Quand l'item « embol » est absent du texte, l'annotation comporte les items suivants : « veine » + « dedans » + ~~« tumeur »~~ si et seulement si présents au sein de la même phrase. Le caractère sain d'une veine est annoté, lorsqu'elle appartient au tissu péri-tumoral. L'absence d'invasion par la tumeur d'une structure vasculaire est annotée. Le caractère normal d'une paroi de veine est annoté. Le caractère non tumoral d'une thrombose veineuse est annoté. L'item « tumeur » n'est pas annoté.	CR 774907 - PAS_D_EMBOLS_VASCULAIRES Infiltration tumorale de la paroi des veines sus-hépatiques médiane et gauche sans atteinte de la lumière veineuse. CR 2063806 94 La PAS_D_EMBOLS_VASCULAIRES tumeur enserre et infiltre le pédicule hépatique droit, refoulant focalement sans l'envahir la branche 95 droite de la veine porte et engainant une section canalaire de grand calibre. La capsule hépatique est CR 15797 Il n'y a PAS_D_EMBOLS_VASCULAIRES ni embole ni engainement péri nerveux. CR 46086 : 69 > PAS_D_EMBOLS_VASCULAIRES Invasion vasculaire tumorale macroscopique : non. CR 889494 : 73 capsule hépatique, sans l'ulcérer. La tranche de section chirurgicale hépatique mesure 14x8 cm. La 74 tumeur vient PAS_D_EMBOLS_VASCULAIRES au contact de la veine sus hépatique gauche, sans l'envahir. La prolifération tumorale 91 La prolifération tumorale vient PAS_D_EMBOLS_VASCULAIRES au voisinage de la veine sus-hépatique gauche, sans l'envahir CR 556983 : Il n'a PAS_D_EMBOLS_VASCULAIRES pas été observé d'embole néoplasique intravasculaire. CR 8284334 :

Label	Section textuelle d'intérêt	Définition	Exemples
			80 PAS_D_EMBOLS_VASCULAIRES Absence d'invasion vasculaire PAS_D_ENGAINEMENT_PERINERVEUX ou d'engainement périnerveux tumoraux. CR 43687084 76 PAS_D_ENGAIN_PERINERV Absence d'engainement péri-nerveux PAS_D_EMBOL ou d'invasion vasculaire. 77 Les LIMITE_NON_ATTEINTE limites chirurgicales sont saines DISTANCE_MIN (3mm). 78 La PAS_D_EMBOL veine sus-hépatique moyenne est intacte. CR 2221374 118 grand calibre. Elle PAS_D_EMBOL respecte la paroi de la veine sus-hépatique droite et de la veine cave, qu'elle refoule 119 légèrement sans l'envahir. CR089721 94 La tumeur arrive PAS_D_EMBOL au contact de la veine sus-hépatique sans infiltrer sa paroi. CR 34678 La section PAS_D_EMBOL veineuse portale présente une paroi normale. CR 17790 90 carcinomateuse portale droite reste très « proximale ». On observe des lésions de PAS_D_EMBOL thromboses portales 91 non néoplasiques, partiellement reperméabilisées, œdémateuses, riche en sidérophages. Il n'y a PAS_D_EMBOL pas 92 d'infiltration carcinomateuse de la paroi de l'artère hépatique. CR 2063806

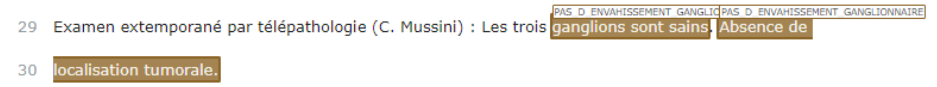
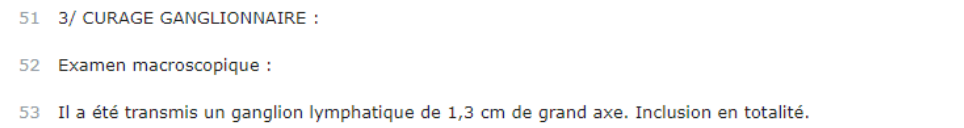
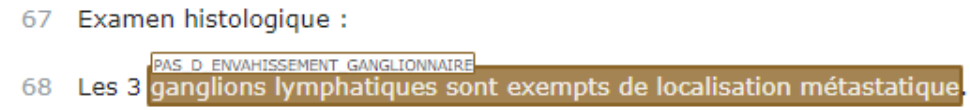
Label	Section textuelle d'intérêt	Définition	Exemples
			94 La tumeur enserre et infiltre le pédicule hépatique droit, refoulant focalement PAS_D_EMBOL sans l'envahir la branche 95 droite de la veine porte et engainant une section canalaire de grand calibre. La capsule hépatique est
Envahissement ganglionnaire	Description microscopique exclusivement	Présence d'un envahissement ganglionnaire par les cellules tumorales : il s'agit d'une catégorisation binaire. Le décompte des ganglions concernés n'est pas annoté. Toute la caractérisation de l'envahissement ganglionnaire peut ne pas être incluse dans l'annotation, c'est-à-dire « ganglion », « caractère envahi », « nature tumorale de l'envahissement » → p. ex., quand elles ne sont pas contenues dans la même phrase →ne pas viser le + petit dénominateur commun sur un plan sémantique, avec annotation des redondances même au sein d'une phrase unique. Par ailleurs, l'abréviation « N+ » peut être annotée en tant que telle.	CR 509698063 107 CURAGE DU PEDICULE : trois ENVAHISSEMENT_GANGLIONNAIRE ganglions sont le siège de métastases comparables à celle de l'adénopathie 108 rétroporte sur cinq ganglions analysés. CR 698063 : 49 Réponse (Dr Amajouegan) : un ENVAHISSEMENT_GANGLIONNAIRE ganglion envahi par un adénocarcinome CR 698063 99 résultat de l'examen extemporané. Il s'agit d'un ENVAHISSEMENT_GANGLIONNAIRE ganglion lymphatique envahi par le contingent moyennement à DIFFERENCIATION peu 100 différencié de la prolifération tumorale qui est essentiellement retrouvée sous la forme de travées de cellules CR 698063 ENVAHISSEMENT_GANGLIONNAIRE (3N+/5N). CR 2063806 131 2/ Curage ganglionnaire droit du pédicule hépatique : 132 - ENVAHISSEMENT_GANGLIONNAIRE Localisation métastatique d'un adénocarcinome dans l'unique ganglion individualisé CR 1757374 : exemple sujet à caution : 163 - ENVAHISSEMENT_GANGLIONNAIRE Localisation ganglionnaire métastatique dans l'unique ganglion individualisé.

Label	Section textuelle d'intérêt	Définition	Exemples
		Le fait que ce soit la tumeur qui envahit le ganglion est sous-entendu et n'est pas annoté quand les items « ganglion » et « envahi » sont explicitement présents dans la phrase. En cas de format semi-structuré (tableur), les annotations ne sont réalisées quand elles sont cohérentes avec une lecture textuelle « horizontale » (CR 631539)	CR 769530 :  CR556983 : annoter « deux d'entre eux »  CR 556983 :  CR 968723 :  CR 199433 :  CR 268984 : 

Label	Section textuelle d'intérêt	Définition	Exemples
			CR 1516591 50 Histologiquement, on trouve sept formations ganglionnaires ; dans l'une d'entre elles, on observe la présence 51 d'une métastase d'un adénocarcinome de type biliaire. CR 744522 82 Il a été retrouvé trois ganglions lymphatiques métastatiques largement infiltrés par une prolifération 83 adénocarcinomateuse. La graisse péri ganglionnaire est également largement infiltrée par la tumeur. CR 1631539

Label	Section textuelle d'intérêt	Définition	Exemples
			226 Localisation Pot Bloc Histologie 227 ganglions 228 Un ganglion lymphatique 229 Bord droit du 230 4 1 D01-D02 métastatique avec rupture 231 pédicule hépatique 232 capsulaire 233 3 ganglions lymphatiques 234 métastatiques avec ruptures 235 Ganglion cœliaque 7 4 G01-G04 236 capsulaires 237 1 ganglion indemne.
Absence d'envahissement	Description microscopique exclusivement	Il existe deux façons d'annoter cet item : - soit d'annoter une expression qui intègre une négation explicite (cf exemple 1 ci-contre)	Ex 1 CR 2063806 101 Le ganglion individualisé dans le pédicule hépatique est exempt de localisation métastatique.

Label	Section textuelle d'intérêt	Définition	Exemples
ganglionnaire		- soit d'annoter directement l'absence d'envahissement ganglionnaire, quand les formes de négation ne sont pas explicites (cf exemple 2). Par ailleurs, il n'y a pas besoin d'annoter le caractère tumoral quand il est précisé dans le texte (cf exemple 3) → annotations en 2 entités car redondance de l'information dans des phrases distinctes. En revanche, au sein d'une même phrase, annoter la redondance au sein de la même annotation (Ex 4 bis). C'est-à-dire qu'une annotation unique ne peut englober 2 phrases à la fois. Par exemple, un « ganglion indemne » suffit, pas besoin d'annoter « ganglion indemne de cellule cancéreuse ». L'information « préservation de l'architecture ganglionnaire » ne suffit	133 3/ Curage ganglionnaire gauche du pédicule hépatique : 134 - PAS_D_ENVAHISSEMENT_GANGLIONNAIRE Absence de localisation métastatique dans les 2 ganglions individualisés Ex 2 CURAGE GANGLIONNIARE DE LA FAUX DE L'ARTERE HEPATIQUE Un volumineux PAS_D_ENVAHISSEMENT_GANGLIONNAIRE ganglion mesurant 3cm, il est indemne. Ex 3 55 Deux PAS_D_ENVAHISSEMENT_GANGLIONNAIRE ganglions lymphatiques indemnes de métastase Ex 4 CR 2063806 34 Le PAS_D_ENVAHISSEMENT_GANGLIONNAIRE ganglion lymphatique d'architecture normale est le siège d'une histiocytose sinusale banale. Ex 4 bis CR 11891 43 Ce curage comporte un N- ganglion lymphatique (2 cm) d'architecture d'ensemble respectée et 44 indemne d'envahissement tumoral. CR 4493725 : 68 1. Vésicule biliaire : Cholécystite chronique lithiasique. Un PAS_D_ENVAHISSEMENT_GANGLIONNAIRE ganglion du col vésiculaire 69 montrant une réaction histiocytaire banale sans cellules tumorale.

Label	Section textuelle d'intérêt	Définition	Exemples
		pas à impliquer l'absence d'envahissement ganglionnaire. En cas de liste « : », l'ensemble est annoté, sauf s'il existe une redondance sémantique (cf exemple 546086 et 8973). Si le sujet « ganglion » n'est pas explicité, il n'est pas annoté. Les sections peuvent être annotées quand elles sont indispensables à la qualification topographique de l'envahissement tumoral (cf CR 8960943) et qu'elles précèdent immédiatement cette qualification. (Si elles sont redondantes avec le contenu de la description, elles ne sont pas annotées.) Autrement, l'annotation de la section n'est pas réalisée, mais seulement le caractère (non) envahi sans topographie. Lorsque la mention « ganglion » est trop éloignée (n mots > 12) de la description	CR 894954  CR 894954   CR 894954  CR 46086 :

Label	Section textuelle d'intérêt	Définition	Exemples
		de l'absence d'envahissement, cette première n'est pas annotée (cf CR 113453408)	75 PAS_D_ENVAHISSEMENT_GANGLIONNAIRE Nombre de gg métastatiques : 0. CR 546086 : annoter le N0 indépendamment 95 > Ganglions régionaux : PAS_D_ENVAHISSEMENT_GANGLIONNAIRE pas de ganglion métastatique (N0). CR 18091 : pas d'annotation réalisée 43 3-Adénopathie latéro-porte 44 Un fragment reçu. 45 Examen extemporané (Dr Benchene) : « saine ». CR 758973 61 4. Origine du tronc cœliaque : 62 1 PAS_D_ENVAHISSEMENT_GANGLIONNAIRE ganglion de 1 cm de grand axe : absence de métastase. CR 22214 : 106 CURAGE ARTERE HEPATIQUE 107 Le prélèvement renferme 3 PAS_D_ENVAHISSEMENT_GANGLIONNAIRE ganglions lymphatiques siège d'une histiocytose sinusale. PAS_D_ENVAHISSEMENT_GANGLIONNAIRE Absence de prolifération 108 tumorale. CR 43687084

Label	Section textuelle d'intérêt	Définition	Exemples
			87 GANGLION DE LA FAUX DE L'ARTERE HEPATIQUE : un ganglion lymphatique histologiquement normal 88 indemne de métastase. 89 90 CURAGE DU PEDICULE HEPATIQUE : six ganglions lymphatiques histologiquement normaux dépourvus de 91 métastases. CR2221374 35 2 / GANGLION DE L'ARTERE HEPATIQUE COMMUNE : 36 Examen macroscopique : 37 Il a été transmis un ganglion de 2 cm de grand axe. 38 Examen extemporané par télépathologie (M. Sebagh) : Ganglion normal, sans tumeur. 39 Examen histologique : 40 L'étude histologique définitive confirme les constatations de l'examen extemporané. Le ganglion 41 d'architecture normale est remanié par une histiocytose sinusale banale. Absence de localisation 42 métastatique. Absence d'engainement périnerveux dans le tissu conjonctivo-adipeux adjacent. CR089721 117 2/ 3/ 4/ 5/ et 7/ Curages ganglionnaires : 118 - Absence d'envahissement ganglionnaire tumoral. CR 1515619

Label	Section textuelle d'intérêt	Définition	Exemples
			56 La reprise du matériel après fixation formolée et inclusion en paraffine (3 et 4) confirme histologiquement 57 l'absence de prolifération tumorale métastatique et la présence d'un ganglion hyperplasique. CR 8960943 105 2. Ganglion de la faux de l'artère hépatique : 106 Pas de tumeur 107 3. Curage de l'artère hépatique : 108 Pas de tumeur. CR 113453408 67 L'un des 4 ganglions comporte sur la coupe paraffine un microfoyer métastatique ; ce foyer métastatique 68 n'était pas présent sur la coupe congelée examinée lors de l'examen extemporané. Les 3 autres ganglions 69 examinés en examen extemporané ainsi que les 7 ganglions de petite taille qui n'avaient pas été examinés 70 lors de l'examen extemporané. sont exempts de localisation métastatique 71 Tous les ganglions sont le siège d'une histiocytose sinusale, d'aspect banal. Quatre des ganglions Et dans le même CR 91 microganglion lymphatique mesurant moins de 0,1 cm de grand axe ; celui-ci est exempt de localisation 92 métastatique.

Label	Section textuelle d'intérêt	Définition	Exemples
pTNM	Description microscopique exclusivement		
Classification TNM	Description microscopique exclusivement	Il s'agit de la version de la classification TNM de référence pour le cas. En effet, les méthodes de classification sont actualisées dans le temps. L'exhaustivité de la caractérisation de la version est annotée, en une annotation unique, sans la mention « pTNM ». La mention « classification » n'est pas annotée.	Classification pTNM [EDITION_TNM UICC 2017] : CR 2063806 : 145 [PTNM pT3N1] (Classification [EDITION_TNM AJCC 2009, 7ème édition]) CR 894954 : 171 pTNM ([EDITION_TNM 8ème version]) : [PTNM pT1bN0] CR 14938173 : 149 Stade TNM ([EDITION_TNM AJCC, 2017, 8ème édition, Intrahepatic Bile Ducts]) : [PTNM ypT4N0]

Annexe 8. Recours aux différents fichiers de règles concernant l'extraction textuelle des entités liées aux critères histopronostiques du cas d'usage du 'Challenge AI for health'

Itération	Date de l'itération	Combinaison des fichiers et du jeu de développement	Version du fichier des règles	Version du fichier de schéma	Version du fichier de terminologie	Version du fichier de négations	Version du jeu de développement
1	23/04/2021	1	1	1	1	1	Interne v1
		2	2	2	2	1	Interne v1
2	06/05/2021	3	3	3	3	2	Interne v1
3	11/05/2021	4	3	3	4	2	Interne v1
4	18/05/2021	5	4	4	5	2	Interne v1
		6	5	4	6	2	Interne v1
5	20/05/2021	7	6	5	7	3	Interne v1
6	05/07/2021	8	6	5	8	4	Interne v1
		9	7	6	9	5	Interne v1
		10	7	6	10	6	Interne v1
7	09/09/2021	11	7	7	11	6	Interne v1
		12	8	8	12	7	Interne v2
		13	8	8	13	7	Interne v2
8	20/09/2021	14	8	9	14	7	Interne v3
		15	8	9	14	8	Interne v3
		16	8	9	15	8	Interne v3
		17	8	9	16	9	Interne v3
9	17/01/2022	18	9	10	17	9	Interne v3
		19	9	11	18	9	Interne v3
		20	10	12	19	9	Interne v3
		21	10	12	20	9	Interne v3
		22	10	12	21	9	Interne v3
		23	10	13	22	9	Interne v3
		24	10	13	23	9	EK

Annexe 9. Performances des algorithmes de vision par ordinateur obtenues pour la prédiction de chaque biomarqueur pronostique histologique, par candidat du Challenge AI for health

Tâches de régression

Candidat 1		2	3	4	5	6
Distance aux marges (mm)						
mse	220.762500	237.212500	236.179381	205.306816	220.664999	212.177272
mae	9.825000	9.475000	10.326624	9.114100	10.554692	8.991028
r2	-0.461937	-0.570872	-0.564031	-0.359586	-0.461291	-0.405084
Taille (mm)						
mse	1625.687500	275.437500	673.281250	2374.406250	1613.310114	484.295835
mae	30.500000	11.625000	20.531250	39.718750	31.817718	16.544800
r2	-0.865691	0.683899	0.227321	-1.724945	-0.851486	0.444207

Tâches de classification

Candidat 1		2	3	4	5	6
Différenciation tumorale						
score F1	0.429435	0.461310	0.243189	0.426012	0.335350	0.486014
précision	0.468750	0.468750	0.281250	0.437500	0.312500	0.531250
Caractère complet de la résection microscopique						
score F1	0.343137	0.673069	0.517557	0.542556	0.381232	0.797855
précision	0.333333	0.666667	0.484848	0.515152	0.363636	0.787879
Emboles vasculaires						
score F1	0.352381	0.491979	0.532749	0.311612	0.395207	0.291811
précision	0.428571	0.500000	0.535714	0.428571	0.392857	0.392857
Engainement périnerveux						
score F1	0.630418	0.703704	0.643031	0.317650	0.547699	0.630418
précision	0.740741	0.703704	0.629630	0.296296	0.518519	0.740741

Candidat 1	2	3	4	5	6
Envahissement ganglionnaire					
score F1	0.519400	0.609053	0.485802	0.654637	0.629630 0.440999
précision	0.629630	0.629630	0.481481	0.666667	0.629630 0.592593
Stade_pTNM_T					
score F1	0.136508	0.389929	0.172466	0.273621	0.229853 0.323003
précision	0.142857	0.380952	0.190476	0.285714	0.190476 0.428571
Stade_pTNM_N					
score F1	0.380000	0.565385	0.431579	0.596000	0.184615 0.390323
précision	0.450000	0.600000	0.450000	0.600000	0.150000 0.55

Table des illustrations

Table des tableaux