RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Drees
Statistique publique
de la santé et des solidarités

DREES MÉTHODES

N° 25 • avril 2026

Sous embargo jusqu'au 9/04/2026 à 06h00

Prédire la suite d'un parcours de soins dans le système national des données de santé

Tristan Haugomat, Aurélia Manns, Gladys Baudet, Judith Abécassis, Gaël Varoquaux, Milena Suarez Castillo

Prédire la suite d’un parcours de soins dans le système national des données de santé

Tristan Haugomat, Aurélia Manns, Gladys Baudet, Judith Abécassis, Gaël Varoquaux, Milena Suarez Castillo

Remerciements : Antoine Sese, Florian Jacquetin, Diane Naouri, Hadrien Le Mer, Léa Hoisnard

Retrouvez toutes nos publications sur : drees.solidarites-sante.gouv.fr

Retrouvez toutes nos données sur : data.drees.solidarites-sante.gouv.fr

SYNTHÈSE

Le **système national des données de santé** (**SNDS**) permet de reconstituer l'historique des parcours de soins de l'ensemble de la population française depuis près de vingt ans, dans un pays où la grande majorité des contacts avec le système de santé génèrent une trace dans le système d'information de l'Assurance Maladie. Les parcours de soins repérables dans le SNDS, en particulier l'enchaînement d'événements significatifs du point de vue de la santé (délivrance de médicament, diagnostic posé à l'hôpital, actes médicaux, etc.), sont largement mobilisés afin d'**identifier des profils de patients** présentant très probablement certaines pathologies, ou à risque de les développer. Ces travaux se fondent sur une expertise au cas par cas, permettant de sélectionner les événements de santé au regard de leurs spécificités et du contexte particulier dans lequel la donnée est collectée puis restituée pour un usage analytique secondaire.

La révolution représentée par les méthodes d'**intelligence artificielle**, notamment dans le domaine du langage, appelle à s'interroger sur la pertinence d'une autre approche. **Purement prédictive**, cette approche capitalise sur la richesse et la dimension des trajectoires de soins et s'inspire des motifs récurrents pour apprendre des régularités susceptibles d'informer sur les risques les plus forts d'une trajectoire individuelle, voire prolonger la trajectoire de façon probabiliste. Ainsi, on pourrait « apprendre l'histoire naturelle des maladies humaines avec des modèles génératifs » (Shmatko, *et al.*, 2025), si tant est d'avoir des données pour ce faire. L'anticipation de l'évolution d'un parcours de soins ouvre des perspectives d'amélioration des prises en charge, par exemple autour d'une prévention plus personnalisée et prédictive.

À l'inverse de l'approche classique, dont l'effort porte sur la sélection de la population d'étude et des variables pertinentes, l'essentiel du prétraitement des données consiste alors à **reconstituer l'enchaînement des événements de santé** sous la forme d'une **chronologie individuelle d'unités sémantiques**. La modélisation, quant à elle, s'inspire des architectures d'apprentissage automatique (par exemple *Word2Vec*) et de réseaux neuronaux ayant fait leurs preuves dans le domaine du langage quand ils ont été appliqués à grande échelle (BERT, GPT, etc.). De nombreux travaux à l'international, relevant du domaine de la recherche, appliquent cette stratégie sur des données de volumétrie importante se prêtant à un tel formalisme (Li, *et al.*, 2020 ; Rasmy, *et al.*, 2021 ; Savcisens, *et al.*, 2024 ; Kraljevic*, et al.*, 2024 ; Shmatko, *et al.*, 2025). À ce titre, le SNDS est une base médico-administrative unique au monde, avec une volumétrie importante, exhaustive à l'échelle de la population d'un pays, des données collectées en routine et non pas sur la base du volontariat, longitudinales, et centralisées. Ainsi, les données du SNDS offrent un cadre privilégié à l'application de ces modélisations, qui n'ont pas encore été pleinement étudiées à très grande échelle (plusieurs dizaines de millions de patients). Néanmoins, les limites du SNDS pour suivre l'état de santé des personnes sont connues. Il s'agit d'un recueil de données conditionné au recours au système de santé, restreint à des données nécessaires à la facturation, et tributaire de la qualité du codage. Les données sont certes massives et informatives sur la santé, mais souvent indirectement, sans données concernant l'exposition à des facteurs de risques majeurs ou données cliniques directes (comme la tension ou les résultats d'analyse biologique).

L'Assurance Maladie propose désormais de mobiliser ses données pour de la prévention personnalisée, notamment via l'espace numérique en santé « Mon Espace Santé ». Néanmoins, si la mise en place de modèles prédictifs en prévention apparaît de plus en plus crédible, de nombreuses incertitudes demeurent pour envisager de les opérationnaliser. **Peut-on modéliser la progression temporelle vers une nouvelle pathologie à partir des événements de soins collectés dans le SNDS**, et uniquement de ceux-ci ? Étant donné l'information disponible, pour quelles pathologies serait-ce crédible ? Quels sont les **biais majeurs**, notamment socio-économiques ? Et jusqu'à quel point peut-on expliquer la modélisation, ce qui peut fortement jouer sur son acceptabilité ? Enfin, **les gains de performance à attendre en augmentant les tailles d'entraînement sont-ils réellement substantiels** quand ils sont comparés à des références bien établies ?

Afin de répondre à ces questions, le premier objectif de cette étude est de quantifier, de manière standardisée, la **capacité (ou valeur prédictive) des signaux d'état de santé disponibles dans le SNDS** à prédire la survenue de chaque pathologie au sein d'un ensemble très large. Pour rendre compte de la richesse des données disponibles, et en s'inspirant de l'ambition des **modèles de fondation visant à la généralité** (par exemple, Kraljevic, *et al.*, 2024 ; Savcisens*, et al.,* 2024 ; Shmatko, *et al*. 2025), quel que soit le problème de prédiction posé, la modélisation inclut toute la population accessible et le socle le plus large possible de variables susceptibles d'apporter une information sur l'état de santé (délivrance de médicaments, diagnostics posés à l'occasion d'une hospitalisation ou d'une affection de longue durée, occurrence de consultations chez un médecin généraliste, spécialiste, ou un autre professionnel de santé, occurrence d'un passage aux urgences, dispositifs médicaux présentés au remboursement, actes de biologie et d'imagerie réalisés, cotation de la dépendance dans le cas d'un passage en soins de suite et de réadaptation ou en hospitalisation à domicile, etc.), soit un référentiel de près de **80 000 événements de santé différents**, chacun représenté par un code, et l'apprentissage statistique fait le tri.

Notre second objectif est **d'évaluer l'apport d'une modélisation fondée sur des architectures de modèles de langage** (inspirés respectivement de *Word2Vec* et BERT), respectivement nommés **les plongements vectoriels de trajectoire** et **BEHRT-SNDS**, qui permettent d'embrasser largement la complexité du SNDS ainsi que sa dimension. La référence choisie est celle d'une sélection de variables issues des algorithmes standards de l'épidémiologie ou de « dires d'experts » (indices de comorbidités de Charlson s'appuyant sur les diagnostics hospitaliers

et *Rx-Risk-V* s'appuyant sur les prescriptions médicamenteuses, ou encore les pathologies repérées dans la cartographie des pathologies de la Caisse nationale de l'Assurance Maladie - Cnam), dans un modèle de *gradient boosting.* Elle entend refléter les performances qui peuvent raisonnablement être atteintes avec une sélection de variables sur « dires d'experts » et des modèles d'apprentissage statistique (*machine learning)* performants désormais classiques.

Pour définir un problème de prédiction suffisamment générique et significatif du point de vue de la santé d'un individu, **notre première tâche de prédiction** est celle du **risque de première hospitalisation dans un ensemble de plus de 180 pathologies** définies par les sous-chapitres de la classification internationale des maladies (CIM-10), complété par une partition plus fine des codes de cancer. La prédiction de la mortalité est également ajoutée : les indices de comorbidités classiques sont usuellement dérivés de cette tâche de prédiction (voir par exemple les travaux autour du *Mortality-Related Morbidity Index* – MRMI – dans le SNDS, Constantinou, ***et al***., 2018). La démonstration de la supériorité de certains modèles sur d'autres, ainsi que l'évaluation de la valeur prédictive des données de parcours au regard d'alternatives expertes, s'effectuent sur ce premier ensemble de cibles.

Le modèle des **plongements vectoriels de trajectoire** se distingue par rapport aux références choisies par une supériorité prédictive globale par rapport aux différentes cibles (en moyenne sur les tâches de prédiction issues des 180 pathologies définies) : pour un échantillon d'entraînement de 10 millions d'individus, ils gagnent en moyenne 9 points d'aire sous la courbe ROC (AUC) stratifiée[1] (l'AUC stratifiée variant de 50 % pour une prédiction non informative, à 100 % pour une prédiction parfaite) par rapport à la seule cartographie des pathologies, passant de 64,2 % à 73,2 %, soit 6,5 points de plus par rapport aux indices de comorbidités, et encore 4 points de plus face au meilleur modèle de référence basé sur les indices de comorbidité associés aux fréquences de consommation. Ce gain n'est pas uniforme : certaines cibles peu fréquentes ou aiguës restent mal prédites, mais la hiérarchie des performances est stable. Les plongements vectoriels de trajectoire sont meilleurs que la cartographie des pathologies pour 97 % des cibles, que les indices de comorbidités pour 96 % des cibles et que les indices de comorbidités associés aux fréquences de consommation pour 80 % des cibles. L'augmentation de la taille d'échantillon apporte un bénéfice clair, en particulier pour le modèle des plongements vectoriels de trajectoire (+8 points d'AUC stratifiée entre 1 million et 10 millions en moyenne sur l'ensemble des cibles contre +4 points en moyenne pour les autres modèles), encore perceptible à 30 millions (+1,5 point entre 10 et 30 millions pour les plongements vectoriels de trajectoire, contre +0,6 point en moyenne pour les autres modèles), suggérant un potentiel d'amélioration lié à l'apprentissage sur de très grands volumes. L'analyse d'explicabilité confirme que ce modèle complexe mobilise des signaux diversifiés, y compris certains codes fins parmi les 80 000 possibles, illustrant la richesse informationnelle de la trajectoire complète.

Le modèle **BEHRT-SNDS** permet de gagner encore en performance, pour au moins deux raisons : il a été pré-entraîné et affiné (*fine-tuned*) sur 50 millions d'individus, et il s'appuie sur une architecture *transformer*. Avec un gain moyen d'AUC stratifiée de près de 2 points par rapport aux plongements vectoriels de trajectoire entraîné sur 30 millions d'individus, il fait mieux que le modèle des plongements vectoriels de trajectoire pour 90 % des cibles, et mieux (AUC stratifiée supérieure) à légèrement moins bien (<0,5 point d'AUC stratifiée de différence) pour 96 % des cibles.

Il apparaît ainsi que **les modèles fondés sur les architectures des modèles de langage** (plongements vectoriels de trajectoire, puis BEHRT-SNDS) **dépassent nettement les modèles de référence construits à partir de « variables expertes »** (cartographie des pathologies, comorbidités, consommations de soins), y compris lorsque ces modèles de référence sont entraînés sur des millions d'individus avec des algorithmes de *machine learning* éprouvés.

Dans un second temps, les **tâches de prédiction sont précisées pour un sous-ensemble de problèmes d'intérêt en santé publique**, afin de décliner les résultats obtenus au plus près de cas d'usage potentiels (ciblage d'une population donnée plutôt que la population générale - par exemple, les patients traités pour une pathologie chronique ; précision de l'événement à prédire, afin de l'adapter aux spécificités de chaque pathologie). Une meilleure définition des tâches de prédiction permet de s'affranchir d'un biais majeur qui émerge de l'ambition de couverture large de l'information disponible par un modèle généraliste : des hypothèses diagnostiques implicites révélées par le parcours (par exemple la présence d'un examen spécifique) peuvent réduire l'intérêt de la prédiction pour le patient, tout en l'améliorant. Par exemple, tant que la prédiction est réalisée en population générale, et l'incidence détectée en milieu hospitalier uniquement, un traitement déjà initié en ville pour une pathologie donnée accroît la probabilité d'observer une hospitalisation chez une personne qui suit ce traitement plutôt qu'une autre qui ne le suit pas ; l'existence de ce traitement suffit en lui-même à prédire correctement une éventuelle hospitalisation, le modèle n'apportant pas de réelle valeur ajoutée. L'explicabilité des modèles permet de mesurer le poids accordé à chacun des événements des trajectoires des patients et de vérifier que le gain de performance prédictive obtenu ne repose pas principalement sur l'inclusion de variables qu'un expert exclurait naturellement car peu pertinentes pour l'usage visé (susceptibles d'introduire des biais ou d'intérêt pratique limité), et dont le pouvoir discriminant éclipserait l'apport d'autres signaux potentiellement prometteurs. Décliner les comparaisons des modèles inspirés du traitement automatique des langues sur ces nouvelles cibles de prédiction affinée permet de s'assurer que le

[1] Dans l'ensemble de la synthèse, l'AUC se réfère à l'AUC stratifiée par âge et sexe, s'interprétant comme la probabilité de discriminer un cas d'un non-cas au sein d'une même classe démographique.

gain de performance persiste sur des situations où la critique précédente ne s'applique a priori que très modérément.

Malgré de forts gains relatifs à l'application de modélisations issues du traitement automatique des langues à l'échelle, **la capacité à prédire la suite d'un parcours reste très hétérogène en fonction des pathologies considérées**. Plus la pathologie fait partie des comorbidités fréquentes, notamment le syndrome métabolique, et des grands enjeux de santé publique, plus les modèles simples reposant sur des variables expertes et des fréquences de consommation suffisent à prédire l'hospitalisation avec des performances comparables aux plongements vectoriels de trajectoire. C'est le cas du diabète de type 1 connu, de l'insuffisance rénale chronique, de l'insuffisance cardiaque ou encore des maladies respiratoires chroniques. Néanmoins, les marges pour mieux prédire existent : le modèle BEHRT-SNDS permet encore de gagner 1,6 point d'AUC stratifiée pour l'insuffisance rénale chronique, 2,0 points d'AUC stratifiée pour l'insuffisance cardiaque et 1,2 point pour les maladies respiratoires chroniques par rapport au meilleur des autres modèles évalués. À l'inverse, lorsque l'événement est aigu, incident ou isolé, tel que l'insuffisance rénale aiguë sans facteurs de risque identifiés (gain de 5,5 points pour le modèle BEHRT-SNDS), la prédiction devient beaucoup plus incertaine et la prise en compte de la trajectoire complète des soins apporte un gain très substantiel de performance. Cette différence illustre la complémentarité entre l'approche experte, performante pour les terrains aux facteurs de risque connus, et les approches de modélisation de trajectoires, plus adaptées aux pathologies complexes.

Passer au tamis de nombreuses cibles de prédiction permet de faire ressortir certaines situations cliniques qui se distinguent par des performances élevées, un gain net grâce à la prise en compte des trajectoires, et une explicabilité non triviale. C'est le cas de l'épilepsie, où les hospitalisations pourraient être anticipées à partir de la séquence de traitements antiépileptiques et de soins précédant la décompensation ; de la maladie de Parkinson, pour laquelle les trajectoires pourraient permettre la prédiction de la perte d'autonomie ou des escalades thérapeutiques ; des maladies hypertensives de la grossesse, dont la prédiction repose peu sur les variables expertes mais s'améliore fortement lorsque le modèle intègre la dynamique des suivis obstétricaux ; et de l'endométriose, où l'analyse temporelle et contextuelle du parcours pourrait permettre d'anticiper les formes nécessitant une prise en charge chirurgicale, avec un gain de près de 10 points d'AUC stratifiée par rapport aux modèles classiques. Dans ces cas, l'information utile à la prédiction ne réside plus dans un événement isolé mais dans la dynamique globale des trajectoires. Si le modèle BEHRT-SNDS ne permet pas d'aller au-delà des performances du plongement vectoriel de trajectoire pour l'épilepsie et la maladie de Parkinson, il accroît encore plus nettement les gains pour l'éclampsie (+8,0 points d'AUC stratifiée) et l'endométriose (+2,4 points d'AUC stratifiée).

Enfin, en ce qui concerne la mortalité – une cible globale qui synthétise l'effet des comorbidités présentes dans les parcours de soins – les prédictions des différents modèles sont meilleures (en termes d'AUC stratifiée) pour les femmes, pour les personnes âgées de 40 à 70 ans et pour les niveaux de vie élevés. **L'analyse des disparités sociodémographiques** gagnerait à être poursuivie et déclinée cas d'usage par cas d'usage. En effet, il apparaît crucial d'étudier les biais socio-économiques et territoriaux de ces modèles prédictifs entraînés à grande échelle, avant que les usages ne se développent. Cela afin de s'assurer que l'utilisation de ces modèles pour orienter une action (par exemple un dépistage organisé) n'induit pas d'iniquités de santé.

SOMMAIRE

INTRODUCTION

Prédire à partir du parcours de soins

La mise en évidence d'un signal prédictif de la suite du parcours de soins d'un patient offre des perspectives multiples, que ce soit pour le patient ou les professionnels de santé : inclusion dans des programmes de dépistage ou d'accompagnement des patients les plus à risque, aide à la décision médicale, etc. Le développement d'un cadre éthique et juridique apparaît nécessaire au cas par cas, ce qui ne peut se faire sans une évaluation approfondie des capacités et biais de ces modèles. Il est aujourd'hui admis que, au-delà de la méthode utilisée, une grande partie des forces comme des biais des algorithmes prédictifs entraînés sur des données massives proviennent des données elles-mêmes. Dès lors, l'augmentation du volume de données d'entraînement peut améliorer les performances, mais aussi reconduire (voire amplifier) les biais présents dans ces données ; c'est pourquoi la plus-value d'entraîner des modèles sur plus d'un million de patients peut encore faire l'objet de débat, que nous cherchons ici à éclairer.

Dans le SNDS…

Aujourd'hui, les données disponibles dans le système national des données de santé (SNDS) sont largement utilisées pour identifier des profils de patients dans les données médico-administratives, permettant ainsi une surveillance passive de l'état de santé de la population, des études rétrospectives de pharmaco-épidémiologie à grande échelle ou, en vie réelle, pour l'inclusion dans des programmes d'accompagnement de l'Assurance Maladie ou encore, de lutte contre la fraude.

L'un des premiers usages est le phénotypage de patients. Par exemple, pour répartir les dépenses de santé entre les grandes maladies chroniques, l'Assurance Maladie produit la cartographie des pathologies (Rachas, *et al.*, 2022) à partir de règles déterministes détaillées pour plus d'une centaine de pathologies, établies par dires d'experts (cliniciens, épidémiologistes, spécialiste du codage), en exploitant jusqu'à 5 ans de données du SNDS (hospitalisation, reconnaissance d'affection de longue durée et consommation de médicament spécifique) pour caractériser de façon binaire l'état de chaque bénéficiaire une année donnée pour chacune des pathologies repérées. Cette approche a également été suivie par l'Institut de recherche et documentation en économie de la santé (Irdes) pour l'identification des situations de handicap (projet RISH). À partir de l'appariement du SNDS avec des données externes apportant la référence pour l'état de santé, comme la cohorte Constances, une autre approche consiste à utiliser un apprentissage statistique supervisé pour identifier des cas, par exemple de diabète (Haneef, *et al.,* 2021 ; Fuentes, *et al*., 2023). Un des objectifs de l'initiative BOAS (bibliothèque ouverte d'algorithmes en santé) est en particulier de développer, valider et partager des algorithmes de ciblage d'une population de patients dans le SNDS. Plus marginalement, certains travaux s'intéressent à la prédiction de l'évolution de certaines maladies, dans des situations très précises impliquant des cohortes de petite taille pour le développement des algorithmes : par exemple le changement de stade de l'asthme, la mortalité à court terme après l'implantation d'un défibrillateur cardiaque implantable ou encore la fragilité chez la personne âgée.

Une expertise conséquente est requise pour construire les algorithmes, au cas par cas. Les données de validation externes souffrent d'une faible dimension, et de limites les éloignant du « *gold standard* » (test de référence) souhaité (données déclaratives, échantillons de volontaires où la prévalence des pathologies recherchées n'est pas toujours représentative). L'approche plus moderne qui pourrait se dessiner consisterait à se fonder sur un modèle généraliste encodant la totalité de l'information des parcours (donc y compris le bruit) en économisant l'expertise initiale de « *feature engineering* » (ou sélection de variables), pour recentrer l'expertise vers l'aval : le « *fine-tuning* » (définition de la cible à prédire ou identifier), la population d'application du modèle (quelles exclusions pour qu'il soit utile), et pour confronter l'explicabilité. Nous illustrons cette approche en l'appliquant dans nos travaux.

Les traces que chaque personne en France laisse dans le système d'information de l'Assurance Maladie reflètent un état de santé sous-jacent, mais aussi un processus de collecte de l'information et un processus de décision médicale. Elles peuvent être considérées comme très informatives par des experts dans le cas d'une chronologie fiable et précise, et dans le contexte où des éléments formels de diagnostics (imageries, résultats de tests biologiques) sont indisponibles, de la présence conjointe d'indices médicaux en lien avec la pathologie d'intérêt : des pathologies chroniques ou graves et aiguës, avec une prise en charge suivant des recommandations consensuelles et structurées (Thurin, *et al.,* 2021). Ainsi, la capacité à prédire la suite d'un parcours dans le SNDS est fortement susceptible de varier d'une pathologie à l'autre.

Les modèles de séquences récents à l'international

Les modèles de fondation (« *Foundation Models* »), c'est-à-dire des modèles d'intelligence artificielle de grande échelle entraînés sur d'immenses volumes de données non annotées au moyen d'un apprentissage autosupervisé, ont marqué un changement de paradigme dans le domaine de l'intelligence artificielle. Dans le domaine du langage, ils permettent de dépasser les modèles conçus sur mesure pour une tâche unique, au profit de modèles plus

généralistes et plus facilement adaptables dans un deuxième temps à de nouvelles tâches. Le modèle BERT (2018) est souvent considéré comme un des premiers modèles de fondation (du type « *transformer encoder* »), soit un des premiers grands modèles pré-entrainés (à apprendre la structure du texte) qui a pu être ré-entrainé sur d'autres tâches (reconnaître la fonction d'un mot, analyse de sentiment, etc.). Le modèle GPT (2018) en est un autre (du type « *transformer decoder* ») se focalisant plus sur l'aspect génératif. *Word2Vec* (2013) a représenté un jalon important vers les modèles de fondation, en offrant des représentations vectorielles de mots (« *embeddings* ») à même de capturer dans une notion de distances les relations sémantiques.

Chaque génération de modèle de langage a, dans le sillage de son succès, donné lieu à une transposition au domaine des représentations de parcours de soins : *Life2Vec*, Med-BERT, BEHRT, Core-BEHRT, CEHR-BERT pour des modèles de types « *encoder* », et *Foresight*, *TransformEHR*, Delphi-2M pour des modèles de types « *decoder* ». Ces modèles ont avant tout une vocation généraliste : bien qu'ils puissent être spécialisés dans un second temps, ils sont d'abord entraînés dans des populations larges, de degré variable de représentativité et de complétude (*tableau 1*), avec un premier objectif d'apprentissage des trajectoires de soins, et le plus souvent soit un réentraînement autour de pathologies spécifiques, soit l'ambition de prédire l'ensemble des événements de la suite du parcours. Il paraît évident que le système national des données de santé, de par son exhaustivité sur la population résidant en France, le chaînage entre la ville et l'hôpital, et la grande variété des informations disponibles au-delà des seuls codes diagnostics hospitaliers, se place en excellente position pour concurrencer les résultats obtenus par ces différents travaux, relativement prometteurs dans leurs cas d'usage respectifs.

Ces travaux restent difficiles à comparer directement, car ils s'appuient sur des jeux de données hétérogènes, dont les caractéristiques et contraintes de collecte varient fortement (cohortes volontaires vs. population générale, couverture plus ou moins complète, sources ville/hôpital, qualité de l'appariement, etc.). Ces différences peuvent non seulement induire des biais distincts qui se répercutent dans les prédictions, mais aussi rendre les performances rapportées difficilement comparables d'un travail à l'autre et nécessiter des choix de modélisation spécifiques. À titre d'illustration, un biais documenté pour le modèle Delphi-2M est celui lié à la collecte partielle des données : (i) un biais d'immortalité qui se manifeste par un saut du risque de mortalité à 40 ans, l'âge du recrutement des volontaires (ii) un biais d'absence d'une source (essentiellement hôpital, avec des diagnostics en moyenne plus graves, ou ville, avec des diagnostics plus communs) du fait d'un défaut d'appariement qui induit qu'un participant qui a déjà une donnée hospitalière voit son risque de nouveau diagnostic hospitalier multiplié par 10 par rapport à un individu sans historique hospitalier.

Au-delà de la question de l'architecture la plus adaptée aux données de parcours de soins, seules des modélisations complexes peuvent prendre en charge l'ensemble des informations disponibles dans le SNDS. La littérature a suivi le sillage des architectures toujours plus performantes du traitement automatique du langage naturel (NLP), sans pleinement démontrer que l'augmentation de la complexité des modèles et du volume de données se traduit systématiquement par un gain prédictif. Néanmoins, la question reste en suspens puisque la dimension des données (nombre de codes d'événements, nombre de patients et de séquences de soins dans l'échantillon d'entraînement) n'a pas encore atteint des volumes comparables à ceux des modèles comme BERT. Dans la première version de ce dernier, le vocabulaire (équivalent du nombre de codes d'événements distincts dans les parcours de soins) est de 30 000 tokens, et le nombre de tokens utilisés au cours de l'entraînement atteint plus de 4 milliards (équivalent du nombre d'événements des séquences de soins accessibles à l'entraînement). Dans la base principale du SNDS, la dimension accessible (en nombre de patients comme en nombre d'événements) est potentiellement bien supérieure aux données internationales, fournissant le cadre dans lequel ces modèles peuvent trouver leur plus forte utilité.

La question de l'utilité de ces modèles reste pour l'heure en débat, pour au moins trois raisons : un manque d'évaluation rigoureuse de leurs bénéfices au-delà de la performance statistique globale, l'absence d'impact clinique à ce jour (Wornow, *et al.,* 2023) et des difficultés liées à l'explicabilité des résultats des modèles. Ainsi, une évaluation rigoureuse des performances statistiques des modèles entraînés sur un entrepôt de données de santé tel que la base principale du SNDS, couplée à une explicabilité détaillée des résultats, permettrait de faire avancer le débat et de le situer par rapport au patrimoine de données françaises.

Nous proposons l'application de deux modélisations inspirées du domaine du traitement automatique du langage naturel aux séquences de soins issues de la base principale du SNDS. La première repose sur un plongement vectoriel de chaque code d'événement, construit à partir de leur cooccurrence (cette étape peut être rapprochée de *Word2Vec*, et appelé Snds2vec dans la suite comme proposé par Doutreligne, *et al*., 2020), puis agrégés temporellement en une trajectoire pour chaque patient, préalable au temps de prédiction. Cette trajectoire correspond à un vecteur de dimension 300, qui est utilisé en entrée d'un modèle de *xgboost (extreme gradient boosting trees)*, une technique similaire aux forêts aléatoires, mais en général plus performante, qui combine plusieurs arbres de décision construits successivement en apprenant à chaque étape des erreurs déjà commises. La deuxième repose sur un modèle BERT (*Bidirectional Encoder Representations from Transformers),* un modèle canonique d'encodeur dans le domaine du traitement automatique du langage naturel, qui a déjà fait l'objet d'extrapolation à la modélisation des séquences de soins de patients, issues de bases administratives médico-économiques, en Angleterre (BEHRT, sur des données issues du *Clinical Practice Research Datalink*), aux États-Unis (Med-BERT, sur des données du *Cerner Health Facts*), ou encore au Danemark, en combinant des données hospitalières et relatives au marché du travail (*Life2Vec*, Savcisens, *et al*., 2024). Les modèles génératifs, plus récents, sont laissés de côté pour des développements ultérieurs. Ils sont plus naturellement orientés vers la génération de la suite de la trajectoire à court terme plutôt que vers des prédictions à horizon fixe, principal problème traité dans nos comparaisons.

Tableau 1 Modèles généralistes appliqués aux parcours de soins et tâches prédictives évaluées

Modèle	Population d'entraînement	Type de données	Type de modèle	Prédiction
BERT (NLP)	3,3 milliards de mots découpés en environ 4,3 milliards de « tokens »	BooksCorpus (800 millions de mots) et Wikipédia Anglais (2 500 millions de mots)	BERT (30 000 tokens, 110 millions de paramètres pour BERT-base, 340 pour BERT-large)	Nombreuses tâches de NLP (après fine-tuning)
BEHRT	1,6 million de patients avec au moins 5 visites dans le CPRD (UK), chaînage ville et hôpital d'un échantillon de médecins généralistes volontaires	Diagnostics (301 groupes de pathologies distinctes) groupés par visite, âge	BERT (301 tokens, 10 millions de paramètres)	301 groupes de pathologies dans la prochaine visite, avec ou sans horizon temporel
Med-BERT	28,5 millions de patients avec au moins 3 codes diagnostics du Cerner Health Facts (600 hôpitaux et cliniques US)	Diagnostics (CIM-9/CIM-10) soit 82 000 codes, groupés par visite, âge	BERT (> 82 000 tokens, 17 millions de paramètres)	Insuffisance cardiaque chez des patients diabétiques ; Cancer du pancréas
CORE-BEHRT	1,8 million de patients d'une région du Danemark avec une procédure chirurgicale entre 2016 et 2022	Diagnostics CIM-10 (niveau 4) ; médicaments (ATC niveau 4), temps	Optimisation BEHRT (17 469 tokens, 8 millions de paramètres)	Mortalité, Traitement de la douleur, infection et 25 pathologies, souvent en sous-population d'âge, pour divers horizons
CEHR-BERT	2,4 millions de patients d'un hôpital à New York ; 184,7 millions d'événements	Conditions, procédure et médicaments	BERT, amélioration de la prise en compte du temps (9 millions de paramètres)	Hospitalisation, mortalité, nouvelle insuffisance cardiaque, ou réadmission pour insuffisance cardiaque
TransformEHR	6,5 millions de patients soignés dans les Hôpitaux Veterans Health Administration (US) entre 2016 et 2019, 1.1 milliards d'événements	Diagnostics (CIM-10), groupés par visite, âge, sexe, groupe ethnique	Transformer	Cancer du pancréas, Gestes auto-infligés chez les patients avec un syndrome de stress post-traumatique ; 10 pathologies communes et 10 pathologies rares
Life2Vec	Population nationale danoise (~6 millions de personnes, 2008 – 2020) issues de registres nationaux danois	Événements de vie : santé (704 diagnostics CIM-10), emploi, revenu, éducation, adresse	BERT (8,4 millions de paramètres)	Mortalité et traits de personnalité
Delphi-2M	400 000 patients volontaires de la UK Biobank, recruté à partir de leurs 40 ans.	Diagnostics (CIM-10), âge, sexe, poids, tabac, alcool	GPT-2 (1 258 tokens, 2 millions de paramètres)	Prolonge la trajectoire dans la typologie des données d'entrée (code diagnostics, mortalité)
Foresight	811 336 patients de 3 hôpitaux londoniens	Données mixtes : notes cliniques pré-structurées, diagnostics, prescriptions	GPT	Prolonge la trajectoire dans la typologie des données d'entrée

Note > Seuls les paramètres disponibles dans les articles référencés sont présentés. **NLP :** Natural Language Processing (traitement automatique du langage naturel)

Un point d'attention important, souvent peu évoqué dans ces travaux, est la capacité des modèles généralistes à apporter une réelle information supplémentaire à un clinicien, plutôt que de simplement apprendre l'étape suivante d'un ensemble d'actions cliniques déjà initiées (Beaulieu-Jones, *et al.,* 2021). Parmi la masse d'information que le modèle peut apprendre, une succession d'événements peut révéler une démarche diagnostique en cours (présence d'un test spécifique), un traitement récurrent amené à se répéter dans le futur ou encore un acte préparatoire à une hospitalisation. Si une personne a déjà été hospitalisée pour un motif, le plus probable pourrait être qu'elle le soit pour le même motif la fois suivante. Ainsi, une partie des performances de BEHRT chutent quand la prédiction est restreinte aux diagnostics dits « incidents », soit ceux qui arrivent pour la première fois dans la trajectoire.

Les biais liés à la manière dont les données sont produites – parce qu'elles dépendent du recours aux soins, des décisions médicales et d'incitations plus ou moins fortes au codage – ne doivent pas être négligés, mais ne relèvent pas nécessairement d'un problème de qualité de données ou de bruit (Agniel, *et al.,* 2018). Il est en effet faux de suggérer que ces éléments n'ont pas de valeur informative alors, qu'au contraire, ils peuvent générer un signal fort en prédiction. Par exemple, un test de laboratoire, même sans le résultat, indique une suspicion clinique, qui, si elle réduit à court terme l'intérêt de prédire la pathologie en question, permet de tenir compte à long terme d'un épisode du parcours qui embarque également indirectement le processus décisionnel des soignants en charge du patient.

Dans nos travaux, plusieurs solutions sont expérimentées : prédire l'incidence, soit la première occurrence d'un code, plutôt que la récurrence des soins, masquer les événements de la période précédant immédiatement le début de la prédiction ; quantifier la perte de performance en étendant l'horizon temporel de prédiction ; enfin, examiner quels codes expliquent le plus les prédictions, pour repérer ceux qui reflètent déjà une démarche diagnostique en cours. Dans une logique de prévention, l'utilisation de ces codes seraient discutables car non révélateurs d'un risque « en amont » mais d'une hypothèse clinique en cours d'exploration (par exemple, un acte de colposcopie dans le cas de la prédiction d'une hospitalisation pour cancer du col de l'utérus).

Surmonter les obstacles techniques

La Drees bénéficie, dans le cadre de son accès permanent au SNDS, d'extractions mises à disposition par la Cnam. Réceptionnées en format de données SAS, les tables sont transformées en format parquet, et aplaties à l'aide de *SCALPEL flattening* (Bacry, *et al*., 2020) pour obtenir une dénormalisation des tables DCIR, MCO, SSR et HAD (application des jointures permettant de passer d'une structure de plusieurs tables gravitant en étoile autour d'une table centrale à une seule table, présentant des redondances mais avec de meilleures performances en lecture). La chaîne de préparation de données présentée dans la suite se situe en aval de ces opérations. Ces opérations nécessitent une infrastructure de calcul puissante, à l'état de l'art et homologuée conforme au référentiel de sécurité du SNDS.

La création du jeu de données SeqNDS, représentant de manière unifiée les séquences d'événements des patients, avec diverses opérations de mise en qualité et un partitionnement par bénéficiaire, a été réalisée à l'aide de Spark. Ce processus a duré environ 40 heures, mobilisant 80 COU et 250 Go de RAM et la base « SeqNDS 2018-2022 » résultante, sous format parquet compressé, pèse 146 Go en stockage.

À partir de ce format de données partitionné, les différentes étapes de *feature engineering* (préparation des variables d'entrée des modèles), de modélisation et d'évaluation ont été optimisées dans une optique d'efficacité — le partitionnement permettant notamment un traitement en mémoire. Ces traitements ont été effectués en Python, avec un usage prédominant des bibliothèques Polars (pour le traitement efficace des DataFrames), xgboost (pour les modèles de gradient boosting) et PyTorch (pour les autres modélisations : régressions logistiques, Singular Value Decomposition approchée, et *deep learning* en combinaison avec la bibliothèque transformers). Ces différentes modélisations ont pleinement tiré parti de la GPU NVIDIA A100 et de ses 80 Go de VRAM dont est équipé le serveur.

L'ensemble de ces outils a permis d'obtenir des temps de traitement raisonnables : de l'ordre de 8 heures pour le calcul des embeddings de Snds2vec ; pour les modélisations avec gradient boosting se basant sur Snds2vec, sur 30 millions d'individus, de l'ordre de 5 heures supplémentaires pour la préparation des données pour l'ensemble des modèles (essentiellement agrégation temporelle avec ACP, cf. *infra*.), puis 18 minutes par cible modélisée avec gradient boosting à partir des données préparées, ces deux temps étant globalement proportionnels à la taille du jeu d'entraînement ; concernant les modèles transformers, le pré-entraînement de BEHRT a duré 12 jours, et le fine-tuning 2 jours.

L'utilisation de codes partagés en open source par les projets de recherche précédents[2] (tous en langage Python, se fondant sur la bibliothèque transformers ou son ancêtre) permet de répliquer les travaux développés par d'autres. Les opérations les plus coûteuses en temps humain ont concerné la préparation des données.

[2] BEHRT : https://github.com/deepmedicine/BEHRT, Life2vec: https://github.com/SocialComplexityLab/life2vec, Delphi-2M: https://github.com/gerstung-lab/Delphi

BASE SEQNDS : UNE VUE INDIVIDUELLE ET LONGITUDINALE SUR LES ÉVÉNEMENTS DE SOINS

Nous partons du SNDS 2018 – 2022, un entrepôt exhaustif mais hétérogène, où les informations cliniques et médico-administratives sont réparties entre plusieurs sous-systèmes (ville, hospitalier MCO/SSR/HAD, médico-social) et codées selon des nomenclatures distinctes (CIM-10, CCAM, NABM, LPP, CIP/UCD/ATC, groupages PMSI, voir *glossaire*). L'objectif méthodologique est de projeter cet ensemble vers un schéma commun, centré patient et ordonné dans le temps, afin de permettre des analyses populationnelles et la modélisation prédictive sans réimplémenter, à chaque étude, des règles spécifiques à chaque source. La préparation des données est détaillée en *annexe 1*, et la typologie des données obtenues (nomenclatures, hiérarchies, sources, etc.) illustrée en *annexe 3*.

Normalisation sémantique des événements

Nous établissons un référentiel de codes unifié qui concatène les différentes nomenclatures dans un dictionnaire commun, hiérarchisé et documenté. Un même concept clinique (par exemple, un principe actif ATC, un acte CCAM ou un diagnostic CIM-10) est rattaché à un identifiant stable, relié à sa description textuelle et à ses parents hiérarchiques.

Tableau 2 Événements disponibles dans la base SeqNDS par grande typologie

Typologie	Type d'événements	Nombre d'événements (millions)	Événements distincts (tokens)	Nombre de bénéficiaires (en millions)	Nombre moyen d'événements par bénéficiaire avec au moins un tel événement
PFS_ACT_NAT	Nature d'actes des professionnels de santé (hors médecin)	8 187,5	29	75,8	107,9
ATC, UCD, CIP	Médicaments	7 265,7	24 872	73,9	98,3
NABM	Actes de biologie	5 509,8	1 092	67,7	81,4
PFS_SPE	Spécialités médicales consultées (médecin)	2 751,6	55	75,1	36,6
CCAM	Actes médicaux	1 905,8	8 237	71,4	26,7
CIM-10	Diagnostics	966,7	16 435	37,9	25,5
LPP	Dispositifs médicaux	892,4	20 786	62,8	14,2
TYP_UM	Type d'unités médicales à l'hôpital	142,5	45	36,3	3,9
GHM	Groupe homogène de malades à l'hôpital (MCO)	131,6	2 625	35,9	3,7
DÉPENDANCE	Cotation de la dépendance (SSR, HAD)	128,3	18	3,3	38,8
	Passage aux urgences	113,2	1	39,3	2,9
GME	Groupe médico-économique (SSR)	34,7	2 057	3,4	10,1
GHPC	Groupe homogène de prise en charge (HAD)	5,6	1 840	0,6	9,8
ETB_CAT	Catégorie d'établissements	1,4	26	1,1	1,3
Total		**28 036,7**	**78 118**	**76,4**	**367,0**

Source > Base SeqNDS, 2018-2022.
Note > ATC : Classification anatomique, thérapeutique et chimique ; UCD : unités communes de dispensation ; CIP : code identifiant des présentations ; NABM : nomenclature des actes de biologie médicale ; CCAM : classification communes des actes médicaux ; CIM-10 : classification internationale des maladies, version 10 ; LPP : liste des produits et des prestations ; PFS_ACT_NAT, PFS_SPE, TYP_UM, ETB_CAT, GHM, DÉPENDANCE, GME, GHPC sont des variables présentes dans le SNDS (SNIIRAM ou PMSI).

Les événements sont représentés par n = 78 118 codes distincts, dans 14 typologies (*tableau 2*). Ils se répartissent en n = 16 435 codes diagnostics distincts (CIM-10), essentiellement en milieu hospitalier[3], n = 21 866 codes représentant des délivrances de médicaments en ville (CIP), n = 3 007 codes représentant les délivrances de certains médicaments à l'hôpital (UCD), n = 20 786 codes représentant les dispositifs médicaux (LPP), n = 8 237 codes représentant des actes médicaux, réalisés en ville ou à l'hôpital (CCAM), n = 1 092 codes représentant des actes de biologie réalisés en ville (NABM), mais aussi 29 codes précisant les professionnels de santé consultés, 55 la spécialité médicale des médecins consultés, 26 codes représentant le type d'ESMS accueillant la personne, et en cas d'hospitalisation, le type d'unité médicale ayant pris en charge le patient (n = 45) et une indicatrice de passage par les urgences. De plus, est également considérée l'information des séjours hospitaliers caractérisant la prise en charge dont a bénéficié le patient, soit le groupe homogène de malades en MCO (n = 2 625), le groupe médico-

[3] Les affections de longue durée (ALD) sont collectées en nomenclature CIM-10, quand elles sont postérieures au début de la période d'observation (1er janvier 2018).

économique en SSR (n = 2 057) ou le groupe homogène de prise en charge en HAD (n = 1 840), ainsi que la cotation de la dépendance en SSR et HAD.

Les événements les plus fréquents dans chacune des typologies sont présentés en *tableau 3* afin d'illustrer quelques exemples de l'unité sémantique utilisée pour représenter les parcours de soins. Dans la suite, un événement est identifié à un code dans ce référentiel, et est entendu comme un « token » dans les modèles de langage.

Tableau 3 Événements les plus fréquents

Code de l'événement	Description
cip : 3 595 583	DOLIPRANE 1 000 mg, comprimé - plaquette(s) thermoformée(s) PVC-aluminium de 8 comprimé(s)
cip : 3474419	KARDEGIC 75 mg, poudre pour solution buvable en sachet-dose - 30 sachet(s)-dose(s) papier aluminium polyéthylène de 153,45 mg
cip : 3461546	DOLIPRANE 2,4 POUR CENT, suspension buvable - 1 flacon(s) en verre brun de 100 ml avec seringue(s) pour administration orale en polyéthylène/polystyrène ou polypropylène/polyéthylène avec fermeture de sécurité enfant
cip : 4153396	DOLIPRANE 1 000 mg, gélule - plaquette(s) thermoformée(s) PVC-Aluminium de 8 gélule(s)
ccam : YYYY600	*Supplément pour archivage numérique d'une mammographie ou d'un examen scanographique ou remnographique*
ccam : HBJD001	*Détartrage et polissage des dents*
ccam : DEQP003	*Électrocardiographie sur au moins 12 dérivations*
ccam : BLQP010	*Examen de la vision binoculaire*
cim10 : Z51.1	Séance de chimiothérapie pour tumeur
cim10 : Z49.1	Dialyse extracorporelle
cim10 : N18.5	Maladie rénale chronique, stade 5
cim10 : I10	Hypertension essentielle (primitive)
dependance : hab4	*Dépendance de type habillage, de niveau 4*
dependance : dpl4	*Dépendance de type déplacement / locomotion, de niveau 4*
dependance : ali2	*Dépendance de type alimentation, de niveau 2*
dependance : cpt2	*Dépendance de type comportement, de niveau 2*
etb_cat : 200	MAISON DE RETRAITE
etb_cat : 354	SERVICE DE SOINS À DOMICILE
etb_cat : 362	ÉTABLISSEMENT DE SOINS DE LONGUE DURÉE
etb_cat : 202	LOGEMENT-FOYER POUR PERSONNES ÂGÉES
ghm : 28Z07Z	*Chimiothérapie pour tumeur, en séances*
ghm : 28Z04Z	*Hémodialyse, en séances*
ghm : 28Z18Z	*Radiothérapie conformationnelle avec modulation d'intensité, en séances*
ghm : 06K04J	*Endoscopie digestive diagnostique et anesthésie, en ambulatoire*
ghpc : 1013	Pansements complexes et soins spécifiques (stomies compliquées) ; Pas de mode de prise en charge associé ; Index de Karnofsky valant 40
ghpc : 1012	Pansements complexes et soins spécifiques (stomies compliquées) ; Pas de mode de prise en charge associé ; Index de Karnofsky valant 50
ghpc : 368	Soins palliatifs ; Pas de mode de prise en charge associé ; Index de Karnofsky valant 30
ghpc : 1011	Pansements complexes et soins spécifiques (stomies compliquées) ; Pas de mode de prise en charge associé ; Index de Karnofsky valant 60
gme : 0509A0	*Coronaropathies (à l'exclusion des coronaropathies avec pontage) , score phy <= 8 , score rr <= 90 - zéro jour*
gme : 2303A1	*Soins palliatifs , score rr <= 60 - niveau 1*
gme : 1903A1	*Toxicomanies avec dépendance , score cog <= 6 - niveau 1*
gme : 0872B1	*Fractures de l'extrémité supérieure du fémur (à l'exclusion des FESF avec implant articulaire) , score phy >= 9 - niveau 1*
lpp : 1187880	PPC, APNÉE SOMMEIL, PATIENT TÉLÉSUIVI (+ DE 112 H), FORFAIT HEBDO 9.TL1
lpp : 1241763	LIT MEDICAL, LIT STANDARD, LOCATION HEBDOMADAIRE, LIT ET ACCESSOIRES
lpp : 2223342	MONTURE , > OU = 18 ANS
lpp : 2264861	OPTIQUE, MONTURE ADULTE DE CLASSE B
nabm : 9005	*FORFAIT DE PRISE EN CHARGE PRE-ANALYTIQUE DU PATIENT*
nabm : 9105	*FORFAIT DE SÉCURITÉ POUR ÉCHANTILLON SANGUIN*
nabm : 1104	*HÉMOGRAMME Y COMPRIS PLAQUETTES (NFS , NFP)*
nabm : 0592	*SANG : CREATININE*
pfs_act_nat : 50	PHARMACIE D'OFFICINE
pfs_act_nat : 24	INFIRMIER
pfs_act_nat : 26	MASSEUR - KINÉSITHÉRAPEUTE
pfs_act_nat : 30	LABORATOIRE
pfs_spe : 01	*MÉDECINE GÉNÉRALE*
pfs_spe : 06	*RADIODIAGNOSTIC ET IMAGERIE MÉDICALE*
pfs_spe : 15	*OPHTALMOLOGIE*
pfs_spe : 33	*PSYCHIATRIE GÉNÉRALE*
typ_um : 53	Autre chirurgie adulte (ou chirurgie indifférenciée adulte)
typ_um : 29	Autres spécialités médicales adultes (non classées ailleurs) ou unité de médecine indifférenciée
typ_um : 21	Hémodialyse en centre pour adulte
typ_um : 43	Unité de chimiothérapie ambulatoire
urgence	*Admission aux urgences*

La source de l'information, qui décrit le contexte dans lequel elle a été recueillie (table et variable), et permettant de faire le lien avec la documentation fournie par la Cnam, a été conservée dans les tables (voir *annexe 3*). Néanmoins, la source de l'information n'est pas prise en compte dans les modélisations, qui ne tiennent compte que des codes des événements. Ainsi, les codes diagnostics CIM-10 sont représentés de la même manière qu'ils procèdent d'un diagnostic hospitalier (par exemple, qui ont pour source : MCO_B__DGN_PAL, soit la table MCO_B et de la variable DGN_PAL) ou d'une affection de longue durée, ALD (source : IR_IMB_R__MED_MTF_COD) et un acte médical décrit par la CCAM est représenté de la même manière qu'il ait été réalisé en ville (source : ER_CAM_F__CAM_PRS_IDE) ou à l'hôpital (source : MCO_A__CDC_ACT).

Description des tables « SeqNDS »

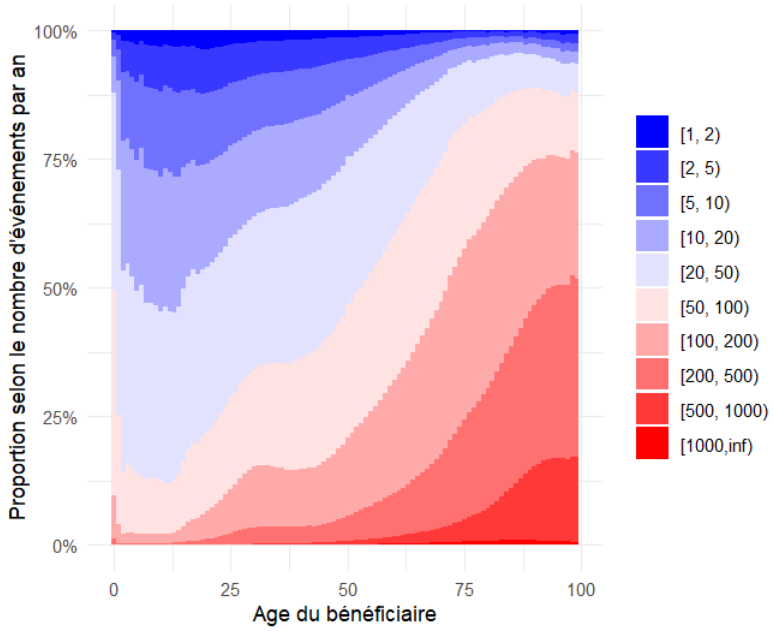
La table « **Bénéficiaires** » contient un enregistrement par personne : l'identifiant, les attributs démographiques minimaux, des repères temporels d'activité (premier et dernier événement observé, volume d'événements) et un identifiant de partition aléatoire facilitant l'échantillonnage au 1/200e.

La table « **Événements** » aligne tous les faits cliniques et médico-administratifs sur un même schéma : identifiant de personne, code standardisé (clé vers le référentiel hiérarchique), source (qui conserve le contexte d'origine) et dates harmonisées permettant l'ordonnancement fin des trajectoires. Chaque ligne est ainsi un jalon temporel interprétable dans le parcours de soins, s'interprétant dans le référentiel hiérarchique.

Cette organisation fait de SeqNDS une base séquentielle centrée patient : elle permet d'extraire des fenêtres d'observation, de décrire des trajectoires et d'alimenter, sans retraitement source spécifique, des analyses descriptives, des modèles tabulaires et des modèles séquentiels.

À l'échelle, la base couvre l'ensemble des 76,4 millions de bénéficiaires sur 2018 – 2022[4] totalisant 28 milliards d'événements, avec des performances pratiques assurées par le partitionnement et la standardisation.

Graphique 1 Nombre d'événements par âge et par an



Lecture > Chaque année, les bénéficiaires de 100 ans sont 17 % à avoir plus de 500 événements dans leur trajectoire.
Source > Base SeqNDS, 2018-2022.

En résumé, nous passons d'un entrepôt multisources et hétérogène à un schéma pivot centré patient, obtenu via (i) la consolidation de l'unité d'analyse (voir *annexe 1* pour l'identification des patients), (ii) la normalisation

[4] Le nombre total de bénéficiaires sur la période 2018-2022 est naturellement supérieur au nombre d'habitants en France à une date donnée, tel qu'estimé par l'Insee (entre 67,0 millions d'habitants au 1er janvier 2018 et 68,2 millions d'habitants au 1er janvier 2023). Par rapport au nombre d'habitants en début de période, le nombre de bénéficiaires potentiels inclut aussi notamment l'ensemble des naissances survenues au cours de la période (3,7 millions), les flux migratoires entrants (2,1 millions), les retraités résidant à l'étranger (1,1 million) et leurs ayants droit ainsi que l'ensemble des personnes de passage sur le territoire français pour moins de 12 mois avec des droits ouverts à l'Assurance Maladie (étudiants internationaux par exemple). De plus, une même personne peut au cours de sa vie être identifiée sous plus d'un pseudonyme dans le SNDS, générant des doubles comptes qui peuvent subsister malgré les précautions prises pour consolider des trajectoires correspondant à une seule et même personne (*annexe 1*, section identification des individus).

sémantique des codes, (iii) l'harmonisation temporelle et (iv) la déduplication et le contrôle de cohérence. Ces étapes produisent les deux tables Bénéficiaires et Événements qui constituent la base SeqNDS, une vue individuelle et longitudinale des soins directement exploitable pour l'épidémiologie, la statistique et la prédiction.

Le nombre d'événements collectés augmente avec l'âge à partir de 12 ans environ (*graphique 1*). À 50 ans, 47 % des bénéficiaires ont plus de 50 événements dans l'année, et à 75 ans, cette proportion monte à 82 %.

Encadré 1 Exemple de séquence autour d'un épisode d'infarctus hospitalisé

Pour illustrer l'information disponible lors d'un séjour hospitalier, sous ce nouveau format, nous prenons l'exemple réaliste du point de vue de la richesse des informations, mais bruité, d'un individu hospitalisé en unité neuro-vasculaire en urgence. L'information sur le groupe homogène de malades indique un séjour en MCO, l'information sur le groupe médico-économique indique un séjour en SSR.

Début : t, Fin : t

Admission aux urgences (urgence)
Désorientation, sans précision (cim10 : R45.1)
Médecine générale (pfs_spe : 01)
Consultation radio-diagnostique et imagerie médicale (pfs_spe : 06)
Remnographie [IRM] du crâne (ccam : ACQN001)
Nature d'actes de professionnel de santé non-médecin : laboratoire (pfs_act_nat : 30)
Supplément pour actes en urgences (nabm : 9005)
(..quelques dizaines d'examens biologiques nabm :..)
UHCD structures des urgences (typ_um : 07A)

Début : t + 1, Fin : t + 6

Infarctus cérébral dû à une embolie (cim10 : I63.4)
Accidents vasculaires intracérébraux (ghm : 01M303)
Soins intensifs en UNV (typ_um : 18)
UNV hors SI (typ_um : 17)
Occlusion et sténose de l'artère (cim10 : I66.0)
(… quelques codes diagnostics cim 10…)
Démence vasculaire mixte, corticole et sous-corticale (cim10 : F01.3)
Médecine générale (pfs_spe : 01)
(…quelques actes ccam...)

Début : t+7 jours, Fin : t + 9 jours

Accidents vasculaires cérébraux autres , score phy >= 9 , score cog >= 5 , score rr <= 90 - niveau 2 (gme : 0148C1)
Accident vasculaire cérébral (cim10 : I64)
(..cotation de la dépendance,…)
(.. code lpp…)

Représentation vectorielle des événements

La dernière étape de transformation des données consiste à obtenir une représentation vectorielle (« *embeddings* », ou plongement vectoriel) pour chacun des codes d'événements, réutilisable en entrée des modélisations de trajectoire. Deux candidats naturels ont été évalués. Le premier, **Snds2vec**, une simplification du modèle « *Word2Vec* » (introduite dans Levy et Goldberg, 2014) s'inscrit dans la lignée des travaux de Doutreligne, *et al*. (2020) avec l'utilisation des cooccurrences des codes dans les parcours de soins afin de représenter les proximités entre les codes. Le second s'appuie sur un modèle de langage pré-entrainé sur des textes médicaux multilingues, « CODER » (Yuan, *et al.*, 2022), qui est appliqué aux libellés descriptifs issus des nomenclatures officielles.

Les cooccurrences des codes d'événements dans les parcours de soins, chez un même bénéficiaire, à moins de 30 jours d'intervalle, sont calculées sur l'ensemble des bénéficiaires de l'échantillon d'apprentissage (50 millions de bénéficiaires).

Pour tout couple de codes d'événements x et y, on calcule $c_{x,y}$, le nombre d'occurrences de ces deux codes chez un même individu dans une fenêtre de 30 jours (décompte des triplets [individu, événement, événement]), ainsi que $c_x = \sum_y c_{x,y}$, et $c = \sum_{x,y} c_{x,y}$.

La matrice suivante, appelée « *Positive Pointwise Mutual Information* » représente le logarithme du rapport de vraisemblance d'observer les deux codes ensemble dans une fenêtre de 30 jours par rapport à ce que l'hypothèse

d'indépendance (produit de leurs fréquences respectives) ne le laisserait supposer (seules les surreprésentations sont conservées) :

$$A_{x,y} = \max\left(\log\frac{c_{x,y}\,c}{c_x c_y}, 0\right).$$

Cette matrice, de dimension $n \times n$, n = 78 118, est diagonalisée afin de réduire sa dimension dans un espace vectoriel de dimension p = 300, associé aux valeurs propres de plus grandes amplitudes. Ainsi, par cette approche inspirée de Word2Vec (Levy et Goldberg, 2014), chaque code d'événement est représenté par un vecteur de dimension p = 300.

L'approche « *Word2Vec* » par plongements vectoriels issus des cooccurrences dans les parcours s'est montrée nettement supérieure à l'approche « CODER » dans les expérimentations avec un gain de 1,3 point d'AUC stratifiée en moyenne sur les cibles de prédiction et les modèles aval (*annexe 2*). Les résultats ne seront présentés dans la suite que pour l'approche « *Word2Vec* ».

La position dans l'espace de dimension p = 300 est supposée refléter une forme de « sens » ou « contenu sémantique ». Afin d'obtenir une représentation visuelle en deux dimensions, nous appliquons un algorithme non linéaire de réduction de la dimension « UMAP, *Uniform Manifold Approximation and Projection for Dimension Reduction* », qui procède à une simplification tout en cherchant à conserver la « similarité » entre les points de l'espace de départ dans l'espace d'arrivée de plus faible dimension. La similarité d'un point i avec un autre tient compte de la taille du voisinage de i et la distance avec l'autre point par rapport à la distance de i avec son plus proche voisin.[5]

Il est alors possible de représenter en deux dimensions l'espace sémantique ainsi créé (*graphique 2*). Si certaines régions sont plutôt occupées par des codes de médicaments, et d'autres par des dispositifs médicaux, les concepts en provenance de nomenclatures différentes peuvent se trouver co-localisés. À titre d'exemple, deux régions sont décrites : (a) l'une spécifique aux infections sexuellement transmissibles et (b) l'autre aux brûlures. La région (a) rassemble des diagnostics CIM-10 liées aux VIH (B20-B24) et *Monkeypox* (B00-B09), à la Syphilis (A50-A64) notamment mais aussi dans d'autres chapitres de la classification des maladies, avec une tumeur : le Sarcome de Kaposi (C46), un sarcome lié à l'infection par le VIH ou encore un trouble mental : la démence de la maladie due au VIH (F02) ; des GHM, GME, GHPC en cohérence, de nombreux codes de la classe ATC « J05 Antiviraux à usage systémique », des codes d'acte de biologie lié à la syphilis, ou VIH ou au *Monkeypox*, et un code de spécialiste « santé publique et médecine sociale ». La région (b) comporte des diagnostics tels que la classe dédiée de la CIM-10 pour les brûlures (T20-T32) mais aussi, bronchite due à des agents chimiques, des émanations des fumées ou des gaz (J68), dermite irritante de contact dû à des produits chimiques (L24.5), coup de soleil (L55), affections atrophiques de la peau (L90), les séquelles de brûlures (T95), les circonstances accidentelles très diverses ayant pu provoquer des brûlures (par exemple, contact avec de l'eau bouillante provenant d'un robinet en X11, lésion auto-infligée en X76, ou des agressions en X97) ou encore greffe de peau (Z94.5). Là encore les GHM et GHE sont en cohérence ainsi que les types d'unités médicales (réanimation et surveillance continue des grands brûlés, adulte ou pédiatrique). Les dispositifs médicaux sont ceux du chapitre 02.1, soit des vêtements compressifs divers, indiqués pour le traitement des grands brûlés. Les actes médicaux recouvrent des pansements chirurgicaux initiaux ou secondaires en lien avec des brûlures ou des greffes cutanées.

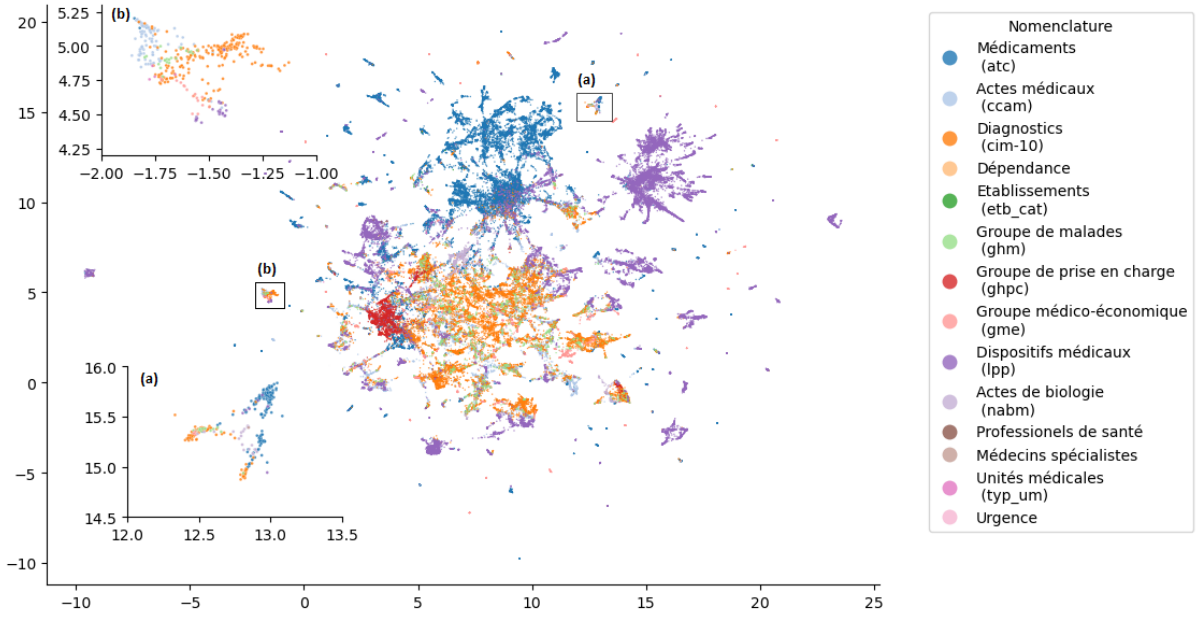
Ainsi, les cooccurrences ont permis de détecter des relations dépassant les nomenclatures hiérarchiques. Il est également notable que les codes diagnostics comme les codes médicaments ne se répartissent pas uniquement par chapitre mais semblent figurer des répartitions dans l'espace sémantique plus complexe, ce qui est moins le cas pour les codes produits de santé (LPP), où les chapitres semblent être plus clairement séparés.

Ces plongements vectoriels (« *embeddings* ») des codes d'événements peuvent être réutilisés en entrée de différentes modélisations. Les résultats des expérimentations en prédiction décrites dans la suite démontrent leur apport informationnel, compétitif par rapport à un modèle BEHRT. Leur usage pouvant largement dépasser des usages dans des modèles prédictifs présentés ci-après, ils sont décrits comme faisant partie de la base SeqNDS. En effet, ils peuvent être utilisés pour visualiser les cooccurrences d'événements dans le SNDS ou encore, après un choix d'agrégation temporelle, représenter les trajectoires de patients pour faire des regroupements de patients similaires (*matching*, approche cas-témoins, *clustering*, etc.).

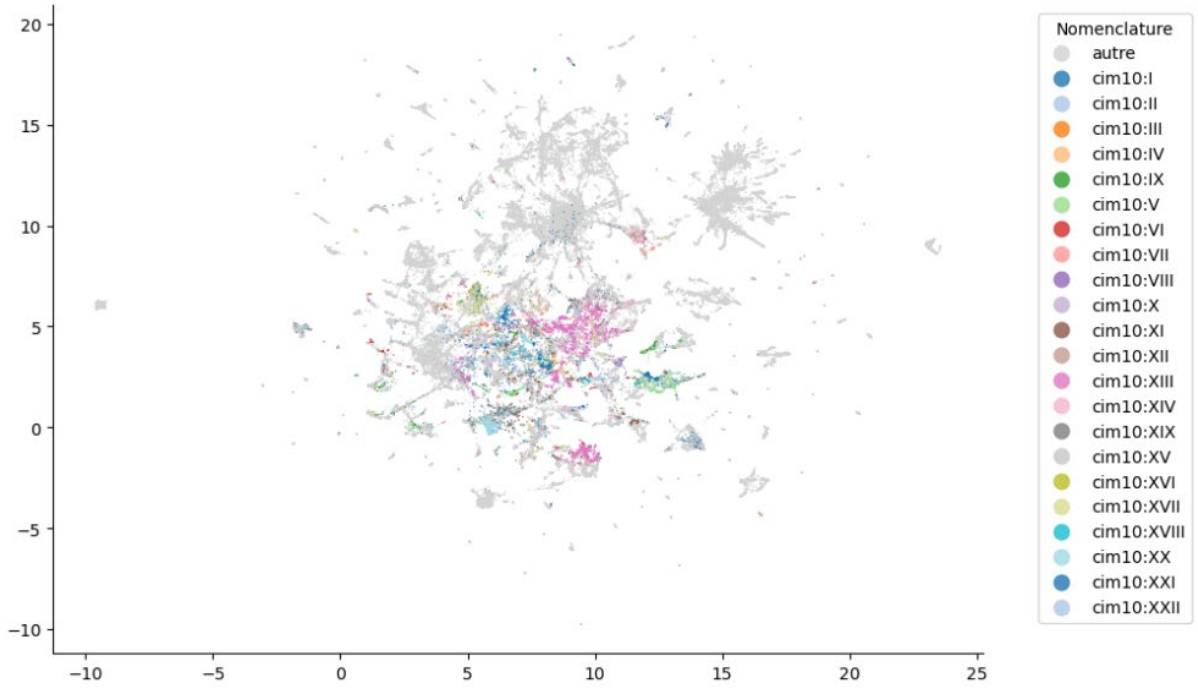
[5] Les paramètres utilisés sont les suivants : nombre de voisins à 15, distance minimale à 0,1 et métrique cosinus.

Graphique 2 Représentations vectorielles des codes Snds2vec

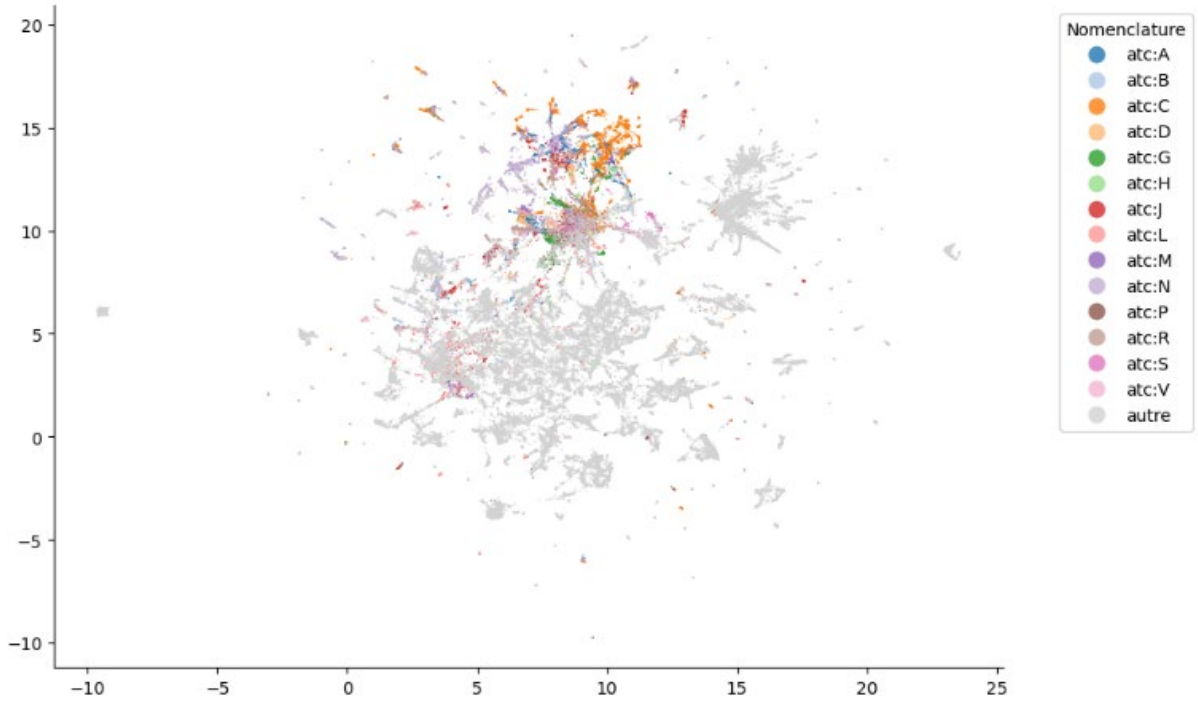
Graphique 2A Par typologie



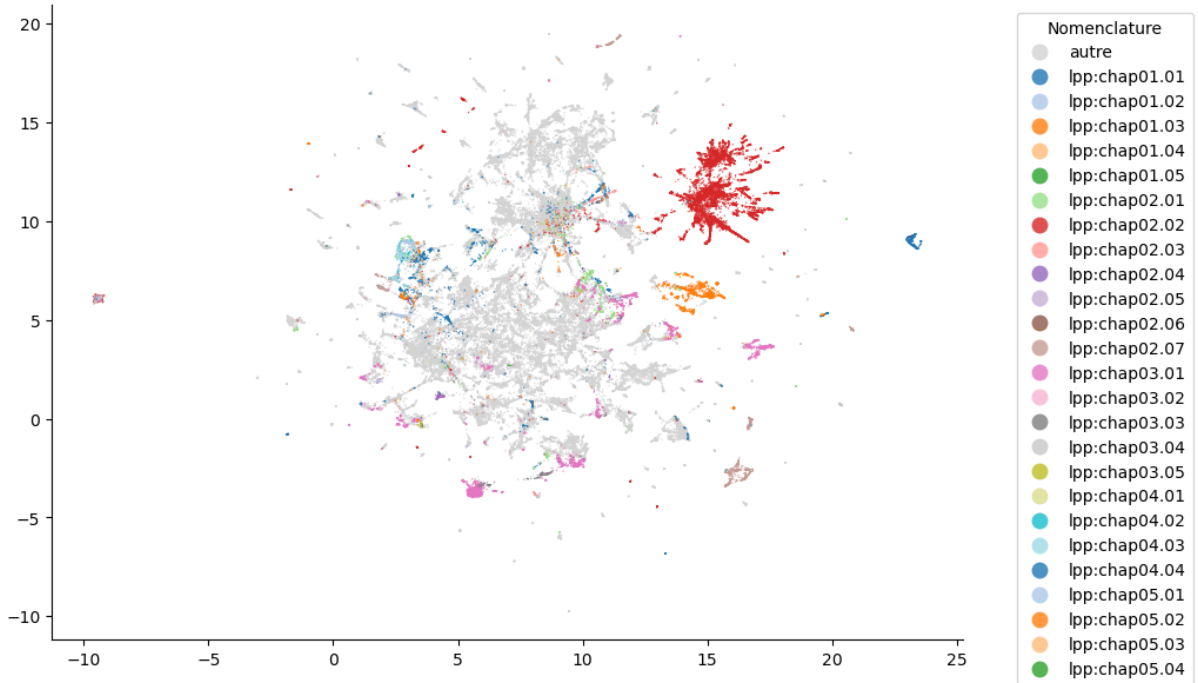
Graphique 2B Par chapitre de la classification internationale des maladies, CIM-10



Graphique 2C **Médicaments : groupe Anatomique (ATC)**



Graphique 2D **Dispositifs médicaux (LPP) par titre (01 : traitements, aides à la vie, aliments et pansements ; 02 : orthèses et prothèses externes ; 03 : DMI, implants et greffons tissulaires d'origine humaine ; 04 : Véhicule pour handicapés physiques ; 05 : DM invasifs non éligibles au titre III.)**



MODÉLISATIONS PRÉDICTIVES

L'objectif est d'évaluer la valeur prédictive des signaux d'état de santé disponibles dans le SNDS, pour un ensemble très large de pathologies et de manière standardisée, ce qui suppose de multiplier les problèmes de prédiction.

Il s'agit de prédire des cibles futures, soit la présence d'un événement ou d'un groupe d'événements dans la trajectoire future sous un certain horizon temporel, à partir de données d'entrée passées. Pour chaque individu, nous tirons un temps de prédiction t_0 au hasard, selon une loi uniforme allant de mi-2018 à mi-2022 et indépendamment des événements. L'entrée regroupe tous les événements antérieurs[6] à t_0, complétés par la date de naissance et le sexe. Ce tirage de t_0 augmente la variabilité, améliore la stationnarité de la prédiction en la forçant à fonctionner dans des contextes temporels différents (par exemple la période des confinements liés à la COVID-19). Il renforce ainsi la robustesse temporelle du modèle pour une validation hors domaine temporel.

Cibles à prédire

Nous considérons plusieurs cibles d'intérêt, qui cherchent à prédire si un événement arrivera avant un horizon donné (3 mois, 6 mois, 1 an ou 2 ans) à compter du temps de prédiction t_0. La mortalité toutes causes sert de référence, susceptible d'être liée à une grande variété d'événements de soins.

Afin d'envisager largement les événements de santé significatifs susceptibles d'affecter un individu la première hospitalisation observée pour une pathologie sert de « tâche patron », définie à partir de la classification internationale des maladies (CIM-10) et déclinée en 182 cibles primaires génériques et 20 cibles secondaires plus affinées. Ce choix vise à se rapprocher de la probabilité d'apparition d'une nouvelle maladie, dans le contexte où les diagnostics observés dans le SNDS le sont en milieu hospitalier pour l'essentiel (si l'on met de côté les affections de longue durée). De plus, l'observation qu'une partie de la performance des modèles comme BEHRT provient de la récurrence des séquences de soins a conforté ce choix (Li, *et al.,* 2020) : le score de précision en moyenne sur les patients[7] de 0,496 (AUC discriminant les codes présents des codes absents, en moyenne sur les patients[8] : 0,954) pour prédire les diagnostics de la prochaine visite médicale est réduit à 0,216 (AUC discriminant les codes présents des codes absents, en moyenne sur les patients : 0,904) quand la prédiction est restreinte à la première incidence des diagnostics dans les données.

Cibles primaires. Nous cherchons d'abord à définir des cibles de prédiction génériques, sous la forme de classes incidentes de pathologies hospitalières couvrant un large spectre de situations cliniques. Pour cela, nous regroupons les diagnostics en 147 classes de la CIM-10 (par exemple J40 – J47, *Maladies chroniques des voies respiratoires inférieures*), estimons la première date d'apparition par classe et excluons un individu pour une classe donnée s'il présente déjà un antécédent dans la classe au temps t_0 (par exemple, un diagnostic en J40 n'est pas considéré comme incident s'il est précédé d'un diagnostic J47). Les sources pour définir les diagnostics incidents, comme les antécédents, sont les diagnostics principaux (DP) et reliés (DR) des séjours en MCO. Pour 34 cas spécifiques (33 types de cancers[9] et 1 cible spécifique aux infections toutes causes, de différents appareils, en rassemblant 423 catégories et sous-catégories de la CIM-10[10]), une granularité différente est retenue, en plus des classes de regroupements de la CIM-10. Cette granularité est déterminée par un expert médical, et les diagnostics utilisés pour le critère d'apparition peuvent différer de ceux utilisés pour le critère d'exclusion (i.e. exclusion de tout antécédent de cancers dans le cas de la prédiction d'un type de cancer donné). En ajoutant la prédiction de la mortalité, les modèles sont donc d'abord évalués sur 182 cibles. En formulant les tâches de prédiction comme dans la littérature (notamment BEHRT et Delphi-2M), cette première famille de cibles permet de (i) comparer les différentes modélisations proposées dans des contextes de prédiction extrêmement variés, (ii) d'indiquer pour quelles hospitalisations les parcours de soins préalables sont le plus susceptibles d'être relié à une hospitalisation à venir, et donc (iii) de faire un premier filtre des pathologies susceptibles de bénéficier de telles modélisations.

Cibles secondaires. Une deuxième famille de cibles est définie en fonction de leur intérêt en santé publique, sans ambition d'exhaustivité. La prédiction peut répondre à un objectif de prévention primaire (comme pour la première

---

[6] La gestion des cas d'incertitudes temporelles dans la date de l'événement, afin de prévenir la fuite d'information est présentée en *annexe :1.*

[7] La part des codes prédits par le modèle qui se révèlent effectivement être dans la prochaine visite du patient, en moyenne sur les patients (voir *glossaire* pour une définition plus précise).

[8] Voir le *glossaire* pour plus de détails.

[9] D'après une liste élaborée par le Dr Diane Naouri et Hadrien Le Mer dans Dépistage du cancer : Is personnes modestes y recourent moins souvent. Drees, *Études et résultats* ,1367

[10] D'après une liste élaborée pour des travaux en cours aux Hôpitaux Universitaires Henri Mondor (APHP) par Dr Melica, Dr Matignon et Dr Hoisnard.

famille de cibles), secondaire (prévenir une aggravation dans le cas d'une pathologie déjà diagnostiquée) ou encore, dans la mesure où le modèle obtenu est explicable, à un objectif d'identification dans les parcours de soins de facteurs pronostics de gravité ou d'errance thérapeutique à même d'éclairer la gestion des parcours de soins.

Représentation des données d'entrée

Afin d'évaluer l'apport des variables représentant l'ensemble du parcours de soins, plusieurs alternatives moins riches sont considérées.

Données classiques

Le modèle **démographique** utilise l'âge exact ainsi que le sexe.

Le modèle « **cartographie** » utilise les 127 indicatrices de pathologies issues de la cartographie des pathologies de la Cnam la plus récente avant t_0.

Le modèle « **comorbidités** » utilise la fréquence des événements entrant dans deux indices de comorbidités classiquement utilisées avec des données administratives, 17 catégories pour l'indice de Charlson (Quan, *et al.*, 2005) et 46 pour le Rx-Risk-V (Pratt, *et al.*, 2018). Ces deux indices sont complémentaires puisque l'indice de Charlson est calculé sur les diagnostics hospitaliers uniquement, et l'indice Rx-Risk-V avec les prescriptions médicamenteuses dans les classes ATC uniquement.

Le modèle « **fréquences de consommation** » utilise les fréquences d'événements dans les 14 grands types d'événements, chacun caractérisé par sa nomenclature, sans en détailler la nature (voir *tableau 2*), c'est-à-dire le nombre de médicaments (resp. actes médicaux, diagnostics, dépendances, séjours en ESMS, GHM, GHPC, GME, produits ou prestations, actes biologiques, natures d'activités de non-médecin, spécialités de médecin, unités médicales, urgences) rapporté à la durée de la trajectoire. Ces variables peuvent être ajoutées au modèle « cartographie » ou au modèle « comorbidités ».

Données intégrant des représentations vectorielles denses des trajectoires des patients

Les **plongements vectoriels de trajectoires** sont obtenus en deux étapes : à partir des plongements vectoriels statiques dits **Snds2vec** (non réentraîné postérieurement) définis *supra*, les plongements vectoriels v_j de codes (dimension n = 300) intervenant dans la trajectoire d'un patient sont sommés avec cinq jeux de pondérations différents, qui introduisent la notion de temps et mettant plus ou moins en avant le court terme :

$$v(\tau) = \sum_j v_j \, e^{-(t_0 - t_j)/\tau} \, .$$

Avec τ, un temps caractéristique variant parmi 5 valeurs (7 jours, 28, 91, 365 et ∞). Ainsi, pour le premier jeu de pondération, seuls les événements dans une proximité de quelques semaines du temps de prédiction t_0 vont compter (au bout d'un peu plus d'un mois, la pondération des événements est inférieure à 1 %), tandis que pour le dernier, l'ensemble des événements sont pris en compte symétriquement. Une trajectoire représentée ainsi est de dimension n = 5 x 300, et une deuxième étape de réduction de dimension par ACP permet de revenir à une dimension n = 300.

Enfin, le modèle BERT prend nativement en entrée des séquences de codes/tokens, et les projette sur des embeddings au cours de l'entraînement. Le modèle BERT (NLP) avait 30 000 tokens, et un jeu de données d'entraînement de l'ordre de 4 milliards de tokens. Le modèle **BEHRT**, transposé aux séquences de soins, ne conservait que 301 tokens et uniquement des codes diagnostics. Tout en reprenant l'architecture proposée, nous utilisons un ensemble beaucoup plus large de codes/tokens : les 78 118 codes présents dans les données et hiérarchisés selon leur nomenclature sont regroupés par élagage en 30 000 codes qui constitueront le vocabulaire utilisé pour le modèle **BEHRT-SNDS**. L'élagage de l'arbre hiérarchique des codes (issu des niveaux des différentes nomenclatures utilisées) est conduit à partir d'un critère entropique sur les fréquences des codes pour regrouper les codes peu fréquents (on fusionne itérativement les codes avec leurs parents de manière à minimiser la perte d'information, au sens de l'entropie de Shannon, sur la distribution des fréquences des codes) tout en contraignant la granularité des codes pour ne pas remonter au-dessus des codes à trois caractères pour la CIM-10 (e.g. J21 Bronchiolite aiguë), au-dessus des codes ATC à trois caractères (e.g. J05 Antiviraux à usage systémique) etc. (voir *annexe 4*). Comme pour le modèle BERT, les séquences sont tronquées aux 512 codes/tokens les plus récents.

Finalement, la dimensionnalité des données d'entrée varie, de quelques dimensions bien choisies (par exemple les 127 indicatrices issues de la dernière cartographie des pathologies avant t_0), à virtuellement 30 000 événements possibles x 512 positions dans la séquence pour le modèle BEHRT.

Tableau 4 Dimensions des différentes représentations des données

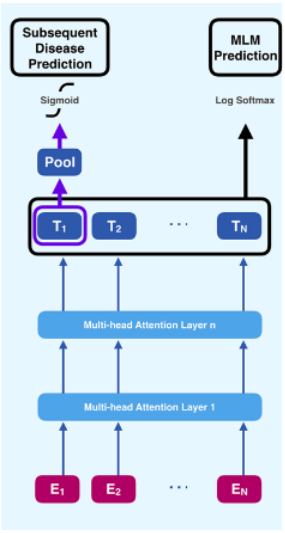
Représentations des données	Dimensions (ajoutées à âge et sexe[11])
BEHRT-SNDS	**30 000 x 512** (78 118 avant élagage x 512 au plus)
Plongements vectoriels des trajectoires (PVT) à partir de Snds2vec	**300** (78 118 x 5 avant réduction de dimension)
Comorbidités et fréquences de consommation	**77** (63 +14)
Cartographie et fréquences de consommation	**141** (127 +14)
Fréquences de consommation	**14**
Comorbidités (Charlson et Rx-Risk-V)	**63** (17 + 46)
Cartographie	**127**
Mortality Related Morbidity Index (MRMI)[12]	**1**
Démographique	**0**

Modélisations retenues

Pour toutes les données d'entrée, sauf celles nécessaires au modèle BEHRT, deux modélisations sont comparées, à données d'entrée fixées : un modèle linéaire et un modèle de gradient boosting ***xgboost***. Étant donné la supériorité manifeste de ce dernier modèle à travers les expérimentations (par rapport au modèle linéaire, gain de 3 points d'AUC stratifiée en moyenne sur l'ensemble des cibles et variantes de données d'entrée avec un entraînement sur 1 million de patients), les résultats présentés dans la suite se restreignent aux modèles de xgboost.

Le modèle **BEHRT** peut être décrit comme l'enchaînement de deux parties : une première partie qui sera commune à l'ensemble des tâches de prédiction, généraliste (*Multi-head Attention Layers*), qui à partir d'une séquence d'embeddings E_1 à E_N en entrée fournit une séquence d'embeddings T_1, …, T_N en sortie (*schéma 1*). Cette couche est pré-entrainée sur de grands volumes de données non labellisés pour fournir une représentation de la séquence T_1, …, T_N qui pourra être réutilisée sur de multiples tâches de prédiction en aval (*Masked Language Model Prediction*). Ensuite, une deuxième partie, dite couche de sortie est définie pour chaque nouvelle tâche de prédiction (par exemple, *Subsequent Disease Prediction*), et les poids associés au réseau de neurones de cette dernière devront être entraînés spécifiquement avec des données labellisées pour cette nouvelle tâche (*fine tuning*).

Schéma 1. **Modèle BEHRT, issu de Li, *et al.* (2020)**



Pour pré-entraîner la partie généraliste ainsi que les embeddings d'entrées, une stratégie « *Masked Language Model* » est utilisée. Le dictionnaire associant un token à un *embedding* est initialisé aléatoirement. Des échantillons de séquences auto-labellisées sont générés en masquant une partie des codes, en les remplaçant par un token dédié ([MASK]) et le modèle est entraîné à compléter la séquence ainsi altérée. La couche de sortie temporaire, spécifique au pré-entraînement répond à un problème de classification multiclasse, qui à chaque position masquée, propose une distribution de la probabilité de chacun des 30 000 tokens/codes composant le vocabulaire. Les poids du modèle sont appris par l'optimisation de la distance *cross-entropy* entre cette distribution et la distribution attribuant une probabilité de 1 aux tokens ayant été masqués.

Pour entraîner la partie spécialisée (« *fine-tuning* »), les poids de la partie généraliste du modèle sont initialisés par les poids obtenus lors du pré-entraînement. L'*embedding* T_1 du premier token qui joue un rôle particulier ([CLS][13]) de représentation de l'ensemble de la séquence (en pratique, seuls les tokens 2 à N correspondent à notre

[11] Dans les travaux de Li, *et al.* (2020) proposant le modèle BEHRT que nous suivons ici, le sexe n'est pas introduit, mais les auteurs montrent comment cette information est indirectement devinée par le modèle. Pour cette variante de données d'entrée, les dimensions sont à comprendre en plus de l'âge.

[12] MRMI : indice de comorbidité proposé par Constantinou, *et al.* (2018), construit à partir de la cartographie des pathologies de la Cnam pour prédire la mortalité à 2 ans.

[13] CLS pour « classification », puisque c'est l'entrée retenue pour les modèles de classification dans le *fine-tuning*.

vocabulaire de codes, le code appris T_1 de ce token joue le rôle de synthèse) est utilisé en entrée d'une nouvelle couche, initialisée aléatoirement, et produit une probabilité pour (i) chacune des 182 cibles primaires considérées, et (ii) chacune des cibles secondaires sur objectif de santé publique, à horizon fixe (1 an). L'entraînement spécialisé permet l'apprentissage des poids de la partie spécialisée, et l'actualisation des poids de la partie généraliste spécifiquement pour la tâche spécialisée. L'avantage de cette modélisation est aussi d'offrir un modèle unifié pour l'ensemble des cibles. Au total, BEHRT-SNDS possède 30 millions de paramètres à entraîner. Les vecteurs E_1 à E_N sont un des résultats de l'entraînement. Leur représentation en deux dimensions via la méthode utilisée pour représenter Snds2vec est fournie en *annexe 4*.

Entraînement

Pour tous les modèles

Le partage entre l'échantillon d'entraînement et de test est construit à partir du mois de naissance, afin de permettre la reproductibilité du partitionnement entre données d'entraînement et de test sur toute autre extraction du SNDS, qui contient toujours cette information. L'échantillon de test est ainsi constitué par les individus nés les mois de janvier, avril, juillet et octobre (soit environ un tiers de l'échantillon), et celui d'entraînement, par ceux nés les autres mois. Ainsi, les individus appartenant à une autre source, **l'échantillon démographique permanent (EDP)**, nés les premiers jours de chaque trimestre, font partie par construction du test et ne sont jamais vu pendant l'entraînement (y compris des *embeddings* statiques).

Cette méthodologie permet d'évaluer les modèles sur l'intégralité de l'échantillon EDP-Santé (appariement de l'EDP et du SNDS), et en particulier de décliner les performances statistiques suivant des caractéristiques indisponibles dans le SNDS, mais présentes dans l'EDP comme le niveau de vie.

Afin d'évaluer le modèle en dehors du domaine temporel d'entraînement, les trajectoires utilisées dans l'entraînement ont un temps de prédiction choisi en excluant la fin de période, qui est laissée à l'évaluation, et qui représente deux fois l'horizon temporel de prédiction plus six mois. Ainsi, pour une prédiction à un an, les temps de prédiction pris au cours de l'entraînement peuvent aller de mi 2018 à mi 2020, et l'horizon de prédiction au cours de l'entraînement peut amener au plus jusqu'à mi 2021. Pour l'évaluation dans l'échantillon de test, il sera possible de distinguer une évaluation dans le domaine temporel (temps de prédiction allant de mi 2018 à mi 2020), mais aussi hors du domaine temporel (temps de prédiction allant de mi 2021 à fin 2021).

Pour les modèles xgboost

Les modèles sont entraînés sur des volumes croissants de patients : 1 million, 10 millions et enfin 30 millions, et évalués sur des échantillons de tests de 30 % la taille de l'échantillon d'entraînement. Les hyperparamètres ont été spécifiquement optimisés pour chaque cible grâce à une optimisation bayésienne (en validant sur le jeu d'entraînement).

Pour le modèle BEHRT

L'échantillon est restreint aux personnes avec des événements sur au moins 5 dates distinctes. L'échantillon d'entraînement représente les 2/3 des bénéficiaires soit environ 50 millions de personnes. Afin de tester le modèle hors de son domaine temporel, les données utilisées pour entraîner le modèle s'arrêtent à mi-juin 2021. On parle d'*epoch* pour chaque passage de l'ensemble des données de ces patients à travers le réseau au cours de l'entraînement. Le modèle est pré-entraîné sur 5 epochs (52 heures par epoch) : il « voit » ainsi 5 fois chaque patient, puis il est *finetuné* sur 2 epochs (25 heures par epoch). L'entraînement de BEHRT suit les principaux choix du modèle original : il est précisément décrit en *annexe 4*, ainsi que les écarts au modèle initial du fait de la spécificité des données (taille du vocabulaire, taille de la population d'entraînement à disposition).

Explicabilité locale

L'enjeu de l'explicabilité locale est de décomposer la prédiction de risque en contributions marginales associés à chacun des événements dans la trajectoire d'une personne. Le risque associé à la personne i se déduit de sa trajectoire via la prédiction, notée $p(S_i)$[14] où S_i représente l'ensemble des événements de la trajectoire. La

[14] En pratique, la prédiction dépend aussi du temps de prédiction, de l'âge et du sexe de l'individu, mais les contributions de ces variables ne sont pas calculées, $p(S_i) = p(t0, a, s, S_i)$

contribution de chacun des événements $j \in S_i$ est calculée, de sorte à obtenir une décomposition du risque en une contribution ϕ_j pour chacun des événements.

$$p(S_i) = \sum_{j \in S_i} \phi_j$$

Une option est d'utiliser les valeurs de Shapley, soit

$$\phi_j = \sum_{Z \subset S_i \setminus \{j\}} \omega(Z) \left(p(Z \cup \{j\}) - p(Z)\right)$$

Où $p(Z \cup \{j\}) - p(Z)$ reflète la modification de la prédiction quand l'événement j s'ajoute à une sous-partie de la trajectoire Z, et la pondération $\omega(Z)$ fait en sorte que chaque sous-partie Z de $S_i \setminus \{j\}$ (l'ensemble des événements de la trajectoire, l'événement j exclu) est représentée de manière équiprobable. Une interprétation est que si l'on tire aléatoirement l'ordre d'arrivée des événements dans la trajectoire, en calculant successivement la prédiction avec de plus en plus d'événements, en moyenne, quand le tour de l'événement j arrive, la prédiction augmente de ϕ_j (ou diminue si ce dernier est négatif). Ainsi, tous les événements de $S_i \setminus \{j\}$ interviennent de manière symétrique dans le calcul, ce qui peut être une hypothèse très forte (par exemple, quand la présence de j est très corrélée à la présence d'un autre code dans la trajectoire, ce qui peut être le cas quand deux médicaments sont souvent prescrits ensemble). De plus, ce calcul nécessite $2^{|S_i|}$ évaluations du modèle (nombre de sous-ensembles de S_i), soit un coût exponentiel en la longueur de la trajectoire, rédhibitoire.

L'alternative adoptée ici se nomme les valeurs de Shapley-Owen, dans une version hiérarchique[15]. Les valeurs de Shapley-Owen permettent d'étudier les interactions entre les événements d'une coalition (ici, un groupe d'événements, par exemple, tous les codes médicaments) en considérant les autres coalitions comme des événements indissociables (tous présents ensemble ou tous absents ensemble). La valeur de Shapley-Owen d'une coalition se répartit en l'ensemble des événements qui la compose. Pour un ensemble de coalitions $\{J\}$ partitionnant l'ensemble des événements,

$$p(S_i) = \sum_J \phi_J \text{ et } \phi_J = \sum_{j \in J} \phi_j \ ,$$

À partir d'un arbre hiérarchique dans lequel sont regroupés les événements en coalitions, les valeurs de Shapley de groupes d'événements (coalitions) représentés par chaque nœud de l'arbre sont calculées successivement en descendant dans l'arbre, jusqu'au niveau de chaque événement. À chaque niveau, il s'agit de répartir la valeur ϕ_J dans le niveau inférieur.

Si l'événement j appartient à la coalition C (par exemple les codes médicaments), et en notant les coalitions restantes $\bar{J} \backslash C$ (par exemple, les codes diagnostics, les codes d'actes médicaux, etc.), la valeur de Shapley-Owen s'écrit :

$$\phi_j^O = \sum_{J \subset \bar{J} \backslash C} \omega(J) \sum_{Z \subset C \setminus \{j\}} \omega(Z) \left(p(J \cup Z \cup \{j\}) - p(J \cup Z)\right)$$

$p(J \cup Z \cup \{j\}) - p(J \cup Z)$ reflète l'apport de j au sein de sa coalition, pour un jeu donné de coalitions restantes, et la pondération $\omega(Z)$ reflète que chaque sous-partie de $C \setminus \{j\}$ (les événements de la coalition de j, l'événement j exclu) est représentée de manière équiprobable, en moyenne sur les possibilités de coalitions J, les coalitions restantes une fois C exclu. Une interprétation est que si l'on tire aléatoirement l'ordre d'arrivée des coalitions dans la trajectoire, puis au sein de chaque coalition, indépendamment, de l'ordre d'arrivée des événements, en calculant successivement la prédiction avec de plus en plus d'événements, en moyenne, quand le tour de l'événement j arrive, la prédiction augmente de ϕ_j. Ainsi, en tirant de manière groupée les événements d'une coalition, on peut rendre compte d'une structure de corrélation pertinente si celle-ci est bien représentée par l'arbre hiérarchique.

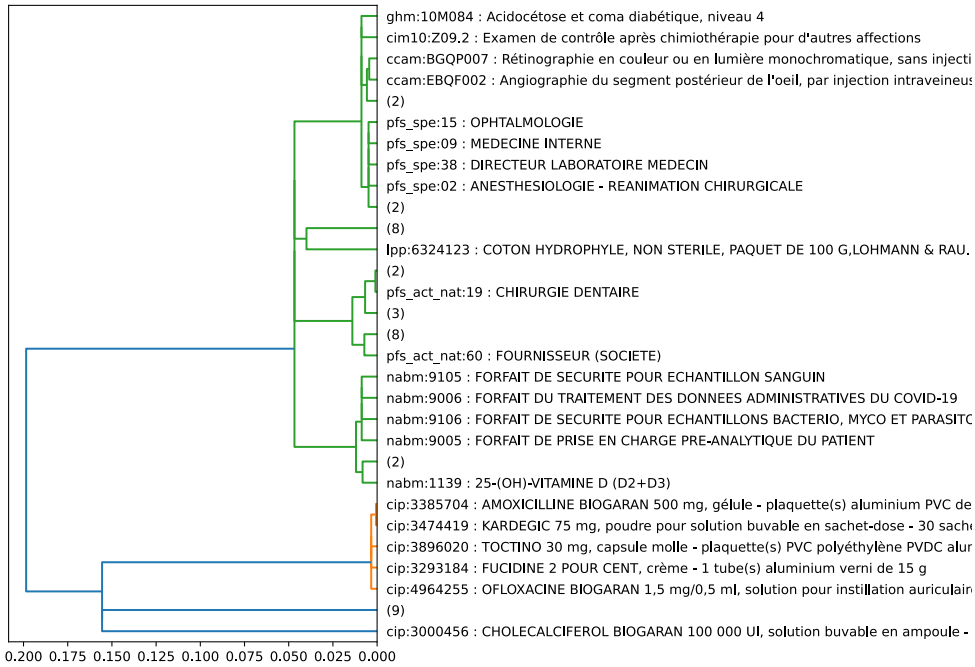
La hiérarchie locale des événements considérée est construite pour être un arbre binaire comprenant l'ensemble des événements de la trajectoire de l'individu, et pour refléter la hiérarchie des nomenclatures utilisées (ATC, CCAM, CIM-10, etc.). Elle est obtenue par classification ascendante hiérarchique (rapprochement deux à deux itératif) en considérant une distance entre deux événements reflétant la proximité dans la hiérarchie (nombre d'arêtes en commun), et en cas d'égalité, la proximité temporelle. Ainsi, les codes appartenant à la même

[15] Obtenues par l'application de la fonction shap.PartitionExplainer — SHAP latest documentation

nomenclature sont regroupés, et d'autant plus classés ensemble qu'ils sont proches dans celle-ci. L'explicabilité obtenue hérite de la structure des nomenclatures utilisées.

L'avantage de cette approche est double, elle permet d'une part d'aligner les coalitions utilisées pour le calcul avec les groupements typiques de données que l'on souhaiterait étudier (e.g. tous les médicaments) et donc d'en avoir une interprétation « avec versus sans ». D'autre part, le gain calculatoire de cette stratégie par rapport aux valeurs de Shapley (quadratique plutôt qu'exponentiel) permet d'obtenir des estimations d'explicabilité locale *pour un individu* en environ 10 secondes.

Schéma 2 Classification ascendante hiérarchique et explicabilité d'une trajectoire individuelle synthétique



Lecture > Une trajectoire individuelle synthétique à laquelle a été appliquée la classification ascendante hiérarchique et les valeurs de Shapley-Owen de l'explicabilité (en valeur absolue) sont représentées. L'explicabilité pour cet individu est donc calculée à chaque nœud d'un arbre ne comprenant que ses propres événements. L'explicabilité est exprimé en part de la prédiction totale, ainsi l'ensemble des événements explique 20 % de la prédiction, le reste étant expliqué par les variables démographiques. L'ensemble des médicaments (CIP) expliquent quant à eux environ 15 % de la prédiction.

Explicabilité globale

L'explicabilité globale est définie, pour chaque groupe d'événements J, comme la moyenne, sur l'ensemble des individus, de la valeur de Shapley-Owen associée à J.

En raison du coût de calcul, l'explicabilité globale du modèle est estimée sur un échantillon d'individus. Cet échantillon est tiré avec une probabilité qui dépend du risque prédit – la quantité que l'on décompose pour l'explicabilité – puis un estimateur d'Horvitz-Thompson des valeurs de Shapley-Owen est utilisé pour obtenir une estimation représentative de la moyenne en population générale des valeurs de Shapley-Owen. Les explicabilités sont présentées sur des échantillons de 1 000 individus. Les erreurs standards associées à l'échantillonnage indiquent que cette taille d'échantillon est suffisante.

VALEURS PRÉDICTIVES DES DONNÉES DU SYSTÈME NATIONAL DES DONNÉES DE SANTÉ

Choix de la métrique. L'**AUC** (aire sous la courbe ROC) peut être lue comme la capacité du modèle à discriminer un cas d'un témoin, c'est-à-dire d'attribuer une probabilité plus élevée au cas qu'au témoin. Une AUC à 0,5 correspond à un classement aléatoire, tandis qu'une valeur proche de 1 indique une très bonne performance du modèle. L'AUC peut être calculée en population générale, ou conditionnellement à certaines caractéristiques comme la classe d'âge et le sexe, variables permettant déjà de fortement discriminer deux individus concernant leur état de santé probable. L'**AUC stratifiée** ainsi obtenue quantifie le pouvoir discriminant supplémentaire par rapport à ces variables. Une AUC stratifiée par la classe d'âge et le sexe peut être interprétée comme la capacité à discriminer un cas d'un témoin parmi deux individus de la même classe démographique.

L'AUC stratifiée est calculée ici au sein de strates d'âge et de sexe, puis agrégée en pondérant chaque strate par le nombre de cas qu'elle contient. Cette pondération présente deux avantages principaux. D'une part, elle permet d'ajuster l'évaluation des performances à la structure du risque observée dans la population : les strates les plus contributrices en termes de cas influencent davantage la mesure globale, ce qui reflète mieux la performance « utile » du modèle en pratique. D'autre part, elle offre une forme de comparabilité entre les différentes cibles, en neutralisant partiellement les biais introduits par des répartitions démographiques hétérogènes. Ainsi, les différences d'AUC stratifiées peuvent être interprétées plus directement comme des différences de performance intrinsèque d'une cible à l'autre, plutôt que comme le reflet de structures de population très contrastées entre cibles.

Prédire la mortalité à partir du système national des données de santé

Nous avons évalué les performances prédictives d'un modèle de xgboost pour estimer le risque de mortalité à différents horizons temporels, allant de 3 mois à 2 ans. Pour cela, nous avons entraîné le modèle à partir de plusieurs types de représentations des données, correspondant à différentes manières de représenter les informations contenues dans les bases médico-administratives (*tableau 5*). Ces modèles sont comparés aux performances atteintes avec l'approche BEHRT-SNDS.

Tableau 5 Différence entre l'AUC globale et l'AUC stratifiée pour la prédiction de la mortalité à 1 an

Représentations des données	AUC globale	AUC stratifiée
BEHRT-SNDS	0,961	0,822
Plongements vectoriels des trajectoires (Snds2vec)	0,958	0,807
Comorbidités et fréquences de consommation	0,953	0,784
Cartographie et fréquences de consommation	0,953	0,781
Fréquences de consommation	0,948	0,760
Comorbidités (Charlson et Rx-Risk-V)	0,949	0,760
Cartographie	0,942	0,729
Mortality Related Morbidity Index (MRMI)	0,913	0,680
Démographique	0,912	[16]

Note > L'échantillon d'entraînement, à l'exception de BEHRT-SNDS, est de taille 30 millions.
Source > Base SeqNDS, 2018-2022.

Sur la mortalité, on observe des différences notables entre l'AUC globale et l'AUC stratifiée, mettant en évidence l'intérêt de cette seconde métrique. Pour l'ensemble des modèles, les AUC globales sont systématiquement plus élevées (de l'ordre de 0,91 à 0,96), alors que les AUC stratifiées varient davantage (de 0,680 à 0,822), révélant de véritables différences de performance selon les modèles.

La prédiction de la mortalité repose en grande partie sur des facteurs fortement corrélés à l'âge et au sexe. Ainsi, un modèle peut facilement atteindre de bonnes performances globales en exploitant ces corrélations, sans nécessairement bien discriminer au sein de chaque groupe démographique. Autrement dit, une partie de la performance brute reflète avant tout la structure démographique du risque, plutôt que la qualité du signal capté par les représentations des données ou la qualité de la modélisation. C'est pourquoi, dans la suite de l'analyse, nous avons choisi de retenir l'AUC stratifiée comme indicateur principal de comparaison : il permet une évaluation de la capacité des modèles à discriminer le risque au sein de groupes démographiquement comparables.

[16] En ce qui concerne le modèle démographique, l'AUC stratifiée est très proche de 0,5 : en effet, la seule information intra-classe démographique (tranche d'âge, sexe) que le modèle peut utiliser pour discriminer deux individus est l'âge exact au sein de sa tranche d'âge.

Performances selon la représentation des données en entrée

Concernant les types de variables utilisées à modélisation xgboost constante, les modèles les moins performants sont ceux qui se basent uniquement sur les variables présentes dans la cartographie des pathologies (AUC stratifiée de 0,73 à 1 an). Il faut néanmoins noter que l'information utilisée dans ce modèle est par nature moins à jour, les variables de la cartographie étant issues d'algorithme visant à caractériser l'année précédent t_0, ce qui peut générer jusqu'à un an de retard par rapport à l'information utilisée par les autres modèles. Les performances sont meilleures pour le modèle comorbidités (AUC stratifiée de 0,76 à 1 an). Ce niveau de performance est globalement comparable à celui obtenu avec un autre type de représentation simple comme représentation des données : les fréquences de consommation de soins, dans les 14 postes de notre typologie, qui comprend par exemple la fréquence des diagnostics hospitaliers, des actes médicaux, ou des remboursements de médicaments (AUC stratifiée de 0,76 de 1 an). Lorsque l'on combine ces fréquences de consommation avec les variables de la cartographie (AUC stratifiée de 0,78 de 1 an) ou avec celles des scores de comorbidité (AUC stratifiée de 0,78 à 1 an), les performances du modèle s'améliorent. Cependant, ce sont toujours des améliorations modestes, suggérant que ces représentations, bien que plus riches, restent limitées pour capturer la complexité des trajectoires de soins. Le modèle le plus performant, et de manière systématique quel que soit l'horizon temporel, est celui qui utilise comme représentation des données les plongements vectoriels de trajectoire, PVT (AUC stratifiée de 0,81 à 1 an).

Le modèle BEHRT-SNDS utilise en entrée des données quasiment équivalentes au modèle PVT. Il diffère néanmoins nettement par la modélisation et l'entraînement. Il est entraîné sur l'ensemble de l'échantillon d'entraînement (50 millions de patients), ce qui représentait une difficulté pour les autres modèles dans notre contexte. C'est le modèle qui atteint les meilleures performances, avec 1,5 point d'AUC stratifiée de plus que le deuxième meilleur modèle.

À titre indicatif, Life2Vec rapporte un *Corrected Matthews Correlation Coefficient* (C-MCC[17]) de 0,41 en seuillant le score de risque à 0,5[18] et une *Area Under the Lift* (AUL[19]) de 0,85 pour la prédiction de la mortalité à 4 ans. En nous restreignant aux individus de 35 à 65 ans (critère d'inclusion de la cohorte de *Life2Vec*), nous relevons pour BEHRT-SNDS à horizon 1 an un MCC de 0,28 en seuillant le score de risque à 0,5, de 0,38 en seuillant optimalement à 0,15, et une AUL de 0,91. Cependant, les sources de données, la profondeur historique et l'horizon temporel étant différents, il ne s'agit pas d'une comparaison directe, mais d'une mise en perspective.

Performances selon la taille de l'échantillon

Les performances des modèles pour la prédiction de la mortalité augmentent de manière importante lorsque la taille de l'échantillon utilisé pour l'entraînement passe de 1 million à 10 millions d'individus (+1,26 point d'AUC stratifiée en moyenne sur les différents modèles pour un horizon 1 an). Ce gain reste observable, mais devient plus modéré, lorsqu'on passe de 10 millions à 30 millions d'individus (+0,25 point). Cela suggère qu'une taille d'échantillon importante améliore la capacité du modèle à capturer des motifs rares ou complexes, mais qu'un effet de plateau peut commencer à apparaître au-delà d'un certain seuil.

Tous les modèles gagnent en performance lors de l'augmentation de la taille de l'échantillon, mais c'est le modèle le plus complexe, PVT, qui en bénéficie le plus, avec une augmentation de 2,03 points d'AUC stratifiée entre 1 million et 30 millions d'individus, contre 1,51 point en moyenne pour les autres modèles.

Compte tenu de sa durée d'entraînement (environ 12 jours), le modèle BEHRT-SNDS, pré-entrainé (5 epochs), puis *fine-tuné* (2 epochs) sur 50 millions d'individus, n'a pas fait l'objet d'expérimentation pour varier la taille de l'échantillon. Néanmoins une partie de sa supériorité peut être attribuée à sa capacité à avoir vu au cours de l'entraînement l'ensemble des séquences des 50 millions d'individus de l'échantillon d'apprentissage.

Performances selon l'horizon

Le modèle se basant sur les PVT permet donc de gagner en performance de prédiction par rapport à ceux utilisant les autres représentations des données. Sur un entraînement sur 30 millions d'individus, PVT permet de gagner entre 6 points d'AUC stratifiée à un horizon 2 ans et 11 points à 3 mois par rapport aux variables de la cartographie

---

[17] Le *Matthews Correlation Coefficient* (MCC) est une métrique basée sur la matrice de confusion. Le MCC vaut 0 pour une prédiction aléatoire et 1 pour une prédiction parfaite. Dans Life2Vec (Savcisens*, et al*., 2024), les données étant partiellement non labellisées, les auteurs ajoutent une correction qui n'est pas nécessaire dans notre cas.

[18] Dans le cas où la prédiction est un score de risque, l'usage est de seuiller ce score à un seuil, les prédictions supérieures étant associées à 1 et inférieures à 0.

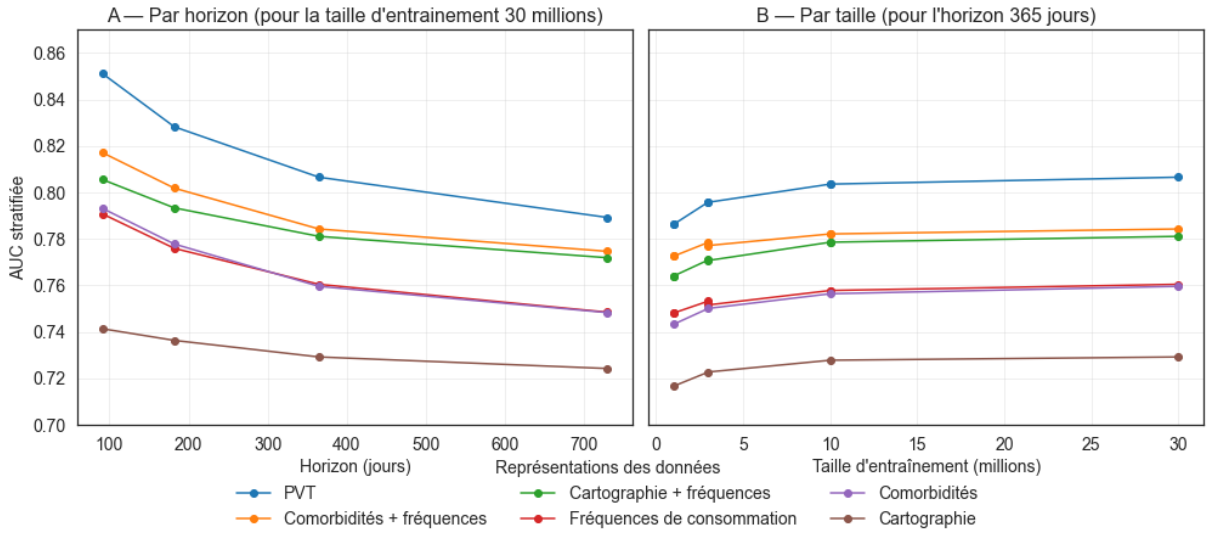
[19] L'AUL est une variante de l'AUC globale, plus adaptée au cas partiellement labellisé. Elle se déduit de l'AUC et de la prévalence : $AUL = AUC - prevalence \times (AUC - 0{,}5)$.

des pathologies. Le gain est moindre lorsque les variables des scores de comorbidités sont cumulées aux fréquences de consommation (modèle le plus performant après PVT), avec +1,8 point à 2 ans et +3,5 points à 3 mois.

Les écarts de performance entre modèles se réduisent à mesure que l'horizon s'allonge. Le modèle PVT atteint une AUC stratifiée de 0,85 à 3 mois, qui diminue progressivement jusqu'à 0,79 à 2 ans. Ce déclin est attendu : les événements proches du point de départ de la prédiction (t_0) sont plus prédictibles, et il devient plus difficile d'anticiper à moyen terme.

En somme, la mortalité est prédite avec une très bonne performance, supérieure à celle obtenue avec les variables contenues dans les indices de comorbidités classiques. Ces résultats suggèrent que les cooccurrences d'événements et leur structuration temporelle apportent une information supplémentaire utile pour estimer le risque à court et moyen terme.

Graphique 3 AUC stratifiées sur l'âge et le sexe, des modèles xgboost selon les représentations des données, la taille de l'échantillon d'entraînement, et l'horizon de prédiction de la mortalité



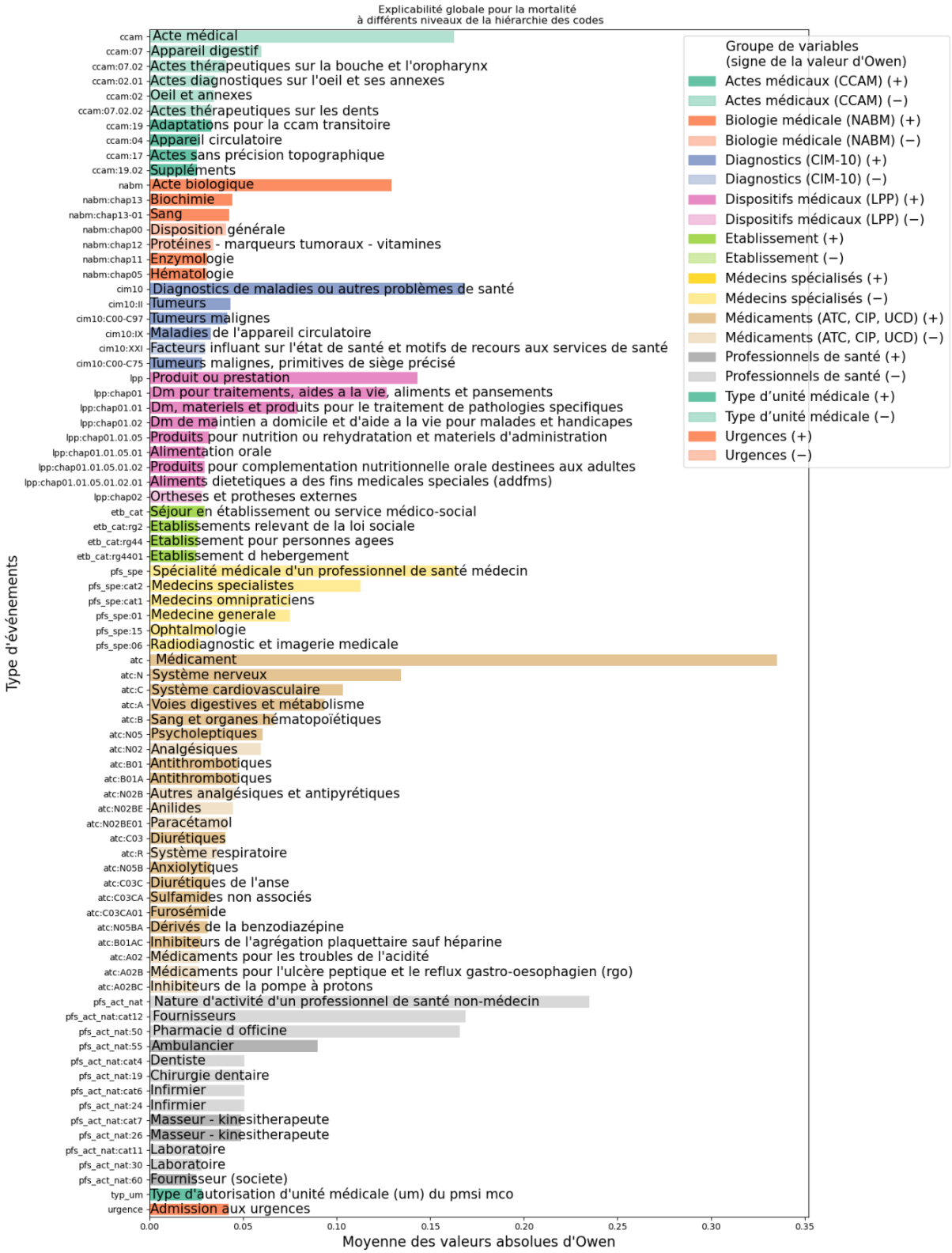
Contribution des événements à la prédiction de la mortalité

L'explicabilité du modèle « PVT » pour la prédiction de la mortalité illustre l'importance des événements dans la trajectoire. La moyenne des valeurs absolues de Shapley-Owen pour le poste médicaments représente 34 % de la prédiction, et la somme des valeurs de Shapley-Owen pour ce poste étant positives, ces événements contribuent dans leur ensemble à accroître le risque estimé de mortalité bien que certains, comme le paracétamol, contribuent négativement (*graphique 4*). Les diagnostics hospitaliers, au premier rang desquels les tumeurs et les maladies cardiovasculaires, contribuent positivement au risque de mortalité estimé, la moyenne des valeurs absolues de Shapley-Owen représentant 17 % de la prédiction. Les natures d'activités de professionnels de santé intervenant, contribuent fortement à la prédiction (moyenne des valeurs absolues de Shapley-Owen de 23 % pour les non-médecins et 16 % pour les spécialités médicales), mais négativement dans leur ensemble, avec quelques exceptions (ambulancier). Il est probable que ces contributions négatives d'une grande partie des soins de ville reflètent la prise en charge plus volontiers hospitalière des personnes à fort risque de mortalité (les deux tiers des décès ayant lieu en établissement de santé[20], comme la prise en charge des patients les plus graves). Contribuent positivement les actes de biologie, les dispositifs médicaux (produits ou prestations), les admissions aux urgences ainsi que les types d'unités médicales. Bien qu'il faille se garder d'une interprétation causale de ces résultats, ils illustrent néanmoins que l'ensemble de la trajectoire intervient bien dans la prédiction, y compris avec des contributions notables de certains codes à des niveaux fins, par exemple avec un accroissement du risque estimé de l'ordre de 3 % pour l'occurrence d'une délivrance de furosémide[21] (atc : C03CA01), ou des contributions négatives qui signent que certains événements sont en défaveur de l'hypothèse de décès pour le modèle.

[20] Études et résultats Drees : Causes de décès en France en 2023 : des disparités territoriales

[21] Le furosémide est un diurétique de l'anse (il agit au niveau de l'anse de Henle dans le rein, où il bloque la réabsorption du sodium et de l'eau) utilisé pour traiter les situations de surcharge hydrosodée (accumulation excessive d'eau et de sel dans l'organisme), notamment liées à une insuffisance cardiaque, rénale ou hépatique. Sa prescription reflète souvent une maladie chronique avancée ou décompensée, ce qui en fait un indicateur indirect de gravité.

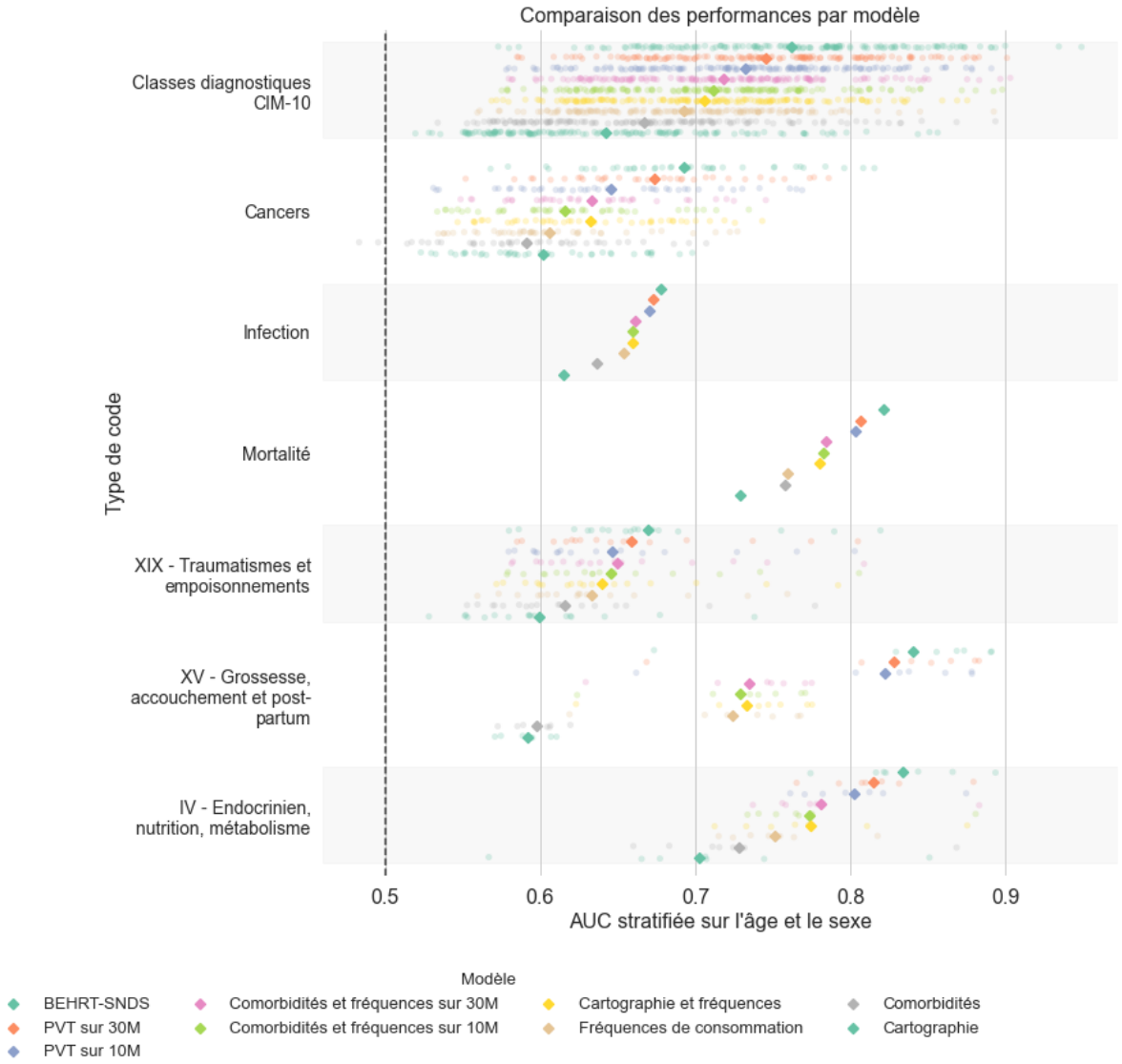
Graphique 4 **Explicabilité globale pour la prédiction du risque de mortalité, à différents niveaux de la hiérarchie des codes d'événements**



Prédire la première hospitalisation en lien avec une pathologie

Les résultats sont à nouveau comparés à l'aide de l'AUC stratifiée selon l'âge et le sexe et pondérée par le nombre de cas positifs pour ce diagnostic (*graphique 5*).

Graphique 5 AUC stratifiées sur l'âge et le sexe des prédictions des différents types de cible à 1 an, selon la représentation des données et la modélisation



Note > L'échantillon d'entraînement, à l'exception de BEHRT-SNDS et sauf indication contraire, est de taille 10 millions.

Entraînés sur 10 millions d'individus pour un horizon de prédiction de 1 an, les modèles basés sur le PVT ont de meilleures performances quel que soit le type de cible (*tableau 6* : AUC stratifiées moyennes : 0,73 pour les 147 classes CIM-10, 0,65 pour les cancers, 0,67 pour les infections). On retrouve le même ordre de performance des modèles selon leur représentation des données que pour la mortalité. Les cancers sont en moyenne moins bien prédits, et avec moins de variabilité étant donné leur plus faible nombre (33). À titre indicatif, Shmatko, *et al*. (2025), obtiennent, pour la prédiction (non incidente) de la prochaine pathologie (1 000 diagnostics CIM-10 collectés en ville comme à l'hôpital), une AUC stratifiée[22] moyenne de 0,76 chez les participants volontaires au sein de UK Biobank, âgés d'au moins 40 ans à l'inclusion ; là encore, les données, les tâches de prédiction et les métriques d'évaluation étant sensiblement différentes, il ne s'agit pas d'une comparaison directe, mais d'une mise en perspective. Ce classement des performances se retrouve quasi systématiquement par chapitre de la CIM-10 (20 chapitres sur 21, la seule exception étant les affections périnatales par définition mal représentées par une trajectoire de soins passée). L'écart est le plus marqué dans un cas très particulier : les hospitalisations en lien avec une grossesse ou un accouchement (gain de 10 points entre PVT et « cartographie et fréquences »). Sans surprise, le parcours de soins d'une femme enceinte est particulièrement informatif sur son futur accouchement, et savoir si le modèle est capable de distinguer plusieurs issues de grossesse parmi des femmes enceintes est une question autrement plus intéressante qui sera abordée dans la deuxième famille de cibles. Ce cas mis à part, les résultats sont informatifs sur la plus ou moins grande capacité des indices de comorbidités ou des algorithmes classiques à

[22] Sans pondérer par le nombre de cas de chaque strate, mais en se restreignant aux personnes entre 50 et 80 ans et aux strates d'âge quinquennale et de sexe avec au moins deux cas.

mobiliser toute l'information disponible. Si dans l'ensemble du chapitre « Endocrinien, nutrition, métabolisme », le modèle PVT est en moyenne plusieurs points d'AUC stratifiée au-dessus des autres, ce n'est pas le cas des hospitalisations pour diabète (classe E10-E14). Dans ce chapitre, l'AUC stratifiée du modèle « cartographie », incluant un top pour repérer les personnes diabétiques et diabétiques insulino-traités (AUC stratifiée : 0,85), une fois ajouté aux fréquences de consommation, atteint une AUC stratifiée équivalente à celui du modèle PVT (0,88) pour prédire leur première hospitalisation. Néanmoins, ce cas est loin d'être majoritaire. Par exemple, pour des pathologies comme la maladie de parkinson, aussi repérée par la cartographie, le modèle PVT (AUC stratifiée : 0,86) dépasse nettement le modèle « cartographie » (AUC stratifiée : 0,79), y compris quand on y ajoute les fréquences de consommation (AUC stratifiée : 0,84). Il pourrait être intéressant de nourrir les réflexions « expertes » pour réviser les algorithmes de repérage de la cartographie des pathologies au regard des événements qui ressortiraient comme déterminants pour la prédiction (explicabilité) sans être inclus dans l'algorithme déterministe, pour les cas où on admet qu'un repérage prédictif de l'hospitalisation est pertinent.

Tableau 6 AUC stratifiées moyennes sur l'âge et le sexe des prédictions des différents types de cible à 1 an, selon les représentations des données et la modélisation

Représentations des données	Classes CIM-10	Cancers	Infections
Cartographie des pathologies	0,642	0,593	0,615
Comorbidités (Charlson et Rx-Risk-V)	0,667	0,582	0,636
Fréquences de consommation	0,692	0,607	0,653
Cartographie et fréquences de consommation	0,706	0,627	0,659
Comorbidités et fréquences de consommation	0,711	0,618	0,659
Plongements vectoriels des trajectoires (Snds2vec), entraînement à 10 millions	0,732	0,653	0,670
Plongements vectoriels des trajectoires (Snds2vec), entraînement à 30 millions	0,746	0,673	0,673
BEHRT-SNDS	**0,765**	**0,706**	**0,678**

Note > L'échantillon d'entraînement, à l'exception de BEHRT-SNDS et sauf indication contraire, est de taille 10 millions.
Source > Base SeqNDS, 2018-2022.

Les chapitres où le modèle PVT est le plus performant, qu'on peut donc considérer comme les chapitres à la prédictibilité la plus forte sont les suivants : les troubles mentaux et du comportement (0,79), les chapitres liés au sang et à l'immunité (0,78), à l'endocrinologie, nutrition et métabolisme (0,80), au système nerveux (0,76), à l'appareil respiratoire (0,78) ou encore à l'appareil génito-urinaire (0,73). Ainsi, ce sont les parcours menant à des hospitalisations dans ces chapitres qui sont le mieux capturés par les données du SNDS.

Lorsque le modèle PVT est entraîné sur un échantillon élargi à 30 millions d'individus, ses performances augmentent encore légèrement, confirmant le gain lié à l'apprentissage sur de grands volumes de trajectoires ; cette tendance se retrouve également avec le modèle BEHRT-SNDS, qui obtient les AUC stratifiées les plus élevées sur l'ensemble des cibles.

Concernant les cancers

Lorsque l'on compare les modèles entraînés sur un échantillon de 10 millions d'individus, on remarque que les meilleures performances sont observées pour les cancers hématologiques (AUC stratifiée moyenne : 0,72) ainsi que pour les cancers du foie (0,70) et du poumon (0,68). À l'inverse, certains cancers sont très mal prédits, comme les cancers thoraciques (0,54) ou les cancers génitaux masculins (0,54). Les modèles basés sur les PVT se distinguent par leurs performances globalement supérieures ou égales aux autres modèles (AUC stratifiée moyenne : 0,65, voir *tableau 5*). À l'inverse, les modèles basés uniquement sur les scores de comorbidité montrent des performances plus faibles. Ces résultats suggèrent que les cancers les plus suivis en ville (comme les hémopathies malignes ou les cancers de la prostate) sont plus facilement prédits par les modèles, tandis que d'autres, plus rares, graves ou soudains (comme les cancers génitaux masculins ou du système nerveux), échappent en grande partie à la détection, même avec des méthodes avancées.

Lorsque le modèle PVT est entraîné sur 30 millions d'individus, ses performances augmentent en moyenne de 1,6 point d'AUC stratifiée, un gain particulièrement marqué pour les cancers rares ou difficilement prédits à 10 millions. Par exemple, pour les cancers endocriniens, l'AUC stratifiée passe de 0,61 à 0,76. Le modèle de comorbidité associé aux fréquences de consommation bénéficie aussi de l'élargissement d'échantillon (+0,8 point en moyenne), mais de façon plus homogène et sans amélioration marquée pour certaines cibles spécifiques. Par exemple, pour les cancers du système nerveux, le PVT gagne près de 9 points (0,59 à 0,68), quand le modèle de comorbidité n'en gagne qu'environ 2,5 (0,57 à 0,60). Enfin, BEHRT-SNDS surpasse nettement l'ensemble des modèles, en particulier pour les cibles les moins bien prédites par les autres (AUC stratifiée 0,68 pour les cancers thoraciques et 0,78 pour les cancers du système nerveux), tout en conservant un avantage plus modéré lorsque les autres modèles performent déjà bien.

Graphique 6 AUC stratifiées sur l'âge et le sexe des prédictions des cancers à 1 an selon les représentations des données et la modélisation

Note > L'échantillon d'entraînement, à l'exception de BEHRT-SNDS et sauf indication contraire, est de taille 10 millions.
Source > Base SeqNDS, 2018-2022.

Concernant les infections

La cible « Infection » regroupe un grand nombre de codes CIM-10 couvrant des pathologies infectieuses très variées. Cette diversité se retrouve dans la variété des tableaux cliniques et des trajectoires de soins associées, avec des parcours patients, des profils cliniques et des facteurs de risque parfois très différents. Elle constitue également la cible comptant le plus de cas positifs (n = 161 187 pour un échantillon de 10 millions d'individus). Cette hétérogénéité peut contribuer à atténuer les performances des modèles, dont les AUC stratifiées des modèles entraînés sur 10 millions d'individus s'échelonnent de 0,62 (modèle basé sur la cartographie des pathologies seule) à 0,67 (PVT). Ces résultats suggèrent que, même avec des jeux de variables d'entrée riches, la prédiction devient plus complexe lorsque les cas regroupés dans une cible suivent des parcours trop dissemblables.

Variabilité des performances d'une cible à l'autre

Tableau 7 AUC stratifiées sur l'âge et le sexe des prédictions de cibles illustratives à 1 an, selon les représentations des données et la modélisation

	Cartographie et fréquences	Cartographie	Comorbidités et fréquences	Comorbidités	Fréquences	Plongements vectoriels	BEHRT-SNDS
Maladies de l'appareil digestif – Appendice	**0,579**	0,530	**0,579**	0,557	0,574	0,578	0,573
Traumatismes – Coude et avant-bras	0,572	0,552	0,579	0,559	0,559	**0,580**	**0,580**
Tumeurs bénignes	0,645	0,562	0,645	0,595	0,635	0,679	**0,711**
Affections épisodiques et paroxystiques	0,712	0,654	0,715	0,685	0,697	0,726	**0,740**
Affections des voies respiratoires supérieures	0,715	0,614	0,731	0,698	0,711	**0,786**	0,694
Grippe et pneumopathie	0,732	0,703	0,732	0,711	0,715	0,738	**0,746**
Sclérose en plaques et autres maladies démyélinisantes	0,808	0,770	0,782	0,672	0,758	0,817	**0,842**
Glaucome	0,787	0,552	**0,898**	0,894	0,777	0,891	0,874
Troubles extrapyramidaux et troubles de la motricité (dont Maladie de Parkinson)	0,837	0,785	0,841	0,804	0,775	0,859	**0,900**
Insuffisance rénale	0,818	0,770	0,845	0,832	0,790	0,849	**0,871**
Diabète sucré	0,875	0,851	0,881	0,868	0,818	0,875	**0,894**

Note > L'échantillon d'entraînement, à l'exception de BEHRT-SNDS, est de taille 10 millions.
Source > Base SeqNDS, 2018-2022.

Les mauvais scores de prédiction concernent principalement des événements aigus, brutaux, souvent peu précédés de signes cliniques visibles dans le SNDS. C'est le cas de pathologies très subites sans parcours préalable (ex. : appendicite se retrouvant dans « Maladies de l'appareil digestif – Appendice ») ou de traumatismes relevant de l'accident (ex. : « Traumatismes, Coude et avant-bras »).

Les performances moyennes concernent un ensemble hétérogène de situations où le signal semble partiel. On y retrouve des pathologies aiguës se greffant sur un terrain connu (ex. : les infections « simples » qui, si résultant en une hospitalisation, signent un terrain fragilisé, comme les « Affections des voies respiratoires supérieures »), ou des classes trop larges où les sous-types sont trop hétérogènes (ex. : « Tumeurs bénignes », « Affections épisodiques et paroxystiques », comprenant à la fois les épilepsies, les migraines et les troubles du sommeil). S'y ajoutent les codes correspondant à des prises en charge chirurgicales programmées, ou à des complications liées aux soins, qui peuvent être identifiables par certains actes ou consultations antérieures, mais restent souvent non spécifiques.

Enfin, les meilleures performances semblent concerner des hospitalisations précédées d'une prise en charge longue et spécifique en ville, combinant biologie, imagerie et traitements ciblés. Ces parcours laissent une trace claire dans le SNDS, notamment lorsqu'ils concernent des pathologies chroniques bien codées (ex. : « Insuffisance rénale », « Diabète », maladies endocriniennes spécifiques). Les modèles tirent aussi profit des cas où le terrain est fortement contributif à la probabilité d'hospitalisation : patients polypathologiques, ou profils avec historique de complications récurrentes (ex : « Grippe et pneumopathie »).

Dans certains cas, on peut poser l'hypothèse qu'une prise en charge en amont, même si elle comporte peu d'événements, serait prédictive si elle implique des examens ou traitements très spécifiques et quasi pathognomoniques. Le cas du glaucome est un bon exemple (traité par collyres bêta-bloqueurs), où la performance atteinte par le

modèle « comorbidités » est très bonne. L'explicabilité obtenue via les valeurs de Shapley-Owen permet de confirmer certaines de ces intuitions : les médicaments anti glaucomateux et myotiques, bêtabloquant (ATC S01ED), traitement spécifique, sont la variable avec la plus forte contribution concernant la première hospitalisation pour glaucome.

En résumé, les bonnes performances (notamment PVT et BEHRT-SNDS sur les cibles les plus prédictibles) semblent reposer sur trois facteurs clés : l'existence d'un signal longitudinal (fréquence et continuité des soins), la spécificité des actes ou traitements, et la cohérence avec un terrain identifiable. Plus ces éléments convergent, meilleure est la performance du modèle. À l'inverse, les événements aigus, les pathologies sans antériorité visible ou les regroupements trop larges limitent fortement la capacité de prédiction.

Variabilité des performances selon les cibles et les représentations des données

Tableau 8 AUC stratifiées sur l'âge et le sexe des prédictions de cibles illustratives à 1 an, selon les représentations des données et la modélisation

	Cartographie et fréquences	Cartographie	Comorbidités et fréquences	Comorbidités	Fréquences	Plongements vectoriels	BEHRT-SNDS
Dermatite et eczéma	0,695	0,628	0,702	0,671	0,696	0,722	**0,742**
Malformations – Appareil génital	0,640	0,582	0,646	0,572	0,647	0,689	**0,767**
Hypertension	0,714	0,683	0,735	0,723	0,673	0,719	**0,738**
Maladies de l'appareil digestif – Foie	0,772	0,730	0,767	0,739	0,731	0,796	**0,827**
Maladies systémiques	0,738	0,677	0,720	0,691	0,697	0,742	**0,762**
Insuffisance rénale	0,818	0,770	0,845	0,832	0,790	0,849	**0,871**
Autres affections dégénératives du système nerveux	0,744	0,704	0,764	0,733	0,693	0,773	**0,834**

Note > L'échantillon d'entraînement, à l'exception de BEHRT-SNDS, est de taille 10 millions.
Source > Base SeqNDS, 2018-2022.

Les performances des modèles varient donc selon la cible considérée, et certaines représentations des données d'entrée semblent mieux adaptées à certaines situations cliniques. On observe ainsi que certains types de données permettent une meilleure prédiction pour certaines pathologies. Pour autant, une hiérarchie assez stable se dessine, quelle que soit la cible. Cela suggère à la fois une forme de spécialisation contextuelle, mais aussi une supériorité générale de certains types de variables en termes de pouvoir prédictif. Le PVT est le plus performant lorsqu'il capte des trajectoires structurées : suivi en ville, examens ou traitements spécifiques, ou hospitalisation faisant suite à un événement récent (chirurgie, complication), et il est par ailleurs le seul modèle (en dehors de BEHRT-SNDS) à inclure l'ensemble des informations de la trajectoire. Le modèle « Comorbidités et fréquences de consommation » fonctionne bien pour les pathologies fréquentes aux traitements bien codés, comme l'insuffisance rénale (AUC stratifiée : 0,85) ou la démence (0,77), où le PVT ne fait que légèrement mieux, ou l'hypertension artérielle (HTA, 0,74), où le PVT fait même moins bien. Cartographie et fréquences de consommation sont utiles pour des pathologies associées à de nombreuses hospitalisations, via un historique hospitalier ou une ALD, comme les maladies du foie (0,77). Le modèle « Fréquences de consommation » agit comme proxy d'une pathologie « active » ou en cours d'exploration, quand les actes sont peu spécifiques (antalgiques, corticoïdes, biologie simple), comme pour les dermatites (0,70) ou maladies systémiques (0,70). Lorsque les variables de la cartographie ou des scores Charlson et Rx-Risk-V ne sont pas accompagnées des fréquences de consommation en entrée des modèles, ceux-ci restent les moins informatifs.

À noter que le modèle BEHRT-SNDS surpasse tous les autres dans plus de 90 % des cibles évaluées. Cette supériorité est particulièrement marquée pour les pathologies rares : par exemple, pour les malformations de l'appareil génital, BEHRT-SNDS atteint une AUC stratifiée de 0,77, contre 0,70 pour le meilleur des autres modèles, soit un écart de 7 points.

Il faut rappeler que les diagnostics analysés ici sont hospitaliers : ils reflètent non pas la pathologie en général, mais sa forme grave. L'hospitalisation pour une pathologie bénigne n'intervient que sur un terrain particulier ou en cas de complication. Ainsi, la prédiction porte moins sur le diagnostic isolé que sur la probabilité qu'il se manifeste sous une forme clinique sévère chez un individu donné. Afin de mieux distinguer les formes aiguës ou complexes des prises en charge chroniques ou programmées, une analyse complémentaire a été réalisée en excluant les séjours hospitaliers sans nuitée de la cible de prédiction. Cette restriction permet d'éliminer les hospitalisations en hôpital de jour, souvent associées à un suivi récurrent ou à des bilans de pathologies chroniques stabilisées, ainsi que les chirurgies ambulatoires planifiées. En théorie, la cible ainsi redéfinie se recentre sur des épisodes plus graves, correspondant majoritairement à des situations aiguës ou à des décompensations de pathologies

chroniques. L'évaluation de la performance des modèles PVT par rapport aux autres n'est que modérément modifiée par ce recentrage sur les hospitalisations plus graves.[23]

En conclusion, il se dessine deux enjeux distincts amenant à redéfinir les cibles de prédiction :

- la prédiction d'un véritable diagnostic incident, qu'on ne connaissait pas en ville, pour permettre de mieux le prévenir (prévention primaire) ou de réduire le délai diagnostique ;
- la prédiction des hospitalisations pour une pathologie donnée, déjà suivie en ville, pour permettre d'en prévenir les complications (prévention secondaire).

Préciser les cibles de prédiction dans un objectif de santé publique

Pour explorer ces deux enjeux, nous avons retenu trois dimensions :

- *L'incidence de cancers.* Nous avons travaillé sur la prédiction des diagnostics incidents de cancers, en appliquant des filtres stricts (masque des événements précédents de peu t_0, absence de cancer antérieur, absence de chimiothérapie antérieure) afin de garantir qu'il s'agissait bien d'un véritable diagnostic nouveau, et de s'affranchir des événements très annonciateurs juste avant l'hospitalisation (bilan pré-opératoire, etc.).
- *Les grands enjeux de santé publique.* Nous avons étudié des maladies chroniques à fort impact en population générale : le diabète, l'insuffisance cardiaque, l'insuffisance rénale, et les maladies respiratoires, en nous concentrant sur la prédiction des hospitalisations chez les personnes déjà traitées, donc sur le risque de complications.
- *Les pathologies aux performances prometteuses.* Enfin, nous avons sélectionné certaines pathologies pour lesquelles les performances de prédiction semblaient particulièrement pertinentes ou prometteuses lors du premier criblage large de 182 cibles. Elles ont été retenues pour leur intérêt en termes de prévention de leur incidence, de leurs complications, ou de leur retard diagnostique.

Incidences des cancers

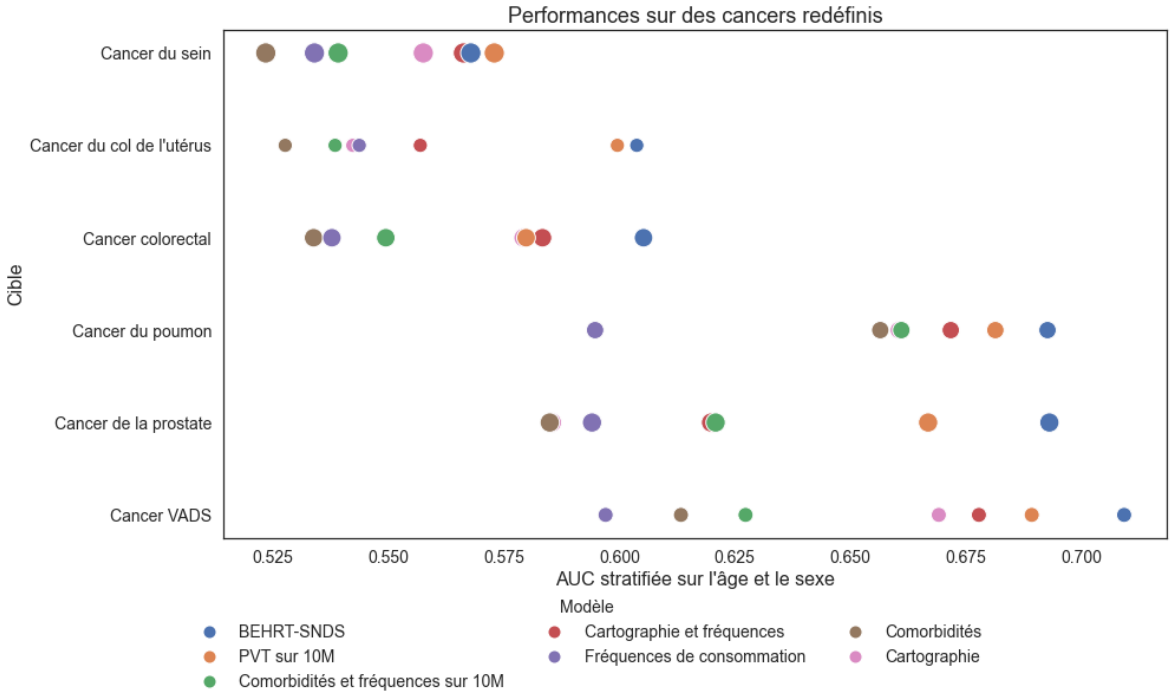
Ici, il s'agit de prédire un diagnostic lorsque celui-ci est formalisé lors de la première hospitalisation pour ce diagnostic principal (DP), ce qui en fait un proxy de l'incidence réelle, observable dans le SNDS, pour les patients qui ne seraient pas diagnostiqués en ville. L'enjeu devient alors de pouvoir prédire, avant même que ce diagnostic soit posé, que le patient est à haut risque de développer cette pathologie, ouvrant la voie à un dépistage ciblé ou à une action préventive déclenchée par signal prédictif. Au-delà des performances globales, l'explicabilité des modèles permet de comprendre quels facteurs observables dans le SNDS contribuent à la prédiction et d'identifier des hypothèses de leviers d'action.

Un masque de 3 mois a été appliqué avant le temps de prédiction t_0 de façon que les événements proches de l'hospitalisation ne puissent pas la faciliter. Aussi, les individus ayant déjà eu des séjours avec un diagnostic de cancer (quel que soit le cancer, en DP ou en DR), une ALD avec un diagnostic de cancer, des chimiothérapies hospitalières, ou des antinéoplasiques en ville ont été exclus.

Ces filtres d'exclusion constituent un argument fort en faveur du caractère incident du diagnostic capté. Ainsi, on prédit ici une incidence hospitalière. Pour certains cancers majoritairement diagnostiqués ou pris en charge en ville (sein, prostate, col de l'utérus), cela correspond plutôt aux formes plus sévères nécessitant une hospitalisation en première intention. En revanche, pour les autres localisations dont le diagnostic se pose habituellement à l'hôpital, cette approche reste proche de l'incidence réelle. Cette approche reste toutefois limitée par la profondeur temporelle des données (cinq ans) : si des antécédents de cancer ou de traitement ont eu lieu avant le début de cette fenêtre, deux situations sont possibles. D'une part, il peut s'agir d'un cancer déjà présent et en cours mais non observé dans les données disponibles, notamment si le t_0 est proche du début de la fenêtre et qu'aucun nouvel événement n'a eu lieu depuis (un biais que l'on limite en imposant au moins six mois d'historique avant t_0). D'autre part, certains cas peuvent correspondre à des rechutes d'un cancer plus ancien, auquel cas l'épisode hospitalier capté reflète malgré tout un nouvel épisode de prise en charge pertinent pour l'analyse prédictive.

[23] Tandis qu'il perd en moyenne 1,5 point d'AUC stratifiée, le modèle « comorbidités et fréquences de consommation » perd 0,5 point et le modèle « cartographie et fréquences de consommation » 0,8. Partant d'une avance de respectivement 2,1 et 2,6 points en moyenne, il reste donc en tête, avec une avance d'au moins 1,1 point. En pratique, cette restriction a fortement réduit le nombre de cas positifs pour certains diagnostics (40 % en moyenne). Dans les cas où la perte de cas positifs est >75 %, on observe une perte marquée d'AUC stratifiée (>3 points pour 30 cibles), parfois jusqu'à une perte totale de pouvoir discriminant, notamment pour les cibles rares ou peu spécifiques. À l'inverse, certaines cibles gagnent en performance (>1 point d'AUC stratifiée pour 26 cibles) probablement du fait d'une population plus homogène, centrée sur des hospitalisations graves. Ce filtre améliore donc la prédiction dans certains cas, mais au prix d'une perte d'information pertinente dans d'autres, notamment les bilans d'extension en hôpitaux de jour, qui peuvent constituer un proxy utile d'incidence dans le cas des cancers. Ces résultats invitent à une réflexion ciblée sur l'usage différencié de ce filtre selon les types de pathologies.

Graphique 7 AUC stratifiées sur l'âge et le sexe, pour la prédiction des cancers à 1 an selon les représentations des données et la modélisation



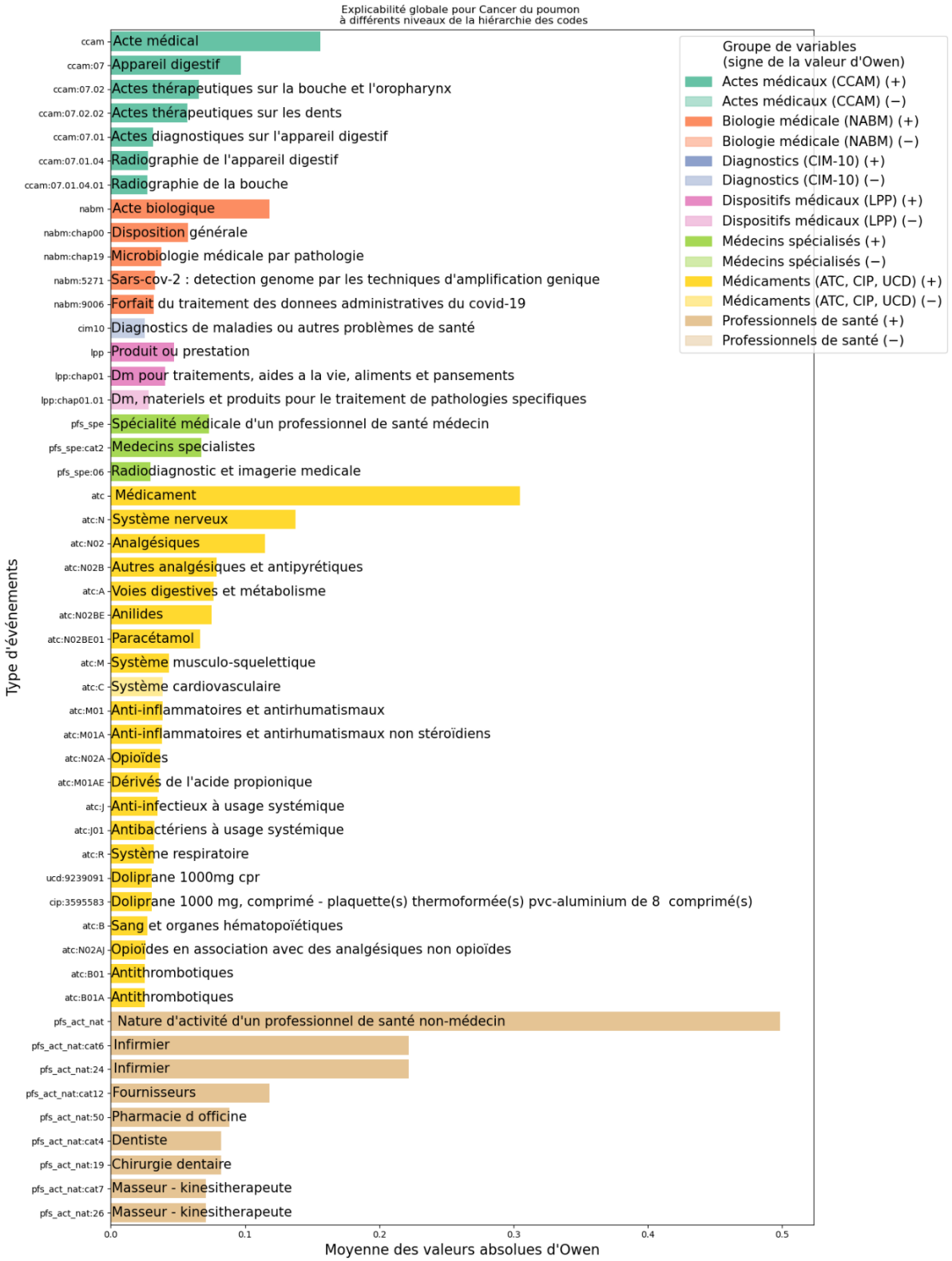
Note > Taille des points proportionnelle au nombre de cas positifs dans l'échantillon de 10 millions d'individus. VADS : voies aérodigestives supérieures.

Sur le *graphique 7*, il est intéressant de noter que les cancers pour lesquels des dépistages organisés sont en place (col de l'utérus, sein, colorectal) sont ceux pour lesquels les modèles de prédiction ont les moins bonnes performances, bien que BEHRT-SNDS apporte un gain notable pour les cancers du col de l'utérus et colorectal. À l'inverse, le cancer du poumon, des voies aérodigestives supérieures (VADS) et de la prostate, qui peuvent plutôt faire l'objet de dépistage à l'échelle individuelle en cas de facteurs de risque, étaient mieux prédits, avec un avantage des modèles PVT et BEHRT-SNDS.

Le cancer du poumon est une bonne illustration. Il est souvent diagnostiqué à l'occasion d'une hospitalisation pour bilan d'extension ou biopsie, précédée d'un scanner thoracique réalisé en ville. Si les performances de prédiction sont correctes (AUC stratifiée de PVT : 0,68, AUC stratifiée de BEHRT-SNDS : 0,69[24]), elles reposent sur des signaux partiels : la comorbidité est bien captée, mais le principal facteur de risque, le tabagisme, n'est pas codé dans le SNDS. Ces patients pourraient alors bénéficier d'un repérage précoce ou d'un parcours de dépistage accéléré, même en l'absence de diagnostic initialement posé. Le même raisonnement s'applique aux cancers des VADS, où le terrain est souvent commun (tabac, alcool). Concernant l'explicabilité de ces modèles, le *graphique 8* montre que les signaux dominants pour le cancer du poumon étaient liés aux médicaments, en particulier les analgésiques, les anti-inflammatoires, les bronchodilatateurs et les antibiotiques. Ces prescriptions traduisent probablement la prise en charge de symptômes respiratoires ou douloureux récents, d'allure infectieuse ou inflammatoire, dans la période précédent le diagnostic. On retrouvait également une contribution notable des actes de recherche du SARS-CoV-2, cohérente avec l'exploration de symptômes respiratoires dans le contexte pandémique ou post-pandémique. Les consultations spécialisées, en particulier pneumologiques, restaient rares dans cette fenêtre temporelle, alors que la majorité des signaux provenaient de recours à des professionnels de santé non médicaux : infirmiers, pharmaciens, kinésithérapeutes, et même orthophonistes (dans des proportions comparables à celles des pneumologues). Ce profil reflète une multiplication d'interactions de proximité pour gérer des symptômes récents et non spécifiques, souvent interprétés initialement comme infectieux ou fonctionnels, et qui n'avaient pas encore conduit à une évaluation spécialisée. Cette succession d'actes non spécifiques, peut également traduire une forme d'errance diagnostique : elle révèle une certaine prédictibilité de ces séquences en attente de l'identification de la pathologie. Dans cette perspective, l'intérêt de ces modèles pourrait aussi être de rendre visibles ces patterns récurrents, qui pourraient à terme contribuer à mieux repérer les situations à risque et à réduire les délais de diagnostic.

[24] Pour le cas de la mise en place d'un dépistage fondé sur un score de risque, l'AUC n'est pas nécessairement la métrique la plus informative, *cf. infra* « comment interpréter la valeur prédictive des modèles ».

Graphique 8 Explicabilité du modèle des plongements vectoriels de trajectoire pour la prédiction du cancer du poumon



Concernant les cancers des VADS, on retrouve un profil d'explicabilité dominé par les médicaments, représentant plus de la moitié de la contribution totale du modèle. Contrairement au poumon, les consultations médicales étaient cette fois largement représentées, avec une place prédominante des médecins généralistes, tandis que la part des consultations infirmières apparaissait bien plus faible. On observait également une contribution plus importante des passages dans les services d'urgences. Ces différences posent la question de l'interprétation du recours aux soins

médical versus paramédical : reflètent-elles des patients en meilleure forme recourant davantage à des actes infirmiers de routine (prises de sang, prélèvements COVID, etc.) et moins aux consultations médicales, ou bien l'inscription dans un parcours de soins primaires de proximité structuré autour des professionnels paramédicaux ? Pour ces deux cancers, il apparaît une contribution significative des actes et consultation de dentisterie, ce qui pourrait être un indicateur de la proximité aux soins, ou de consommation de tabac (facteur de risque commun à ces cancers).

À l'opposé, le cancer du col de l'utérus se heurte à une autre difficulté : le diagnostic est théoriquement posé en ville, dans le cadre du dépistage. Les données montrent que la prédictibilité reste faible, mais avec un apport réel de la prise en compte des trajectoires (+4 points d'AUC stratifiée entre le PVT ou BEHRT-SNDS et le modèle basé sur la cartographie et les fréquences de consommation). Par ailleurs, l'analyse des contributions montre une dispersion plus forte : aucun facteur n'émerge clairement, signe qu'à la fois la population concernée est plus jeune, moins polypathologique ou atteinte de maladie chronique, et n'est pas issue d'un groupe à risque évident. On notait toutefois un signal inattendu lié au lévothyrox (3 %), dont la présence peut relever d'un bruit statistique, ou d'une association liée à la prévalence élevée du traitement substitutif thyroïdien dans la population féminine : l'explicabilité ne traduit pas nécessairement une causalité mais peut inviter à s'interroger et à poursuivre les investigations. Donc même en l'absence d'événement « marqueur », les trajectoires des patientes permettent de prédire en partie l'hospitalisation pour cancer du col (AUC stratifiée PVT : 0,60). Cela pourrait révéler un biais potentiel de la tâche de prédiction : le modèle capterait moins le risque clinique de cancer qu'un profil de faible recours au dépistage et aux soins, associé à un diagnostic plus tardif et donc à une hospitalisation.

Ainsi, l'explicabilité complète l'évaluation prédictive : elle permet non seulement de juger de la robustesse du modèle (si l'on retrouve des facteurs de risque connus, des marqueurs de proximité aux soins, des événements triviaux attendus dans le parcours de soins préalable à l'hospitalisation…), mais aussi d'ouvrir la discussion sur l'usage en pratique (repérage précoce, ciblage de campagnes, adaptation de parcours de soins).

Enjeux de santé publique

Pour une pathologie donnée, la prédiction en population générale de chaque classe de la classification internationale des maladies CIM-10 n'est pas toujours le plus pertinent d'un point de vue santé publique. Par exemple, il peut être assez aisé d'associer un risque élevé d'hospitalisation avec un DP de diabète en population générale, lorsque le patient est traité par insuline depuis des années. L'enjeu n'est donc plus de prédire une incidence de la maladie, mais ses complications qui amènent à une hospitalisation.

Tableau 9 Redéfinition des pathologies cibles

Pathologie	Problématiques ciblées	Population	Définition de l'incidence	Masque temporel
Diabète (DT1 / DT2)	(i) Diagnostic initial sévère (DT1 ou DT2) sans prise en charge préalable en ville (ii) Décompensations ou escalades thérapeutiques	(i) Patients naïfs : non traités (ii) Patients traités : hypoglycémiants oraux (ATC A10, hors A10BX06, A10A3) ou insuline (A10A3)	1er diagnostic E10 – E11 en DP/DR (ii) Exclusion : séjours < 1 jour	(i) 1 mois (ii) 3 mois
Insuffisance rénale aiguë ou chronique (IRA N17/ IRC N18)	(i) IRA incidente (ii) IRA sur terrain IRC (iii) complications IRC	(i) Patients naïfs : non traités par hypoglycémiants, antihypertenseurs (C02, C03, C07) ou néphroprotecteurs (C08, C09) et sans antécédant IRC (ii) Patients traités ou avec antécédant IRC (iii) Patients traités	(i), (ii) 1er diagnostic N17 en DP/DR (iii) 1er diagnostic N18 en DP/DR (iii) Exclusion : séjours < 1 jour	(i), (ii) 1 mois (iii) 3 mois
Insuffisance cardiaque	Décompensation menant à la première hospitalisation	Patients traités par digitaliques (C01B, C01C) ou antihypertenseurs (C03, C07, C09)	1er diagnostic hospitalier I11, I13, I50 en DP/DR Exclusion : séjours < 1 jour	3 mois
Maladies chroniques des voies respiratoires inférieures	Décompensation ou adaptation thérapeutique	Patients traités par bronchodilatateurs / anti-asthmatiques (R03)	1er diagnostic hospitalier J40 – J47 en DP/DR - Exclusion : séjours < 1 jour	3 mois

Tableau 10 AUC stratifiées sur l'âge et le sexe des prédictions des cibles redéfinies à 1 an, selon les représentations des données et la modélisation

Pathologie	Nombre de cas et taille de l'échantillon d'entraînement (hors BEHRT-SNDS)	Cartographie et fréquences	Cartographie	Comorbidités et fréquences	Comorbidités	Fréquences de consommation	PVT	BEHRT-SNDS
Diabète de type 1 connu	2 241 / 749 946	0,830	0,815	0,820	0,719	0,806	0,832	**0,840**
Diabète de type 1 non connu	857 / 10 000 000	0,533	0,552	0,539	**0,559**	0,532	0,536	0,551
Diabète de type 2 connu	9 340 / 760 235	0,699	0,683	0,692	0,658	0,666	0,717	**0,735**
Diabète de type 2 non connus	1 576 / 10 000 000	0,688	0,640	0,684	0,638	0,660	0,698	**0,707**
Insuffisance cardiaque	26 372 / 2 919 674	0,742	0,727	0,747	0,741	0,684	0,760	**0,780**
Insuffisance rénale aiguë avec facteurs de risque	4 764 / 234 284	0,760	0,727	0,783	0,772	0,735	0,791	**0,811**
Insuffisance rénale aiguë sans facteurs de risque	694 / 10 000 000	0,694	0,638	0,690	0,650	0,683	0,694	**0,749**
Insuffisance rénale chronique avec facteurs de risque	2 909 / 3 254 833	0,825	0,765	0,879	0,867	0,796	0,887	**0,904**
Maladie respiratoire chronique	11 071 / 2 186 070	0,778	0,755	0,798	0,787	0,698	0,805	**0,819**

Note > Les données d'entrée sont préparées pour chaque délai et pour toutes les cibles pour plus de 10 millions d'individus, puis filtrées en fonction des critères énoncés en *tableau 9*. La taille d'entraînement peut varier d'une cible à l'autre en fonction, d'une part de la fréquence de la population cible en population générale, et d'autre part de la population initialement préparée au moment où l'entraînement est lancé.

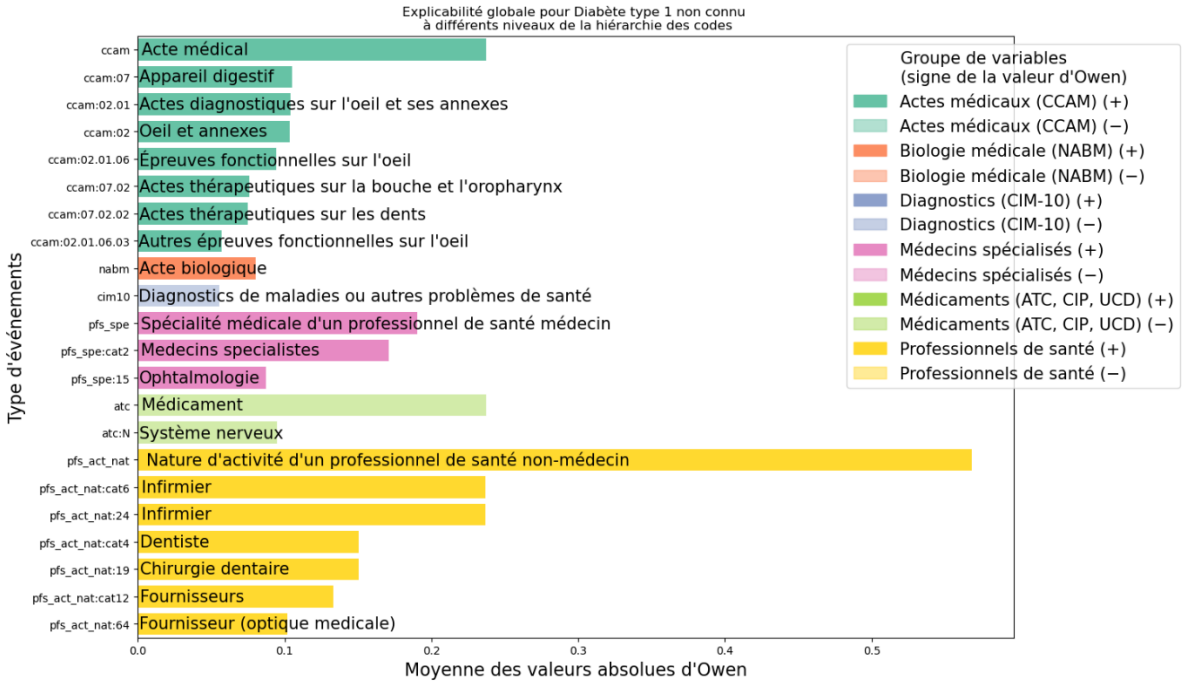
Nous avons donc affiné les tâches de prédiction, en travaillant sur des sous-populations ciblées et des définitions d'incidence adaptées, de façon à éprouver les capacités prédictives sur des problèmes *a priori* plus complexes. Renoncer à un modèle en population générale implique souvent de s'appuyer sur un échantillon d'apprentissage de taille plus réduite (le modèle PVT, en s'appuyant sur les embeddings de cooccurrences, embarque une information apprise en population générale, comme le modèle BEHRT-SNDS qui est pré-entrainé en population générale).

Nous avons ainsi sélectionné un ensemble de pathologies d'intérêt, en raison de leur importance en santé publique, et de la possibilité de définir des filtres cliniquement pertinents. Pour chacune, des critères d'inclusion spécifiques ont été appliqués afin de restreindre la population étudiée, puis une définition précise de l'incidence a été adoptée. Pour certaines pathologies, les séjours de moins d'un jour, correspondant en grande partie à des hospitalisations de jour pour le suivi ou des traitements par voie veineuse, ont été exclus car ils n'indiquaient pas un changement dans l'évolution de la maladie. Les données correspondant aux événements survenus juste avant le temps de prédiction t_0 ont été masquées pour certaines pathologies de façon à ne pas être influencé par les bilans préopératoire ou pré-admission (masque d'1 mois), ou par l'accélération des soins avant la décompensation, dont l'équipe soignante du patient est témoin (masque de 3 mois). En effet, à ce stade, la prédiction perd son intérêt : les signaux cliniques sont déjà visibles pour les soignants. Les principaux choix sont résumés dans le *tableau 9*. Globalement, les performances prédictives reflètent des profils très contrastés entre pathologies incidentes et pathologies chroniques déjà connues (*tableau 10*). Dans le cas du diabète de type 1 non connu, les performances sont faibles (AUC stratifiée moyenne 0,54, n = 857, où n désigne dans la suite le nombre de cas dans la population

d'entraînement[25]) : l'événement est souvent brutal et peu anticipable, et seuls des signaux de terrain ou des recours aux soins discrets sont captés. En revanche, pour des affections également incidentes comme le diabète de type 2 non connus (0,67 ; n = 1 576) ou l'insuffisance rénale aiguë sans facteurs de risque identifiés (0,69 ; n = 694), les performances sont meilleures : bien que le diagnostic soit incident, ces situations semblent se développer sur un terrain sous-jacent détectable, que les modèles parviennent partiellement à capturer. Notamment, pour l'insuffisance rénale aiguë sans facteurs de risque, le modèle BEHRT-SNDS dépasse le PVT d'environ 5 points d'AUC stratifiée, alors même qu'il s'agit de la cible présentant le plus petit nombre de cas positifs dans l'échantillon d'entraînement. Ce résultat suggère que BEHRT-SNDS conserve une capacité de généralisation supérieure dans les contextes où les signaux sont rares et peu redondants.

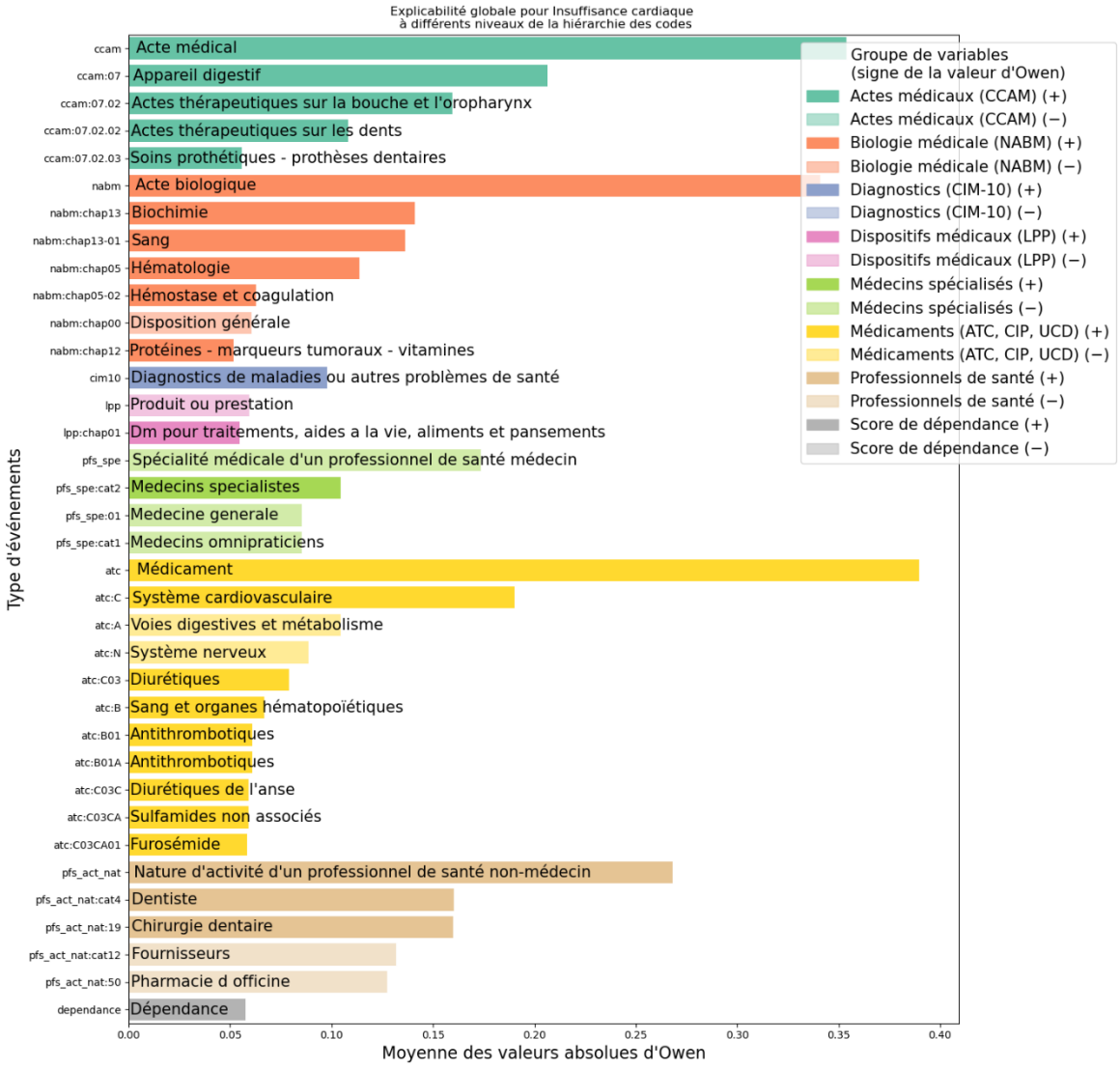
Lorsque la maladie est déjà suivie en ville ou associée à un terrain chronique, les performances deviennent nettement plus élevées : diabète de type 1 connu (0,81, n = 2 241), diabète de type 2 connu (0,69, n = 9 340), insuffisance cardiaque (0,74, n = 26 372), maladies respiratoires chroniques (0,78, n = 11 071) ou insuffisance rénale chronique (0,85, n = 2 909). Dans ces contextes cependant, les modèles experts captent déjà une grande partie du signal et le modèle PVT n'apporte qu'un gain modeste (souvent inférieur à 2 points d'AUC stratifiée). L'écart se creuse en revanche davantage avec BEHRT-SNDS, qui ajoute encore 1 à 2 points d'AUC stratifiée à celle de PVT. L'explicabilité des modèles appliqués aux pathologies chroniques met en évidence plusieurs constantes. D'une part, les modèles s'appuient toujours sur des éléments attendus de suivi : biologie, dispositifs médicaux d'autosurveillance, examens spécialisés et polythérapie, qui traduisent le poids du terrain chronique. D'autre part, apparaissent de façon récurrente des recours moins spécifiques, comme le dentiste ou l'ophtalmologie, qui semblent fonctionner comme des proxys de proximité au système de soins ou de conditions socio-économiques, surtout lorsque le signal clinique est peu marqué (diabète non connu, IRA sans FDR). Dans les pathologies les plus souvent associées à d'autres pathologies chroniques (IRC, insuffisance cardiaque), les marqueurs de dépendance et de fragilité viennent compléter ces trajectoires. Cette dualité souligne que, dans les maladies chroniques, l'essentiel du signal est déjà porté par les variables expertes et les fréquences de consommation, alors que dans les situations incidentes ou hétérogènes, c'est la dynamique des trajectoires et possiblement le profil de recours aux soins qui permet de capter une information supplémentaire. Des exemples sont illustrés *graphique 9* et *graphique 10*.

Graphique 9 Explicabilité du modèle des plongements vectoriels de trajectoire pour la prédiction du diabète de type 1 non traité en ville



[25] Les effectifs modestes observés pour certaines pathologies s'expliquent par la définition retenue des cas incidents, limitée aux diagnostics posés à l'hôpital sans traitement préalable identifié. Cette approche cible donc des formes plus sévères, souvent révélées à l'occasion d'un épisode aigu (comme un coma acido-cétosique pour le diabète de type 1), et non l'ensemble des cas incidents survenus en population, ce qui explique à la fois les nombres réduits et la nature plus grave des situations captées.

Graphique 10 Explicabilité du modèle des plongements vectoriels de trajectoire pour la prédiction de l'insuffisance cardiaque



Autres cibles aux performances prometteuses

Au-delà des cancers et des grandes maladies chroniques, certaines pathologies se sont démarquées lors du premier criblage exploratoire des 182 diagnostics étudiés. Elles ne relèvent pas nécessairement d'un fardeau de santé publique massif, mais présentent des caractéristiques cliniques ou organisationnelles qui rendent la prédiction particulièrement utile. Qu'il s'agisse de maladies souvent diagnostiquées avec retard, de situations cliniques à fort risque de complications évitables, ou encore de trajectoires de soins marquées par des passages itératifs à l'hôpital, ces pathologies offrent un terrain propice pour tester l'apport des modèles prédictifs (PVT et BEHRT-SNDS). Leur étude permet ainsi d'illustrer la diversité des cas d'usage envisageables, allant de l'anticipation d'une incidence encore méconnue à la prévention d'événements aigus survenant au cours d'un suivi chronique.

La même méthodologie que pour les pathologies précédentes de filtrage de la population ciblée et d'application d'un masque temporel spécifique a été appliquée, détaillée dans le *tableau 11*.

Tableau 11 Redéfinition des pathologies cibles

Pathologie	Problématiques ciblées	Population	Définition de l'incidence	Masque temporel
Maladie de Parkinson	Perte d'autonomie nécessitant hospitalisation ou escalade thérapeutique	Patients traités par antiparkinsoniens (N04)	1er diagnostic hospitalier G20 – G22 en DP/DR - Exclusion : séjours < 1 jour	3 mois
Maladies hypertensives de la grossesse - Éclampsie	Complication gravidique hypertensive	Patientes identifiées par présence d'échographies obstétricales dans le parcours (JQQM002, JQQM015)	1er diagnostic hospitalier O12 – O16 en DP/DR	1 mois
Épilepsie	Décompensations ou réajustements thérapeutiques	Patients traités par anti-épileptiques (N03)	1er diagnostic hospitalier G40 – G41 en DP/DR - Exclusion : séjours < 1 jour	1 mois
Maladies inflammatoires chroniques de l'intestin (MICI)	Décompensation ou réajustement thérapeutique	Patients hospitalisés avec antécédent K50 – K52 ou traités par Salicylés (ATC A07EB01, A07EC01-A07EC03), budésonide (ATC A07EA06), méthotrexate (ATC L01BA01, L04AX03), thiopurines (ATC L01BB02, L04AX01), anti-TNF (ATC L04AB02- L04AB06), vedolizumab (ATC L04AA33) ou ustekinumab (ATC L04AC05)	Diagnostic hospitalier K50 – K52 en DP/DR - Exclusion : séjours < 1 jour	1 mois
Endométriose	Formes nécessitant une prise en charge chirurgicale ou un diagnostic par coelioscopie	Patientes ayant eu une imagerie pelvienne (échographie ccam:08.01.02 ou IRM ZCQJ001, ZCQJ002, ZCQJ004 ou ZCQJ005)	1er diagnostic hospitalier N80 en DP/DR	1 mois

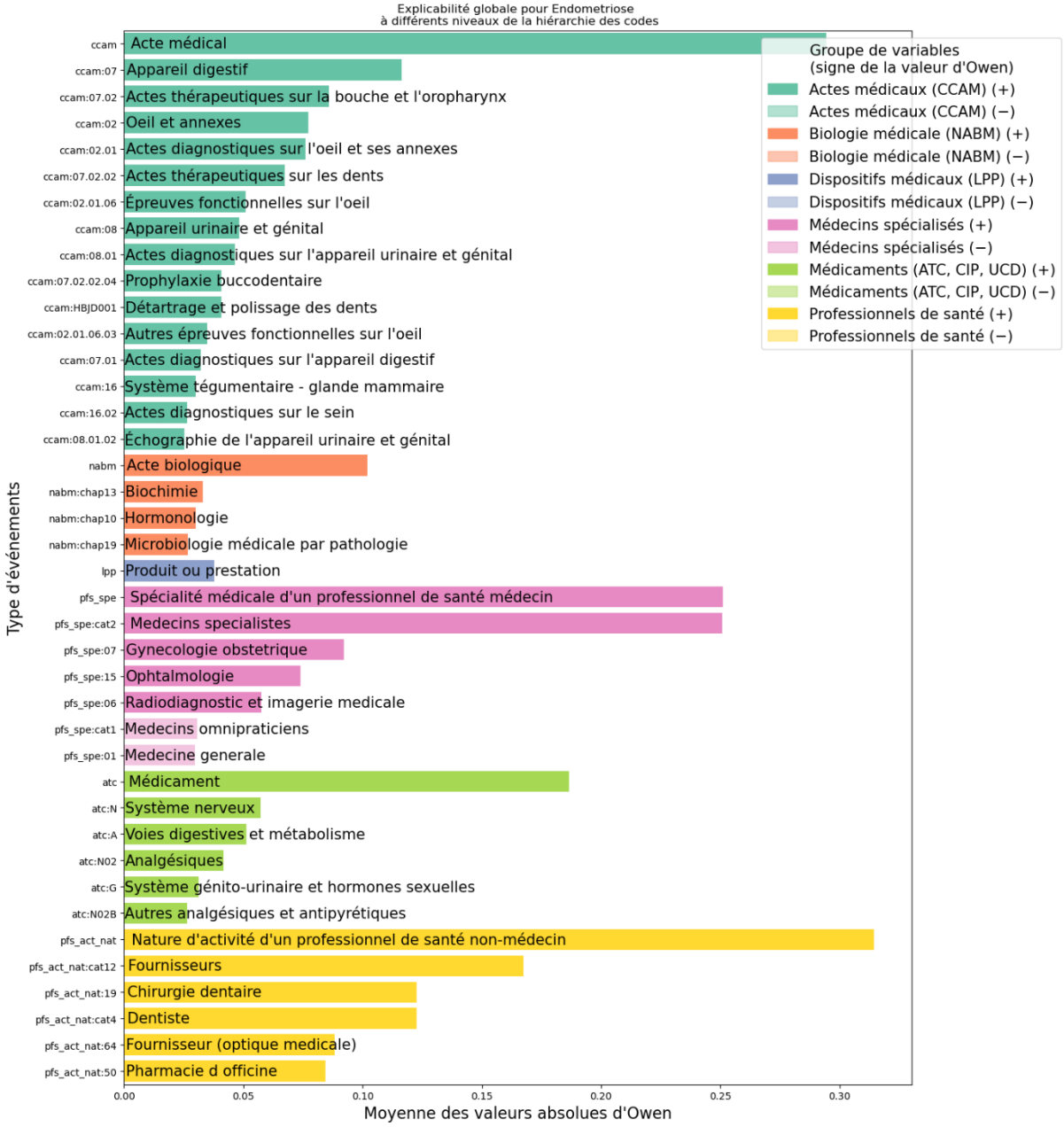
Tableau 12 AUC stratifiées sur l'âge et le sexe des prédictions des cibles redéfinies à 1 an, selon les représentations des données et la modélisation

Pathologie	Nombre de cas dans l'échantillon d'entraînement	Cartographie et fréquences	Cartographie	Comorbidités et fréquences	Comorbidités	Fréquences de consommation	PVT	BEHRT-SNDS
Parkinson	2 062 / 260 337	0,784	0,755	0,784	0,754	0,690	**0,810**	0,799
Maladies hypertensives de la grossesse - Éclampsie	307 / 53 731	0,694	0,662	0,697	0,613	0,688	0,804	**0,885**
Épilepsie	5 109 / 388 000	0,749	0,721	0,731	0,710	0,672	**0,761**	0,737
Maladies inflammatoires chroniques de l'intestin (MICI)	2 646 / 283 625	0,802	0,769	0,764	0,718	0,691	0,800	**0,810**
Endométriose	3 117 / 1 528 108	0,660	0,594	0,672	0,656	0,641	0,779	**0,804**

Note > Les données d'entrée sont préparées pour 10 millions d'individus, puis filtrées en fonction des critères énoncés en *tableau 11*. Ainsi, la taille d'entraînement peut varier d'une cible à l'autre en fonction de la fréquence de la population cible en population générale.

Dans les pathologies comme la maladie de Parkinson, voir *tableau 12*, (AUC stratifiée moyenne 0,77, n = 2 062) les modèles permettent d'anticiper des hospitalisations traduisant probablement un déclin du patient (perte d'autonomie, escalade thérapeutique, difficultés de maintien à domicile). De la même façon, dans l'épilepsie (0,73, n = 5 109), les hospitalisations prédites correspondent plus vraisemblablement à un état de mal épileptique[26] ou à un mauvais contrôle de la maladie. Ceci ouvre alors la voie à des perspectives d'action à la fois individuelles (réajustement thérapeutique, renforcement du suivi à domicile) et systémiques (planification des ressources, coordination ville – hôpital – médico-social).

Graphique 11 Explicabilité du modèle des plongements vectoriels de trajectoire pour la prédiction de l'endométriose



Certaines pathologies spécifiques, comme les MICI, montrent d'excellentes performances (AUC stratifiée 0,80), y compris avec des modèles experts, du fait d'une codification très caractéristique des hospitalisations, traitements et ALD associées. À l'inverse, pour les pathologies hypertensives de la grossesse, les indices génériques et les fréquences de consommation (AUC stratifiée 0,70) se révèlent insuffisants, et seuls les plongements vectoriels

[26] Un état de mal épileptique est une crise d'épilepsie prolongée ou une succession de crises rapprochées sans retour à la conscience entre elles.

(0,80) et BEHRT-SNDS (0,89) atteignent des performances élevées, probablement en captant la dynamique des suivis obstétricaux. L'explicabilité du modèle PVT révèle un poids important des traitements et dispositifs médicaux utilisés dans le diabète dans la prédiction des pathologies hypertensives de la grossesse. Cela peut s'expliquer par le fait que le diabète figure parmi les comorbidités classiques associées à ces complications, mais les modèles avancés semblent ici capter plus finement des situations de déséquilibre ou de mauvais contrôle glycémique, traduisant un surrisque réel chez ces patientes.

L'endométriose illustre un autre profil : les modèles experts sont limités (AUC stratifiée ~0,66 – 0,67), tandis que le modèle PVT apporte un gain net (0,78, soit +11 points), dépassé par BEHRT-SNDS (0,80, soit +13 points). L'explicabilité du modèle PVT illustrée, le *graphique 11* suggère que ce n'est pas un événement isolé mais la dynamique des trajectoires qui porte l'information : on retrouve des prescriptions variées (contraceptifs, antalgiques, traitements associés), mais aussi une forte contribution inattendue des soins dentaires (12 % consultations, 8 % actes) et ophtalmologiques (7 % chacun), peut-être davantage marqueurs d'accès aux soins ou de profil sociodémographique que déterminants cliniques spécifiques. L'hospitalisation prédite correspond probablement à une intervention chirurgicale, et ces signaux ouvrent des pistes pour mieux orienter et organiser les filières spécialisées. Cette succession de recours non spécifiques avant la formalisation du diagnostic, peut également traduire une phase d'errance diagnostique, et les schémas repérés par le modèle pourraient constituer un signal exploitable pour mieux repérer, en amont, les patientes présentant des parcours de soins fragmentés ou prolongés avant la prise en charge spécialisée. Le raisonnement ici est différent de celui des pathologies réellement incidentes : le modèle est entraîné parmi les femmes ayant déjà initié une démarche diagnostique, identifiée notamment par la réalisation d'imageries pelviennes, pour prédire celles qui iront jusqu'à l'hospitalisation. L'intérêt est donc d'identifier les trajectoires menant à la chirurgie, afin de mieux organiser le parcours de soins en amont, de faciliter l'accès et l'organisation des filières spécialisées, et, le cas échéant, de réduire les situations d'errance ou de retard diagnostique. Le diagnostic retenu ici étant hospitalier, il ne couvre qu'une partie des formes d'endométriose : beaucoup de patientes ont un diagnostic déjà posé et traité médicalement, avec une chirurgie différée ou évitée, tandis que d'autres n'auront jamais d'hospitalisation et échappent ainsi à cette définition.

En résumé, les modèles prédictifs (PVT et BEHRT-SNDS) apparaissent robustes pour les hospitalisations dans les pathologies chroniques et bien codifiées, alors que les diagnostics plus aigus restent plus difficiles à anticiper. Les variables expertes et les fréquences de consommation suffisent dans certains cas (MICI, DT1 connu), mais dans d'autres (épilepsie, endométriose, pathologies hypertensives de la grossesse), seule la dynamique fine des trajectoires apporte un signal discriminant supplémentaire.

Comment interpréter la valeur prédictive des modèles ?

L'AUC stratifiée constitue notre mesure principale de discrimination, car elle évalue la capacité du modèle à distinguer correctement les individus à risque en neutralisant les effets d'âge et de sexe. Elle renseigne ainsi sur la qualité intrinsèque du signal appris. Cependant, cette métrique présente aussi des limites : lorsqu'un événement est rare, l'AUC tend mécaniquement à être plus élevée pour une même utilité pratique, car il est plus facile de distinguer les nombreux individus négatifs. À l'inverse, pour des cibles fréquentes ou mal codifiées, même une AUC autour de 0,65 – 0,70 peut traduire une valeur informationnelle importante. Un gain absolu de quelques points d'AUC (1 à 2 points) peut en outre correspondre à des milliers de cas mieux classés, surtout sur des populations de grande taille.

Pour juger la valeur réelle d'un modèle, il faut compléter cette lecture statistique par des métriques qui traduisent son impact opérationnel. Nous avons développé plusieurs métriques complémentaires (détaillée dans le **Glossaire**) permettent d'appréhender cette valeur d'usage, notamment dans une optique de dépistage :

1. En fixant le rappel à 50 % (détecter 50 % des cas), le ratio de population à dépister entre le modèle démographique et le modèle évalué mesure le gain d'efficacité du ciblage : il indique le gain de réduction de la population à dépister nécessaire pour détecter 50 % des cas.
2. En fixant la taille de la population à dépister de sorte que le modèle démographique atteigne un rappel de 10 % (le modèle de base détecte 10 % des cas), le ratio de rappel entre le modèle évalué et le modèle démographique mesure également le gain d'efficacité du ciblage : il indique le gain en nombre de cas détectés pour une même proportion de population dépistée.

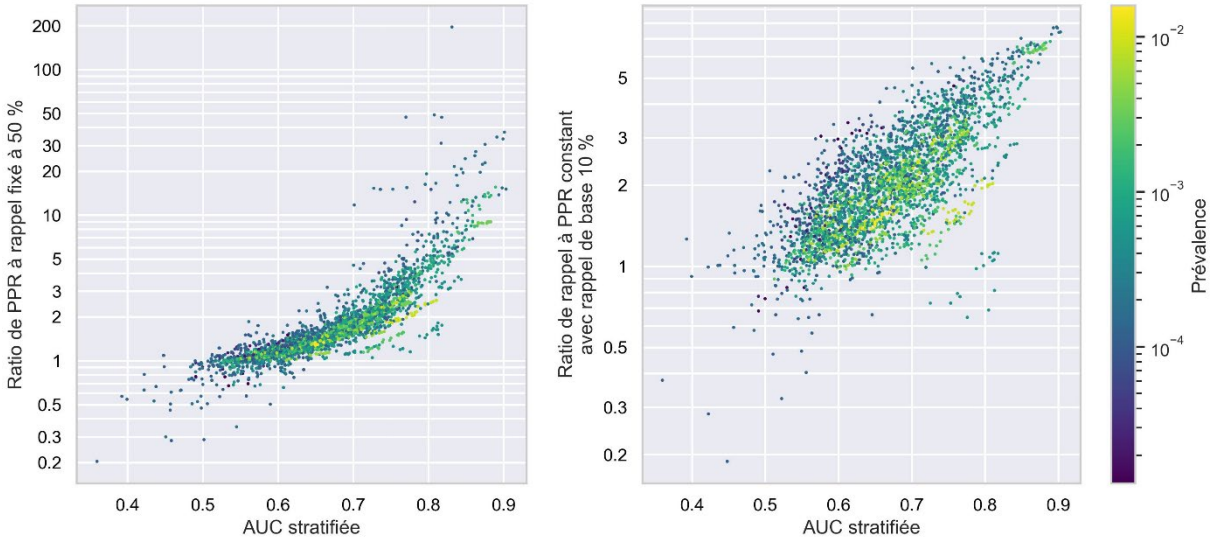
Ces métriques traduisent différentes dimensions de la valeur prédictive : statistique, lorsqu'il s'agit d'évaluer la qualité du signal appris ; opérationnelle, lorsqu'on mesure l'impact sur le ciblage ou la priorisation des interventions ; décisionnelle, lorsqu'on relie la performance à une stratégie de politique publique. Cependant, elles reposent sur des hypothèses de seuil ou de prévalence propres à chaque scénario, ce qui rend leur comparaison systématique difficile.

Cependant, on observe que 92 % (resp. 91 %) des classements de modèles selon l'AUC stratifiée sont conservés, en classant plutôt par la 1re (resp. la 2e) métrique ci-dessus. Ainsi l'AUC stratifiée reste un bon indicateur de l'impact opérationnel d'un modèle. La *graphique 12* illustre le lien entre l'AUC stratifiée et les deux métriques présentées ici.

L'exemple du cancer du poumon illustre le cas où un modèle peut présenter une AUC stratifiée modérée tout en apportant un gain concret de santé publique. En effet, le modèle PVT atteint une AUC stratifiée de 0,68 mais permet, par rapport à un modèle démographique, de réduire de 30 % la population à cibler pour identifier la moitié des cas, et de détecter 2,5 fois plus de cas à effort constant (ratio de rappel = 2,5 à population ciblée fixée par le rappel du modèle démographique = 10 %). Ce décalage entre performance statistique et impact opérationnel illustre la nécessité d'interpréter la valeur prédictive au regard de son usage et de sa finalité.

Cette lecture pragmatique soulève une question essentielle : à qui profite réellement la performance prédictive ? Un modèle peut mieux sélectionner les individus à dépister et repérer davantage de cas pour un même effort, mais rien ne garantit que ces gains soient équitablement répartis. Il est possible qu'il identifie plus efficacement les personnes déjà proches du système de soins, mieux suivies ou socialement favorisées, et qu'il contribue, involontairement, à creuser les inégalités de santé plutôt qu'à les réduire. La section suivante explore précisément cette dimension, en mobilisant l'échantillon EDP-Santé pour analyser comment les performances se distribuent selon les caractéristiques individuelles, sociales et territoriales.

Graphique 12 Relation entre métriques complémentaires orientées dépistage et AUC stratifiée selon l'âge et le sexe



Lecture > Chaque point représente une tâche de prédiction : son abscisse correspond à l'AUC stratifiée et son ordonnée à la métrique complémentaire : « ratio de population à dépister » dans le graphique de gauche, « ratio de rappel » dans le graphique de droite.

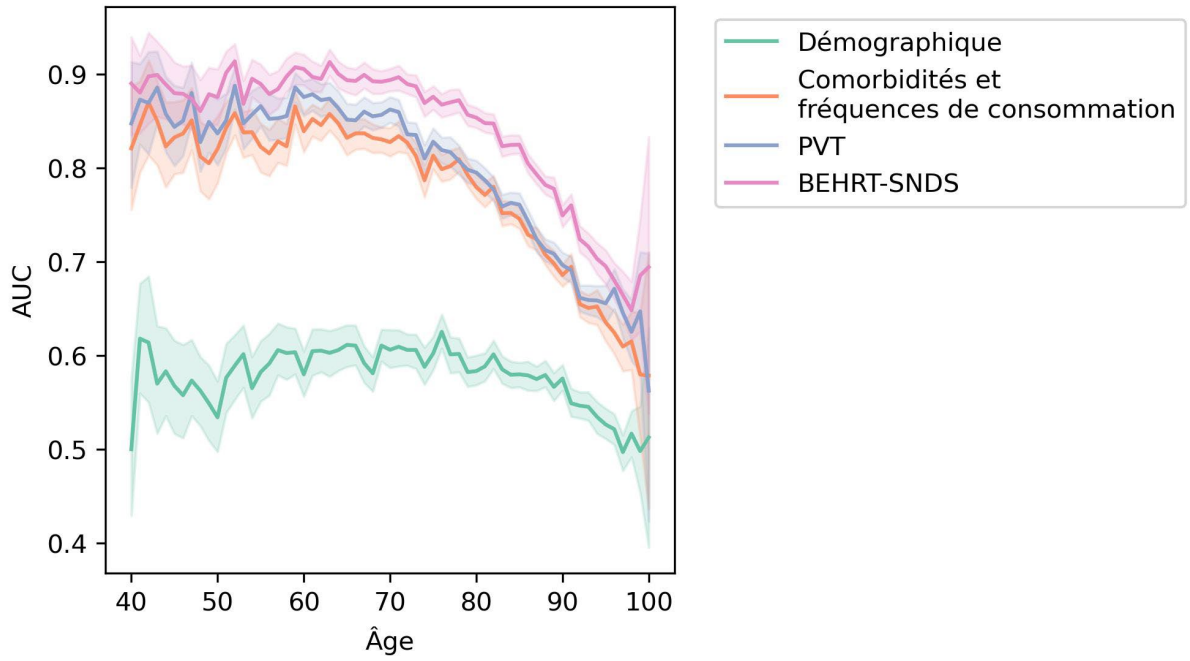
POUR QUI PRÉDIT-ON BIEN ?

Une question importante est la performance de ces modèles selon les caractéristiques des individus observées dans le SNDS (par exemple, sexe, région de résidence) mais aussi des dimensions non observées dans les données, par exemple le niveau de vie. Afin de préciser si les performances en prédiction peuvent varier d'un individu à l'autre, nous mobilisons l'échantillon démographique permanent de l'Insee apparié au SNDS (EDP-Santé). La transformation des données décrite dans la partie II est appliquée à l'extraction du SNDS 2008-2022 disponible pour les individus de l'échantillon démographique permanent, qui font partie par construction de l'échantillon « test ». L'enrichissement dans le temps des données de l'EDP implique que l'ensemble des sources n'est pas disponible depuis le début de la période.

Le cas de la prédiction de la mortalité est étudié dans un premier temps comme un cas universel et recouvrant des parcours très divers. Néanmoins, il est évident qu'étudier l'équité des algorithmes, par exemple dans l'optique d'un ciblage de patients pour un accompagnement personnalisé, impliquerait au préalable de bien définir la politique publique en question, ses objectifs et d'en déduire le problème de prédiction associé le plus pertinent, ainsi que la métrique. Cette section illustre qu'il est alors envisageable, grâce à l'EDP-Santé, d'évaluer les différences de performances prédictives entre les individus via l'AUC, bien que d'autres métriques puissent être plus adaptées à une politique publique donnée (par exemple, pour le ciblage d'une population fondée sur son risque, une métrique plus directement lisible pourrait être le *likelihood* ratio, LR+ ou une version du *Net Clinical Benefit*).

Pour 1,5 million d'individus EDP, âgés de plus de 40 ans en 2018, représentatifs de l'ensemble de la population résidant en France de ces âges (35,4 millions), entre un et deux temps de prédiction aléatoires sont tirés entre mi 2018 et mi 2022, pour atteindre 2,83 millions d'observations. Les principaux modèles entraînés sur le SNDS entre 2018 et 2022 (sur 30 millions d'individus pour le modèle de comorbidités et PVT, et sur 50 millions pour BEHRT-SNDS) sont appliqués à l'ensemble de la trajectoire disponible entre le 1er janvier 2010 (première année de disponibilité du niveau de vie) et t_0, ce qui est rendu possible grâce à un historique de soins plus profond dans l'EDP-Santé que dans l'extraction utilisée pour cette étude, au prix d'une sortie du domaine d'entraînement. Pour environ 45 300 observations, un décès est observé dans l'année suivant t_0.

Graphique 13 AUC stratifiée en fonction de l'âge pour la prédiction de la mortalité



Note > EDP-Santé, transformation SeqNDS. AUC calculée pour chaque âge. Intervalles de confiance calculés par *bootstrap*.

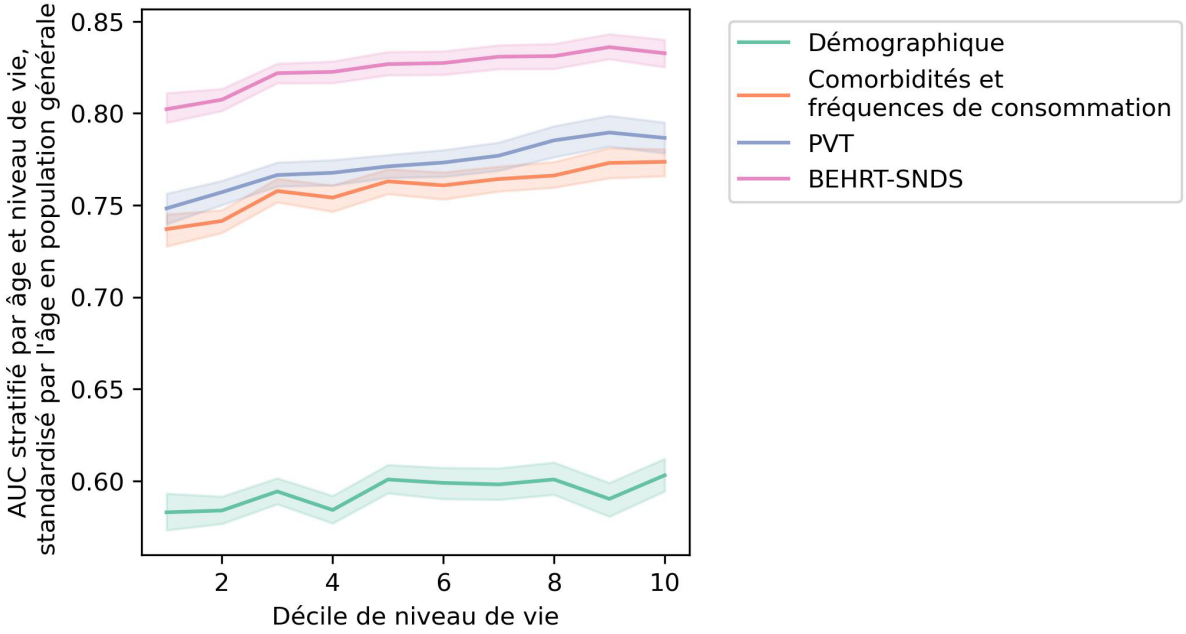
Les AUC globales atteignent respectivement 0,908 et 0,914. Les deux modèles sont plus performants pour les individus âgés entre 40 et 70 ans, avant de décroître continûment jusqu'à 100 ans (*graphique 13*), probablement car la population âgée est plus homogène au regard de son risque de décès que ne l'est la population d'âge intermédiaire, où les différentiels d'état de santé peuvent être plus conséquents. En stratifiant par âge et sexe, et en repondérant par le nombre de décès de chaque strate, les AUC stratifiées atteignent 0,755 et 0,770, soit des performances moindres de plusieurs points d'AUC à celles obtenues sur l'échantillon test de l'exhaustif, probablement

du fait de la sortie du domaine d'entraînement (historique plus long ; évolutions de la qualité et des pratiques de codage, des sources disponibles, des nomenclatures…), ainsi que d'un champ d'évaluation différent (résidents pour l'EDP-Santé vs bénéficiaires de l'assurance maladie pour le SNDS), malgré des trajectoires plus longues dans l'EDP-Santé, susceptibles d'apporter davantage d'information.

Les deux modèles sont plus performants pour les femmes de 2,3 (PVT) et 2,7 (comorbidités et fréquences de consommation) points d'AUC et de 1,8 (PVT) et 2,8 (comorbidités et fréquences de consommation) points d'AUC stratifiées par âge et sexe, après standardisation sur la structure d'âge globale des décès pour neutraliser l'effet de l'âge plus avancé que les femmes atteignent en moyenne. C'est peut-être le fait de leur plus grande proximité avec le système de santé et d'un plus fort recours qui génère des trajectoires plus claires et interprétables, des différentiels de morbidités avec des prévalences plus fortes de certaines pathologies mieux captées par le SNDS et reliées à la mortalité ou encore le moindre rôle des morts violentes (causes externes) dans la mortalité des femmes, difficilement prédictibles à partir du recours aux soins. Par région, les AUC stratifiées par âge et région, après standardisation sur la structure d'âge globale des décès afin de neutraliser les différences de structure d'âge entre les régions, indiquent des écarts limités entre les régions dans les performances du modèle PVT, si l'on met à part Hauts-de-France, Corse et les départements et régions d'outre-mer (DROM) : il y a moins d'un point d'AUC de différence entre les meilleures performances en Occitanie et en Île-de-France et les moins bonnes dans le Grand Est et les Pays de la Loire. Cependant, par rapport au Grand Est, constituant le bas de ce classement resserré, les Hauts-de-France sont un point en dessous et les DROM 3,6 points en dessous.

Enfin, les performances statistiques sont de 3,8 points supérieures pour les 10 % des niveaux de vie les plus élevés par rapport aux 10 % des niveaux de vie les plus pauvres (mesurées par l'AUC stratifiée par âge et niveau de vie et standardisée par la structure d'âge des décès dans la population générale, afin de raisonner à âge équivalent entre les niveaux de vie) pour le modèle PVT. En calculant des intervalles de confiance par bootstrap[27], il apparaît que ces différences sont bien significatives (*graphique 14*).

Graphique 14 AUC stratifiée par niveau de vie, à structure d'âge équivalente entre les déciles, pour la prédiction de la mortalité



Note > EDP-Santé, transformation SeqNDS. AUC calculée pour chaque tranche d'âge quinquennale et décile de niveau de vie en 2018, et standardisée par niveau de vie en prenant pour pondération le nombre de décès par tranche d'âge quinquennale en population générale. Les variables expliquant que le modèle démographique atteint une AUC >0,5 sont le sexe et l'âge exact conditionnellement à la tranche d'âge quinquennale. Intervalles de confiance calculés par *bootstrap*.

[27] Tirage avec remise parmi les 2,83 millions d'observations réitérés 500 fois, recalcul des métriques sur chacun des échantillons bootstrappés, et intervalle de confiance empirique à 95 % de ces dernières

CONCLUSION

Malgré l'absence d'informations cliniques précises, les signaux d'état de santé disponibles dans la base principale du SNDS ont un pouvoir prédictif significatif sur la suite du parcours de soins. Les performances atteintes en prédiction sont très hétérogènes d'une pathologie à l'autre, et recouvrent des situations variées, dont certaines apparaissent pertinentes dans une perspective de santé publique à l'échelle populationnelle.

Les informations contenues dans le SNDS, entendues comme un énorme corpus d'apprentissage des régularités des parcours de soins, et donc indirectement de l'histoire des pathologies, du recours aux soins, et des prises en charge sous-jacentes, paraissent sous-exploitées au regard des méthodes d'intelligence artificielle les plus modernes.

Les performances obtenues en appliquant les modélisations issues du traitement automatique des langues sont significativement supérieures par rapport à des approches « dires d'experts », limitées dans leur capacité à intégrer l'ensemble des informations disponibles, et par rapport à des approches machine learning standard, limitées dans leur capacité à intégrer l'ensemble de la population d'apprentissage accessible et à tirer parti des séquences. Par le caractère généraliste des modélisations les plus modernes, en capacité d'intégrer l'ensemble du parcours de soins pour en prédire la suite, l'expertise en santé publique nécessaire se déplace de l'amont de la modélisation (sélection des variables pertinentes) vers l'aval (explicabilité, définition des objectifs de santé publique et des populations d'intérêt pour le *fine-tuning*) où le travail d'interprétation reste complexe.

BEHRT-SNDS, le meilleur modèle appris dans nos travaux, n'a pas fait l'objet d'une optimisation particulière : nous avons délibérément privilégié une architecture éprouvée et reproductible, représentative de la famille des modèles Transformer et largement réutilisée comme base de variantes plus récentes, afin de fournir un point de référence solide. Pourtant depuis BERT (2018) et BEHRT (2020), de nombreux progrès ont été réalisés pour tirer le meilleur parti de ces méthodes, laissant présager qu'il reste une marge significative d'apprentissage, à données d'entrée fixées. Nos résultats permettent une première estimation de la valeur prédictive de l'usage secondaire des données collectées à grande échelle, en routine, sur les parcours, valeur qui risque de continuer de s'accroître, étant donné les progrès techniques récents et à venir, et l'intégration de nouvelles données dans le SNDS.

Il apparaît également crucial d'étudier les biais socio-économiques et territoriaux de ces modèles prédictifs (PVT et BEHRT-SNDS) entraînés à grande échelle, avant que les usages ne se développent. En effet, s'il pouvait s'agir d'un formidable outil pour mieux cibler des populations à risque dans les politiques de santé, les risques associés sont encore peu explorés. Des performances prédictives croissantes avec le niveau de vie sont de nature à accroître les inégalités de santé, en ciblant mieux les plus aisés et les personnes les plus proches du système de santé. Dans le même temps, en ciblant mieux les plus à risque, et en capturant indirectement les comportements de recours aux soins et le surrisque qui peut résulter d'un éloignement du système de santé, le déploiement de ces modèles pourrait aussi contribuer à en résorber une partie.

POUR EN SAVOIR PLUS

Agniel, D., Kohane, I. S., Weber, G. M. (2018, avril). Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. Bmj, 361.

Bacry, E., *et al.* (2020, septembre). SCALPEL3: a scalable open-source library for healthcare claims databases. International Journal of Medical Informatics, 141, 104203.

Beaulieu-Jones, B. K., *et al.* (2021, mars). Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? NPJ digital medicine, 4(1), 62.

Constantinou, P., *et al.* (2018, novembre). Two morbidity indices developed in a nationwide population permitted performant outcome-specific severity adjustment. Journal of Clinical Epidemiology, 103, 60-70.

Doutreligne, M., *et al.* (2020, mars). Snds2vec, représentations continues pour les concepts médicaux du Système national des données de santé. Revue d'Épidémiologie et de Santé Publique, 68, S35.

Fuentes, S., *et al.* (2023, juin). Identifying type 1/type 2 diabetes in medico-administrative database to improve health surveillance, medical research and prevention in diabetes: Algorithm development and application. Diabetes Epidemiology and Management, 10, 100137.

Haneef, R., *et al.* (2021, septembre). Use of artificial intelligence for public health surveillance: a case study to develop a machine Learning-algorithm to estimate the incidence of diabetes mellitus in France. Archives of Public Health, 79(1), 168.

Kraljevic, Z., *et al.* (2024, avril). Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. The Lancet Digital Health, 6(4), e281-e290.

Levy, O., Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. Advances in neural information processing systems, 27.

Li, Y., *et al.* (2020, avril). BEHRT: transformer for electronic health records. Scientific reports, 10(1), 7155.

Odgaard, M., *et al.* (2024). CORE-BEHRT: A Carefully Optimized and Rigorously Evaluated BEHRT. Machine Learning for Healthcare Conference. PMLR, 2024.

Pratt, N. L., *et al.* (2018, mars). The validity of the Rx-Risk comorbidity index using medicines mapped to the anatomical therapeutic chemical (ATC) classification system. BMJ open, 8(4), e021122.

Quan, H., *et al.* (2005, novembre). Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Medical care, 43(11), 1130-1139.

Rachas, A., *et al.* (2022, septembre). The economic burden of disease in France from the National Health Insurance Perspective: the healthcare expenditures and conditions mapping used to prepare the French social security funding act and the public health act. Medical care, 60(9), 655-664.

Rasmy, L., *et al.* (2021, mai). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine, 4(1), 86.

Savcisens, G., *et al.* (2023, décembre). Using sequences of life-events to predict human lives. Nature Computational Science, 4(1), 43-56.

Shmatko, A., *et al.* (2025, septembre). Learning the natural history of human disease with generative transformers. Nature, 1-9.

Thurin, N. H., *et al.* (2021, mai). Intra-database validation of case-identifying algorithms using reconstituted electronic health records from healthcare claims data. BMC Medical Research Methodology, 21(1), 95.

Wornow, M., *et al.* (2023, juillet). The shaky foundations of large language models and foundation models for electronic health records. Npj digital medicine, 6(1), 135.

Yuan, Z., *et al.* (2022, février). CODER: knowledge-infused cross-lingual medical term embedding for term normalization. Journal of biomedical informatics, 126, 103983.

Yang, Z., *et al.* (2023, novembre). TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. Nature communications, 14(1), 7857.

GLOSSAIRE

SNDS

MCO — Médecine, Chirurgie, Obstétrique

Secteur du PMSI regroupant les séjours d'hospitalisation en médecine, chirurgie ou obstétrique. Les données MCO alimentent la construction des Groupes Homogènes de Malades (GHM).

SSR — Soins de Suite et de Réadaptation

Secteur du PMSI dédié aux patients nécessitant des soins de rééducation, de réadaptation ou de réinsertion après une phase aiguë (ex. post-AVC, rééducation orthopédique). Les séjours en SSR sont caractérisés par le Groupe Médico-Économique (GME).

HAD — Hospitalisation à Domicile

Forme d'hospitalisation où les soins sont réalisés au domicile du patient, mais avec la même intensité et coordination qu'à l'hôpital. Champ spécifique du PMSI (HAD) avec ses propres groupes homogènes de prise en charge (GHPC).

CIM-10 — Classification Internationale des Maladies, 10e révision

Système international (OMS) servant à coder les diagnostics médicaux. En France, c'est la base officielle pour coder les motifs d'hospitalisation dans le PMSI.

DP / DR — Diagnostic Principal / Diagnostic Relié

Le diagnostic principal (au format CIM-10) est le motif des soins qui justifient l'hospitalisation dans le PMSI MCO, il est, si nécessaire, précisé par un diagnostic relié (au format CIM-10).

ALD — Affections de Longue Durée

Maladie dont la gravité et/ou le caractère chronique nécessite un traitement prolongé, et ouvre droit à des remboursements de soins plus favorables pour les soins en lien avec cette maladie. Le type d'ALD est repéré par le code diagnostic CIM-10 de demande d'ALD.

CCAM — Classification Commune des Actes Médicaux

Nomenclature des actes techniques médicaux réalisés par les professionnels de santé (actes chirurgicaux, gestes techniques, examens…), et sert de référence pour la facturation des actes médicaux.

NABM — Nomenclature des Actes de Biologie Médicale

Nomenclature des examens de biologie (analyses de sang, urine, microbiologie, etc.), qui sert de référence pour la facturation des actes de biologie.

LPP — Liste des Produits et Prestations Remboursables

Répertoire des dispositifs médicaux, prothèses, orthèses, pansements, etc., pris en charge par l'Assurance Maladie.

Nomenclatures des médicaments

CIP — Code Identifiant de Présentation

Code à 7 ou 13 chiffres attribué à chaque présentation commerciale d'un médicament (ex. conditionnement). Sert à identifier précisément une boîte vendue en pharmacie.

UCD — Unité Commune de Dispensation

Code standardisé correspondant à l'unité minimale de dispensation d'un médicament (ex. un comprimé, une ampoule…). Permet d'uniformiser la gestion et la facturation des médicaments en établissement de santé.

ATC — *Anatomical Therapeutic Chemical Classification*

Classification internationale (OMS) des médicaments selon : leur organe ou système cible (niveau anatomique) ; leur indication thérapeutique ; et leur structure chimique.

Groupages PMSI

Processus automatisé qui transforme les informations codées (diagnostics, actes, durées, âge, etc.) en groupes homogènes servant à la tarification (T2A). Ce groupage dépend du champ PMSI (MCO, SSR, HAD).

GHM — Groupe Homogène de Malades

Unité de classification du PMSI MCO : chaque séjour est affecté à un GHM en fonction de critères médicaux et de sévérité.

GME — Groupe Médico-Économique

Unité de classification utilisée dans le champ SSR, regroupant les séjours similaires sur les plans médical et économique.

GHPC — Groupe Homogène de Prise en Charge

Catégorie du PMSI HAD, définissant les types de prises en charge homogènes selon le motif d'hospitalisation à domicile (ex. soins palliatifs, chimiothérapie, suivi de grossesse).

Autres nomenclatures spécifiques au SNDS

PFS_SPE — Spécialités médicales consultées (médecin)

PFS_ACT_NAT — Nature des actes des professionnels de santé (hors médecin)

TYP_UM — Type d'unité médicale à l'hôpital

ETB_CAT — Catégorie d'établissements

DEPENDANCE — Cotation de la dépendance (SSR, HAD)

Métriques

Matrice de confusion — Pour une cible à prédire dont les valeurs valent 0 ou 1, à partir d'une prédiction binaire dont les valeurs possibles sont également 0 ou 1, on note :

- TP (*True Positive*) : le nombre de cibles dont la valeur réelle est 1 et dont la prédiction vaut 1 ;
- FP (*False Positive*) : le nombre de cibles dont la valeur réelle est 0 et dont la prédiction vaut 1 ;
- TN (*True Negative*) : le nombre de cibles dont la valeur réelle est 0 et dont la prédiction vaut 0 ;
- FN (*False Negative*) : le nombre de cibles dont la valeur réelle est 1 et dont la prédiction vaut 0 ;
- $prev = (TP + FN)/\ (TP + FN + TN + FP)$ (prévalence) : Le taux de positifs réels parmi l'ensemble de la population ;
- $PPR = (TP + FP)\ /\ \ (TP + FN + TN + FP)$ (*Predicted Positive Rate*) : Le taux de prédiction positif parmi la population ;
- $TPR = TP\ /\ (TP + FN)$ (*True Positive Rate* / rappel) : Le taux de prédiction positif parmi les réellement positifs ;
- $FPR = FP\ /\ (TN + FP)$ (*False Positive Rate*) : Le taux de prédiction positif parmi les réellement négatifs ;

Dans le cas où la prédiction est un score de risque, l'usage est de seuiller ce score à un seuil, les prédictions supérieures étant associées à 1 et inférieures à 0, on obtient ainsi une famille de matrices de confusion dépendantes du seuil.

AUC — L'AUC (aire sous la courbe ROC) peut être lue comme la capacité du modèle à discriminer un cas d'un témoin, c'est-à-dire d'attribuer une probabilité plus élevée au cas qu'au témoin. Une AUC à 0,5 correspond à un classement aléatoire, tandis qu'une valeur proche de 1 indique une très bonne performance du modèle. Précisément, la courbe ROC est la courbe croissante du rappel (TPR) en fonction de FPR obtenue en faisant varier le seuil. L'AUC est l'aire sous cette courbe prise entre 0 et 1.

AUC globale — Utilisation la plus classique de l'AUC que nous appelons AUC globale pour éviter toute confusion, pour chaque cible l'AUC est calculée sur la population générale.

AUC stratifiée — Version de l'AUC que nous utilisons pour évaluer nos modèles à âge et sexe « contrôlés », variables permettant déjà de fortement discriminer deux individus concernant leur état de santé probable. Pour chaque cible, l'AUC est calculée pour chaque classe d'âge (classes de 4 ans) et le sexe, puis moyennée en pondérant chaque strate par le nombre de cas qu'elle contient.

AUC discriminant les codes présents des codes absents, en moyenne sur les patients — Version de l'AUC utilisée dans BEHRT original (Li, *et al.*, 2020) et y étant noté « AUROC ». Pour chaque patient, l'AUC est calculée le long des cibles, puis les AUC obtenues sont moyennées le long des patients. Elle peut s'interpréter, patient par patient, comme la probabilité que le modèle sache discriminer un code présent dans les prochains mois d'un code absent, probabilité qui est ensuite moyennée sur l'ensemble des patients. La fonction sklearn.metrics.roc_auc_score avec l'argument « *average='samples'* » est utilisée.

Score de précision en moyenne sur les patients — Métrique principale utilisée dans BEHRT original (Li, *et al.*, 2020) et y étant noté « APS » pour *average precision score*. Pour chaque patient de l'échantillon test, plusieurs codes diagnostics peuvent apparaître dans la prochaine visite à prédire, et le modèle attribue un score de probabilité à chacun des codes diagnostics possibles. À seuil s donné prédire la présence du code (score>s), la précision représente la part des codes prédits par le modèle qui sont effectivement présents pour ce patient, et le rappel la part des codes du patient qui sont prédits comme étant présents par le modèle. Le score de précision au niveau du patient est un résumé de la courbe « précision-rappel », courbe empirique paramétrée par le seuil. C'est la somme pondérée de la précision atteinte à plusieurs seuils, avec des pondérations reflétant l'accroissement du rappel avec la diminution du seuil (ce qui diminue la précision du modèle). Enfin, le score de précision moyenne est obtenu en prenant la moyenne du score de précision au niveau du patient sur l'ensemble des patients de l'échantillon de test. La fonction *sklearn.metrics.average_precision_score* avec l'argument « average='samples' » est utilisée.

AUL — Variante de l'AUC globale utilisée dans *Life2Vec* (Savcisens, *et al.*, 2024), plus adaptée au cas partiellement labellisé. Précisément, la courbe Lift est la courbe croissante du rappel (TPR) en fonction du taux de prédiction positive PPR obtenue en faisant varier le seuil. L'AUL est l'aire sous cette courbe prise entre 0 et 1. Elle se déduit de l'AUC et de la prévalence : $AUL = \mathrm{AUC} - \mathrm{prev} \times (\mathrm{AUC} - 0{,}5)$.

MCC — *Matthews Correlation Coefficient*, métrique utilisée dans Life2Vec (Savcisens, *et al*, 2024), basée sur la matrice de confusion (les scores sont seuillés à 0,5 dans l'article de Life2Vec). Le MCC vaut 0 pour une prédiction aléatoire et 1 pour une prédiction parfaite. Précisément :

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

Ratio de PPR à rappel fixé à 50 % — Métrique comparant l'utilité de 2 modèles, le modèle à évaluer et un modèle de base (baseline). Chaque modèle est seuillé en choisissant un seuil par modèle pour que les rappels TPR soient de 50 % (autre valeur possible), la métrique d'évaluation est alors le PPR du modèle de base divisé par le PPR du modèle à évaluer. Cette métrique indique le gain de réduction de la population à dépister nécessaire pour détecter 50 % des cas.

Ratio de rappel à PPR constant avec rappel de base 10 % — Métrique comparant l'utilité de 2 modèles, le modèle à évaluer et un modèle de base (baseline). Le modèle de base est seuillé pour avoir un rappel (TPR) de 10 % (autre valeur possible), le modèle à évaluer est seuillé pour avoir le même PPR que le modèle de base. La métrique d'évaluation est alors le rappel (TPR) du modèle à évaluer divisé par celui du modèle de base. Cette métrique indique le gain en nombre de cas détectés pour une même proportion de population dépistée.

Ratio de rappel à PPR constant avec rappel d'évaluation à 50 % — Métrique comparant l'utilité de 2 modèles, le modèle à évaluer et un modèle de base (baseline). Le modèle à évaluer est seuillé pour avoir un rappel (TPR) de 50 % (autre valeur possible), le modèle de base est seuillé pour avoir le même PPR que le modèle à évaluer. La métrique d'évaluation est alors le rappel (TPR) du modèle à évaluer divisé par celui du modèle de base. Cette métrique indique le gain en nombre de cas détectés pour une même proportion de population dépistée.

Modélisation

Entraînement / Evaluation — Les données disponibles sont partagées en deux parties, le jeu de test constitué des individus nés les mois de janvier, avril, juillet ou octobre et le jeu d'entraînement constitué des autres individus. Les poids des modèles d'apprentissage statistique sont ajustés pour minimiser la loss de prédiction sur le jeu d'entraînement (ou une sous-partie) puis des métriques sont évaluées sur le jeu de test (ou une sous-partie).

Temps t_0 et horizon de prédiction — Nécessaire pour définir la tâche de prédiction, pour chaque individu un temps de prédiction t_0 est choisi aléatoirement, la tâche de prédiction est alors de prédire à partir des événements passés (ayant eu lieu avant t_0) si un événement cible va se produire avant l'horizon (entre t_0 et t_0 + horizon).

Feature engineering — Transformation des données brutes en variables utiles aux modèles : filtres de sous-population, création de features de décomptes, calculs d'*embeddings*, agrégats temporels, fenêtrage temporel, padding de séquence, masquage d'information, etc. Il s'agit de la partie la plus coûteuse en temps de développement humain.

Embedding — Plongement des valeurs d'une variable dans l'ensemble des vecteurs à d coordonnées, $\mathbb{R}^d$, où d est la dimension de l'*embedding*. Ce plongement abstrait est construit pour que la distance vectorielle des embeddings représente la « proximité » des valeurs plongées.

Explicabilité — Ensemble de méthodes pour comprendre les prédictions. Globale (quels facteurs importent en moyenne) vs locale (pour un individu). Les explicabilité présentées se basent sur une variante des valeurs de Shapley.

Loss — Fonction d'erreur minimisée à l'entraînement, choisie pour être alignée avec les objectifs de minimisation et pour avoir de bonnes propriétés en optimisation. Les loss utilisées dans ce projet se basent sur l'entropie croisée (log loss).

Gradient boosting — Famille de modèles additifs d'arbres appris séquentiellement pour corriger les erreurs des arbres précédents. Implémentations utilisées basée sur xgboost avec choix des principaux paramètres (spécifiques à chaque modélisation) obtenus par optimisation bayésienne.

Termes spécifiques à BEHRT (au *deep learning*)

Tokens — Unités les plus fines composant les séquences en entrée des modèles transformers. Dans le cas du traitement du langage naturel, par exemple pour les modèles BERT et GPT, il s'agit de découpages de mots, avec en moyenne, de l'ordre de 1,5 token par mot, dépendant du type de texte. Dans le cas de ce projet il s'agit des codes obtenus après regroupement hiérarchique en 30 000 codes (CIM-10, ATC, etc.), avec adjonction de tokens spéciaux [CLS] (représentant l'ensemble de la séquence) et [MASK] (token masqué).

Pré-entraînement et fine-tuning — Un modèle de transformer commence par être pré-entraîné sur une tâche auto-supervisée (*Masked Language Modeling* pour BEHRT) lui permettant d'assimiler un maximum de patterns génériques du SNDS. Il est ensuite fine-tuné sur nos tâches d'intérêt avec la loss associée.

Batch — Un batch est un lot d'exemples traité de taille « *batch size* », à partir duquel le gradient de la loss est calculé permettant une mise à jour itérative des poids.

Optimiseur, learning rate, warm-up — L'optimiseur est la méthode mettant à jour les poids du modèle à partir du gradient. AdamW est l'optimiseur classiquement utilisé pour les transformers, il possède un paramètre important « *learning rate* » contrôlant l'amplitude des mises à jour. Le mécanisme de warm-up amène linéairement le *learning rate* de 0 à une valeur fixée.

Epoch — Un passage complet sur le jeu d'entraînement, découpé en batchs.

Annexe 1. Préparation des données

Unité d'analyse et population d'étude

À partir du référentiel des bénéficiaires du SNDS, nous construisons un identifiant bénéficiaire permettant de rassembler les événements correspondant à un même individu, en excluant les cas ambigus, et en consolidant trois attributs fondamentaux : mois/année de naissance, sexe et le cas échéant, date de décès. Cette consolidation, décrite ci-après, garantit une identification précise à travers toutes les sources. Une des conséquences est l'exclusion des jumeaux de même sexe, indissociables dans les données hospitalières chaînées au référentiel des bénéficiaires. Chaque identifiant est ensuite affecté à 200 partitions aléatoires, ce qui facilite l'échantillonnage représentatif et le traitement par lot des parcours de soins. La population finale correspond aux bénéficiaires avec au moins un événement de soins sur la période considérée (n = 76,4 millions de bénéficiaires[28]).

Encadré A1 Identification des individus

L'algorithme d'identification correspond à celui utilisé pour l'identifiant de la cartographie des pathologies de la Cnam (« ID CARTO »), avec une étape supplémentaire filtrant les bénéficiaires où l'identifiant obtenu peut encore prêter à ambiguïté dans le sens où il identifie potentiellement plusieurs personnes distinctes.

Identifiants disponibles dans le SNDS :
BEN_NIR_ANO : identifiant univoque, hachage du NIR du bénéficiaire.
BEN_NIR_PSA : identifiant correspondant au hachage du NIR de l'ouvrant droit, du sexe et de la date de naissance du bénéficiaire. Un même individu pouvant être l'ayant droit de plusieurs personnes (par exemple, chacun de ses parents et lui-même), cet identifiant peut être multiple pour une même personne, et dans un cas particulier, les jumeaux de même sexe ayant droit d'une même personne (un parent), un même identifiant peut correspondre à deux individus distincts. Le rang gémellaire (BEN_RNG_GEM) peut permettre de distinguer ces cas.

Par défaut, les prestations en ville sont systématiquement identifiées par le couple (BEN_NIR_PSA, BEN_RNG_GEM), et à l'hôpital par BEN_NIR_PSA.

1. Ne sont conservés que les bénéficiaires dont l'identifiant est basé sur un NIR non fictif (mais sont conservés les NIR provisoires) et ayant eu au moins une consommation de soins dans le DCIR à partir de 2018.

2. L'identifiant « ID CARTO » associé à un (BEN_NIR_PSA, BEN_RNG_GEM), clé identifiante dans le référentiel des bénéficiaires, est défini comme suit, étape par étape :
a) Si un BEN_NIR_ANO associé à ce (BEN_NIR_PSA, BEN_RNG_GEM) existe : **BEN_NIR_ANO**
b) Sinon, si BEN_NIR_PSA correspond à un cas où le NIR de l'ouvrant-droit est le NIR du bénéficiaire, ainsi que BEN_NIR_PSA est bien spécifique à un individu unique, on peut ignorer le rang gémellaire :
b.1) si un BEN_NIR_ANO est associé à ce BEN_NIR_PSA (et sans tenir compte du rang gémellaire comme à l'étape a) : **BEN_NIR_ANO**
b.2) sinon, **BEN_NIR_PSA**

c) Sinon, la concaténation **BEN_NIR_PSA || BEN_RNG_GEM**

3. Pour chaque BEN_NIR_PSA, le nombre d'« ID CARTO » est calculé : si ce nombre est supérieur à 1, les identifiants « ID CARTO » associés sont exclus : il s'agit en majorité de jumeaux de même sexe (par exemple, chacun avec son BEN_NIR_ANO). Ils sont exclus pour permettre d'éviter de confondre leur trajectoire en une seule, notamment car ils sont indissociables l'un de l'autre dans les données hospitalières s'ils ne sont pas leur propre ouvrant droit.

[28] Le nombre total de bénéficiaires sur la période 2018-2022 est naturellement supérieur au nombre d'habitants en France à une date donnée, tel qu'estimé par l'Insee (entre 67,0 millions d'habitants au 1er janvier 2018 et 68,2 millions d'habitants au 1er janvier 2023). Par rapport au nombre d'habitants en début de période, le nombre de bénéficiaires potentiels inclut aussi notamment l'ensemble des naissances survenues au cours de la période (3,7 millions), les flux migratoires entrants (2,1 millions), les retraités résidant à l'étranger (1,1 million) et leurs ayants droit ainsi que l'ensemble des personnes de passage sur le territoire français pour moins de 12 mois avec des droits ouverts à l'Assurance Maladie (étudiants internationaux par exemple).

Définition des dates des événements

Pour chaque sous-système du SNDS, nous extrayons les événements minimaux interprétables suivants : l'identifiant bénéficiaire, le code du référentiel, la source (qui fournit le contexte, par la table et la variable dont l'information a été extraite : diagnostic principal vs. associé, acte, délivrance, etc.) et les dates. Les dates sont harmonisées en intervalles, de façon à prévenir la fuite d'information[29] lors d'une prédiction : une borne inférieure et supérieure de début, et une borne de fin. La borne supérieure de début est introduite afin de traiter les cas d'incertitude temporelle (par exemple, un diagnostic sur la durée d'un séjour) et prévenir la fuite d'information. Pour la grande majorité des événements (par exemple les délivrances de médicaments), l'ensemble de ces trois dates correspond, mais un soin particulier est apporté à la définition des trois concepts de date pour les séjours en MCO, SSR, HAD, en ESMS, ainsi que les périodes d'ALD.

À une date donnée t_0, l'ensemble des événements passés (resp. futurs) est défini comme l'ensemble des événements dont la borne supérieure (resp. inférieure) de début d'événement est strictement antérieure (resp. postérieure ou égale) à la date t_0. Par exemple, pour un diagnostic hospitalier, on considère que la borne supérieure du début de l'événement (le moment où le diagnostic est posé) est la date de sortie. Si le temps de prédiction correspond au milieu du séjour, l'ensemble des événements relatifs au séjour sera ainsi filtré, en considérant qu'ils n'appartiennent pas au passé.

La date de fin d'événement, quant à elle, ne sert pas à définir le passé mais plutôt à définir la date de l'événement, en la tronquant à t_0, ce qui est utile pour les informations relatives à des périodes longues (ALD, séjour en ESMS…) : après avoir filtré les événements du passé, si le temps de prédiction intersecte la période entre la date de début de l'événement et la date de fin de l'événement, pour qualifier le caractère actuel de l'information (ex. : la personne est encore en ALD), l'événement est daté au temps de prédiction (« tronqué »), plutôt qu'en date de début de l'événement. Pour un diagnostic en ALD, la date de début est connue avec précision : la borne supérieure et la borne inférieure de début sont identiques et correspondent à la date de début d'ALD. Ainsi, dès lors que le temps de prédiction est postérieur à cette date, cet événement fait bien partie du passé et la prédiction en tient compte. La borne de fin est égale à la date de fin d'ALD : si le temps de prédiction intervient avant la fin de l'ALD, l'événement « ALD » est considéré comme actif à t_0 et est alors daté (tronqué) au temps de prédiction.

Des filtres amont spécifiques à chaque source, fondés sur la documentation officielle, écartent les enregistrements hors périmètre, manifestement erronés ou redondants (par exemple, chevauchements connus entre ville et hôpital ou séjours de transfert).

Déduplication, contrôles de cohérence et projection pivot

Les tables événementielles dont l'unité d'observation est l'événement, décrit par [individu, date de début min, date de début max, date de fin, source, code d'événement] sont ensuite dédupliquées. Les doublons stricts sont supprimés ; les recouvrements inter-sources ne sont conservés que lorsqu'ils apportent une information clinique distincte (par exemple, diagnostics de séjour vs. diagnostics d'unité médicale, dont la sémantique diffère). Des contrôles transversaux vérifient la cohérence interne (dates de naissance/décès, validité du sexe, existence d'un code du référentiel et d'une source valable). Une fois ces contrôles passés, toutes les sources sont projetées vers un format pivot commun à deux tables.

[29] La "fuite d'information" est le fait de laisser le modèle utiliser, au moment où l'on prétend prédire, des informations qui ne sont en réalité connues qu'après (ou pendant) l'événement à prédire, ce qui gonfle artificiellement les performances.

Annexe 2. Comparaison de deux méthodes d'*embeddings* de codes

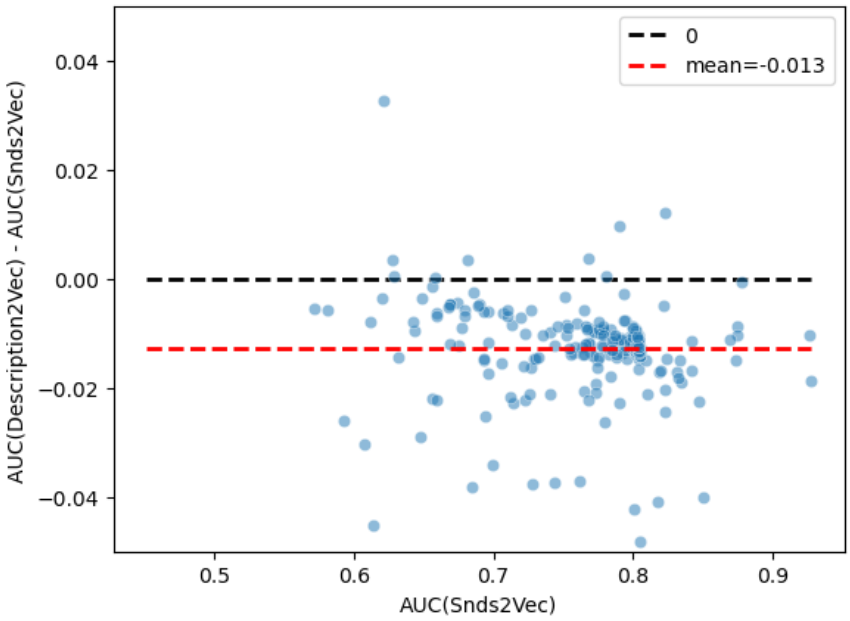
Deux candidats d'*embeddings* statistiques ont été testés pour représenter les codes d'événements sous forme de vecteurs :

1. **Cooccurrences** (`Snds2vec`) :
Réduction de la dimension de la matrice de cooccurrence des événements à moins de trente jours d'écart.

2. **Description NLP** (`Description2vec`) :
« CODER », un modèle de langage pré-entrainé spécialisé dans le domaine médical, appliqué aux libellés des événements dans les nomenclatures officielles.

En faisant varier la méthode d'embedding, en gardant les données d'entrées, les cibles et les autres paramètres de modélisation égaux par ailleurs, sur les plus de 180 cibles pour prédire une première hospitalisation, avec un entraînement sur 5 millions de patients et différentes variations, les AUC stratifiées indiquent une supériorité quasi-systématique, d'en moyenne de 1,3 point d'AUC de la première alternative sur la deuxième (graphique ci-dessous).

Graphique A1 Comparaison des résultats obtenus par xgboost entraîné sur un jeu de 5 000 000 de bénéficiaires, en utilisant les *embeddings* Snds2vec versus des *embeddings* issus des descriptions des codes, toutes choses égales par ailleurs



Lecture > Chaque point correspond à une cible, l'abscisse correspond à l'AUC stratifiée obtenue à partir des embeddings Snds2vec, en ordonnée le gain (la perte si négative) d'AUC stratifiée obtenue en utilisant plutôt les embeddings issus de CODER appliqué aux descriptions des codes. En moyenne, on observe une perte d'AUC stratifiée de 1,3 point.

Annexe 3. Typologie, hiérarchie, sources

Chaque événement collecté est décrit par un code issu d'une codification spécifique à sa source (par exemple, CIM-10 pour les diagnostics, CIP pour les médicaments en pharmacie). Ces diverses codifications possèdent des hiérarchies qui permettent d'analyser les événements à différents niveaux de granularité. Voici, à titre d'exemple, un code CIP (boîte de médicaments) suivi de tous ses codes parents :

- 3595583 (CIP) : DOLIPRANE 1 000 mg, comprimé - plaquette(s) thermoformée(s) PVC-aluminium de 8 comprimés
- 9239091 (UCD) : DOLIPRANE 1 000 mg CPR
- N02BE01 (ATC) : Paracétamol
- N02BE (ATC) : Anilides
- N02B (ATC) : Autres analgésiques et antipyrétiques
- N02 (ATC) : Analgésiques
- N (ATC) : Système nerveux

Comme on peut également le voir dans cet exemple, chaque code est associé à une description textuelle. Ces hiérarchies, ainsi que les descriptions associées, jouent un rôle essentiel dans la fiabilisation des données, la compréhension et l'analyse des événements et modèles (par exemple pour l'explicabilité) et dans les modélisations (en permettant par exemple de réduire le nombre de codes en « remontant » la hiérarchie).

Il s'est donc avéré utile de constituer un référentiel unique regroupant les différents référentiels disponibles en *données ouvertes*. Ce référentiel, comprenant environ 150 000 codes (y compris les codes parents issus des hiérarchies), contient une description textuelle de chaque code ainsi que la liste de ses ancêtres hiérarchiques.

Le *tableau A1* explicite la hiérarchie (ensemble des ancêtres) pour des codes d'événements de chaque typologie (sauf « urgence », sans ancêtre), codes dont les descriptions sont présentées en *tableau A2*. Le *tableau A3* liste, pour chaque typologie, les sources (table et variable) utilisées pour collecter les événements.

Tableau A1 Position de codes d'événements dans les hiérarchies

Code	Hiérarchie	Ancêtres						
lpp:1103570	lpp	lpp:chap01	lpp:chap01.01	lpp:chap01.01.03	lpp:chap01.01.03.01	lpp:chap01.01.03.01.01	lpp:chap01.01.03.01.01.01	lpp:1103570
cim10:Z04.801	cim10	cim10:XXI	cim10:Z00-Z13	cim10:Z04	cim10:Z04.8	cim10:Z04.801		
cim10:I64	cim10	cim10:IX	cim10:I60-I69	cim10:I64				
ccam:LFHC001	ccam	ccam:12	ccam:12.01	ccam:12.01.06	ccam:LFHC001			
nabm:1104	nabm	nabm:chap05	nabm:chap05-01	nabm:1104				
cip:3595583	atc	atc:N	atc:N02	atc:N02B	atc:N02BE	atc:N02BE01	ucd:9239091	cip:3595583
pfs_act_nat:28	pfs_act_nat	pfs_act_nat:cat9	pfs_act_nat:28					
pfs_spe:15	pfs_spe	pfs_spe:cat2	pfs_spe:15					
typ_um:53	typ_um	typ_um:chirurgie	typ_um:53					
etb_cat:200	etb_cat	etb_cat:rg2	etb_cat:rg44	etb_cat:rg4401	etb_cat:200			
dependance:cpt2	dependance	dependance:cpt	dependance:cpt2					
ghm:28Z07Z	ghm	ghm:28	ghm:28Z	ghm:28Z07	ghm:28Z07Z			
gme:0841A1	gme	gme:08	gme:0841	gme:0841A_	gme:0841A1			
ghpc:1012	ghpc	ghpc:mp09	ghpc:mp0900	ghpc:1012				

Tableau A2 Signification des codes d'événements en Tableau A1

Code	Description
lpp:1103570	AUTOCONTRÔLE DU GLUCOSE INTERSTITIEL, LECTEUR, ABBOTT, FREESTYLE LIBRE.
cim10:Z04.801	Examen et mise en observation pour polysomnographie
cim10:I64	Accident vasculaire cérébral, non précisé comme étant hémorragique ou par infarctus
ccam:LFHC001	Biopsie osseuse et/ou discale de la colonne vertébrale, par cœlioscopie ou par rétropéritonéoscopie
nabm:1104	HÉMOGRAMME Y COMPRIS PLAQUETTES (NFS , NFP)
cip:3595583	DOLIPRANE 1 000 mg, comprimé - plaquette(s) thermoformée(s) PVC-aluminium de 8 comprimé(s)
pfs_act_nat:28	ORTHOPHONISTE
pfs_spe:15	OPHTALMOLOGIE
typ_um:53	Autre chirurgie adulte (ou chirurgie indifférenciée adulte)
etb_cat:200	MAISON DE RETRAITE
dependance:cpt2	Dépendance de type comportement, de niveau 2
ghm:28Z07Z	Chimiothérapie pour tumeur, en séances
gme:0841A1	Arthroses du genou avec implant articulaire , score phy <= 8 - niveau 1
ghpc:1012	Pansements complexes et soins spécifiques (stomies compliquées) ; Pas de mode de prise en charge associé ; Index de Karnofsky valant 50

Tableau A3. Liste des sources (tables et variables) pour chaque typologie

Typologie	Type d'événements	Sources (table__variable)
PFS_ACT_NAT	Nature d'actes des professionnels de santé (hors médecin)	ER_PRS_F__PSE_ACT_NAT MCO_FBSTC__EXE_SPE MCO_FCSTC__EXE_SPE SSR_FBSTC__EXE_SPE SSR_FCSTC__EXE_SPE
ATC, UCD, CIP	Médicaments	ER_PHA_F__PHA_PRS_C13 ER_UCD_F__UCD_UCD_COD HAD_MED__UCD_UCD_COD HAD_MEDAPAC__UCD_UCD_COD HAD_MEDATU__UCD_UCD_COD HAD_MEDCHL__UCD_UCD_COD MCO_MED__UCD_UCD_COD MCO_MEDAPAC__UCD_UCD_COD MCO_MEDATU__UCD_UCD_COD MCO_MEDTHROMBO__UCD_UCD_COD MCO_FHSTC__UCD_UCD_COD SSR_MED__UCD_UCD_COD SSR_MEDAPAC__UCD_UCD_COD SSR_MEDATU__UCD_UCD_COD SSR_FHSTC__UCD_UCD_COD
NABM	Actes de Biologie	ER_BIO_F__BIO_PRS_IDE MCO_FLSTC__NABM_COD SSR_FLSTC__NABM_COD
PFS_SPE	Spécialités médicales consultées	ER_PRS_F__PSE_SPE_COD MCO_FBSTC__EXE_SPE MCO_FCSTC__EXE_SPE SSR_FBSTC__EXE_SPE SSR_FCSTC__EXE_SPE

CCAM	Actes médicaux	ER_CAM_F__CAM_PRS_IDE HAD_A__CCAM_COD MCO_A__CDC_ACT MCO_FMSTC__CCAM_COD SSR_CCAM__CCAM_ACT SSR_FMSTC__CCAM_COD
CIM-10	Diagnostics	IR_IMB_R__MED_MTF_COD HAD_B__DGN_PAL HAD_D__DGN_ASS HAD_DMPA__DGN_ASS_MPA HAD_DMPP__DGN_ASS_MPP MCO_B__DGN_PAL MCO_B__DGN_REL MCO_D__ASS_DGN MCO_UM__DGN_PAL MCO_UM__DGN_REL SSR_B__ETL_AFF SSR_B__FP_PEC SSR_B__MOR_PRP SSR_D__DGN_COD
LPP	Dispositifs médicaux	ER_TIP_F__TIP_PRS_IDE MCO_DMIP__TIP_PRS_IDE MCO_FPSTC__TIP_PRS_IDE
TYP_UM	Type d'unités médicales à l'hôpital	MCO_UM__AUT_TYP_UM
GHM	Groupe Homogène de Malades à l'hôpital (MCO)	MCO_B__GRG_GHM
DEPENDANCE	Cotation de la dépendance (SSR, HAD)	HAD_B__AVQ_ALI HAD_B__AVQ_CON HAD_B__AVQ_CPT HAD_B__AVQ_HAB HAD_B__AVQ_LOC HAD_B__AVQ_REL SSR_B__ALI_DEP SSR_B__CON_DEP SSR_B__CPT_DEP SSR_B__DPL_DEP SSR_B__HAB_DEP SSR_B__REL_DEP
Urgence	Passage aux urgences	ER_PRS_F__PRS_NAT_REF:urgence MCO_B__ENT_PRV:urgence MCO_UM__AUT_TYP_UM:urgence MCO_FBSTC__ACT_COD:urgence MCO_FCSTC__ACT_COD:urgence
GME	Groupe Médico-économique (SSR)	SSR_B__GRG_GME
GHPC	Groupe Homogène de Prise en Charge (HAD)	HAD_GRP__PAP_GRP_GHPC
ETB_CAT	Catégorie d'établissements	IR_ESM_R__ESM_CAT_COD

Source > Base SeqNDS, 2018-2022.

Annexe 4. Entraînement du modèle BEHRT

Notre approche a été de s'éloigner le moins possible de l'implémentation de Li, *et al*. (2020) tout en procédant aux adaptations nécessaires pour tenir compte des particularités de nos données.

Implémentation.

Li, *et al.* (2020) ont ouvert leur code qui est disponible ici : https://github.com/deepmedicine/BEHRT

L'implémentation retenue ici se fonde sur ces développements. Le code étant relativement ancien et s'appuyant sur l'ancêtre de la bibliothèque Python **transformers**, il a été nécessaire de migrer la base de code afin de lui permettre de tourner dans un environnement de data science récent.

La recherche des meilleurs hyperparamètres n'est pas réitérée, et les hyperparamètres retenus dans BEHRT original sont repris (*Hidden size* = 288, Layers = 6, *Attention Heads* = 12*, Intermediate Size* = 512).

Écart par rapport à Li, et al. (2020)

L'unité temporelle correspond à la date de l'événement, contre la visite (ville ou hôpital) rassemblant l'ensemble des codes diagnostic dans l'article original. En cohérence, le critère d'inclusion dans le jeu d'entraînement de BEHRT est adapté : filtre sur 5 dates distinctes d'événements contre 5 visites (chez le médecin de ville ou à l'hôpital) pour l'article original, ce qui permet d'avoir un échantillon beaucoup plus représentatif tout en maintenant le critère de séquence suffisamment longue de tokens pour justifier l'apprentissage.

Les embeddings de segments alternés A/B ne sont pas repris. Issus de BERT (NLP), qui distinguait les phrases entre elles, ils permettent dans BEHRT de distinguer une visite de la suivante, information *a priori* redondante dans notre cas puisque nous disposons de la position dans la séquence (embedding positionnel conservé en prenant pour unité temporelle la date). Le concept de « visite » s'applique naturellement aux séjours hospitaliers mais moins à l'ensemble de notre typologie d'événements. Pour la même raison, l'ajout de tokens « [SEP] » entre séjours n'est pas repris.

La longueur des séquences a été fixée à 512, contre 64 dans l'original, en cohérence avec les longueurs typiques observées dans SeqNDS.

La taille du vocabulaire a été fixée à 30 000, contre 301 dans l'original. Le passage du vocabulaire initial de 78 118 codes de SeqNDS à 30 000 se fait ainsi :

- on conserve tous les codes des catégories suivantes : PFS_ACT_NAT (29 codes distincts), NABM (1 092), PFS_SPE (55) TYP_UM (45), DEPENDANCE (18), urgence (1), ETB_CAT (26) ;
- on conserve tous les codes ATC à 7 caractères (1 906) ;
- on conserve tous les codes CIM-10 à 3 caractères (2 056) ;
- on conserve tous les codes de regroupement (codes non terminaux) issus de la hiérarchie CCAM (1 413) ;
- on conserve tous les codes de regroupement (codes non terminaux) issus de la hiérarchie LPP (875) ;
- les codes restants sont ensuite itérativement remplacés par leur code parent jusqu'à obtenir 30 000 codes, en choisissant les remplacements minimisant la perte d'entropie de Shannon ; ceci ajoute ainsi certains codes CIP (2 642), UCD (7 789), CIM-10 (3 979 codes à plus de 3 caractères), CCAM (2 104 codes terminaux), LPP (3 731 codes terminaux), GHM (1 226), GME (858), GHPC (155).

On conserve les mêmes hyperparamètres du réseau de neurones, mais l'augmentation de la taille du vocabulaire donne lieu à un modèle comprenant 30 millions de paramètres plutôt que 10 millions dans l'original.

Pour s'aligner avec nos autres modèles (PVT, BEHRT-SNDS, comorbidités…) et obtenir une comparaison juste, les mêmes restrictions du jeu d'entraînement pour la tâche de pré-entraînement et de fine-tuning sont appliquées, à savoir la restriction aux individus nés au cours de l'un des neuf mois spécifiques et aux événements antérieurs à mi-2021 (voir section **Entraînement**).

Pour la tâche de pré-entraînement (même tâche de MLM), la volumétrie de patients beaucoup plus importante (50 millions dans notre échantillon d'entraînement contre 1,6 millions pour l'original), justifie de n'entraîner que sur 5 epochs (52h par epoch) et non sur 100. En particulier, le modèle se sera entraîné sur plus de *batch* que le modèle original. Le *learning rate* a été choisi pour optimiser la *loss* (par rapport à ce nombre d'*epoch* fixé et un *batch size* inchangé de 256), à 0,0001, contre 0,00003 pour l'original.

Pour la tâche de fine-tuning, nous avons choisi une tâche similaire à la tâche dite « T3 » de l'original, consistant à prédire l'ensemble de nos cibles d'intérêts simultanément (décès, hospitalisation avec diagnostic incident selon le découpage de la CIM-10, infection, cancers, et cibles spécifiées par rapport aux enjeux de santé publique). La différence principale, en dehors des différences entre les cibles, est l'ajout de critères d'inclusion spécifiques à chaque cible pour définir la population « éligible », les « *log loss* » n'étant calculées que sur ces individus. Le learning rate a été choisi pour optimiser la loss, à 0,00002 (avec un *warm-up* de 10 000 batches).

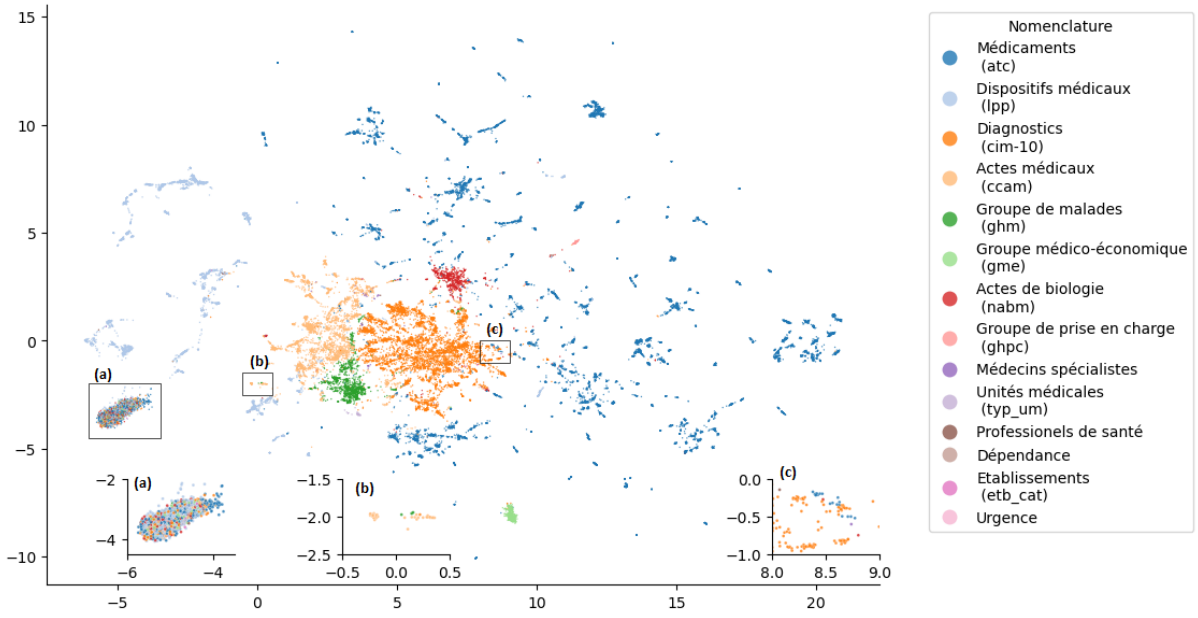
Représentation de l'espace sémantique des embeddings d'entrée du modèle BEHRT (après pré-entraînement et fine-tuning)

La position dans l'espace de dimension p = 288 est supposée refléter une forme de « sens » ou « contenu sémantique » utile au modèle BEHRT-SNDS dans ses tâches de prédiction, et, contrairement à Snds2vec, ces représentations sont apprises au cours de l'entraînement, et non pas fixes. Afin d'obtenir une représentation visuelle en deux dimensions, nous appliquons le même algorithme non linéaire de réduction de la dimension « UMAP, Uniform Manifold Approximation and Projection for Dimension Reduction » que celui appliqué pour représenter Snds2vec en *graphique A2*.

Alors même que la hiérarchie de codes et leur provenance d'une nomenclature particulière ne sont pas une information fournie au modèle, l'espace sémantique issu de l'entraînement reflète globalement une répartition dans l'espace par grande typologie (diagnostics, dispositifs médicaux, médicaments, actes médicaux…). Si les codes en provenance de deux nomenclatures différentes sont moins souvent co-localisés que pour Snds2vec, c'est néanmoins le cas de certaines régions, dont nous tirons deux exemples. La région (b) figure pour l'essentiel des actes liés à la radiothérapie à l'hôpital (irradiations ccam : YYYY047, et plus largement les actes de radiothérapie, champs fixes, 19.1.10.1, ou de préparation à la radiothérapie, ccam : ZZMK018, ou des gestes complémentaires thérapeutiques 18.2.17.2 dans le contexte d'une irradiation externe ccam : ZZML002), mais aussi le contexte via le groupe homogène de malades (Séances de préparation à une irradiation externe 28Z19, 28Z20 et 28Z22). La zone (c) figure des diagnostics CIM-10 ainsi que des médicaments en lien avec la santé mentale (troubles mentaux et du comportement liés à l'utilisation de substances psychoactives, schizophrénie, troubles de l'humeur ; lithium indiqué dans les troubles bipolaires, médicaments pour sevrer l'alcoolisme…) mais aussi les consultations de psychiatres et des actes de biologies de dosage de lithium et phosphatases alcalines.

Enfin, la zone (a) rassemble plus de 3 000 codes très hétérogènes ; elle est pour l'essentiel composée des codes les moins fréquents, que le modèle distingue moins facilement.

Graphique A2 Représentations vectorielles des codes issus de BEHRT-SNDS



Drees Méthodes
N° 25 • avril 2026

Prédire la suite d'un parcours de soins dans le système national des données de santé

Directeur de la publication
Thomas Wanecq

Responsable d'édition
Valérie Bauer-Eubriet

ISSN
2740-3564