RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Drees
DIRECTION DE LA RECHERCHE, DES ÉTUDES, DE L'ÉVALUATION ET DES STATISTIQUES

DREES MÉTHODES

N° 8 • mars 2023

Les statistiques provisoires sur les causes de décès en 2018 et 2019

Une nouvelle méthode de codage faisant appel à l'intelligence artificielle

François Clanché et Nirintsoa Razakamanana (DREES), Elise Coudin et Aude Robert (CépiDc-Inserm)

STATISTIQUE PUBLIQUE
La DREES fait partie du Service statistique public piloté par l'Insee.

Inserm CépiDc

Les statistiques provisoires sur les causes de décès en 2018 et 2019

Une nouvelle méthode de codage faisant appel à l'intelligence artificielle

François Clanché et Nirintsoa Razakamanana (DREES), Elise Coudin et Aude Robert (CépiDc-Inserm)

Le travail retracé par ce document a été réalisé par Nirintsoa Razakamanana (DREES) dans le cadre d'un groupe de travail rassemblant Elise Coudin, Aude Robert, Haris Medjahed (CépiDc), François Clanché (DREES) et Rémi Flicoteaux (AP-HP). Les auteurs remercient pour leur relecture et précieuses suggestions Walid Ghosn, Diane Martin, Zina Hebbache (Inserm CépiDc), Grégoire Rey (Inserm-France Cohortes), Vianney Costemalle, Fabrice Lenglart, Elisabeth Ferry-Lemonnier, Mathilde Gaini et Diane Naouri (DREES)

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

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

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Drees
DIRECTION DE LA RECHERCHE, DES ÉTUDES, DE L'ÉVALUATION ET DES STATISTIQUES

DREES MÉTHODES

N° 8 • mars 2023

Synthèse
Les statistiques provisoires sur les causes de décès en 2018 et 2019

Une nouvelle méthode de codage faisant appel à l'intelligence artificielle

François Clanché et Nirintsoa Razakamanana (DREES), Elise Coudin et Aude Robert (CépiDc-Inserm)

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

STATISTIQUE PUBLIQUE
La DREES fait partie du Service statistique public piloté par l'Insee.

Inserm CépiDc

SYNTHÈSE

Dans le cadre du Projet de rattrapage et de refonte de la statistique des causes de décès la DREES a élaboré, en étroite collaboration avec le CépiDc de l'Inserm, une méthode de codage des textes des certificats médicaux de décès faisant appel à l'intelligence artificielle.

Grâce à ce travail, il a été possible de diffuser en mars 2023 des statistiques provisoires relatives aux causes de décès des années 2018 et 2019.

Ce document présente les motivations et le contexte de ce travail (spécificités de la source, conjoncture spécifique de l'année 2022, impossibilité d'utiliser des approches statistiques habituelles), détaille les choix de méthode réalisés (utilisation des modèles de type *Transformers* pour traduire des textes en séquences de codes, choix des causes initiales) et la façon dont la méthode a été validée (prédictions et analyse de performance sur les données de 2016 et 2017). Il présente enfin le dispositif appliqué aux données de 2018 et 2019 et analyse ses premiers résultats,

Les prédictions réalisées sur les données de 2016 et 2017 (années pour lesquelles on peut comparer les causes prédites par le modèle avec les codes effectivement attribués par les codeurs du CépiDc), ainsi que la plausibilité des données de 2018 et 2019 (par comparaison de tendance avec celles de 2015 à 2017 et celles de 2020) montrent la bonne capacité de traitement de cette approche. Elles mettent aussi en lumière ses limites.

L'IA combinée avec le codage automatique sur règles Iris/Muse prédisent correctement la cause initiale au niveau de la CIM-10 dans 93,0 % des cas, et 95,4 % lorsque l'on se concentre sur le regroupement de causes diffusé par Eurostat (*European short-list*). Cependant, la qualité de la prédiction n'est pas homogène selon les causes. En particulier, certaines causes de décès sont soit moins systématiquement repérées par l'IA, soit trop systématiquement prédites : il s'agit notamment des maladies infectieuses (tuberculose, VIH/SIDA), des maladies du sang, de l'appareil digestif, des intoxications accidentelles, des accidents, des homicides et d'autres causes externes. Ces cas couvrent donc certains effectifs très faibles, des causes de type « autres » mais aussi certains cas d'intérêt spécifique en santé publique à suivre de près.

Ce travail novateur devra être poursuivi dans plusieurs directions pour finalement mettre à disposition des données définitives sur les années 2018 et 2019, grâce à des compléments d'expertise et des adaptations des modèles là où c'est encore nécessaire. Ensuite, le CépiDc envisage une approche très similaire pour élaborer dans les mois qui viennent les données de 2021 puis de 2022, puis une intégration à moyen terme dans son système de production régulier.

Ce travail fournit un exemple supplémentaire des apports de l'intelligence artificielle dans la production de la statistique publique sur une tâche de classification, tâche pour laquelle la performance de ces méthodes est suffisamment bonne et prouvée pour pouvoir envisager des utilisations régulières.

SOMMAIRE

INTRODUCTION

La statistique sur les causes des décès, codifiée par l'Organisation mondiale de la santé (OMS), est établie en France par le CépiDc de l'Inserm. La production de cette statistique a connu ces dernières années des difficultés importantes qui se sont traduites par des retards de publication au regard du règlement statistique européen qui couvre ce domaine et des besoins de nombreux utilisateurs

En décembre 2022, la France a transmis à Eurostat des données provisoires sur les causes de décès pour les années 2018 et 2019. Cette transmission revêt un caractère particulier : pour environ 38 % de ces données la séquence des causes de décès et la cause initiale du décès ont été prédites par une approche d'intelligence artificielle, impliquant des algorithmes de *deep learning* entraînés à classer du texte dans la classification internationale des maladies (CIM 10) à partir des textes et des codes des certificats de décès des années précédentes.

Cette approche d'intelligence artificielle (IA) a été développée par la DREES en étroite collaboration avec le CépiDc dans le cadre du Projet de refonte du processus de production des statistiques sur les causes de décès[1].

L'approche d'intelligence artificielle s'appuie et étend les travaux de Falissard (2021) et Falissard et al. (2022) menés au CépiDc ces dernières années qui montrent que cette tâche de codage, de classification, peut être assimilée à une tâche de traduction linguistique. Les algorithmes de type *transformers* sont particulièrement performants sur ce type de tâche dépassant les performances atteintes par les méthodes d'apprentissage statistique classiques testées lors de campagnes dédiées (voir Névéol et al 2018). Les travaux menés au Cépidc n'avaient cependant pas été jusqu'à l'entraînement d'algorithmes complets, de la cause par ligne, à la séquence de causes et à la détermination de la cause initiale et surtout n'avaient pas été jusqu'à composer les échantillons d'apprentissage et de test dans une optique de prédiction de l'année suivante. Ce travail les étend donc et les complète en ce sens.

Ce document présente, en partie I, les éléments de contexte relatifs à cette tâche de classification des textes remplis par les médecins certificateurs en codes de la classification internationale des maladies, ainsi que les raisons pour lesquelles une approche statistique standard ne résout pas le problème. La deuxième partie décrit la méthode retenue et analyse sa performance sur les deux années 2016 et 2017 utilisées comme test. La troisième partie se concentre sur la prédiction des années 2018 et 2019, et, s'appuyant sur l'analyse du test sur 2016 et 2017 et des comparaisons de tendances avec les années précédentes et suivante, relève les catégories de causes pour lesquelles la méthode retenue est performante et celles pour lesquelles il faut s'attendre au niveau populationnel à des sous- ou des surestimations, qu'il conviendra de corriger autant que faire se peut dans la version définitive des fichiers.

[1] Projet engagé en février 2022 suite à un rapport Igésr-IGAS-IGInsee d'avril 2021, qui vise à résoudre le retard structurel de production de la statistique nationale des causes de décès et à améliorer durablement les conditions de production de ces statistiques. Ce travail ne constitue qu'une partie du Projet de refonte, dont les objectifs sont multiples et les axes d'action variés.

UN BESOIN STATISTIQUE SPECIFIQUE

Les statistiques sur les causes de décès sont établies, conformément aux recommandations de l'Organisation mondiale de la santé, sur la base du texte du certificat rédigé par le médecin qui authentifie le décès. Les autorités statistiques qui traitent ces données médico-administratives doivent analyser ce texte en suivant les règles de codage définies par l'OMS et classer, c'est-à-dire coder / associer aux textes des certificats les codes des causes de décès tels qu'ils sont décrits dans la nomenclature de la Classification internationale des maladies (CIM version 10).

Les autorités statistiques doivent aussi à partir de ces causes codées déterminer la cause initiale de décès (*underlying cause*), c'est-à-dire la cause ou la circonstance en cas de mort violente, à l'origine du processus morbide ayant entraîné le décès. Elles suivent pour cela aussi les règles de détermination fournies par l'OMS, conçues pour assurer autant que possible une homogénéité des traitements dans le temps et dans l'espace afin notamment de réaliser des comparaisons internationales.

Le CépiDc n'avait pas entamé le codage des décès des années 2018 et 2019 quand la crise sanitaire du Covid-19 l'a poussé à concentrer ses moyens sur les décès de l'année 2020. Afin, d'une part, de respecter les engagements statistiques internationaux de la France et d'autre part d'élaborer une première estimation de la situation de la mortalité en France avant la crise sanitaire, le Projet a développé une méthode de prédiction des causes initiales de décès pour les certificats des années 2018 et 2019 qui n'avaient pas fait l'objet d'une codification automatique

La statistique sur les causes de décès

La statistique la plus courante sur les causes de décès consiste en une répartition de l'ensemble des décès selon leurs causes initiales. On appelle cause initiale d'un décès la maladie, ou les circonstances en cas de mort violente, à l'origine du processus morbide ayant entraîné le décès. À titre d'exemple, la répartition par cause initiale des décès survenus en France en 2020 est présentée dans le tableau 1.

Tableau 1 • Principales causes initiales de décès en 2020

Cause de mortalité	Hommes	Femmes	Ensemble
Toutes causes	**334 034**	**333 462**	**667 496**
Tumeurs	**95 562**	**75 244**	**170 806**
dont tumeur maligne de l'œsophage	2 736	894	3 630
dont tumeur maligne du côlon. rectum et anus	9 110	8 087	17 197
dont tumeur maligne du pancréas	6 285	6 191	12 476
dont tumeur maligne de la trachée. des bronches et du poumon	21 021	9 914	30 935
dont tumeur maligne du sein	208	12 800	13 008
Maladies cardio-neurovasculaires	**63 630**	**71 133**	**134 763**
dont cardiopathies ischémiques	19 089	12 003	31 092
dont autres maladies du cœur	20 891	27 170	48 061
dont maladies cérébrovasculaires	13 157	17 955	31 112
dont autres maladies cardio-neurovasculaires	10 493	14 005	24 498
Covid-19	**35 077**	**34 161**	**69 238**
Maladies de l'appareil respiratoire	**20 285**	**18 426**	**38 711**

dont pneumonie	5 663	5 896	11 559
dont maladies chroniques des voies respiratoires inférieures	5 933	4 160	10 093
Maladies du système nerveux et des organes des sens	**15 315**	**22 301**	**37 616**
dont maladie d'Alzheimer	4 970	13 274	18 244
dont maladie de Parkinson	3 980	3 032	7 012
Maladies endocriniennes. nutritionnelles et métaboliques	**10 643**	**12 955**	**23 598**
dont diabète sucré	6 062	6 202	12 264
Troubles mentaux et du comportement	**9 937**	**15 451**	**25 388**
dont démence	5 902	12 694	18 596
Maladies de l'appareil digestif	**13 394**	**11 581**	**24 975**
Maladies de l'appareil génito-urinaire	**5 832**	**6 260**	**12 092**
Maladies infectieuses et parasitaires	**5 346**	**5 709**	**11 055**
Maladies du sang et des organes hématopoïétiques et certains troubles du système immunitaire	**1 251**	**1 550**	**2 801**
Symptômes et états morbides mal définis	**30 600**	**37 172**	**67 772**
Causes externes de morbidité et mortalité	**23 302**	**16 731**	**40 033**
dont accidents de transport	1 662	482	2 144
dont suicides et lésions auto-infligées	6 737	2 249	8 986
dont chutes accidentelles	4 237	4 836	9 073

Note > Les causes de décès sont présentées par grands chapitres et catégories de la CIM10.
Champ > Décès de personnes résidant et décédées en France en 2020.
Source > CépiDc, 2020.

La statistique européenne dans ce domaine est définie par le règlement européen EU n°328/2011[2]. Les pays fournissent des causes initiales classées au niveau le plus fin de la version 10 de la CIM, mais Eurostat ne les publie que sous un format plus agrégé de 86 postes appelé *European short list*[3].

[2] Le règlement européen n°328/2011 d'avril (L_2011090FR.01002201.xml (europa.eu) stipule que chaque pays doit envoyer à Eurostat, au plus tard 24 mois apres la fin de l'année de référence, les effectifs de décès classés par cause initiale. Le règlement donne la possibilité soit d'envoyer à Eurostat des données individuelles (micro data), soit des données agrégées. Mais le niveau de détail des données agrégées (sexe, âge quinquennal, région du décès, région ou pays de résidence, cause détaillée, mois du décès) est tel que l'habitude de la France (et de nombreux pays) est d'envoyer un fichier de micro-données.

[3] *L'Europan short list* est une nomenclature de diffusion statistique des données sur les causes de décès construite par regroupement de positions de la CIM 10. L'ensemble des causes sont réparties en 70 catégories élementaires et, avec les différents regroupements, ce sont 86 items qui sont systématiquement diffusés. Cette nomenclature est présentée sur l'espace consacré aux nomenclaures sur le site internet d'Eurostat : Europa - RAMON - LST_NOM_DTL (https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=COD_2012&StrLanguageCode=EN&IntPcKey=&StrLayoutCode=HIERARCHIC&IntCurrentPage=1). Presque systématiquement, les regroupements de causes de décès diffusées pour diffuser les données en France correspondent à des postes de cette nomenclature.

La partie médicale des certificats de décès

Les données permettant d'établir les causes de décès sont tirées des certificats de décès.

Chaque décès survenu sur le territoire national donne lieu à la rédaction, par un médecin, d'un document officiel attestant du décès : le certificat de décès. La forme de ce document est définie réglementairement[4] et sa partie médicale se conforme aux prescriptions établies par l'OMS. Un certificat de décès validé est nécessaire pour fermer un cercueil, établir l'acte d'état civil de décès et procéder à une inhumation.

Le certificat de décès comprend deux volets, l'un administratif et l'autre médical (voir annexe 1).

Son premier volet, nominatif, est de nature administrative et comprend, entre autres, la commune de décès, l'état civil du défunt (nom, prénoms, date et lieu de naissance, adresse), la date et l'heure de la mort, ainsi que des informations nécessaires à la délivrance de l'autorisation de fermeture du cercueil et à la réalisation des opérations funéraires. Il est utilisé par la mairie du lieu de décès pour rédiger l'acte d'état civil du décès, acte lui-même transmis à l'Insee par la mairie. L'Insee utilise l'acte de décès pour mettre à jour le Répertoire d'identification des personnes physiques (RNIPP), mais aussi pour établir les statistiques de décès.

Le second volet du certificat, non nominatif, a un caractère médical et confidentiel. Il est anonyme et ne comporte ni le nom, ni les prénoms de la personne décédée. Il renseigne en revanche sur la commune de décès, la commune de domicile, la date de naissance du défunt, son sexe et la date de décès. Puis le médecin y décrit les « maladies ou affections morbides ayant directement provoqué le décès » et la « séquence morbide » qui y a conduit (partie I), ainsi que, s'il y a lieu, les « autres états morbides, facteurs ou états physiologiques ayant contribué au décès » (partie II). Le texte rempli dans la partie I et les facteurs ajoutés dans la partie II constitue les **causes de décès**. Le médecin certificateur consigne également des informations sur le lieu du décès (domicile, milieu hospitalier,), l'état de grossesse, les circonstances apparentes de décès (mort subite, accident, suicide, décès lors d'une activité professionnelle…), ainsi qu'une éventuelle demande de recherche médicale ou médico-légale de la cause du décès.

La partie relative aux causes de décès est remplie sous la forme d'un texte libre, selon le principe que, sa rédaction engageant la responsabilité personnelle du praticien, celui-ci est libre d'établir et de rédiger son diagnostic de décès.

Le volet médical du certificat est cacheté (pour le format papier) par le médecin pour en garantir le caractère confidentiel vis-à-vis des proches et des services de la mairie. Il n'est consultable que par les médecins de l'agence régionale de santé (ARS), puis les spécialistes du CépiDc.

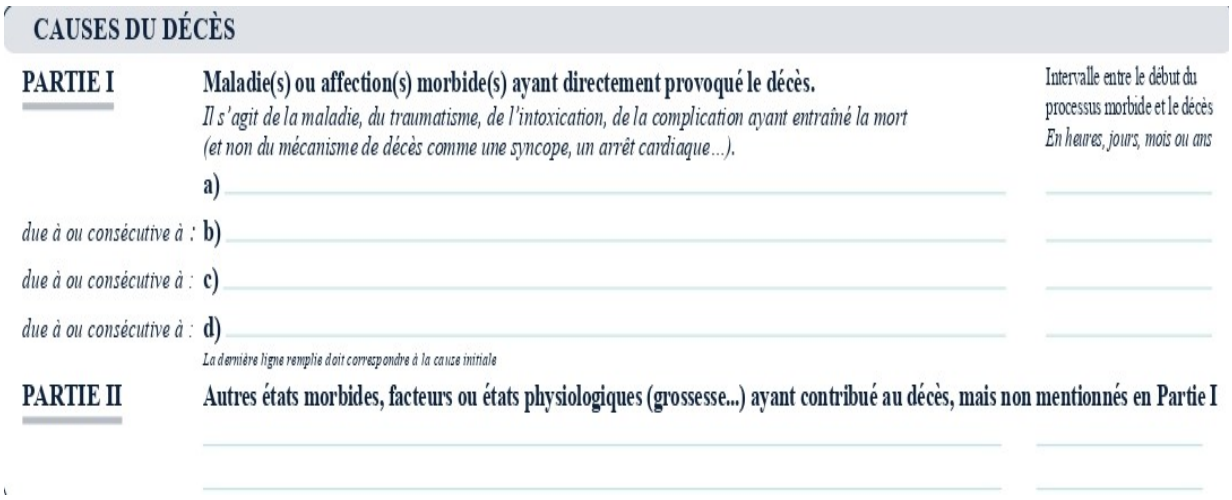
Le texte rempli par le médecin décrit, dans la majorité des cas,

- dans les 4 premières lignes (partie I) le processus morbide, c'est-à-dire l'enchaînement des causes qui a directement provoqué le décès. Normalement la première ligne correspond à la cause immédiate, et la dernière ligne doit décrire la cause initiale. Le médecin est invité à remplir une cause par ligne, chaque cause indiquée sur une ligne étant « due » à celle indiquée sur la ligne suivante.
- dans les lignes 5 et 6 (partie II) les autres états morbides, facteurs ou états physiologiques (grossesse, …) ayant contribué au décès, mais non mentionnés dans la partie I parce que non impliquées directement dans l'enchaînement des causes ayant conduit au décès.

Ce modèle (Parties I et II, relation de conséquence des lignes entre elles) est international et les règles de codage de l'OMS reposent sur cette structure (voir annexe 3 pour quelques exemples).

[4] Articles L2223-42 (https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000023711931/2022-03-14/) et R2213-1 du Code général des collectivités locales. L'Arrêté du 17 juillet 2017 relatif aux deux modèles du certificat de décès (NOR PRMX1720890A) détaille le contenu et la forme du certificat. Il précise notamment qu'un certificat spécifique doit être utilisé pour les décès d'enfants de moins de 28 jours (modèle de certificat néonatal). Dans la mesure où l'usage de ce modèle est quantitativement très minoritaire (environ 2 000 décès par an), même s'il est essentiel pour les politiques de santé périnatale, son traitement statistique n'a pas été abordé dans ce travail

Extrait du modèle de certificat de décès



CAUSES DU DÉCÈS

PARTIE I — Maladie(s) ou affection(s) morbide(s) ayant directement provoqué le décès.
Il s'agit de la maladie, du traumatisme, de l'intoxication, de la complication ayant entraîné la mort (et non du mécanisme de décès comme une syncope, un arrêt cardiaque...).

Intervalle entre le début du processus morbide et le décès
En heures, jours, mois ou ans

a) _____
due à ou consécutive à : b) _____
due à ou consécutive à : c) _____
due à ou consécutive à : d) _____
La dernière ligne remplie doit correspondre à la cause initiale

PARTIE II — Autres états morbides, facteurs ou états physiologiques (grossesse...) ayant contribué au décès, mais non mentionnés en Partie I

En pratique, la première ligne de la partie I est quasi systématiquement toujours remplie, la deuxième, 3 fois sur 4, et la troisième 4 fois sur 10 (cf. annexe 2). Dans un tiers des cas, le médecin certificateur va indiquer une ou plusieurs informations dans la partie II[5].

La longueur du texte par certificat est très variable : en moyenne 7 mots, 63 caractères, avec une moitié des certificats entre 4 et 9 mots, un quart de certificats très -parfois trop- courts (moins de 4 mots), et un autre quart plus riche.

On ne s'intéresse pas dans ce document à la façon dont ces certificats sont rédigés, collectés, ni aux travaux consistant à saisir les contenus rédigés sur papier[6] ni à vérifier l'exhaustivité de la source. On s'intéresse uniquement à la façon dont ce texte va être transcrit en codes dans la nomenclature internationale des maladies, la version 10 de la CIM.

Les règles de codage des causes de décès

Afin d'être analysées quantitativement, ces textes doivent faire l'objet d'une codification, dont les règles sont définies de façon précise par l'Organisation mondiale de la santé (OMS) dans les trois volumes de la Classification internationale des maladies (CIM).

La nomenclature actuellement en vigueur de la classification internationale des maladies, la CIM10[7] comporte 11 800 positions réparties en 17 grands chapitres (décrites dans le volume 1 de la classification), un index (volume 3), ainsi que des règles pour définir les causes de décès possibles, et choisir, parmi elles, la cause initiale suivant le modèle international du certificat de décès (volume 2). Même si on code depuis 2011 en CIM10, les versions de la classification utilisées et les règles de codage ont connu des révisions importantes sur certains chapitres notamment en 2016 et 2019.

Tous les postes de la CIM ne constituent pas une cause initiale de décès[8]. En 2017, en France métropolitaine, 2 800 positions différentes de la CIM ont été mobilisées pour décrire une cause initiale de décès, avec une forte concentration : les 450 postes qui apparaissent au moins 100 fois représentent 95 % des occurrences. À l'autre extrémité, 1500 codes apparaissent moins de 10 fois dans l'année (ce qui rend leurs analyses statistiques non significatives) et pèsent pour 0,8 % du total.

[5] L'ensemble des indications quantitatives décrivant le degré de remplissage de certificats sont tirées de l'analyse des certificats rédigés en 2016.

[6] le certificat peut être rempli électroniquement ou sur papier.

[7] La dernière version en française de la nomenclature se trouve ici ICD-10 Version:2008 (who.int), mais seule la version anglaise tient compte des évolutions récentes liées au Covid : ICD-10 Version:2019 (who.int) https://icd.who.int/browse10/2019/en#/U00-U49. En effet les versions de la classification utilisées ainsi que les règles de codage peuvent être adaptées chaque année, avec des révisions parfois importantes sur certains chapitres, notamment en 2016 et 2019

[8] Les règles d'analyse et de codage l'OMS considèrent que certaines maladies ne peuvent pas, par nature, constituer à elles seules des causes initiales de décès. C'est le cas par exemple d'une verrue virale, d'une lésion, ou d'une otite externe. Par ailleurs, dans les faits, il y a des causes possibles qu'on ne retrouve pas dans la liste des causes effectivement mentionnées une année donnée.

Le travail de codage consiste donc en 2 étapes :

1. Repérer dans le texte du médecin toutes les expressions de la CIM mentionnées pour décrire le processus morbide ainsi que les autres causes mentionnées en partie II du volet médical notamment

En moyenne l'analyse complète d'un texte de certificat (partie I + partie II) permet de repérer 3,6 causes au sens « expressions de la CIM ». La situation la plus fréquente est celle où il y a trois causes mentionnées (22 %). Les situations où il n'y a qu'une seule cause (14 %) et deux causes (20 %) ne sont pas rares.

La difficulté de cette étape tient dans la différence, dans de nombreux domaines, entre la langue utilisée par le médecin, ses abréviations, et les titres des catégories de la CIM.

Le tableau qui suit donne quelques exemples, simples et fréquents, du transcodage nécessaire entre les expressions effectivement trouvées dans les certificats et celles de la classification. Dans certains cas s'y rajoutent les fautes d'orthographe (ici carconome pour carcinome) et les abréviations (ici oap Oedeme aigu pulmonaire et bpco Bronchopneumopathie chronique obstructive)[9].

Texte rédigé par le médecin	Code CIM 10	Titre de la catégorie de la CIM
arret cardiaque	I46.9	Arrêt cardiaque, sans précision
Bronchopneumopathie	J18.0	Bronchopneumopathie, sans précision
cancer du colon	C18.9	Tumeur maligne du côlon sans précision
ischémie aigue des membres inférieurs	I74.3	Embolie et thrombose des artères des membres inférieurs
ischémie mésentérique	K55.0	Troubles vasculaires aigus de l'intestin
carconome hepatique	C22.9	Tumeur maligne du Foie, sans précision
Oap	I50.1	Insuffisance ventriculaire gauche
Bpco	J44.9	Maladie pulmonaire obstructive chronique, sans précision

Par ailleurs, les médecins ne suivent pas toujours l'invitation à ne mentionner qu'une cause par ligne. Avoir plusieurs causes par ligne complique l'interprétation des liens de conséquence entre les causes. En moyenne, on trouve 1,38 expressions par ligne, avec une grande différence entre la partie I et la partie II du certificat.

Dans les lignes 1 à 4 (Partie I) la moyenne est de 1,25 (80 % de lignes avec une seule cause, 15 % avec 2 causes et 5 % avec plus de 3 causes. Ces situations de multiplicité des causes sur une même ligne empêchent souvent le codage automatique. Dans la partie II du certificat, il y a en moyenne plus de deux causes, avec un moindre impact sur le codage automatique car il n'y a pas de notion de type « consécutif à » dans cette partie du texte.

[9] Afin de coder autant que possible les textes de façon automatique, le CépiDc maintient aussi un index francophone de 157 000 expressions où chaque terme est relié à un ou plusieurs codes de la CIM. Il s'agit de l'index (volume 3) de la CIM traduit et enrichi. Il réalise également des opérations de standardisation automatique pour gérer des synonymes, les abréviations, repérer des expressions régulières, supprimer des termes bloquant le codage et ainsi « nettoyer » le texte. Cet index et ces règles de standardisation sont mobilisés par le système expert de règles Iris pour aboutir à un codage systématique soit complètement automatique, soit en guidage de l'équipe de codage. Voir l'encadré « Iris/Muse »

Ci-dessous quelques exemples de multiplicité des causes sur une même ligne.

Texte rédigé par le médecin	Titres et codes des différentes causes repérées
choc cardiogénique sur oap	Choc cardiogénique (R57.0), Insuffisance ventriculaire gauche (I50.1)
cancer du poumon metastatique	Tumeur maligne Bronche ou poumon, sans précision (C349), Tumeur maligne secondaire d'autres sièges non précisés (C799)
adk vésicule biliare avec carcinose péritonéale	Tumeur maligne de la vésicule biliaire (C23), Tumeur maligne secondaire du rétropéritoine et du péritoine (C78.6)
surinfection sévère de bpco post-tabagique et sur silicose L	Maladie pulmonaire obstructive chronique, sans précision (J44.9), Troubles mentaux et du comportement liés à l'utilisation de tabac (F179) Pneumoconiose due à d'autres poussières de silice (J62.8)

2. Puis choisir parmi ces causes la cause initiale, les autres étant appelées « causes associées ».

L'étape suivante consiste à déterminer la cause initiale du décès. Il s'agit d'appliquer un ensemble de règles définies par l'OMS (volume 2) et qui reposent sur la plausibilité de l'enchaînement des causes entre elles, tel que l'a décrit le médecin ainsi que sur des règles spécifiques convenues pour mettre en avant certaines pathologies dont il y a intérêt en santé publique de suivre les évolutions, voir Rey (2016).

Un médecin déclare en partie I du volet médical que le décès d'un homme âgé 65 ans, est dû à une « défaillance multiviscérale consécutive à une insuffisance rénale chronique elle-même due à une fibrillation auriculaire », et, en partie II, que la personne était paraplégique. De toutes ces indications, l'opération de codage va retenir 4 causes dans la nomenclature internationale : R688 « Autres symptômes et signes généraux précisés », N189 « Insuffisance rénale chronique, sans précision », I489 « Fibrillation et flutter auriculaires, sans précision » et G822 « Paraplégie, sans précision ».

Au vu de la présentation du médecin certificateur de ces affections, de la logique médicale de leur enchaînement causal et des règles définies par l'OMS, on considère que la cause initiale du décès, celle qui a déclenché la séquence conduisant au décès, est I489 « Fibrillation et flutter auriculaires, sans précision », les autres causes sont considérées comme des causes associées.

Le CépiDc[10], comme la plupart des organismes producteurs de cette statistique, utilise un système expert pour analyser les certificats, repérer les causes citées par le médecin certificateur (standardisation, index) et choisir la cause initiale (règles déterministes et systématiques). Ce système expert est mobilisé soit pour aboutir à un codage purement automatique, soit pour guider l'équipe de codage. Il s'agit du logiciel Iris/MUSE (voir encadré).

Encadré 1 • Iris/Iris Muse

Ce logiciel de codage international utilisé pour le codage automatique des causes de décès est un système-expert composé de deux modules/éléments. Le premier est dépendant de la langue utilisée et donc maintenu par chaque pays (avec des échanges avec les pays de même langue) le second est international. Ce logiciel est maintenu par l'institut Iris. Le CépiDc, qui a fortement contribué à la mise en place de ce logiciel, l'utilise depuis 2011.

Le premier module permet de transformer du texte en une séquence de codes. Il est composé de deux éléments :

- des règles de standardisation. En France, il s'appuie sur environ 1000 expressions régulières permettant de gérer les synonymes, les abréviations, de supprimer des termes qui bloquent le codage,

- un index, c'est-à-dire un dictionnaire de termes utilisés par les médecins certificateurs et traduits en codes de la CIM. En France cet index comporte 157 000 expressions ou chaque terme est relié à un ou plusieurs codes de la CIM10. C'est ce

[10] Pour plus de détails sur la production des causes de décès voir la documentation les Statistiques sur les causes de décès de A à Z sur le site du CépiDc. https://www.cepidc.inserm.fr/qui-sommes-nous/les-statistiques-sur-les-causes-medicales-de-deces-de-z

module qui permet de définir les codes de la séquence de causes. Il s'agit de l'index (volume 3) de la CIM traduit en français par les pays francophones et enrichi par le CépiDc.

Le deuxième module (MUSE) est l'ensemble des règles de choix et de priorité de l'OMS, des règles de codages qui permet notamment de déterminer la cause initiale. Il s'agit de tables de décision, de la formalisation algorithmique des instructions incluses dans le volume 2 de la CIM-10. Il y a environ 30 millions de règles implémentées (pour capter les relations entre des paires de codes de la CIM). Ces tables de décision ont d'abord été développées pour le système ACME (le précédent), « traduites » de la CIM 8 vers la CIM 9 puis vers la CIM 10. Elles sont désormais maintenues par le système Muse. Chacun des modules est mis à jour régulièrement par les pays utilisateurs en ce qui concerne les deux premiers éléments et par l'Institut Iris en fonction des mises à jour officielles annuelles de l'OMS (voir le tableau ci-dessous).

Année de décès	Version de la CIM 10 (révisions majeures seulement)	Version du logiciel Iris et des tables de décision
2011	2011	ACME tables-Y2011S2 et Styx et Iris
2012	2012	ACME tables-Y2012S1 et Iris 4.0.52
2013	2013	ACME tables-Y2013S1, Iris 4.3.0
2014	2014	ACME tables-Y2014S2 et Iris 4.3.0
2015	2015	ACME tables-Y2015S1 et Iris 4.5.6
2016	2016	Iris 5.4.0, Muse specV2016SR10
2017	2016	Iris 5.4.4, Muse specV2018SR10
2018	2016	Iris 5.5.0, Muse 2.6
2019	2019	Iris 5.6.0, Muse 2.7.1
2020	2019	Iris 5.7.0, Muse 2.8

Versions de la CIM-10 et versions du logiciel de codage automatique utilisé par le CépiDc. Source : metadata

https://ec.europa.eu/eurostat/cache/metadata/EN/hlth_cdeath_simscd_fr.htm

La situation des données de 2018 et 2019 en 2022

Les travaux de paramétrage réalisés par le CépiDc ont permis à Iris/MUSE de coder automatiquement la cause initiale pour 64 % des certificats médicaux relatifs aux décès de 2018 et 2019 qui lui sont parvenus. Pour les autres, un regard humain expert était nécessaire : c'est ce que l'on appelle le codage « manuel ».

Le CépiDc n'a pas eu les moyens de coder les certificats nécessitant un regard manuel pour les décès des années 2018 et 2019. En 2021 et en 2022, il s'est concentré sur l'année 2020, ce qui a permis de publier les données relatives à cette année très particulière en décembre 2022. C'est le Projet de rattrapage et de refonte du processus de production des statistiques des causes de décès qui s'est concentré sur le codage des 447 800 décès restant non codés de 2018 et 2019.

L'objectif était donc de coder la « Cause initiale de décès » ainsi que la séquence des causes pour les 36 % de certificats non encore codés de 2018 et 2019, soit 447 800 certificats, avec un intérêt spécifique sur la cause initiale, celle requise par le règlement européen.

La démarche statistique d'évaluation de la méthode

Elle se fait à deux niveaux, en mobilisant la base de test (données de 2016 et 2017 codées manuellement), complétée de l'ensemble des données de 2016 et 2017 codés automatiquement par le système expert lorsque l'on fait l'analyse au niveau populationnel :

-Au niveau individuel sur la base de test : pour chaque certificat de la base de test, on compare la prédiction du modèle avec la « vraie » valeur, celle issue de la codification manuelle. Le nombre de codes détaillés de la CIM 10 étant très important et ses effectifs très déséquilibrés, on regarde cette cohérence/incohérence au niveau plus regroupé de la shortlist européenne (critère de diffusion d'Eurostat).

-L'objectif premier de ce travail étant statistique, on compare aussi au niveau populationnel les répartitions des causes initiales prédites par le modèle à celles réellement observées (toujours sur la base de test). Pour cela on se concentre sur les causes codées au niveau de la shortlist européenne. Pour mesurer la significativité des écarts entre effectifs prédits et effectifs réellement observés, on fait l'approximation que les effectifs des décès dus à chacune des causes pris une à une suivent une loi de poisson et on teste si l'effectif prédit a une forte probabilité d'être issu de la même loi que l'effectif observé[11]. Cette méthode a l'avantage de mettre en évidence, cause par cause, les cas que l'algorithme peine à prévoir ou prévoit en excès.

Cette approche est complétée par une démarche de type Kolmogorov-Smirnov (test d'égalité de distributions) sur l'ensemble de la répartition des causes (au niveau shortlist toujours). La statistique du Khi² de ce test mesure en effet la distance entre la répartition des causes prédites et la répartition des causes observées. Elle fournit une métrique synthétique permettant de comparer la performance de plusieurs méthodes de prédiction entre elles.

La nécessité d'une méthode originale

Le Projet a au départ envisagé une méthode classique d'imputation de valeurs manquantes pour compléter les codages automatiques de 2018 et 2019, comme il est d'usage dans les traitements statistiques d'enquêtes ou de données administratives dont les données sont « imparfaites ».

En s'appuyant sur les informations connues pour tous les certificats (variables auxiliaires), notamment l'âge et le sexe de la personne décédée et les circonstances de son décès, on a testé des méthodes d'imputation aléatoires (hot deck) en s'appuyant sur des certificats « donneurs » déjà codés. Ces méthodes sont décrites en annexe 4.

À l'aune des critères d'évaluation décrits plus haut, cette approche s'est révélée insuffisante dans notre situation car :

- La majorité de l'information permettant de déterminer la variable d'intérêt (la cause initiale du décès) n'est disponible que dans les textes du certificat de décès : il est très difficile de retrouver la véritable cause sans l'apport des textes, même avec un sous-ensemble de certificats « donneurs » déjà codés important. Un échantillon de donneurs important et représentatif permet certes de reconstituer une distribution statistique correcte, mais rarement d'attribuer individuellement la bonne cause au bon certificat.
- Et surtout parce que les populations de certificats codés automatiquement par le système de règles Iris Muse et la population des certificats que Iris/Muse ne sait pas coder sont trop différentes pour que l'une serve de donneurs pour l'autre. En effet, le système expert code des certificats plutôt simples, c'est-à-dire décrit avec des termes du dictionnaire, avec des relations logiques claires : il n'y a aucune raison que ces certificats soient représentatifs de l'ensemble de la mortalité. Enfin le système expert s'empêche même de coder lorsque le certificat relève d'un cas où l'OMS demande explicitement un regard humain : dans ce cas, évidemment, l'approche est infructueuse.

Le CépiDc avait de son côté, depuis quelques années investi au travers de collaborations et de data-challenge sur les méthodes de sciences des données. Le CépiDc avait proposé un jeu de données composé des textes des certificats et de leurs codes à plusieurs compétitions entre équipes de recherche en traitement automatique du langage dans le cadre des campagnes CLEF eHealth entre 2016 et 2018 pour repérer l'ensemble des causes en CIM dans le texte, et les équipes ont pu tester l'apport de diverses méthodes de sciences des données (méthodes lexicales, classifieur SVM, réseaux de neurones, analyse sémantique (topic modelling), modèle hybride…). La conclusion en a été que certaines méthodes de science des

[11] Pour chaque effectif, on compare la valeur « de base » (celui de la distribution réelle) avec la valeur « de comparaison » (issue de l'imputation d'une parti des résultats individuels), avec la macrofonction suivante :

```
Test <- function (base,comp)
 {   p<- (poisson.test(x=comp, r=base)$p.value) cut(p,breaks=c(0,0.01,0.05,0.1,0.2,0.3,1),labels=c("*****","****","***","**","*",""))   }
```

Le résultat se lit ainsi : ***** : on a moins de 1 % de chances de se tromper si annonce que le 2d effectif n'est pas issu de la même distribution : le modèle n'est vraisemblablement pas correct. **** on a moins de 5 % de chances de se tromper son annoncé un changement de distribution, *** 10% de chances, ** 20% , * 30%,…. Si aucune étoile, cela signifique qu'on a plus de 30 % de chances de se tromper si on dit que les deux valeurs ne sont pas tiéres de la même distribution : la valeur estimée est probablement issue de la même loi que la valeure réelle.

données étaient bien performantes pour ce type de tâche (voir Névéol et al 2017 et 2018 ainsi que Robert et al 2019). Cette conclusion a été confortée par les performances des algorithmes *transformers* développés par Louis Falissard[12] dans sa thèse de doctorat au CépiDc, et ceux entraînés sur des jeux de données bien plus importants.

Les travaux menés au Cépidc n'avaient cependant pas été jusqu'à l'entraînement d'algorithmes complets, de la cause par ligne à la séquence de causes et à la détermination de la cause initiale, et surtout n'avaient pas été jusqu'à composer les échantillons d'apprentissage et de test dans une optique de prédiction de l'année suivante. Ce travail, engagé après un avis favorable des datasicentists du laboratoire de recherche-développement en méthodes statistiques de l'Insee (le SSP Lab), les étend donc et les complète en ce sens.

[12] Voir Falissard (2021) et Falissard et al. (2022).

LA MÉTHODE CHOISIE

La méthode choisie relève du traitement automatique du langage (TAL ou *natural language processing* en anglais) puisque c'est principalement du texte qui est analysé. Elle relève aussi de l'apprentissage profond (*deep learning*) car ce sont des réseaux de neurones un peu compliqués qui sont mobilisés. Elle s'inscrit ainsi donc dans le cadre de l'apprentissage automatique (*machine learning*) ou apprentissage statistique, ici supervisé, dont on suit la démarche pour estimer le modèle et évaluer sa performance, et dont on utilise le vocabulaire[13].

Les « algorithmes » (c'est-à-dire les modèles) sont « entraînés », c'est-à-dire que les nombreux poids/paramètres qui les composent sont estimés sur un jeu de données que l'on appelle la base d'entraînement. Une fois entraîné, l'algorithme appliqué à des données va prédire une séquence de codes ou un code. La prédiction correspond aux « *fitted values* », dans le langage classique en statistique. Les performances de l'algorithme sont alors testées sur un autre jeu de données[14] (test) en comparant les prédictions auxquelles aboutit l'algorithme avec les vraies valeurs des séquences de codes ou du code. Puis l'algorithme retenu est finalement appliqué aux données que l'on souhaite coder.

La proximité avec une traduction

Coder la séquence des causes de décès peut s'envisager comme une tâche de traduction automatique des langues. En effet, il est censé y avoir un enchaînement logique dans la description du processus morbide qui fait penser à une « phrase » et l'enjeu est de traduire cette description en une série de codes de la CIM, série qui peut elle aussi être considérée comme une phrase. C'est ce raisonnement qui a conduit Falissard (2021) à appliquer des algorithmes de *deep learning* utilisés dans la traduction linguistique, algorithmes dits « phrase à phrase » (seq2seq) au cas pratique du codage des causes de décès. Au lieu d'apprendre à la machine à passer d'une phrase dans une langue à une phrase ayant le même sens dans une autre langue en lui donnant à analyser des phrases déjà traduites, on lui demande de passer d'une séquence de textes médicaux à une séquence de codes CIM, en utilisant les données déjà codées des années antérieures (et le codage automatique par système expert Iris/Muse des années récentes).

Dans ce domaine, les algorithmes à l'état de l'art aujourd'hui sont les *transformers*[15]. Ils ont été proposés par Vashani et al. (2017). Falissard et al. (2022) les appliquent au cas de la codification des causes de décès. Leur spécificité par rapport aux autres réseaux mobilisés auparavant pour les tâches de traduction réside dans le fait qu'ils sont beaucoup plus faciles à entrainer/estimer que d'autres réseaux, qu'ils requièrent pour ce faire moins de données « annotées » (c'est-à-dire de données déjà « codées » utilisées dans l'entraînement), moins de capacité de calcul (même s'ils requièrent une machine avec une GPU), que leurs calculs sont hautement parallélisables, qu'ils sont disponibles en librairies *opensource* (ici on utilise Keras et Tensorflow) et donc relativement facilement implémentables en pratique. Ces gains de rapidité/facilité par rapport aux autres algorithmes de traduction viennent notamment du fait que les *transformers* ont une façon particulière de modéliser les liens entre les mots dans la phrase, laquelle permet de tenir compte de l'ordre des expressions, leurs co-présences, répétitions etc. Il s'agit du mécanisme d'« attention », (« multi-headed attention »). Ce mécanisme d'« attention » permet aussi de réduire le temps de traitement par rapport aux autres modèles classiques de traduction (par exemple les RNNs, Recurrent neural networks).

Principes généraux de l'approche

L'approche d'estimation/ prédiction suit les étapes classiques du traitement automatique des langues et de l'apprentissage automatique.

Il s'agit tout d'abord de mettre en forme les données, c'est-à-dire définir la façon dont le texte (la suite des mots des certificats), et ce que l'on cherche en sortie (les codes) vont venir en entrée des modèles. C'est la partie *datapipeline/ feature engeneering* (1). En effet, les textes sont des données non structurées, contrairement à des variables dont on connaît *a priori* les modalités

[13] Pour une présentation générale des modèles d'apprentissage automatique et des exemples d'utilisation en statistique publique, voir Babet. D, Deltour. Q, Faria. T, Himpens. S « Les réseaux de neurones appliqués à la statistique publique : méthodes et cas d'usages » (Février 2023).

[14] Cette séparation des jeux de données vient du fait que l'on cherche à éviter le « sur-apprentissage » c'est-à-dire l'adaptation trop fine des paramètres/poids au jeu d'entraînement qui va nuire aux capacités de l'algorithme à prédire d'autres données.

[15] Voir Ashish Vaswani Noam Shazeer Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polosukhin « Attention is all you need », NIPS (2017). https://arxiv.org/abs/1706.03762?context=cs

possibles. Il faut donc les transformer, les adapter et les structurer. Par analogie avec une démarche habituelle de modélisation statistique ou économétrique de type $Y = F(X)$, cette étape correspond aux choix des *X* et de *Y*. Les *X*, ici seront une multitude d'indicatrices indiquant si un mot donné apparaît dans la phrase (matrice document terme où chaque colonne représente l'occurrence ou non d'un mot du vocabulaire et chaque ligne un certificat). Sa dimension est très importante : quelques millions de lignes, plusieurs dizaines de milliers de colonnes.

Ensuite, on spécifie l'architecture globale du modèle (2). Si on continue l'analogie avec la modélisation classique, il s'agit de définir la forme de la fonction *F,* laquelle fait intervenir de multiples compositions de fonctions auxquelles sont rattachés des poids et des paramètres à estimer. Ici cette fonction *F* suit un modèle de *transformer*. Les diverses compositions de fonctions dont est fait ce type de modèle permettent de réduire la dimension des *X* en tenant compte des proximités lexicales des mots en rapport avec la tâche de traduction (deux mots apparaissant dans le même certificat et dans des phrases aboutissant au même code vont être proches), en tenant compte de la position des mots dans la phrase et du rapport qu'ils ont entre eux selon leur position dans la phrase (attention).

L'estimation des paramètres intervenant dans le modèle (poids) se fait lors de l'entraînement (3). La phase d'entraînement calcule donc les paramètres de la fonction *F*, c'est-à-dire les coefficients qui optimisent, dans le jeu d'entraînement, les relations entre les *X* et les *Y*. Une partie des compositions de fonctions n'est pas connue à l'avance et dépend d'autres paramètres (hyperparamètres) qui sont déterminés, également lors de l'entraînement, par une mesure de performance du modèle sur un sous-échantillon de validation. Il y a donc des allers-retours entre la partie 2 et la partie 3.

Les étapes 1 à 3 ne peuvent s'effectuer que sur les certificats déjà codés (apprentissage supervisé). L'étape (4) consiste à prédire c'est-à-dire calculer, pour chaque certificat restant à coder l'estimateur de Y en appliquant la fonction F estimée aux valeurs des X pour ce certificat.

On décrit enfin (5) les bases ou les jeux de données sur lesquels l'algorithme est estimé/entraîné, puis le jeu sur lequel on teste sa performance en comparant sa prédiction avec le codage réel auquel a abouti l'équipe de codage. Dans l'approche retenue, ce sont les données de 2016 et de 2017 codées manuellement qui vont servir de base de test, puisqu'il s'agit de données « annotées », ou « labellisées » c'est-à-dire que l'on connaît pour ces données le résultat du codage manuel et que l'on peut donc le comparer au résultat prédit par l'IA. On analyse donc la performance de l'approche sur cet échantillon.

La partie suivante détaille les grandes lignes de ces étapes. Sa lecture n'est pas indispensable pour ceux qui se contentent de l'idée générale de l'approche. Ceux-ci sont invités à se rendre directement au point (5) décrivant les bases d'apprentissage ou (6) donnant quelques exemples pratiques.

Les spécifications du modèle en grandes lignes

Entre juin et octobre 2022, le Projet a expérimenté plusieurs modèles mais le temps imparti et la lourdeur des traitements des données ont limité les possibilités d'expérimentations multiples. On se concentre sur les résultats finaux.

1- Datapipeline/ feature engineering

Les séquences en entrée des modèles sont les concaténations des textes écrits sur chaque ligne du certificat de décès séparés par le label de la ligne. On y ajoute certaines variables de contexte comme le sexe, le groupe d'âge du défunt et l'année du décès[16]. Le fait d'intégrer l'année de décès notamment permet *a priori* de capter des différences de codage (par exemple, des changements dans les règles ou dans leur application) selon les années. Les séquences se terminent par le label « cause_initiale » qui indique que l'on demandera en plus de la séquence des codes au modèle à prédire la cause initiale.

Par exemple, pour une femme de 85 ans, décédée en 2016 et dont le certificat mentionnait :

Ligne 1	insuffisance respiratoire aiguë
Ligne 2	Leucostase
Ligne 3	crise blastique
Ligne 4	Lmc

On obtient : « femme 85 ans 2016 ligne cause1 insuffisance respiratoire aigue lignecause2 leucostase lignecause3 crise blastique lignecause4 lmc cause_initiale ».

[16] Les seules variables additionnelles prises en compte par le modèle en plus du texte sont dont le sexe, l'année de décès et le groupe d'âge. On pourra plus tard compléter ces variables avec les informations sur les types de certificats (deux modèles de certificats de décès co-existent depuis 2018), le mode de codage (manuel ou automatique) et par les informations supplémentaires que contient le nouveau modèle de certificat notamment sur les circonstances apparentes du décès, ou encore celles sur les durées/dates associées aux pathologies qui permettent notamment de mieux déterminer les séquelles.

Les séquences que l'on attend en sortie font intervenir les codes à 4 positions de la CIM10 pour chaque cause repérée, aussi séparés par les indicateurs de ligne, et complétés du sexe, du groupe d'âge et de l'année de décès. À la fin de la séquence de sortie, après le séparateur « cause_initiale » est indiqué le code CIM de la cause initiale.

On mobilise ensuite un tokenizer (TextVectorisation inclus dans la librairie Tensorflow) pour découper les séquences en bouts appelés « tokens ». Les tokens sont constitués pour la séquence d'entrée des mots qui se retrouvent dans le texte du certificat (il n'y a pas d'utilisation de n-grammes de caractères ni de bi ou tri-grammes de mots) ainsi que des séparateurs, et pour la séquence de sortie de codes de la CIM et des séparateurs. Le corpus des mots en entrée, ensemble des tokens rencontrés au moins une fois, comprend 79000 tokens (en l'occurrence, les tokens sont des mots) différents et celui en sortie 6000 tokens (ici, les token sont des codes) différents.

2- Architecture du modèle

Le modèle est spécifié comme un *transformer*. L'architecture des *transformers* est de type encodeur/décodeur *(encadré 2).*

- L'encodeur s'applique sur la séquence d'entrée. Il comporte une couche de plongement lexical (« *embedding* ») qui va représenter les *tokens* dans un espace vectoriel captant leur proximité linguistique et ainsi réduire la dimension du problème, s'y ajoute un plongement de la position du mot dans la séquence (*positional embedding*), ainsi que la couche utilisant le mécanisme d'attention qui peut contenir plusieurs têtes (8 têtes dans notre cas) et qui permet de tenir compte des liens des *tokens* entre eux dans la séquence. Les scores d'attention obtenus sont ensuite traités par un réseau de neurones appelé « *feed forward* », ce qui va constituer les sorties de l'encodeur.
- Le décodeur comprend les mêmes étapes que l'encodeur mais s'applique à la séquence de sortie (la séquence de codes, celle que l'on cherchera à prédire, outputs dans le schéma). Cela vise à apprendre (estimer les poids/paramètres nécessaires) de façon supervisée.
- Les sorties du décodeur seront traitées par une couche linéaire et seront ensuite normalisées par une couche de softmax pour obtenir les probabilités associées aux prédictions (une probabilité par token en sortie).

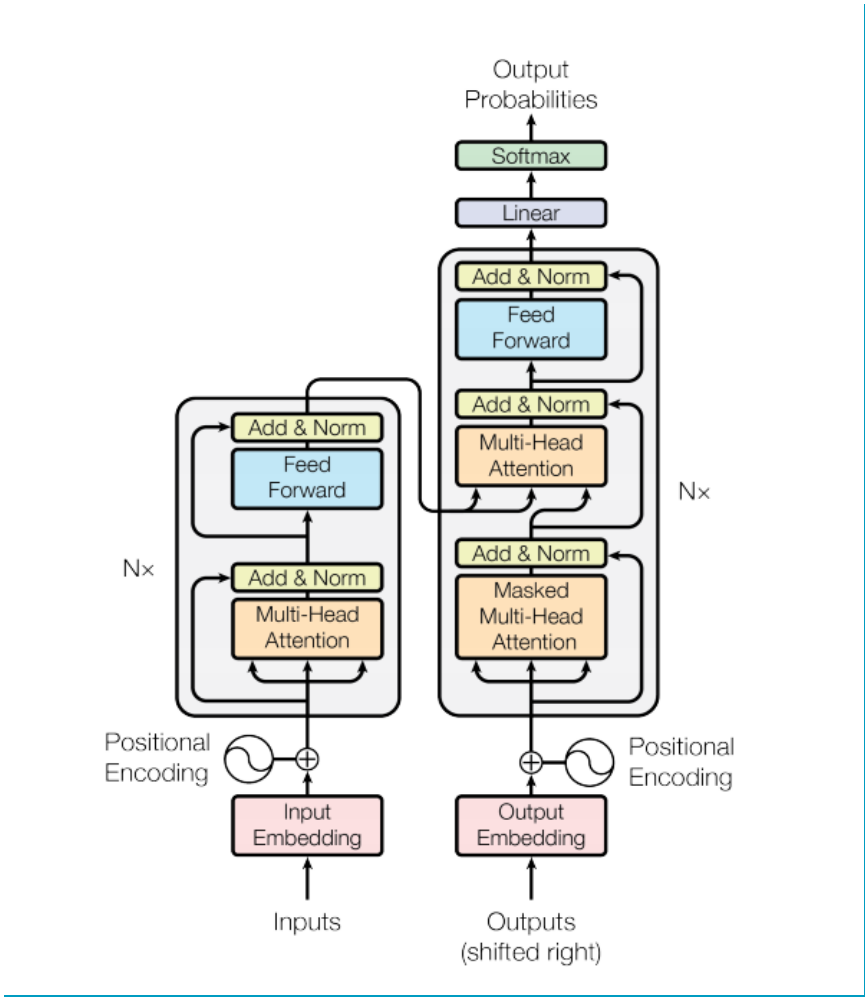
3- Entraînement

L'entraînement, c'est-à-dire l'estimation, du modèle consiste en un problème de minimisation d'une fonction de coût (l*oss fonction*) où l'on va calculer les poids ou les paramètres qui permettent de minimiser les écarts entre les séquences prédites par le modèle et les séquences données par la base d'entraînement.) La fonction de coût utilisée ici est une forme d'entropie croisée adaptée aux problèmes dits « sparse categorical cross-entropy », car chaque token est encodé en un seul chiffre numérique (au lieu de l'encodage One-hot classique). L'optimisateur mobilisé est Adam.

En pratique, on utilise les librairies open source Tensorflow et la couche Keras pour écrire l'algorithme. Le modèle a été écrit entièrement avec l'utilisation des fonctions API de Keras qui facilite à la fois la maitrise et la maintenance du programme dans le temps. Il est ainsi possible, grâce à cette flexibilité, de créer d'autres modèles en se basant sur le modèle de base du transformer (c'est le cas du modèle de classification dans la partie Oversampling). Les hyper-paramètres sur le nombre d'itérations, la taille du dictionnaire des mots, la profondeur des « réseaux de neurones » sous -jacents (nombre de couches), la dimension du modèle global, le nombre couches avant de décoder, sont décrites en annexe 6.

Le modèle a été entraîné sur une machine avec 48Go de RAM en GPU (une Nvidia RTX A6000). La durée de l'entraînement est d'environ 4 jours pour 3 millions de certificats.

Encadré 2 • Architecture du modèle transformer



Source : Vaswani (2017).

4- Prédiction

Les prédictions sont réalisées par une approche dite « greedy search ». C'est une approche itérative selon laquelle la prédiction de chaque token/ code dans la séquence entre dans la prédiction du token/code suivant : on applique au texte le modèle issu de l'étape précédente mot après mot, en choisissant la séquence de codes qui a la plus forte probabilité de correspondre[17].

En plus de la prédiction proprement dite, l'algorithme donne en sortie la probabilité que la cause soit correctement prédite ainsi que l'écart de probabilité entre le code le plus probable et le second code le plus probable. Ces informations ne sont pas utilisées à ce stade, mais elles pourront à terme entrer dans des indicateurs de confiance sur la prédiction qui pourront être mobilisés pour cibler les certificats pour lesquels on est confiant dans la prédiction de l'IA et ceux pour lesquels on demandera un codage manuel.

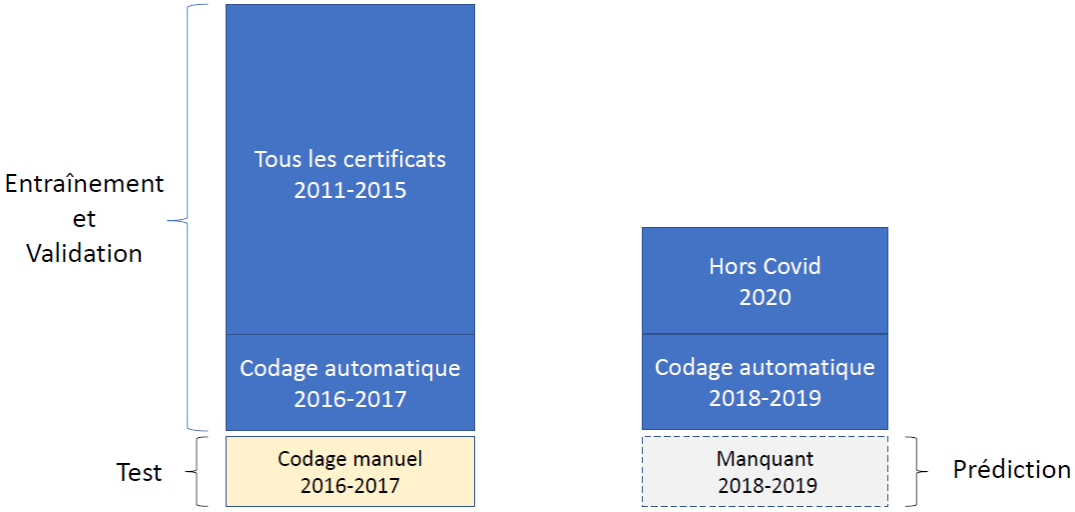
La prédiction de 487 000 certificats prend 3 jours de traitement sur une seule machine (avec une GPU).

[17] On a aussi implémenté des méthodes dites de « beam search » dans cette étape de prédiction. Cette approche permet de regarder en globalité la probabilité d'une séquence complète prédite de plusieurs mots et de prendre les séquences les plus probables à la fin de l'itération selon un paramètre de nombre à définir. Par exemple avec un paramètre beam_width=5, il va prendre les 5 séquences les plus probables à chaque étape de la prédiction et il va aussi sortir les 5 séquences finales les plus probables. Cette approche a fait l'objet de premiers tests, mais ils ne sont pas montrés beaucoup plus efficaces que ceux de la « prédiction simple », alors que la méthode est beaucoup plus gourmande en temps de calcul. Elle a donc été laissée de côté au moins provisoirement.

5- Les bases d'apprentissage

Pour prédire les données 2018 et 2019 qui auraient dû être codées manuellement, on prendra dans la base d'entraînement et de validation, tous les certificats de décès codés (manuellement ou via le système expert Iris/Iris muse) de 2011 à 2017 et 2020 (hors Covid) ainsi que les certificats de décès des années 2018 et 2019 codés automatiquement par le système expert Iris/Muse.

Pour tester la performance du modèle et de l'approche, on a réalisé dans un premier temps un apprentissage sur l'ensemble des décès des années 2011 à 2015 ainsi que les décès de 2016 et 2017[18] codés automatiquement par le système expert Iris/Muse (soit 3,2 millions de certificats au total). On a considéré alors comme base de test les certificats de 2016 et 2017 qui n'ont pas pu être codés automatiquement (430 000 certificats). Pour ces certificats on peut comparer la prédiction de l'algorithme avec les codes attribués par le codage au CépiDc. Les bases d'apprentissage sont ainsi décrites dans le schéma ci-dessous.



6- Quelques exemples

Pour reprendre notre exemple de la page 14 et en se concentrant sur la séquence de causes (sans précision de la cause initiale), ce cas a été codé ainsi par le codage manuel au CépiDc :

	Code Cim	Pour information, titre de la rubrique CIM
Ligne 1	j960	Insuffisance respiratoire aiguë
Ligne 2	i898	Autres atteintes non infectieuses précisées des vaisseaux et des ganglions lymphatiques
Ligne 3	d758	Autres maladies précisées du sang et des organes hématopoïétiques
Ligne 4	c921	Leucémie myéloïde chronique

L'output attendu, fourni au modèle comme traduction attendue est donc « femme 85 ans 2016 lignecause1 j960 lignecause2 i898 lignecause3 d758 lignecause4 c921 causeinitiale c921 »

En ce qui concerne la prédiction de la séquence des causes, le modèle a retrouvé exactement la liste des codes choisis manuellement par le CépiDc dans de nombreux cas.

Autre exemple, pour un homme de 85 ans décédé en 2016, le texte du certificat était : « lignecause1 arrêt cardiorespiratoire ; lignecause2 embolie pulmonaire ; lignecause3 thrombose veineuse omi ; lignecause6 trachéostomie traumatisme crânien démence causeinitiale embolie pulmonaire » et le modèle a prédit : « lignecause1 r092 lignecause2 i269 lignecause3 i829 r600 lignecause6 z930 s069 f03 causeinitiale XXX », ce qui est très exactement le résultat du codage réalisé manuellement au CépiDc.

[18] Pour les années 2016 et 2017, 100 % des décès ont été codés, soit de façon automatique (pour 58,3 %), soit « à la main » (pour 42,7 % donc).

Parfois la prédiction du modèle est imparfaite, comme dans le cas de cet homme de 75 ans décédé en 2017 :

	Texte	Codes Cim trouvés (Modélisation)	Codes Cim attendus (CépiDc)
Ligne 1	sepsis streptocoque	a419 b955	a409
Ligne 2	détresse respiratoire	j960	j960
Ligne 3	démence vasculaire	f019	f019

Les lignes 2 et 3 ont été correctement modélisées : J960 « Insuffisance respiratoire aigüe » et F019 « Démence vasculaire sans précision », mais pas la 1re. Le modèle n'a pas su traduire l'expression « sepsis streptocoque » en « Septicémie à streptocoques », code a409[19]. Il a prédit des codes proches mais séparés, A419 « Autres sepsis » et B955 « Streptocoques non précisés » ; et donc faux dans ce cas. Rappelons que le corpus de tokens de la séquence d'entrée est composé des mots pris un à un, on peut penser qu'ajouter des bigrammes de mots pourrait permettre de régler ce type d'erreur. Notons aussi que le modèle n'est pas contraint à prédire un seul code par ligne, et qu'au contraire le fait qu'il en prédise plusieurs rend compte de ce qu'il rencontre dans les données en entrée.

Au total, sur l'échantillon de test (les données qui ont été codées manuellement pour les années 2016 et 2017), la version du modèle finalement utilisée, prédit correctement 77,7 % des séquences de causes (hors partie spécifiant la cause initiale, sur lequel on reviendra par la suite). Ceci signifie que dans 77,7 % des situations, l'ensemble des causes du certificat est bien retranscrit, et dans le bon ordre.

Intéressons-nous désormais aux causes prises une à une. Pour rappel, il y a en moyenne 3,6 causes par certificat. On constate que 95,9 % (et 92,6 % lorsque sont mis en entrée les textes bruts, non nettoyés par les batchs automatiques ou l'équipe du CépiDc) des causes prédites par le modèle sont correctes (indicateur de précision), et que 95,7 % (90,8 % pour les textes bruts) des causes attendues ont effectivement été prédites par le modèle (indicateur de rappel). La moyenne harmonique de ces deux indicateurs, appelée F-Mesure, est de 95,7 % (91,4 % en texte brut).

À l'issue de cette étape, dont les résultats sont très encourageants, il reste à déterminer, parmi les causes repérées, laquelle est la cause initiale.

[19] Dans ce cas, on considérera que l'indicateur de Précision (proportion de codes attendus parmi les codes trouvés) est de P= 2/4=50%, que l'indicateur de Retour (proportion de codes trouvés parmi les codes attendus) est de R=2/3=0,66, et la F-Mesure (moyenne harmonique des 2 précédentes) est 2PR/(P+R)=0.57.

Les méthodes de sélection de la cause initiale

Plusieurs démarches ont été entreprises pour prédire la cause initiale. Les trois qui se sont révélées les plus efficaces et qui ont été retenues pour la suite des travaux sont les suivantes[20].

Prédiction « Keras 4 »

La première méthode consiste à utiliser comme prédiction de la cause initiale directement la prédiction donnée par le modèle dans la démarche décrite plus haut. En effet, le dernier élément attendu de la séquence de codes concerne la cause initiale.

Par exemple, pour la phrase en input « annee2016 femme age85ans lignecause1 état de choc anurie lignecause2 lymphome lignecause6 cancer colique causeinitiale », la séquence à prédire est « annee2016 femme age85ans lignecause1r579 r34 lignecause2 c859 lignecause6 c189 causeinitiale c859 », signifiant qu'outre les codes correspondant à l'ensemble des expressions médicales on demandait au modèle d'apprendre à repérer la cause initiale, ici C859 (lymphome).

Sur l'échantillon de test (données 2016 et 2017 qui devaient être codées manuellement), 82,1 % des prédictions de la cause initiale sont correctes au niveau de la CIM-10 à 4 positions, et dans 88 % des cas si l'on regroupe les codes dans les 86 catégories de la « short list » d'Eurostat.

Si l'on rajoute maintenant les certificats codés automatiquement par le système expert Iris/Muse (63 % de l'ensemble des certificats de décès en 2016 et 2017) et que l'on suppose que ces certificats codés automatiquement par le système expert Iris/Iris Muse sont systématiquement correctement codés, on arrive à un taux codage correct de la cause initiale pour 93 % des décès au niveau de la CIM-10 à 4 positions et de 95 % des décès au niveau des regroupements de la short-list européenne. Pour autant cette méthode comporte des biais dans le repérage de causes spécifiques, comme le montre la répartition complète par causes sur laquelle elle débouche (annexe7).

Prédiction « Iris-Muse »

Cette seconde méthode consiste à déterminer la cause initiale en appliquant le module MUSE d'Iris (le système expert qui choisit la cause initiale parmi les causes figurant sur le certificat) à la séquence des causes prédites par le modèle décrit ci-dessus (hors cause initiale évidemment). L'avantage de cette approche est qu'en appliquant le système expert Iris-Muse, on s'attend à satisfaire les règles de priorité de l'OMS, y compris dans leurs évolutions. On sera ainsi capable de déterminer la cause initiale dans une version précise (et précisée) de la CIM-10, voire de changer de version, et ce n'est pas l'IA qui la choisira. Par ailleurs, le système expert Iris/Muse a été programmé pour mobiliser des informations complémentaires figurant sur le certificat notamment sur le lieu et les circonstances de décès ainsi que les dates/durées des pathologies (requises pour déterminer les séquelles par exemple). En revanche, comme tout système expert de règles, il ne conclut pas toujours, notamment lorsque plusieurs codes sont possibles. Il ne conclut pas non plus pour certains décès sensibles pour lesquels l'OMS demande explicitement un regard humain.

Sur l'échantillon de test, le modèle expert MUSE code la cause initiale dans 84 % des cas. Sur ces 84 % des cas, la proportion de codage correct de cette cause initiale est de 89 %, un peu plus que ce donne le modèle Keras sur ces mêmes certificats (85 %). Deux raisons peuvent expliquer pourquoi MUSE n'atteint pas 100 % sur ces certificats : Tout d'abord, Keras4 a pu prédire une séquence de causes avec erreurs, lesquelles se reportent ensuite sur la détermination par MUSE de la cause initiale. Ensuite, il peut y avoir des erreurs de programmation dans le système de règles lesquelles sont souvent corrigées à la main lors du passage par l'équipe de codage.

[20] On a testé, notamment, deux autres approches:

- Réitérer la même logique que les étapes précédentes (Transformers et prédiction simple), en mettant dans le décodeur du transformer non pas une séquence de codes mais un code unique, celui de la cause initiale : les outputs ont la forme « femme 85 ans elec 2016 c921 », avec donc un seul code CIM. Après 4 jours de traitement, on aboutit à 77 % de causes correctement prédites, mais à trop de statistiques catégorielles imprécises. Par ailleurs, cette approche ne fournit pas les causes associées.
- dans une approche de classification visant directement la cause initiale, toujours avec un transformer, en choisissant à chaque étape les probabilités les plus élevées de cause initiale. Comme la précédente, cette approche ne fournit pas les causes associées.son temps de traitement est plutôt court (2,5 jours au total) et ses résultats plutôt meilleurs que ceux du modèle précédent (82,1 % des causes initiales sont prédites correctement). mais les erreurs ne sont pas aléatoirement réparties, et les écarts sur la répartition statistique restent importantes : une dizaine de causes sur 86 sont estimées avec trop d'écart à la réalité.

Muse ne propose pas de cause initiale[21] pour les 16 % restants. Dans ces cas-là, on utilise la cause initiale proposée par le modèle Keras 4. La précision de la prédiction sur ce sous-échantillon est de 74,8 %. En combinant ces deux approches, on obtient **donc 85,3 % de codage exact de la cause initiale au niveau de la CIM-10 à 4 positions** (avec les textes bruts, ce taux est ramené à 82,2 %), contre 82,1 % dans le modèle précédent.

Si on prend comme critère les capacités à bien classer le code dans la short-list européenne, 89,8 % (87,7 % avec les textes bruts) des décès sont correctement positionnés (contre 87,6 % pour le modèle précédent), et si l'on suppose que les 63 % codés automatiquement sont correctement traités on aboutit à **95,8 % de bonne affectation** sur l'ensemble des décès, ce qui est légèrement supérieur aux performances du modèle précédent.

Malgré ces améliorations, la répartition au niveau populationnel des catégories prédites est encore trop éloignée de celle des codes réels (annexe 6), avec un Khi² encore trop élevé.

Prédiction « Oversampling »

Cette 3e prédiction résulte d'un traitement spécifique pour les 16 % des certificats qui n'ont pas été arbitrés par Muse à l'étape précédente. L'idée alors est d'entraîner un *transformer* en surreprésentant ce type de certificat dans la base d'entraînement de façon à mieux adapter sa capacité de prédiction à ces cas particuliers. La prédiction s'obtient en appliquant alors *ce transformer* pour les 16 % des certificats que Muse ne sait pas traiter.

Cette nouvelle base d'apprentissage spécifique à ce traitement est construite en sélectionnant les certificats dont les textes sont les plus proches de ceux qui « posent problème » à MUSE grâce à un modèle de classification binaire (Oui/non) lui aussi à base de *transformer*[22] : Sur les 3,2 millions de certificats de la base initiale, on en a ainsi sélectionné 420 000.

Sur les cas qui lui sont soumis, l'approche « oversampling » permet de prédire parfaitement 68 % des causes initiales en CIM à 4 positions (et 77 % % pour la shortlist européenne). Sur l'ensemble des certificats, ce modèle est un peu moins souvent précis qu'Iris/Muse combiné à Keras, mais cette approche réduit le Khi2 du test de Kolmogorov-Smirnov d'égalité des distributions des causes en population entre celles prédites et celles véritablement codées, car elle améliore la précision dans certaines catégories spécifiques plutôt mal codées par les deux autres modèles.

Finalement, on dispose donc de trois modèles, dont les efficacités globales sont résumées dans le tableau 2.

Tableau 2 • Indicateurs de précision du codage des différents modèles

Taux de prédiction exacte de la CIM 10 à 4 chiffres :

	Keras4	Iris-Muse	« Oversampling »
Sur l'ensemble	92,6 %	93,9 %	93,6 %
Sur le codage manuel	81,9 %	85,3 %	84,1 %

Taux d'affectation dans la bonne catégorie (70+16 postes) :

	Keras4	Iris-Muse	« Oversampling »
Sur l'ensemble	94,8 %	95,8 %	95,7 %
Sur le codage manuel	87,6 %	89,8 %	89,6 %

[21] Ou plus précisément, MUSE peut proposer une valeur tout en indiquant qu'elle doit être revalidée par un opérateur car d'autres sont probables.

[22] Une alternative reposant sur de l'analyse sémantique (topic modelling) avec une méthode de latent dirichlet allocation [LDA] combinée à une classification (boosting et forêt aléatoire) est moins performante.

Khi2 de la répartition en 70 catégories (* 2 années) :

	Keras4	Iris-Muse	« Oversam-pling »
Sur l'ensemble	465	666	611
Sur le codage manuel	783	1077	988

Note > Ces indicateurs ont été calculés avec des textes nettoyés : ils surestiment légèrement les performances atteignables avec des textes bruts.
Lecture > Les scores du Khi² additionnent les écarts normalisés entre les statistiques attendues pour chacune des catégories de décès sur les deux années et leurs valeurs réelles connues. Plus le score est faible, moins il y a d'écart entre les deux distributions et donc plus le modèle est statistiquement efficace
Champ > Certificats des années 2016 et 2017 codés par IA.

Ces trois modèles n'ont pas la même efficacité relative selon le type de causes.

Pour chacun des 3 modèles[23], on observe les capacités de bonne classification individuelles (précision) : par exemple pour 100 cas de tuberculose prédites par le modèle Keras4, 92,8 % correspondent effectivement à des tuberculoses. Ce taux est parfois très élevé, ce qui montre la confiance que l'on peut accorder à la prédiction. On a également calculé les « rappels », qui montrent la capacité du modèle à repérer les causes : par exemple sur 100 vrais cas de décès dus à la tuberculose, le modèle Keras4 n'en repère que 74,5 %. Iris/Muse fait mieux (84,1 %).

En comparaison avec les « vraies » distributions complètes de 2016 et 2017, selon la catégorie, ce n'est pas toujours le même modèle qui semble avoir la meilleure performance. Dans les cas de sous-estimation systématique, il peut être pertinent de privilégier le modèle qui donne le meilleur « rappel ». L'annexe 7 décrit les répartitions des causes prédites pour chacun des modèles et les compare à la répartition attendue des causes.

Le choix d'une synthèse des modèles

Au vu des résultats précédents, on a fait le choix de combiner les 3 modèles, en fonction de leurs avantages relatifs selon les différentes classes de la CIM-10 :

1 - Pour les maladies infectieuses, en général mal repérées, on utilise l'*oversampling* pour imputer la tuberculose (01.1) et le Sida (01.2), car ils ont un bon « rappel » sans dégrader la précision. Pour le reste c'est Keras4 le plus efficace.

2 - Sur toutes les tumeurs, Keras4 est le plus précis dans les répartitions statistiques et les autres ne font pas mieux en prédiction individuelle.

3 - Pour les maladies du sang, Iris est meilleur que Keras4 et l'*oversampling* n'apporte rien

4- Pour les maladies endocriniennes, c'est Keras4 le plus précis en statistiques, et les autres ne font pas mieux en individuel.

5 et 6 : Sur les causes mentales et nerveuses, c'est Iris qui s'impose, sauf pour les pharmacodépendances (5.3) que Keras4 sous estime alors qu'Iris est plus précis

7 : sur le circulatoire, léger avantage à Keras4

8 : sur les maladies respiratoires, l'*oversampling* est assez efficace pour repérer les catégories détaillées (8.1, 8.2, et 8.3), qui sont plutôt difficiles à repérer. En revanche pour le poste « autres » (8.4) on préfère s'appuyer sur Keras4.

9 : sur les pathologies digestives, Iris est un peu plus souvent précis.

10 : sur les maladies de la peau, l''*oversampling* a un léger avantage

11: sur les articulations, avantage à Iris à tous points de vue

12 : pour les pathologies génito urinaires, l'*oversampling* est efficace car il repère des cas rares (bon rappel)

13 : Pour les effets des grossesses, à défaut de mieux, c'est le l'*oversampling* le plus efficace, ou plutôt le moins inefficace des 3 modèles.

14 : sur le périnatal, Iris est incontestablement meilleur de Keras, et l'*oversampling* apporte peu.

15 : sur les origines congénitales, catégorie toujours sous-repérée, on pourrait faire appel au meilleur en « Rappel », à savoir l'*oversampling*.

[23] Chiffres non reproduits en annexe car trop volumineux

16 : Autres : Mis à part sur la mort subite du nourrisson (16.1) où l'*oversampling* est très performant, Iris s'impose à tous points de vue

17 - Causes externes : c'est globalement Keras4 qui s'impose.

L'application de ces règles a cependant, à l'usage, plusieurs limites :

Il arrive qu'un même certificat soit concerné par deux des règles de choix précédentes (par ex un individu est codé en tumeur par Keras4 et par une maladie du sang par Iris). Pour régler ce problème, on a ordonné l'algorithme de synthèse, les règles les 1eres étant les plus susceptibles d'être remises en cause. On a globalement classé en dernier les règles concernant les catégories les plus souvent sous-estimées.

Inversement que certains ne soient concernés par aucune des règles édictées ci-dessous : il a donc fallu décider d'une modélisation choisie « par défaut » : celle de Keras4, la plus homogène.

Au final, malgré ces choix, quelques catégories restent insuffisamment repérées et statistiquement sous estimées. On leur a donc appliqué des traitements particuliers, après avoir vérifié en détail que cela ne conduisait par à trop d'imputations par erreur:

Pour la tuberculose (1.1) et les homicides (17.3) : pour arriver à une estimation correcte, on a retenu tous les cas repérés par l'un OU l'autre des 3 modèles. Pour les hépatites virales (2.2), les pathologies liées à la grossesse (13) et liées à des origines congénitales (15), on les retient en cause initiale si elles apparaissent au moins une fois dans la série des causes associées prédite par Keras4 (mais non sélectionnées comme causes initiales)[24].

Le texte de l'algorithme de synthèse qui s'avère le plus efficace figure en annexe 6.

Cette combinaison n'améliore pas la performance (précision/rappel) de la prédiction au niveau individuel. En revanche, la distribution statistique issue de cette prédiction minimise nettement le Khi² des écarts aux distributions réelles, et la rend plus pertinente pour établir des estimations statistiques provisoires[25].

Tableau 3 • Indicateurs de précision du codage du modèle de Synthèse

Taux de prédiction exacte de la CIM 10 à 4 chiffres :

	Keras4	Iris-Muse	« Oversampling »	Synthèse
Sur l'ensemble	92,6 %	93,9 %	93,6 %	**93,0 %**
Sur le codage manuel	81,9 %	85,3 %	84,1 %	**83,4 %**

Taux d'affectation dans la bonne catégorie (70 postes) :

	Keras4	Iris-Muse	« Oversampling »	Synthèse
Sur l'ensemble	94,8 %	95,8 %	95,7 %	**95,4 %**
Sur le codage manuel	87,6 %	89,8 %	89,6 %	**88,9 %**

[24] Respectivement pour la catégorie 2.2 les causes associées dont les codes CIM commencent par B18 et B19, pour la classe 13 les codes en O8 et O9 et pour la classe 15 les codes Q8 et Q9.

[25] Cette amélioration peut être due à un effet de sur-apprentissage, le choix des règles spécifiques et la détermination du modèle combiné ayant été réalisés sur les données de test 2016 2017.

Khi2 de la répartition en 70 catégories (* 2 années) :

	Keras4	Iris-Muse	« Oversampling »	Synthèse
Sur l'ensemble	465	666	611	**173**
Sur le codage manuel	783	1077	988	**336**

Note > Indicateurs établis avec des textes nettoyés.
Champ > Certificats des années 2016 et 2017 codés par IA.

Évaluation de la performance du modèle de synthèse combiné

Les tableaux en annexe 8 montrent les résultats en termes de distribution statistique pour chacune des catégories de la short list européenne, sur l'ensemble des décès, pour les deux années 2016 et 2017. Pour la majorité des catégories les écarts ne sont pas significatifs, mais certains effectifs par catégories ont encore tendance à surestimer la réalité (autres maladies infectieuses, maladies du sang, d'autres troubles mentaux et comportementaux, autres maladies de l'appareil respiratoire, intoxications accidentelles), d'autres à les sous-estimer (tuberculose, tumeurs malignes de la cavité buccale, Maladie de Hodgkin et lymphomes, malformations congénitales, autres maladies du système musculo-squelettique et du système génito-urinaire, accidents de transport, autres causes externes de mortalité).

L'annexe 9 « Précision individuelle du modèle de synthèse » permet de mesurer, catégorie par catégorie, le degré de confiance que l'on peut avoir dans les imputations au niveau individuel, à travers les indicateurs de précision et de rappel. Il est établi sur l'agrégation des données de 2016 et 2017. Les 3 premières colonnes correspondent uniquement aux données prédites par l'approche IA, les trois suivantes à l'ensemble des certificats (IA et codage automatique sur règles). La précision correspond au % correctement prédits par l'approche IA parmi les prédictions d'une catégorie donnée, le rappel[26] correspond au % correctement prédit par le modèle parmi toutes données réellement codées dans la catégorie donnée.

Ce tableau met en évidence les catégories qui doivent être examinées avec prudence. Les catégories associées à une mesure F inférieure à 90 %, pour lesquelles la performance du modèle est donc moindre : la tuberculose, le VIH/sida, l'hépatite virale, les maladies du sang, la toxicomanie, d'autres troubles mentaux et comportementaux, les maladies de la peau, la polyarthrite rhumatoïde, d'autres maladies du système musculo-squelettique, du système génito-urinaire, les complications de grossesses, malformations congénitales, empoisonnement accidentel, autres accidents, homicide, événement d'intentions indéterminées, autres causes externes.

- Les catégories prédites de façon insuffisamment précises, sont notamment :
    - Des catégories à petits effectifs (inférieurs à 500 décès par an)
    - Des catégories « autres », qui sont probablement les conséquences de plusieurs types d'imprécisions par sous-classes, qu'on n'a pas eu le temps d'analyser
    - Des catégories pour lesquelles les règles de codage ont évolué en 2016 et 2017, ce qui n'est pas appris dans le modèle testé ici (VIH, hépatites[27] et maladies du sang[28])
    - Enfin certaines catégories assez importantes, à la fois numériquement et en termes de santé publique notamment les maladies infectieuses (sur estimation) et les maladies de l'appareil digestif (sous-estimation).

[26] Exemple de lecture de ce tableau : quand le modèle choisit la modalité « Tuberculose », c'est effectivement cette cause qui a été choisie par le codeur du CépiDc dans 93 % des cas (indicateur de précision). Et quand la véritable cause est « Tuberculose », le modèle le trouve dans 86 % des cas (indicateur de rappel). On est ici dans la situation où la précision est inférieure au rappel, ce qui est cohérent avec la sous-estimation tendancielle de cette modalité. Pour la grande majorité des causes la précision, comme le rappel, sont supérieurs à 90 %.

[27] Le codage du VIH est un codage complexe qui se fait majoritairement manuellement. De plus les règles de codage du VIH et des hépatites a été modifié dans le volume 2 publié en 2016. Il s'agit d'une des plus grosses modifications des règles de codage de ces dernières années (faite en vue de préparer le passage à la CIM-11). Les nouvelles règles sont appliquées en codage manuel dès l'année 2016 avec une mise en place progressive et complète dès le codage de l'année 2017. L'effet attendu de la modification des règles était d'augmenter les effectifs des décès dus à des hépatites et au VIH/Sida. Le modèle ayant appris sur 2015 et précédentes n'est a priori pas capable de reproduire ces changements de règles.

[28] Les codes associés aux maladies du sang sont souvent modifiés manuellement car moins bien gérés par le système de règles Iris. Il s'agit des codes de thrombopénie, d'anémie, de myélodysplasie (liste non exhaustive). En 2017, il y a eu une modification du codage (règle locale) dans ce chapitre. En effet, dans certains cas des anticoagulants passent d'un code Y à un code D. Cette modification ne peut pas avoir été prédite par le modèle entraîné essentiellement sur les données 2015 et précédentes.

Ces catégories sont explicitement mentionnées dans la documentation qui a accompagné la diffusion des données, voir l'annexe « Report on provisional 2018 and 2019 CoD data partly coded with deep learning » accessible sur la page des metadata françaises de cette statistique européenne (https://ec.europa.eu/eurostat/cache/metadata/EN/hlth_cdeath_simscd_fr.htm). Les catégories repérées devront faire l'objet d'une attention particulière dans les développements futurs.

Notons par ailleurs, que d'un point de vue méthodologique, l'approche de combinaison retenue n'est pas parfaitement rigoureuse car elle revient à mettre des hyper-paramètres dans un "sur modèle", et ces hyper-paramètres sont testés sur les mêmes données (les années 2016 et 2017) que celles sur lequel ils ont été calibrés. Cette approche s'est imposée car on ne dispose pas de la distinction codage automatique/codage manuel pour les années antérieures à 2016. On aurait pu construire les règles de priorité uniquement sur l'année 2016, puis les tester sur 2017, mais on aurait eu moins de classes statistiques pour tester (86 au lieu de 2*86). On tiendra compte de cette remarque dans la suite de nos travaux, en "automatisant" le choix entre les 3 modèles.

L'ELABORATION DES DONNÉES PROVISOIRES SUR LES CAUSES DE DECES DE 2018 ET 2019

Codage des données 2018 et 2019

La méthode décrite dans la partie précédente, testée et optimisée sur les données 2016-2017, a été appliquée aux données de 2018 et 2019, afin de proposer un code de cause initiale pour les décès non codés automatiquement pour ces deux années, à savoir 218 769 certificats pour 2018 et 229 068 pour 2019.

Les deux algorithmes Keras4 et oversampling ont été réentraînés sur une base d'entraînement comprenant l'ensemble des certificats codés de 2011 à 2017, les résultats des codages automatiques de 2018 et 2019, ainsi que l'ensemble du codage de 2020 (automatique et manuel), duquel on a cependant exclu les décès dus au Covid et 15 % de certificats utilisés comme test. La base d'apprentissage contient donc 5,2 millions d'individus (contre 3,4 pour le travail sur 2016 et 2017), ce qui ne peut qu'améliorer les performances du dispositif[29].

Pour tenir compte du fait que les textes des certificats de 2018 et 2019 non encore codés n'avaient pas été corrigés des scories, fautes orthographiques et autres imperfections formelles qui sont habituellement supprimées à l'occasion du codage manuel, ce sont les textes « bruts » des certificats des années précédentes qui sont pris en compte dans la base d'apprentissage. Cela impacte la variété du vocabulaire utilisé, ainsi que certains paramètres de précision interne des modèles (*accuracy*), mais pas la capacité d'analyse de MUSE (qui ne repose que sur les codes), et les tests sur l'échantillon test de 2020 montrent que l'efficacité globale du dispositif n'en est pas affectée.

Après application de la modélisation Keras4 et du système expert Iris/Muse, 17 % des certificats ont fait l'objet de l'estimation spécifique par « oversampling ». On a ensuite appliqué l'algorithme de choix entre les 3 modèles décrits plus haut, ce qui a permis d'attribuer une cause initiale à tous les certificats.

Dans le fichier détaillé qui sera mis à la disposition des utilisateurs, la mention de la source d'imputation finalement choisie est indiquée par le code de la variable « typecodage ». La répartition, au final, de la répartition des codes choisis entre les 3 modèles est la suivante :

Tableau 4 • Part des différents modèles dans le codage de synthèse

	Source Keras4	Source Iris-Muse	Source « Over-sampling »	Règles spécifiques	Total
Code	2	3	4	5	
2018	155 221	47 554	15 202	792	218 769
2019	163 155	49 402	15 659	852	229 068
Total	318 376	96 956	30 861	1 644	447 837
Parts	71,1 %	21,6 %	6,9 %	0,4 %	100,0 %

À noter que parmi les 318 400 certificats où c'est le code issu du modèle « Keras4 » qui a été choisi, le modèle « Iris-Muse » faisait la même prédiction neuf fois sur dix (288 500 situations). Pour 64 % des certificats, le code est donc choisi par les deux modèles.

[29] Voir les annexes dans les pages suivantes pour les détails des paramètres de ces modèles

Constitution de la base de données des causes de décès provisoires

Une fois rajoutés les certificats qui ont été codés automatiquement (Code 1) et les décès pour lesquels le CépiDc n'a jamais reçu de certificats (Code 9, Cause initiale codée en R99), on obtient comme répartition technique :

Tableau 5 • Répartition de l'ensemble des certificats par type de codage

	Code	annee2018	annee2019	Total	Parts
Codage automatique	1	375 855	369 619	745 474	61%
Codage par IA		218 769	229 068	447 837	37%
dont	2	155 221	163 155	318 376	26%
	3	47 554	49 402	96 956	8%
	4	15 202	15 659	30 861	3%
	5	792	852	1 644	0%
Absence de certificat	9	15 472	15 160	30 632	3%
Total		610 096	613 847	1 223 943	100%

Les données codées provisoirement pour 2018 et 2019 ont été envoyées à Eurostat fin décembre 2022. Lors des vérifications automatiques d'Eurostat, le CépiDc a été amené à en corriger un peu moins de 300[30].

Il résulte de ces travaux une répartition par cause initiale de l'ensemble des décès de 2018 et de 2019, que l'on peut mettre en regard de celles des années 2015 à 2017 et de 2020, disponibles depuis décembre 2020[31].

Afin de permettre les comparaisons temporelles et de mettre ces chiffres en perspective avec les publications décrivant les causes de mortalité en 2020[32], on a mené les analyses uniquement sur les personnes résidant et décédant en France (métropole + DROM), ce qui conduit à exclure des analyses 2120 certificats en 2018 et 1428 en 2019.

En annexe 10 sont présentées les séries d'effectifs de décès de 2015 à 2020 pour chacune des 86 catégories.

L'annexe 11 présente les séries de taux standardisés de mortalité par causes, sur les catégories les plus fréquentes. On a indiqué par la mention « Risques modèle repéré » les lignes pour lesquelles les tests sur 2016 et 2017 avaient fait apparaître une tendance systématique à la sur- ou à la sous-estimation.

[30] Il s'agissait de tenir compte d'incohérences entre les codes et les sexes ou les tranches d'âge, ou encore d'évolutions dans les listes de causes admises selon les années. Aucune des modifications manuelles effectuées n'a remis en cause la pertinence de la modélisation par IA : Les corrections portaient principalement sur des codes relevant d'anciennes versions de la CIM-10 que celle en cours que l'IA avait prédites.

[31] Données disponibles sur le site internet d'Eurostat (https://ec.europa.eu/eurostat/databrowser/explore/all/popul?lang=fr&subtheme=hlth.hlth_cdeath&display=list&sort=category) et sur celui du CépiDc (https://www.cepidc.inserm.fr/causes-medicales-de-deces/causes-de-deces-en-2020).

[32] Naouri, D., Fouillet, A., Ghosn, W., Coudin, E. (2022, décembre. Covid-19 : troisième cause de décès en France en 2020, quand les autres grandes causes de décès baissent. DREES, Études et Résultats, 1250. Et Fouillet A, Ghosn W, Naouri D, Coudin E. Covid-19 : troisième cause de décès en France en 2020, quand les autres grandes causes baissent. Bull Épidémiol Hebd. 022;(Cov_16):2-15.

Première analyse des résultats provisoires

Les tumeurs sont la principale cause de décès en France en 2018 et 2019. Comme les années précédentes, elles représentent 28 % des décès. Parmi les tumeurs malignes, les cancers dont la mortalité est la plus élevée sont respectivement ceux de la trachée, du poumon et des bronches (48 décès pour 100 000 habitants), du colon, du rectum et de l'anus (27), du pancréas (17), du sein (17) et de la prostate (17). Les données de 2018 et 2019 confirment les nettes tendances à la baisse de la mortalité due aux cancers du poumon, du colon et moins marquées pour le sein et la prostate. En revanche, les cancers du pancréas font chaque année de plus en plus de victimes. Dans l'ensemble, les données relatives à l'année 2020, en baisse dans ce domaine, apparaissent conformes aux tendances des années précédentes.

Tableau 6 • Estimations des taux standardisés de mortalité pour les principaux groupes de causes de décès

	2015	2016	2017	Estim. 2018 *	Estim. 2019 *	2020	* Risques modèle repérés
Toutes causes	**890,6**	**867,4**	**861,7**	**850,0**	**837,7**	**899,0**	
Covid-19	0	0	0	***0***	***0***	92,9	
Maladies infectieuses et parasitaires	16,6	15,2	16,3	***15,6***	***16,0***	14,7	surestimation
Tumeurs	269,3	266,9	262,0	256,7	253,0	247,6	
dont tumeur maligne de l'œsophage	6,5	6,3	6,2	5,9	5,9	5,5	
dont tumeur maligne du côlon. rectum et anus	27,5	27,4	26,8	25,4	24,9	24,3	
dont tumeur maligne du pancréas	17,0	17,3	17,2	17,4	17,6	17,8	
dont tumeur maligne de la trachée, des bronches et du poumon	53,1	51,7	50,1	48,9	47,9	46,9	
dont tumeur maligne du sein	16,8	16,9	16,8	16,6	16,1	16	
Maladies du sang et des organes hémato-poïétiques et certains troubles du système immunitaire	3,2	3,3	3,6	***4,5***	***4,4***	3,7	surestimation
Maladies endocriniennes. nutritionnelles et métaboliques	31,8	30,4	30,4	30,1	30,7	30,8	
Dont diabète sucré	18,4	17,4	17	16,4	16,5	16,5	
Troubles mentaux et du comportement	35,8	35,0	33,7	35,6	34,6	30,8	
Dont démence	26,3	25,6	24,4	25,6	24,4	21,2	
Maladies du système nerveux et des organes des sens	53,4	53,1	52,8	51,8	49,8	47,8	
Dont Maladie de Parkinson	9,7	10,2	10,2	10,1	9,9	9,9	
Dont Maladie d'Alzheimer	26,8	26,2	25,2	24,0	22,0	20,5	

Maladies cardio-neurovasculaires	212,9	204,3	198,4	192,8	183,1	175,1	
Dont cardiopathies ischémiques	53,2	50,4	49,1	47,3	45,4	43,9	
Dont autres maladies du cœur	77,6	74,1	72,5	70,8	65,0	60,5	
Dont maladies cérébrovasculaires	46,0	44,9	42,9	42,2	41,3	39,7	
Dont autres maladies cardio-neurovasculaires	36,1	34,9	33,9	32,5	31,3	31	
Maladies de l'appareil respiratoire	65,3	61,2	63,9	63,0	61,4	52,7	
Dont grippe	2,7	1,4	3,4	3,1	3,7	1,2	
Dont pneumonie	20,0	19,2	19,3	19,3	18,7	15,3	
Dont maladies chroniques des voies respiratoires inférieures	18,6	17,6	17,6	17,5	16,8	14,4	
Maladies de l'appareil digestif	36,4	35,9	35,0	***33,4***	***33,2***	34,3	*sous-estimation*
Maladies de l'appareil génito-urinaire	15,5	14,9	15,5	15,4	16,1	16,1	
Symptômes et états morbides mal définis	79,6	78,2	81,2	83,9	87,5	86,2	
Causes externes de morbidité et mortalité	59,7	58,0	58,1	***57,5***	***57,7***	56,2	sous-estimation
dont accidents de transport	5,0	5,0	4,8	4,1	4,0	3,3	sous-estimation
dont chutes accidentelles	11,4	11,2	11,5	12,2	12,0	11,7	sous-estimation
dont suicides et lésions auto-infligées	14,8	13,9	13,4	14,1	13,6	14,1	

Après les tumeurs, la deuxième grande catégorie de causes de décès est celle des maladies cardio-neurovasculaires, responsables de 23 % des décès en 2019, ce qui est légèrement inférieur aux années précédentes. L'orientation à la baisse des taux standardisés de mortalité observée depuis deux décennies se confirme en 2018 et 2019, tant pour les différentes catégories de pathologies cardiaques que pour les maladies cérébrovasculaires Les taux observés en 2020 semblent se situer dans les tendances des années 2015-2019.

Les pathologies de l'appareil respiratoire représentaient en 2019 7 % du total des décès. Les taux de mortalité standardisés ont légèrement diminué entre 2015 et 2019, mais la baisse observée entre 2019 et 2020 est particulièrement forte, et très probablement liée aux gestes barrières lors de la crise sanitaire du Covid-19, comme évoqué dans Fouillet et al 2022, Naouri et al. 2022.

La situation est sensiblement la même pour les troubles mentaux et du comportement (4 % du total des décès) : après une légère baisse tendancielle pendant les années 2015-2019, l'année 2020 marque une rupture.

Parmi les maladies du système nerveux et des organes du sens (6 % des décès), seule la mortalité par maladie d'Alzheimer (22 décès pour 100 000 habitants en 2019) connaît une baisse régulière.

Les deux dernières grandes familles de pathologies pour lesquelles ces données provisoires apportent des éléments assez précis, celles des maladies endocriniennes, nutritionnelles et métaboliques (4 % des décès, dont 2 % pour le seul diabète sucré), et les maladies de l'appareil génito-urinaire (2 %), les chiffres de 2018 et 2019 ne font pas apparaître d'évolutions notables par rapport au passé, et l'année 2020 s'avère très similaire aux précédentes.

En revanche, du fait de la performance insuffisante de l'IA sur ces catégories, on ne peut pas établir de diagnostic clair sur les évolutions, en 2018 et 2019, des décès dus à des maladies infectieuses et parasitaires (2 %), aux maladies de l'appareil digestif (4 %) ni sur causes externes de mortalité (accidents, suicides, homicides), qui représentent au total entre 6 et 7 % des décès. Pour cette dernière catégorie, à la fois le changement de modèle de certificat de décès et de meilleures collectes des volets médicaux via le retour des instituts médicaux légaux influent aussi les tendances de ces années.

CONCLUSION

Ce travail novateur, imposé par le contexte d'un retard de production important mais rendu possible par un travail de recherche impulsé par le CépiDc, témoigne des apports potentiels importants des développements récents de l'intelligence artificielle pour aider à la codification massive de textes complexes et très hétérogènes.

Il a permis d'aboutir à des résultats suffisamment fiables pour être diffusés sur la situation de la mortalité en France juste avant la crise sanitaire du Covid-19.

Pour autant, l'approche par l'intelligence artificielle ayant rendu possible la production de ces chiffres devra être rapidement complétée par plusieurs types de travaux avant de pouvoir prétendre à régler structurellement les difficultés de codage des causes de décès :

- Une qualité suffisante sur l'ensemble des familles de causes ne sera possible que si une partie des certificats de décès des années 2018 et 2019 (dont il faut établir les caractéristiques) font l'objet d'un regard humain expert, pour confirmer ou corriger les propositions des modèles et aboutir à des données définitives permettant les usages habituels de la source ;
- Pour que la méthode puisse être appliquée pour une partie des certificats durant les années à venir, il faudra continuer à coder manuellement un volume significatif –et représentatif- des certificats, afin que la base d'apprentissage tienne compte des nouvelles pathologies, à commencer par le Covid-19.

Par ailleurs, on devra analyser les raisons pour lesquelles certains certificats ont été parfaitement compris par l'IA alors que le système Iris n'arrive pas à les lire : peut-être que cela permettra d'améliorer le paramétrage, que l'on sait perfectible, d'Iris en langue française. Enfin, les algorithmes devront être adaptés assez rapidement pour anticiper le passage de la CIM 10 à la CIM 11 dans quelques années. Des travaux en parallèle à ceux-ci et toujours en cours tendent à montrer que des approches très similaires pouvaient s'appliquer pour traduire le texte du certificat en texte propre ou élément de l'index, plus facilement transposable à une nouvelle nomenclature.

POUR EN SAVOIR PLUS

Babet. D, Deltour. Q, Faria. T, Himpens. S (2023). « Les réseaux de neurones appliqués à la statistique publique : méthodes et cas d'usages », Insee, Documents de travail, N° M2023-01 https://www.insee.fr/fr/statistiques/6801092

Boulat, T., et al. (2019, juillet). **Principales évolutions de la mortalité par cause sur la période 2000-2016 en France métropolitaine**. Bull Epidémiol Hebd., 29-30, pp. 576-584. http://beh.santepubliquefrance.fr/beh/2019/29-30/2019_29-30_1.html

Eurostat (2012). Liste européenne succincte pour les causes de décès.

Falissard, Louis « Epidémiologie profonde : méthodes d'apprentissage profond et leurs applications sur des bases de données médicoadministratives », Louis Falissard, thèse de doctorat, 2021 https://urldefense.com/v3/__https://tel.archives-ouvertes.fr/tel-03402715/document__;!!FiWPmuqhD5aF3oDTQnc!xbIJUJphGcmWDJeUIMvbH_B3zITBe-4_6NwY7VL-KgcV7geUs9XaDqlsYbze9CU3YEsGauwGx4U$

Falissard, Louis, Morgand, Claire,Ghosn, Walid, Imbaud, Claire, Bounebache, Karim and Rey, Grégoire. (2020). **Neural translation and automated recognition of ICD-10 medical entities from natural language: Algorithm Development and Validation** (Preprint). JMIR Medical Informatics. https://pubmed.ncbi.nlm.nih.gov/35404262/

Fouillet A, Ghosn W, Naouri D, Coudin E. Covid-19 : troisième cause de décès en France en 2020, quand les autres grandes causes baissent. Bull Épidémiol Hebd. 022;(Cov_16):2-15.

Gouvernement français (2017, août). Arrêté du 17 juillet 2017 relatif aux deux modèles du certificat de décès.

https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000035388290#:~:text=Arr%C3%AAt%C3%A9%20du%2017%20juillet%202017%20relatif%20aux%20deux%20mod%C3%A8les%20du%20certificat%20de%20d%C3%A9c%C3%A8s,-NOR%20%3A%20PRMX1720890A&text=Il%20est%20institu%C3%A9%20%C3%A0%20compter,partir%20du%20vingt%2Dhuiti%C3%A8me%20jour.

Naouri, D., Fouillet, A., Ghosn, W., Coudin, E. (2022, décembre). **Covid-19 : troisième cause de décès en France en 2020, quand les autres grandes causes de décès baissent**., DREES, Décembre 2022,DREES, Études et Résultats, 1250.

Névéol, A., Robert, A., Grippo, F., Morgand, C., Orsi, C., Pelikan, L., Ramadier, L., Rey, G., Zweigenbaum, P. (2018) **CLEF eHealth 2018 Multilingual Information Extraction Task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian**, *CEUR workshop proceedings*, https://ceur-ws.org/Vol-2125/invited_paper_18.pdfNévéol A, Anderson RN, Cohen KB, Grouin C, Lavergne T, Rey G, et al. CLEF eHealth 2017 Multilingual Information Extraction task overview: ICD10 coding of death certificates in English and French. In: CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS. 2017.

Rey, G. (2016, octobre)**. Les données des certificats de décès en France : processus de production et principaux types d'analyse**. La Revue de médecine interne, 37, pp. 685-693. https://www.hal.inserm.fr/inserm-03677899/document

Robert A, Baghdadi Y, Zweigenbaum P, Morgand C, Grouin C, Lavergne T, et al. (2019) **Développement et application de méthodes de traitement automatique des langues sur les causes médicales de décès pour la santé publique**. Bull Epidémiol Hebd. 2019;(29-30):603-9. http://beh.santepubliquefrance.fr/beh/2019/29-30/2019_29-30_5.html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I.. (2017) **"Attention is all you need." Paper presented at the meeting of the Advances in Neural Information Processing Systems**. https://arxiv.org/abs/1706.03762?context=cs

World Health Organization (2020, avril). **International Guidelines for Certification and Classification (Coding) of COVID-19 as Cause of Death** https://www.who.int/publications/m/item/international-guidelines-for-certification-and-classification-(coding)-of-covid-19-as-cause-of-death

Sites Internet :

CépiDc

les statistiques sur les causes médicales de décès de A à Z - https://www.cepidc.inserm.fr/qui-sommes-nous/les-statistiques-sur-les-causes-medicales-de-deces-de-z

le système de codage Iris https://www.cepidc.inserm.fr/causes-medicales-de-deces/systeme-de-codage-automatique-iris

la Classification internationale des maladies - https://www.cepidc.inserm.fr/causes-medicales-de-deces/classification-internationale-des-maladies-cim

Interrogation des données sur les causes médicales de décès - https://opendata-cepidc.inserm.fr/

Grandes causes de décès en 2020 et tendances récentes - https://www.cepidc.inserm.fr/donnees-et-publications/grandes-causes-de-deces-en-2020-et-tendances-recentes

DREES

Les causes de décès, DataDrees (2023) - https://data.drees.solidarites-sante.gouv.fr/pages/accueil/

Eurostat

Causes of death French metadata https://ec.europa.eu/eurostat/cache/metadata/EN/hlth_cdeath_simscd_fr.htm

Database Health\ Causes of death - https://ec.europa.eu/eurostat/web/health/data/database

Annexe 1. Modèle de certificat de décès

DÉPARTEMENT : ☐☐☐

CERTIFICAT DE DÉCÈS conforme à l'arrêté du 17 juillet 2017

VOLET ADMINISTRATIF À remplir par le médecin ayant constaté le décès

Je soussigné(e) M. _____, docteur en médecine, certifie que le décès de la personne désignée ci-dessous, est réel et constant.
(Nom lisible en majuscules) *(voir au verso 1)*

Date et heure (réelle ou estimée) de la mort : ____/____/____ à ____ h ____

À défaut (impossibilité à établir), date et heure du constat de décès : ____/____/____ à ____ h ____

INFORMATIONS D'ÉTAT CIVIL

COMMUNE DE DÉCÈS : .....

Code postal ☐☐☐☐☐

NOM : .....

NOM de jeune fille, *le cas échéant* : _____

Prénoms : _____

Date de naissance : ____/____/____ Sexe : ☐ M ☐ F

Domicile : _____

.....

_____

RÉSERVÉ À LA MAIRIE
Numéros à reproduire au verso.

N° d'acte	N° d'ordre du décès

INFORMATIONS FUNÉRAIRES *Cocher chaque ligne par oui ou par non*

Obstacle médico-légal *(voir au verso 2)* : ☐ oui ☐ non
Même en ce cas, renseigner au mieux l'ensemble du certificat de décès.

Obligation de mise en bière immédiate *(voir au verso 5)* :
– dans un cercueil hermétique : ☐ oui ☐ non
– dans un cercueil simple : ☐ oui ☐ non

Obstacle aux soins de conservation *(voir au verso 5)* : ☐ oui ☐ non

Obstacle au don du corps à la science *(voir au verso 5)* : ☐ oui ☐ non

Recherche de la cause du décès demandée (ou demande en cours) par prélèvement, examen ou autopsie médicale *(voir au verso 3)* : ☐ oui ☐ non
Si transport de corps nécessaire, délai de *(voir au verso 3)* : ☐ 48 h ☐ 72 h

Présence identifiée, au moment du décès, d'une prothèse fonctionnant au moyen d'une pile *(voir au verso 4)* : ☐ oui ☐ non
Si prothèse présente, enlèvement de prothèse déjà effectué par le médecin : ☐ oui ☐ non

SIGNATURE À _____, le .....
et cachet obligatoire du médecin

Volet original à détacher et à conserver dans la mairie du lieu de décès

Réf. 503 101 - **Berger-Levrault** (1710)

VOLET MÉDICAL À remplir et à clore par le médecin ayant constaté le décès – *Renseignements confidentiels et anonymes*

INFORMATIONS RELATIVES AU DÉFUNT

Commune de décès : Code postal :

Commune de domicile : Code postal :

Date de décès : ☐ date réelle OU ☐ constatée

Date de naissance :

Sexe : ☐ masculin ☐ féminin

CAUSES DU DÉCÈS

PARTIE I **Maladie(s) ou affection(s) morbide(s) ayant directement provoqué le décès.**
Il s'agit de la maladie, du traumatisme, de l'intoxication, de la complication ayant entraîné la mort (et non du mécanisme de décès comme une syncope, un arrêt cardiaque...).

Intervalle entre le début du processus morbide et le décès
En heures, jours, mois ou ans

a) _____

due à ou consécutive à : **b)** _____

due à ou consécutive à : **c)** _____

due à ou consécutive à : **d)** _____

La dernière ligne remplie doit correspondre à la cause initiale

PARTIE II **Autres états morbides, facteurs ou états physiologiques (grossesse...) ayant contribué au décès, mais non mentionnés en Partie I**

INFORMATIONS COMPLÉMENTAIRES *(cocher la case appropriée pour chaque point)*

LIEU DU DÉCÈS
☐ Domicile (du défunt ou autre) ☐ Établissement de santé public
☐ EHPAD, maison de retraite ☐ Établissement de santé privé
☐ Voie publique ☐ Établissement pénitentiaire
☐ Autre lieu ou indéterminé

MORT SUBITE S'agit-il d'un décès brutal et inattendu, évocateur de mort subite* ?
☐ oui ☐ non ☐ ne sait pas
* décès non traumatique (adulte, enfant, nourrisson) avec mode de survenue brutal (en moins d'une heure ou probablement) et inattendu (exclusion des maladies chroniques au stade terminal)

CIRCONSTANCES APPARENTES DU DÉCÈS
☐ Mort naturelle ☐ Faits de guerre
☐ Accident ☐ Complications de soins médicaux, chirurgicaux
☐ Suicide ☐ Investigations en cours
☐ Atteinte à la vie d'autrui ☐ Indéterminées

EN CAS DE MORT VIOLENTE (accidentelle, délictuelle, suicidaire, criminelle)
Précisez le lieu de survenue de l'événement déclencheur :
☐ Domicile ☐ Lieu de sport ☐ Voie publique
☐ Commerce ☐ Local industriel, chantier ☐ Exploitation agricole
☐ Établissement accueillant du public ☐ Autre lieu ou indéterminé

GROSSESSE La femme décédée était-elle enceinte ?
☐ non, pas au cours de l'année précédant le décès ☐ pas au moment du décès, mais grossesse terminée depuis 42 jours ou moins ☐ pas au moment du décès, mais grossesse terminée depuis plus de 42 jours et moins d'1 an
☐ oui, au moment du décès ☐ ne sait pas
La grossesse a-t-elle contribué au décès ? ☐ oui ☐ non ☐ ne sait pas

ACTIVITÉ PROFESSIONNELLE Le décès est-il survenu lors d'une activité professionnelle* ?
☐ oui ☐ non ☐ ne sait pas
* toute activité source de revenu (y compris au domicile), les trajets domicile-travail, les déplacements professionnels, etc.

RECHERCHE DE LA CAUSE DU DÉCÈS
Une recherche de la cause du décès a-t-elle été demandée ?
☐ oui, recherche médicale ☐ oui, recherche médico-légale ☐ non
Si oui, un volet médical complémentaire sera établi ultérieurement par le médecin ayant réalisé le diagnostic des causes de décès

SIGNATURE *Nom lisible et cachet obligatoire du médecin*

Ce volet n'est destiné qu'aux personnes autorisées pour des motifs de santé publique

Le certificat peut être saisi électroniquement à l'adresse suivante **https://sic.certdc.inserm.fr**

Annexe 2. Statistiques sur le remplissage des volets médicaux des certificats

	Taux de remplissage des lignes du certificat					
	Ligne 1	Ligne 2	Ligne 3	Ligne 4	Ligne 5	Ligne 6
Tous les certificats	96 %	76 %	39 %	14 %	34 %	0 %
À coder manuellement	97 %	81 %	50 %	21 %	47 %	1 %

	Répartition du nombre total de lignes remplies					
	1	2	3	4	5	6
Tous les certificats	20,1 %	29,8 %	28,4 %	16,2 %	5,2 %	0,3 %
À coder manuellement	9,8 %	25,8 %	32,2 %	22,7 %	9,0 %	0,5 %

	Nombre de caractères dans le texte du certificat					
	Min	Q1	Médian	Mean	Q3	Max
Tous les certificats	1	34	56	63,5	84	642
À coder manuellement	3	54	79	86,02	110	645

	Nombre de mots dans le texte du certificat					
	Min	Q1	Median	Mean	Q3	Max
Tous les certificats	1	4	6	6,99	9	71
À coder manuellement	1	6	8	9,119	12	71

	Répartition des certificats selon le nombre de causes repérées à l'issue du codage										
	1	2	3	4	5	6	7	8	9	10	Plus de 10
Tous	14,5 %	20,4 %	22,0 %	16,9 %	10,7%	6,5%	3,9%	2,2%	1,3%	0,7%	0,9%
Codage manuel	4,0 %	12,6 %	19,3 %	19,6 %	15,5%	10,8%	7,2%	4,5%	2,7%	1,6%	2,2%

Annexe 3. Exemples de séquences morbides décrites dans les certificats de décès

La ***cause initiale*** est indiquée en gras pour information, même si ce n'est pas le médecin certificateur qui la définit.

PARTIE I : Maladie(s) ou affection(s) morbide(s) ayant directement provoqué le décès		
	a)	arrêt respiratoire
due ou consécutive à :	b)	métastases multiples
due ou consécutive à :	c)	***cancer du poumon***
due ou consécutive à :	d)	
PARTIE II : Autres états morbides ayant contribué au décès		
		Diabète

PARTIE I : Maladie(s) ou affection(s) morbide(s) ayant directement provoqué le décès		
	a)	Coma
due ou consécutive à :	b)	traumatisme cranien
due ou consécutive à :	c)	***accident de la circulation***
due ou consécutive à :	d)	
PARTIE II : Autres états morbides ayant contribué au décès		

PARTIE I : Maladie(s) ou affection(s) morbide(s) ayant directement provoqué le décès		
	a)	détresse respiratoire
due ou consécutive à :	b)	***état grippal***
due ou consécutive à :	c)	insuffisance cardiaque
due ou consécutive à :	d)	
PARTIE II : Autres états morbides ayant contribué au décès		
		Diabète

PARTIE I : Maladie(s) ou affection(s) morbide(s) ayant directement provoqué le décès		
	a)	**pancréatique aiguë**
due ou consécutive à :	b)	anorexie, déshydrratation
due ou consécutive à :	c)	arrêt cardio-respiratoire
due ou consécutive à :	d)	
PARTIE II : Autres états morbides ayant contribué au décès		

PARTIE I : Maladie(s) ou affection(s) morbide(s) ayant directement provoqué le décès		
	a)	cachexie
due ou consécutive à :	b)	sarcome de Kaposi
due ou consécutive à :	c)	
due ou consécutive à :	d)	
PARTIE II : Autres états morbides ayant contribué au décès		
		Sida

PARTIE I : Maladie(s) ou affection(s) morbide(s) ayant directement provoqué le décès		
	a)	pneumonie
due ou consécutive à :	b)	fausse route
due ou consécutive à :	c)	
due ou consécutive à :	d)	
PARTIE II : Autres états morbides ayant contribué au décès		
		Alzheimer

Annexe 4. Méthode testée d'imputation par hot deck de la cause initiale

La méthode habituelle pour compléter les valeurs non connues d'une variable quand on travaille sur une population exhaustive et de taille importante est l'imputation des valeurs manquantes, qui permet de conserver parmi les modalités imputées même variabilité information que celle des « donneurs ». Les données estimées étant qualitatives et non ordonnées, il est difficile d'imaginer des imputations économétriques. On a donc testé une méthode de type hot deck, en choisissant les donneurs en fonction des variables auxiliaires.

Mise au point de la méthode d'imputation sur un échantillon aléatoire de certificats

Afin de construire le modèle d'imputation le plus robuste, une première étape consiste à tirer un échantillon aléatoire[33] parmi l'ensemble des certificats de décès de l'année 2016, avec le package sample de R, et de mettre fictivement à vide (NA) la variable d'intérêt (la cause initiale) pour les autres individus. On cherche alors à imputer une valeur pour ces derniers, et à comparer la distribution obtenue avec celle de la population initiale (cf. les mesures de qualité décrites page 9). Cela permet de savoir si, et à quelle condition, on pourrait obtenir une statistique de bonne qualité en codant uniquement une partie des certificats et en imputant les causes des autres. À noter que, dans cet exercice, on tire aléatoirement aussi bien des certificats codés automatiquement que manuellement.

On dispose pour cela, pour tous les individus, d'un certain nombre de variables descriptives, normalisées et codées dans pratiquement tous les cas. Il s'agit d'informations sur le défunt (sexe, âge, lieu de naissance et de résidence, groupe social[34], statut matrimonial) et sur le décès (date, lieu et type de lieu, circonstances apparentes du décès[35], commune).

La Package R Simputation propose plusieurs méthodes, plus ou moins sophistiquées, depuis le traditionnel hot-deck séquentiel en prenant la dernière variable remplie dans un fichier bien classé, jusqu'à des méthodes plus sophistiquées utilisant les k plus proches voisins ou des classifications. La recommandation du département des méthodes statistiques de l'Insee est d'utiliser une imputation aléatoire par classe.

La modélisation la plus efficace consiste à imputer les valeurs manquantes de la variable Causimp en fonction de du croisement des variables annexes sexe, tranche d'âge, type de lieu de décès, région de résidence et état matrimonial[36]:

```
HD2<- impute_rhd(HD2, Causimp ~ sexe + Tage + lieudc + REG + etamat, backend ="VIM")
```

Avec un échantillon de 150 000 individus codés, et les autres imputés, on obtient une distribution insuffisante sur la plupart des postes de la nomenclature (et une statistique du Khi^2 supérieure à 640 et une p-value quasiment nulle). En revanche si on considère 400 000 décès comme codés (ce qui est important, certes, mais de l'ordre de grandeur de ce qui est codé automatiquement), et qu'on impute par Hot deck le reste des certificats, on parvient à une distribution redressée beaucoup plus proche de la réalité (Khi^2 = 34, p-value=1), avec une grande majorité des postes de la nomenclature convenablement estimés (voir les résultats en annexe 5 colonnes « Redressement d'un échantillon aléatoire »).

À noter cependant que :

- Selon l'échantillon tiré, les catégories de la nomenclature qui sont surreprésentées ou sous-représentées changent : le modèle n'est donc pas stable, et on ne peut pas, dans l'immédiat, être certain de son efficacité pour repérer des évolutions atypiques d'une année sur l'autre.
- La qualité individuelle de l'imputation est médiocre : seuls 7 % des codes imputés sont individuellement les bons. Ceci limite sérieusement les usages possibles des données.

---

[33] Des essais ont également été réalisés en stratifiant cet échantillon, en particulier selon l'âge des personnes décédées ou le lieu de décès : les performances ne sont pas significativement améliorées.

[34] Les données dans ce domaine sont très succinctes. En particulier, on ne dispose que d'une catégorie sociale à un chiffre, et uniquement pour les personnes actives au moment de leur décès.

35 Depuis le certificat mis en place en 2018, le médecin certificateur indique, sous forme de case à cocher, si le décès semble dû à : une mort naturelle, un accident, un suicide, une atteinte à la vie d'autrui… On disposera donc d'une variable auxiliaire supplémentaire, à partir de 2018, très utile pour repérer les décès autres que les morts naturelles. Mais cette variable n'existait pas dans les bulletins antérieurs à 2017 : on aura donc comme seule base d'apprentissage « parfaite » l'année 2020.

36 On a essayé d'enrichir ce modèle avec des technique d'analyse textuelle des causes rédigées (Topic Model), mais sans grand succès : ni le Xhi^2 ni la proportion de bons codages ne sont améliorés par rapport à cette modélisation.

Application de la méthode dans la situation des certificats de 2016 non codés automatiquement

Ce modèle a ensuite été appliqué dans la situation où les certificats non codés ne sont pas choisis aléatoirement, mais sont ceux qui n'ont pas été effectivement codés automatiquement parmi les décès de 2016, ce qui correspond à notre situation réelle.

Comme le montre l'annexe 5 (colonnes « Redressement du codage automatique »), la distribution est alors très fortement biaisée et tout à fait insuffisante : les effectifs de beaucoup de catégories sont très mal prédits, le khi² est très élevé (14 8000) et la p-value nulle.

L'explication de cet échec, c'est que contrairement à la situation expérimentale décrite plus haut, où on infère des données à partir d'un échantillon aléatoire, la probabilité d'être codé automatiquement n'est pas du tout indépendante de la variable d'intérêt, à savoir la cause initiale de décès. La 2eme colonne du tableau de l'annexe 5 montre que la probabilité pour un certificat d'être codé automatiquement varie fortement selon la famille de cause. Le redressement par Hot deck ne suffit pas à corriger ce biais.

Même avec des redressements par hot deck, qui ont convenablement fonctionné avec un échantillon aléatoire, **il est impossible de redresser la répartition pour se rapprocher de la distribution réelle.**

Annexe 5. Simulation des résultats de la précision statistique avec des méthodes de hot deck

Titre du groupe de causes	Effectif réel, codé à 100 %	Taux de codage automatique	Redressement d'un échantillon aléatoire			Redressement du codage automatique		
			Effectif simulé	Écart		Effectif simulé	Écart	
1.1 Tuberculose	407	12%	382	*	-6%	83	*****	-80%
1.2 SIDA (maladie VIH)	344	18%	329		-4%	123	*****	-64%
1.3 Hépatites virales	593	41%	575		-3%	439	*****	-26%
1.4 Autres maladies infectieuses et parasitaires	9 206	57%	9 207		0%	9 198		0%
2.1.1 Tumeurs malignes de la lèvre, de la cavité buccale et du pharynx	3 947	44%	3 989		1%	3 139	*****	-20%
2.1.2 Tumeur maligne de l'œsophage	3 912	49%	3 883		-1%	3 494	*****	-11%
2.1.3 Tumeur maligne de l'estomac	4 621	58%	4 683		1%	4 747	***	3%
2.1.4 Tumeur maligne du côlon, rectum et anus	18 075	56%	18 165		0%	17 885	**	-1%
2.1.5 Tumeur maligne du foie et des voies biliaires intrahépatiques	8 814	59%	8 753		-1%	9 362	*****	6%
2.1.6 Tumeur maligne du pancréas	11 328	64%	11 300		0%	13 132	*****	16%
2.1.7 Tumeur maligne du larynx	1 073	47%	1 114	*	4%	856	*****	-20%
2.1.8 Tumeur maligne de la trachée, des bronches et du poumon	31 959	59%	32 053		0%	34 755	*****	9%
2.1.9 Mélanome malin de la peau	1 755	48%	1 711	*	-3%	1 557	*****	-11%
2.1.10 Tumeur maligne du sein	12 985	58%	12 953		0%	13 099		1%
2.1.11 Tumeur maligne du col de l'utérus	808	49%	824		2%	702	*****	-13%
2.1.12 Tumeur maligne d'autres parties de l'utérus	2 846	53%	2 790	*	-2%	2 713	****	-5%
2.1.13 Tumeur maligne de l'ovaire	3 505	61%	3 486		-1%	3 727	*****	6%
2.1.14 Tumeur maligne de la prostate	9 043	59%	9 143	*	1%	9 154	*	1%
2.1.15 Tumeur maligne du rein	3 601	54%	3 525	*	-2%	3 386	*****	-6%
2.1.16 Tumeur maligne de la vessie	5 361	57%	5 469	**	2%	5 336		0%
2.1.17 Tumeur maligne du cerveau et du système nerveux central	3 976	57%	3 855	***	-3%	4 107	****	3%
2.1.18 Tumeur maligne de la thyroïde	382	44%	392		3%	287	*****	-25%
2.1.19 Maladie de Hodgkin et lymphomes	4 909	49%	5 041	***	3%	4 318	*****	-12%
2.1.20 Leucémie	6 040	53%	6 025		0%	5 779	*****	-4%
2.1.21 Autres tumeurs malignes des tissus lymphoïde et hématopoïétique	3 449	53%	3 456		0%	3 198	*****	-7%

2.1.22 Autres tumeurs malignes	21 811	49%	21 857		0%	18 957	*****	-13%
2.2 Tumeurs non-malignes (bénignes et incertaines)	7 543	47%	7 620		1%	6 025	*****	-20%
3. Maladies du sang et des organes hématopoïétiques et certains troubles du système immunitaire	2 305	43%	2 320		1%	1 761	*****	-24%
4.1 Diabète sucré	11 889	54%	11 739	**	-1%	10 165	*****	-15%
4.2 Autres maladies endocriniennes, nutritionnelles et métaboliques	9 432	56%	9 345		-1%	8 838	*****	-6%
5.1 Démence	19 769	66%	19 931	*	1%	20 596	*****	4%
5.2 Abus d'alcool (y compris psychose alcoolique)	2 582	54%	2 510	**	-3%	2 243	*****	-13%
5.3 Pharmacodépendance, toxicomanie	231	32%	205	***	-11%	139	*****	-40%
5.4 Autres troubles mentaux et du comportement	3 460	51%	3 422		-1%	2 956	*****	-15%
6.1 Maladie de Parkinson	6 649	66%	6 567		-1%	7 184	*****	8%
6.2 Maladie d'Alzheimer	21 122	70%	21 131		0%	22 835	*****	8%
6.3 Autres maladies du système nerveux et des organes des sens	11 168	54%	11 165		0%	10 591	*****	-5%
7.1.1 Infarctus aigu du myocarde	14 163	60%	14 175		0%	14 220		0%
7.1.2 Autres cardiopathies ischémiques	19 077	55%	19 022		0%	17 540	*****	-8%
7.2 Autres maladies du cœur	53 344	65%	53 432		0%	55 568	*****	4%
7.3 Maladies cérébrovasculaires	32 343	55%	32 325		0%	30 409	*****	-6%
7.4 Autres maladies de l'appareil circulatoire	25 233	52%	25 080		-1%	21 514	*****	-15%
8.1 Grippe	966	53%	940		-3%	872	*****	-10%
8.2 Pneumonie	13 332	71%	13 422		1%	16 167	*****	21%
8.3.1 Asthme	936	61%	926		-1%	971	*	4%
8.3.2 Autres maladies chroniques des voies respiratoires inférieures	10 445	58%	10 260	***	-2%	10 436		0%
8.4 Autres maladies de l'appareil respiratoire	15 757	69%	15 702		0%	18 214	*****	16%
9.1 Ulcère gastro-duodénal	870	40%	818	***	-6%	650	*****	-25%
9.2 Cirrhoses, fibroses et hépatites chroniques	6 941	55%	6 890		-1%	7 045	*	1%
9.3 Autres maladies de l'appareil digestif	16 453	49%	16 597	*	1%	14 142	*****	-14%
10. Maladies de la peau et du tissu cellulaire sous-cutané	1 491	42%	1 547	**	4%	1 077	*****	-28%
11.1 Arthrite rhumatoïde et ostéoarthrite	567	38%	581		2%	346	*****	-39%
11.2 Autres maladies du système ostéo-articulaire, des muscles et du tissu conjonctif	3 599	31%	3 618		1%	1 973	*****	-45%

12.1 Maladies du rein et de l'uretère	7 592	59%	7 533		-1%	7 561		0%
12.2 Autres maladies de l'appareil génito-urinaire	2 555	41%	2 582		1%	1 773	*****	-31%
13. Complications de grossesse, accouchement et puerpéralité	40	13%	42		5%	10	*****	-75%
14. Certaines affections dont l'origine se situe dans la période périnatale	1 510	15%	1 469	*	-3%	1 009	*****	-33%
15. Malformations congénitales et anomalies chromosomiques	1 701	36%	1 718		1%	1 394	*****	-18%
16.1 Syndrome de la mort subite du nourrisson	177	8%	168		-5%	61	*****	-66%
16.2 Causes inconnues ou non précisées	11 595	76%	11 847	****	2%	25 404	*****	119%
16.3 Autres symptômes et états morbides mal définis	28 196	85%	28 095		0%	37 558	*****	33%
17.1.1 Accidents de transport	3 285	38%	3 394	***	3%	2 375	*****	-28%
17.1.2 Chutes accidentelles	7 831	30%	7 736	*	-1%	3 800	*****	-51%
17.1.3 Noyade et submersion accidentelles	966	64%	951		-2%	1 053	*****	9%
17.1.4 Intoxications accidentelles	1 810	31%	1 882	***	4%	970	*****	-46%
17.1.5 Autres accidents	13 816	40%	13 832		0%	9 179	*****	-34%
17.2 Suicides et lésions auto-infligées	8 626	55%	8 641		0%	7 955	*****	-8%
17.3 Homicides	326	20%	328		1%	111	*****	-66%
17.4 Événements dont l'intention n'est pas déterminée	795	9%	836	**	5%	121	*****	-85%
17.5 Autres causes externes de morbidité et mortalité	1 393	12%	1 327	***	-5%	287	*****	-79%
Total	**578 631**	**58%**	**578 631**			**578 631**		

Note > Voir annexe 4.
Lecture > Les degrés de signification des écarts de comptage proviennent de tests d'égalité supposant que les fréquences réelles sont distribuées selon une loi de Poisson : * pval<.3, ** pval<.2, *** pval<.1, ****pval<.05, ***** pval<.01.
Champ > Ensemble des certificats médicaux de 2016 dont le CépiDc a reçu le volet médical.

Annexe 6. Paramètres techniques des modèles d'intelligence artificielle utilisés et algorithme de synthèse entre les 3 modèles

1a : Hyperparamètres des modèles pour le test 2016-2017

-> Les modèles keras4 et Oversampling (TL)
inp_vocab_size = 30 000 avec des textes propres et 67 555 avec des textes bruts
tar_vocab_size = 6 000 avec des textes propres et 5 792 avec des textes bruts
batch_size = 200
buffer_size = 5 000
d_model = 514
latent_dim = 2 048
num_heads = 8
num_layers = 1
dropout = 0.1
epoch = 100

Optimizer : Adam
Loss : Sparse categorical crossentropy
Metric pour le training: Accuracy
Metrics pour le test : Precision, Recall, F_measure

-> Le modèle de classification : pour rappel, ce modèle sert à créer la base d'apprentissage pour le TL à partir des caractéristiques (textes) des certificats qui ont été rejetés par Iris-Muse.
Ce modèle binaire est composé d'une partie du transformer (encoder) + une fonction qui permet la classification.

inp_vocab_size = 30 000 avec des textes propres et 67 555 avec des textes bruts

batch_size = 6 000 avec des textes propres et 5 792 avec des textes bruts
buffer_size = 5000
d_model = 514
latent_dim = 2048

num_heads = 8
num_layers = 1
dropout = 0.1
epoch = 100

Optimizer : Adam
Loss : Binary crossentropy
Metric pour le training: Accuracy

1b : Détails sur les bases d'apprentissage pour le test 2016-2017

Pour Keras4

Base d'apprentissage :

- Population : tous les certificats de 2011 – 2015, les certificats en codage automatique de 2016 – 2017.
- Effectif : 3 447 459 certificats

Base de test :

- Population : les certificats en codage manuel de 2016-2017.
- Effectif : 487 047 certificats

Pour le oversampling

Base d'apprentissage :

- Population : tous les certificats 2011-2017 que le modèle de classification à détecter.
- Effectif : 422 456 certificats (sur les 3 447 459 certificats)

Base de test :

- Population : tous les certificats 2016-2017 qui ont été rejetés par iris (Status rejected).
- Effectif : 78 753 certificats de 2016 et 2017

Remarque : les identifiants ayant comme mention « Pas de certificat » ne sont pas inclus dans ces bases.

2a : Hyperparamètres des modèles pour la production des données 2018 - 2019

-> Les modèles keras4 et Oversampling (TL)
inp_vocab_size = 78 982 (correspond à la taille du vocabulaire créer par le tokenizer sur les causes en texte brut)
tar_vocab_size = 5 935(correspond à la taille du vocabulaire créer par le tokenizer sur les codes CIM10)
batch_size = 200
buffer_size = 5 000
d_model = 514
latent_dim = 2 048
num_heads = 8
num_layers = 1
dropout = 0.1
epoch = 100

Optimizer : Adam
Loss : Sparse categorical crossentropy
Metric pour le training: Accuracy
Metrics pour le test : Precision, Recall, F_measure

-> Le modèle de classification : pour rappel, ce modèle sert à créer la base d'apprentissage pour le TL à partir des caractéristiques (textes) des certificats qui ont été rejetés par Iris-Muse.
Ce modèle binaire est composé d'une partie du transformer (encoder) + une fonction qui permet la classification.

inp_vocab_size = 78982
batch_size = 200
buffer_size = 5000
d_model = 514
latent_dim = 2048

num_heads = 8
num_layers = 1
dropout = 0.1
epoch = 100

Optimizer : Adam
Loss : Binary crossentropy
Metric pour le training: Accuracy

2b : Détails sur les bases d'apprentissage pour la production des données 2018 - 2019

Pour Keras4
Base d'apprentissage :

- Population : tous les certificats de 2011 – 2017, les certificats en codage automatique de 2018 – 2019, et 85 % des certificats de 2020 hors Covid.
- Effectif : 5 173 106 certificats

Base de test :

- Population : 15 % des certificats de 2020 hors Covid obtenu après un tirage aléatoire simple
- Effectif : 87 951 certificats

Base de prédiction :

- Population : les certificats non codés de 2018 et 2019
- Effectif : 447 837 certificats

Pour le transfer learning

Base d'apprentissage :

- Population : tous les certificats 2011-2020 que le modèle de classification a détectés.
- Effectif : 1 877 493 certificats (sur les 5 173 106 certificats)

Base de prédiction :

- Population : tous les certificats 2018-2019 qui ont été rejetés par iris (Status rejected).
- Effectif : 76 230 certificats de 2018 et 2019

Remarque : les identifiants ayant comme mention « Pas de certificat » ne sont pas inclus dans ces bases.

3 : Algorithme de synthèse entre les 3 modèles

```
# phase 1 : application des modèles dominants dans chaque classe
        Synt <- ifelse (TL_86 % in % c("17.3") ,  iris_tl2, NA)
        Synt <- ifelse (IRIS_86 % in % c("05.1","05.2") ,  iris_keras4,  Synt)
        Synt <- ifelse (KERAS4_86 % in %
       c("01.4","02.2","04.1","04.2","07.1","07.2","07.3","07.4","08.4","17.1","17.3","17.4","17.5")  ,  keras4,  Synt)
        Synt <- ifelse (IRIS_86 % in % c("06.1","06.2","06.3","09.1","09.2","09.3","11.1","11.2","16.2","16.3") ,  iris_keras4,
       Synt)
        Synt <- ifelse (TL_86 % in % c("01.2","08.1","08.2","08.3","10  ","12.1","12.2","13  ","15  ","16.1") ,  iris_tl2,  Synt)
        Synt <- ifelse (IRIS_86 % in % c("05.3","14  ") ,  iris_keras4,  Synt)
        Synt <- ifelse (KERAS4_86 % in % c("01.3","02.1","05.4","17.2")  ,  keras4,  Synt)
        Synt <- ifelse (IRIS_86 % in % c("03  ") ,  iris_keras4,  Synt)

# phase 2 : traitement spécifique des 5 catégories sous estimées
       # variables spécifiques :
        Tub <- ifelse( KERAS4_86 =="01.1",1,0) +ifelse(  IRIS_86 =="01.1",1,0)
        Hep <- str_count( keras_code, "b18")+ str_count( keras_code,"b19")
        Gros <- str_count( keras_code, "o8")+ str_count( keras_code,"o9")
        Cong <- str_count( keras_code, "q8")+ str_count( keras_code,"q9")
        Hom <- ifelse( KERAS4_86 =="17.3",1,0) +ifelse(  IRIS_86 =="17.3",1,0)  + ifelse(  TL_86 =="17.3",1,0)
       # imputations correspondantes :
        Synt <- ifelse (Tub > 0 , "B909",  Synt)
        Synt <- ifelse (Gros> 0 , "O95",   Synt)
        Synt <- ifelse (Cong> 1 , "Q600",   Synt)
        Synt <- ifelse (Hom > 0 , "X99",   Synt)
        Synt <- ifelse (is.na( Synt) & Hep > 0, "B182",  Synt)

# phase 3 : imputation par défaut pour les décès concernés par aucune des règles précédentes :
 Synt <- ifelse (is.na( Synt),  iris_keras4,  Synt)
```

Annexe 7. Distributions des causes prédites selon les modèles d'IA et comparaison avec la distribution des causes observée

	Effectif attendu	Keras		Iris-Muse		Oversampling		Synthèse	
1.1 Tuberculose	**360**	*****	-21,4	*****	-13,6	****	-11,9	***	**-8,9**
1.2 SIDA (maladie VIH)	**281**	*****	-27,0	*****	-23,8		-3,6	*****	**-15,3**
1.3 Hépatites virales	**348**		-4,9	*****	-17,5	*****	-16,1		**5,2**
1.4 Autres maladies infectieuses et parasitaires	**3995**	*****	4,2	*****	10,5	*****	9,7	*****	**7,3**
2.1.1 Tumeurs malignes de la lèvre, de la cavité buccale et du pharynx	**2220**	***	-4,0	**	-3,2	***	-4,1	***	**-3,7**
2.1.2 Tumeur maligne de l'œsophage	**1976**		-0,6	*	-2,6	*	-2,7		**-0,5**
2.1.3 Tumeur maligne de l'estomac	**1944**		0,1		-0,8		-1,0		**0,3**
2.1.4 Tumeur maligne du côlon, rectum et anus	**7882**		0,5	*	-1,3	*	-1,4		**0,8**
2.1.5 Tumeur maligne du foie et des voies biliaires intrahépatiques	**3586**		0,3		-1,5		-1,4		**0,9**
2.1.6 Tumeur maligne du pancréas	**4081**		-0,4	****	-3,9	****	-3,5		**-0,2**
2.1.7 Tumeur maligne du larynx	**568**		-4,2		-3,3		-3,9		**-4,0**
2.1.8 Tumeur maligne de la trachée, des bronches et du poumon	**13012**		0,3	*****	-2,3	*****	-2,5		**0,2**
2.1.9 Mélanome malin de la peau	**904**		0,0		0,1		-0,1		**0,6**
2.1.10 Tumeur maligne du sein	**5513**		-0,1		-0,8		-0,4		**0,2**
2.1.11 Tumeur maligne du col de l'utérus	**416**		0,2		-1,4		-2,4		**-0,2**
2.1.12 Tumeur maligne d'autres parties de l'utérus	**1335**		0,1		-1,1		-0,7		**0,4**
2.1.13 Tumeur maligne de l'ovaire	**1370**		-0,9		-2,5		-0,7		**-0,9**
2.1.14 Tumeur maligne de la prostate	**3706**		-0,3		-1,3		-1,7		**0,2**
2.1.15 Tumeur maligne du rein	**1673**		0,8		-1,7		-0,8		**1,3**
2.1.16 Tumeur maligne de la vessie	**2304**	*	-2,4		-2,0		-0,8		**-2,0**
2.1.17 Tumeur maligne du cerveau et du système nerveux central	**1724**	***	-4,1	*****	-7,0	*****	-7,1	**	**-3,9**
2.1.18 Tumeur maligne de la thyroïde	**214**		-4,7		-4,7		-4,2		**-5,1**
2.1.19 Maladie de Hodgkin et lymphomes	**2519**	****	-4,0	*****	-6,0	*****	-6,0	****	**-5,0**
2.1.20 Leucémie	**2820**		-1,1	*****	-5,7	*****	-5,9	*	**-2,3**
2.1.21 Autres tumeurs malignes des tissus lymphoïde et hématopoïétique	**1625**		-1,5	**	-3,3	**	-3,5		**-2,0**

2.1.22 Autres tumeurs malignes	**11042**		0,8		0,5		0,2	**	**1,5**
2.2 Tumeurs non-malignes (bénignes et incertaines)	**3980**	**	2,1	***	2,8	*	1,9	*	**1,7**
3. Maladies du sang et des organes hématopoïétiques et certains troubles du système immunitaire	**1319**	*****	-13,9	*****	13,1	*****	17,8	*****	**13,0**
4.1 Diabète sucré	**5482**		1,0	*****	-4,3	*****	-5,5		**-1,0**
4.2 Autres maladies endocriniennes, nutritionnelles et métaboliques	**4112**		0,0	*****	4,4	****	4,0		**0,4**
5.1 Démence	**6758**	*****	7,2	*****	7,2	*****	7,4	*	**1,5**
5.2 Abus d'alcool (y compris psychose alcoolique)	**1200**	*****	10,8	****	6,5	****	6,1		**-0,8**
5.3 Pharmacodépendance, toxicomanie	**156**	*****	-20,5		-2,6		0,0		**-5,1**
5.4 Autres troubles mentaux et du comportement	**1683**	**	-3,7	*****	8,0	*****	7,2	*****	**6,4**
6.1 Maladie de Parkinson	**2266**		1,8		-0,5		-0,7		**-1,1**
6.2 Maladie d'Alzheimer	**6330**	*	1,5		1,0		0,9		**0,5**
6.3 Autres maladies du système nerveux et des organes des sens	**5135**	****	-3,3	*****	4,6	*****	4,5		**1,1**
7.1.1 Infarctus aigu du myocarde	**5675**	****	3,1	*****	4,2	*****	4,1	****	**2,6**
7.1.2 Autres cardiopathies ischémiques	**8611**		0,2	*	-1,1	****	-2,1		**-1,0**
7.2 Autres maladies du cœur	**18836**		-0,4	**	-1,1		0,1		**-0,6**
7.3 Maladies cérébrovasculaires	**14512**		0,4		0,8	****	-1,9		**0,2**
7.4 Autres maladies de l'appareil circulatoire	**12073**	*	1,0		-0,3	*	-1,0		**-0,2**
8.1 Grippe	**450**		-1,1		-2,4		-0,9		**-1,3**
8.2 Pneumonie	**3878**	**	2,6	*****	-4,5	**	-2,1		**-0,6**
8.3.1 Asthme	**367**		-2,7		1,4		2,2		**3,3**
8.3.2 Autres maladies chroniques des voies respiratoires inférieures	**4401**	**	-2,0	*	-1,6	**	-2,4		**-0,7**
8.4 Autres maladies de l'appareil respiratoire	**4942**	*****	5,5	*****	10,1	*****	11,0	*****	**5,1**
9.1 Ulcère gastro-duodénal	**526**	***	-7,8		-2,3	****	-8,6		**-4,6**
9.2 Cirrhoses, fibroses et hépatites chroniques	**3113**	**	2,4	*	1,9		1,3		**-0,5**
9.3 Autres maladies de l'appareil digestif	**8373**	*	1,3		-0,3	**	-1,6	*****	**-3,2**
10. Maladies de la peau et du tissu cellulaire sous-cutané	**863**	*****	-11,8		3,1		0,1		**2,3**
11.1 Arthrite rhumatoïde et ostéoarthrite	**351**	**	8,3		4,8		5,4		**3,4**
11.2 Autres maladies du système ostéo-articulaire, des muscles et du tissu conjonctif	**2486**	****	5,1	**	-2,7	*****	-5,4	*****	**-6,6**
12.1 Maladies du rein et de l'uretère	**3112**	*****	-11,7	*****	-8,4		0,3	**	**2,3**

12.2 Autres maladies de l'appareil génito-urinaire	**1503**		-2,6	*	-2,7	*	-2,9		**-1,1**
13. Complications de grossesse, accouchement et puerpéralité	**35**	*****	-62,9	*****	-62,9	*	-20,0		**-2,9**
14. Certaines affections dont l'origine se situe dans la période périnatale	**1289**	***	-5,2		2,6		2,2		**1,7**
15. Malformations congénitales et anomalies chromosomiques	**1094**	*****	-21,8	*****	-21,8	*****	-19,7	*****	**-14,1**
16.1 Syndrome de la mort subite du nourrisson	**163**		-4,3		0,6		1,8		**0,0**
16.2 Causes inconnues ou non précisées	**2814**	*****	13,3	*****	7,2	*****	6,9	*****	**6,5**
16.3 Autres symptômes et états morbides mal définis	**4126**	**	-2,0		-0,9		-1,1	*	**-1,9**
17.1.1 Accidents de transport	**2036**	**	-3,4	***	-3,8	**	-3,6	***	**-4,1**
17.1.2 Chutes accidentelles	**5516**	***	-2,4	*****	-5,2	*****	-3,9	****	**-3,2**
17.1.3 Noyade et submersion accidentelles	**344**		-2,3		-1,7		-2,9		**-3,5**
17.1.4 Intoxications accidentelles	**1251**	****	7,1	****	5,8		-0,8	***	**5,1**
17.1.5 Autres accidents	**8344**	**	1,8	*****	-3,1		-0,9		**0,6**
17.2 Suicides et lésions auto-infligées	**3854**	****	-3,2	*****	-5,5	*****	-4,3	***	**-3,1**
17.3 Homicides	**260**	*****	-16,9		-6,2	*	-7,7		**-3,8**
17.4 Événements dont l'intention n'est pas déterminée	**727**	***	6,878	*****	16,2	*****	15,7	***	**6,7**
17.5 Autres causes externes de morbidité et mortalité	**1223**	*****	-32,5	*****	62,6	*****	61,9	**	**-4,7**
Total	**242987**								

Lecture > Les degrés de signification des écarts de comptage proviennent de tests d'égalité supposant que les fréquences réelles sont indépendantes ligne à ligne et distribuées selon une loi de Poisson : * pval<.3, ** pval<.2, *** pval<.1, ****pval<.05, ***** pval<.01.
Champ > Certificats codés par IA car non codés automatiquement pour les décès de l'année 2016.

Annexe 8. Comparaison entre la distribution des causes prédites par le modèle de synthèse et la distribution des causes observée en 2016 et 2017

	annee2016				annee2017			
	Attendu	Prédit	Test	Ecart	Attendu	Prédit	Test	Ecart
1.1 Tuberculose	407	375	**	-7,9	404	380	*	-5,9
1.2 SIDA (maladie VIH)	344	301	****	-12,5	245	266	**	8,6
1.3 Hépatites virales	593	611		3,0	786	751	*	-4,5
1.4 Autres maladies infectieuses et parasitaires	9 206	9 498	*****	3,2	10 222	10 660	*****	4,3
2.1.1 Tumeurs malignes de la lèvre, de la cavité buccale et du pharynx	3 947	3 864	**	-2,1	3 812	3 769		-1,1
2.1.2 Tumeur maligne de l'œsophage	3 912	3 903		-0,2	3 876	3 871		-0,1
2.1.3 Tumeur maligne de l'estomac	4 621	4 626		0,1	4 628	4 642		0,3
2.1.4 Tumeur maligne du côlon, rectum et anus	18 075	18 138		0,3	18 041	18 072		0,2
2.1.5 Tumeur maligne du foie et des voies biliaires intrahépatiques	8 814	8 846		0,4	8 584	8 667		1,0
2.1.6 Tumeur maligne du pancréas	11 328	11 318		-0,1	11 507	11 498		-0,1
2.1.7 Tumeur maligne du larynx	1 073	1 050		-2,1	1 002	982		-2,0
2.1.8 Tumeur maligne de la trachée, des bronches et du poumon	31 959	31 979		0,1	31 487	31 504		0,1
2.1.9 Mélanome malin de la peau	1 755	1 760		0,3	1 773	1 775		0,1
2.1.10 Tumeur maligne du sein	12 985	12 995		0,1	13 055	13 081		0,2
2.1.11 Tumeur maligne du col de l'utérus	808	807		-0,1	822	813		-1,1
2.1.12 Tumeur maligne d'autres parties de l'utérus	2 846	2 851		0,2	2 918	2 921		0,1
2.1.13 Tumeur maligne de l'ovaire	3 505	3 493		-0,3	3 559	3 565		0,2
2.1.14 Tumeur maligne de la prostate	9 043	9 049		0,1	9 229	9 264		0,4
2.1.15 Tumeur maligne du rein	3 601	3 623		0,6	3 622	3 668		1,3

2.1.16 Tumeur maligne de la vessie	5 361	5 315		-0,9	5 157	5 104		-1,0
2.1.17 Tumeur maligne du cerveau et du système nerveux central	3 976	3 908	*	-1,7	4 106	4 033	*	-1,8
2.1.18 Tumeur maligne de la thyroïde	382	371		-2,9	420	413		-1,7
2.1.19 Maladie de Hodgkin et lymphomes	4 909	4 783	***	-2,6	4 960	4 867	**	-1,9
2.1.20 Leucémie	6 040	5 976		-1,1	6 173	6 085	*	-1,4
2.1.21 Autres tumeurs malignes des tissus lymphoïde et hématopoïétique	3 449	3 416		-1,0	3 242	3 165	**	-2,4
2.1.22 Autres tumeurs malignes	21 811	21 982	*	0,8	22 191	22 271		0,4
2.2 Tumeurs non-malignes (bénignes et incertaines)	7 543	7 610		0,9	7 616	7 654		0,5
3. Maladies du sang et des organes hématopoïétiques et certains troubles du système immunitaire	2 305	2 477	*****	7,5	2 585	2 603		0,7
4.1 Diabète sucré	11 889	11 833		-0,5	11 967	11 898		-0,6
4.2 Autres maladies endocriniennes, nutritionnelles et métaboliques	9 432	9 450		0,2	10 228	10 030	***	-1,9
5.1 Démence	19 769	19 870		0,5	19 673	20 585	*****	4,6
5.2 Abus d'alcool (y compris psychose alcoolique)	2 582	2 572		-0,4	2 466	2 440		-1,1
5.3 Pharmacodépendance, toxicomanie	231	223		-3,5	192	187		-2,6
5.4 Autres troubles mentaux et du comportement	3 460	3 567	***	3,1	3 622	3 782	*****	4,4
6.1 Maladie de Parkinson	6 649	6 625		-0,4	6 833	6 842		0,1
6.2 Maladie d'Alzheimer	21 122	21 154		0,2	20 976	21 017		0,2
6.3 Autres maladies du système nerveux et des organes des sens	11 168	11 226		0,5	11 813	11 657	**	-1,3
7.1.1 Infarctus aigu du myocarde	14 163	14 313	*	1,1	14 115	14 322	***	1,5
7.1.2 Autres cardiopathies ischémiques	19 077	18 991		-0,5	19 135	19 076		-0,3
7.2 Autres maladies du cœur	53 344	53 236		-0,2	53 805	53 722		-0,2
7.3 Maladies cérébrovasculaires	32 343	32 373		0,1	31 902	32 497	*****	1,9

7.4 Autres maladies de l'appareil circulatoire	25 233	25 211		-0,1	25 297	25 217		-0,3
8.1 Grippe	966	960		-0,6	2 509	2 472		-1,5
8.2 Pneumonie	13 332	13 307		-0,2	13 942	13 771	**	-1,2
8.3.1 Asthme	936	948		1,3	917	932		1,6
8.3.2 Autres maladies chroniques des voies respiratoires inférieures	10 445	10 412		-0,3	10 771	10 769		0,0
8.4 Autres maladies de l'appareil respiratoire	15 757	16 011	****	1,6	16 703	17 013	****	1,9
9.1 Ulcère gastro-duodénal	870	846		-2,8	867	858		-1,0
9.2 Cirrhoses, fibroses et hépatites chroniques	6 941	6 925		-0,2	6 799	6 810		0,2
9.3 Autres maladies de l'appareil digestif	16 453	16 183	****	-1,6	16 593	16 258	*****	-2,0
10. Maladies de la peau et du tissu cellulaire sous-cutané	1 491	1 511		1,3	1 630	1 549	****	-5,0
11.1 Arthrite rhumatoïde et ostéoarthrite	567	579		2,1	580	549	*	-5,3
11.2 Autres maladies du système ostéo-articulaire, des muscles et du tissu conjonctif	3 599	3 435	*****	-4,6	3 434	3 470		1,0
12.1 Maladies du rein et de l'uretère	7 592	7 664		0,9	8 124	8 158		0,4
12.2 Autres maladies de l'appareil génito-urinaire	2 555	2 538		-0,7	2 754	2 477	*****	-10,1
13. Complications de grossesse, accouchement et puerpéralité	40	39		-2,5	41	35		-14,6
14. Certaines affections dont l'origine se situe dans la période périnatale	1 510	1 532		1,5	1 693	1 649	*	-2,6
15. Malformations congénitales et anomalies chromosomiques	1 701	1 547	*****	-9,1	1 645	1 532	*****	-6,9
16.1 Syndrome de la mort subite du nourrisson	177	177		0,0	141	141		0,0
16.2 Causes inconnues ou non précisées	11 595	11 779	***	1,6	12 763	12 785		0,2
16.3 Autres symptômes et états morbides mal définis	28 196	28 118		-0,3	29 820	29 687		-0,4
17.1.1 Accidents de transport	3 285	3 202	**	-2,5	3 146	3 099		-1,5
17.1.2 Chutes accidentelles	7 831	7 655	****	-2,2	8 308	8 295		-0,2
17.1.3 Noyade et submersion accidentelles	966	954		-1,2	925	929		0,4

17.1.4 Intoxications accidentelles	1 810	1 874	**	3,5	1 733	1 831	****	5,7
17.1.5 Autres accidents	13 816	13 863		0,3	14 328	13 894	*****	-3,0
17.2 Suicides et lésions auto-infligées	8 626	8 507	*	-1,4	8 406	8 351		-0,7
17.3 Homicides	326	316		-3,1	284	292		2,8
17.4 Événements dont l'intention n'est pas déterminée	795	844	***	6,2	1 112	873	*****	-21,5
17.5 Autres causes externes de morbidité et mortalité	1 393	1 336	**	-4,1	1 535	1 454	****	-5,3
Total	**578 631**	**578 631**			**589 549**	**589 549**		

Lecture > Les degrés de signification des écarts de comptage proviennent de tests d'égalité supposant que les fréquences réelles sont distribuées selon une loi de Poisson : * pval<.3, ** pval<.2, *** pval<.1, ****pval<.05, ***** pval<.01.
Champ > Ensemble des certificats médicaux de 2016 et 2017.

Annexe 9. Performances du modèle de synthèse IA évaluée sur les décès de 2016 et 2017

	Certificats codés IA			Ensemble des certificats		
	Précision	**Rappel**	**F-Mesure**	**Précision**	**Rappel**	**F-Mesure**
1.1 Tuberculose	92 %	85 %	**88 %**	93 %	86 %	**89 %**
1.2 SIDA (maladie VIH)	75 %	71 %	**73 %**	79 %	76 %	**78 %**
1.3 Hépatites virales	70 %	69 %	**70 %**	83 %	82 %	**83 %**
1.4 Autres maladies infectieuses et parasitaires	76 %	83 %	**79 %**	89 %	92 %	**91 %**
2.1.1 Tumeurs malignes de la lèvre, de la cavité buccale et du pharynx	96 %	93 %	**95 %**	98 %	96 %	**97 %**
2.1.2 Tumeur maligne de l'œsophage	97 %	97 %	**97 %**	99 %	98 %	**99 %**
2.1.3 Tumeur maligne de l'estomac	97 %	97 %	**97 %**	99 %	99 %	**99 %**
2.1.4 Tumeur maligne du côlon, rectum et anus	96 %	97 %	**97 %**	98 %	99 %	**99 %**
2.1.5 Tumeur maligne du foie et des voies biliaires intrahépatiques	94 %	96 %	**95 %**	98 %	98 %	**98 %**
2.1.6 Tumeur maligne du pancréas	98 %	98 %	**98 %**	99 %	99 %	**99 %**
2.1.7 Tumeur maligne du larynx	95 %	91 %	**93 %**	97 %	95 %	**96 %**
2.1.8 Tumeur maligne de la trachée, des bronches et du poumon	97 %	97 %	**97 %**	99 %	99 %	**99 %**
2.1.9 Mélanome malin de la peau	95 %	95 %	**95 %**	97 %	98 %	**97 %**
2.1.10 Tumeur maligne du sein	96 %	97 %	**96 %**	98 %	99 %	**98 %**
2.1.11 Tumeur maligne du col de l'utérus	97 %	95 %	**96 %**	98 %	98 %	**98 %**
2.1.12 Tumeur maligne d'autres parties de l'utérus	97 %	97 %	**97 %**	98 %	99 %	**98 %**
2.1.13 Tumeur maligne de l'ovaire	98 %	97 %	**97 %**	99 %	99 %	**99 %**
2.1.14 Tumeur maligne de la prostate	94 %	95 %	**95 %**	98 %	98 %	**98 %**
2.1.15 Tumeur maligne du rein	94 %	96 %	**95 %**	97 %	98 %	**98 %**
2.1.16 Tumeur maligne de la vessie	97 %	95 %	**96 %**	99 %	98 %	**98 %**
2.1.17 Tumeur maligne du cerveau et du système nerveux central	98 %	94 %	**96 %**	99 %	97 %	**98 %**
2.1.18 Tumeur maligne de la thyroïde	97 %	93 %	**95 %**	98 %	96 %	**97 %**
2.1.19 Maladie de Hodgkin et lymphomes	95 %	91 %	**93 %**	98 %	96 %	**97 %**
2.1.20 Leucémie	95 %	92 %	**94 %**	98 %	96 %	**97 %**

2.1.21 Autres tumeurs malignes des tissus lymphoïde et hématopoïétique	95 %	91 %	**93 %**	97 %	96 %	**97 %**
2.1.22 Autres tumeurs malignes	92 %	93 %	**93 %**	96 %	97 %	**96 %**
2.2 Tumeurs non-malignes (bénignes et incertaines)	87 %	88 %	**88 %**	93 %	94 %	**93 %**
3. Maladies du sang et des organes hématopoïétiques et certains troubles du système immunitaire	64 %	68 %	**66 %**	79 %	82 %	**80 %**
4.1 Diabète sucré	86 %	85 %	**85 %**	93 %	93 %	**93 %**
4.2 Autres maladies endocriniennes, nutritionnelles et métaboliques	79 %	78 %	**78 %**	91 %	90 %	**91 %**
5.1 Démence	85 %	92 %	**88 %**	95 %	97 %	**96 %**
5.2 Abus d'alcool (y compris psychose alcoolique)	86 %	85 %	**86 %**	94 %	93 %	**94 %**
5.3 Pharmacodépendance, toxicomanie	79 %	75 %	**77 %**	87 %	84 %	**85 %**
5.4 Autres troubles mentaux et du comportement	76 %	82 %	**79 %**	88 %	91 %	**90 %**
6.1 Maladie de Parkinson	95 %	94 %	**94 %**	98 %	98 %	**98 %**
6.2 Maladie d'Alzheimer	95 %	96 %	**96 %**	99 %	99 %	**99 %**
6.3 Autres maladies du système nerveux et des organes des sens	87 %	86 %	**86 %**	94 %	93 %	**94 %**
7.1.1 Infarctus aigu du myocarde	88 %	91 %	**90 %**	95 %	97 %	**96 %**
7.1.2 Autres cardiopathies ischémiques	88 %	87 %	**88 %**	95 %	94 %	**94 %**
7.2 Autres maladies du cœur	87 %	87 %	**87 %**	96 %	96 %	**96 %**
7.3 Maladies cérébrovasculaires	86 %	88 %	**87 %**	94 %	94 %	**94 %**
7.4 Autres maladies de l'appareil circulatoire	84 %	83 %	**84 %**	92 %	92 %	**92 %**
8.1 Grippe	97 %	94 %	**95 %**	99 %	97 %	**98 %**
8.2 Pneumonie	87 %	85 %	**86 %**	96 %	96 %	**96 %**
8.3.1 Asthme	87 %	90 %	**88 %**	95 %	96 %	**96 %**
8.3.2 Autres maladies chroniques des voies respiratoires inférieures	90 %	90 %	**90 %**	96 %	96 %	**96 %**
8.4 Autres maladies de l'appareil respiratoire	78 %	82 %	**80 %**	93 %	95 %	**94 %**
9.1 Ulcère gastro-duodénal	89 %	86 %	**88 %**	94 %	92 %	**93 %**
9.2 Cirrhoses, fibroses et hépatites chroniques	93 %	93 %	**93 %**	97 %	97 %	**97 %**
9.3 Autres maladies de l'appareil digestif	89 %	86 %	**87 %**	95 %	93 %	**94 %**

10. Maladies de la peau et du tissu cellulaire sous-cutané	80 %	77 %	**78 %**	89 %	87 %	**88 %**
11.1 Arthrite rhumatoïde et ostéoarthrite	84 %	81 %	**82 %**	90 %	89 %	**89 %**
11.2 Autres maladies du système ostéo-articulaire, des muscles et du tissu conjonctif	76 %	74 %	**75 %**	84 %	82 %	**83 %**
12.1 Maladies du rein et de l'uretère	79 %	81 %	**80 %**	91 %	92 %	**92 %**
12.2 Autres maladies de l'appareil génito-urinaire	85 %	77 %	**81 %**	92 %	86 %	**89 %**
13. Complications de grossesse, accouchement et puerpéralité	75 %	67 %	**71 %**	78 %	72 %	**75 %**
14. Certaines affections dont l'origine se situe dans la période périnatale	94 %	94 %	**94 %**	95 %	94 %	**95 %**
15. Malformations congénitales et anomalies chromosomiques	87 %	76 %	**81 %**	92 %	84 %	**88 %**
16.1 Syndrome de la mort subite du nourrisson	93 %	93 %	**93 %**	93 %	93 %	**93 %**
16.2 Causes inconnues ou non précisées	92 %	95 %	**93 %**	98 %	99 %	**98 %**
16.3 Autres symptômes et états morbides mal définis	94 %	92 %	**93 %**	99 %	99 %	**99 %**
17.1.1 Accidents de transport	97 %	94 %	**95 %**	98 %	96 %	**97 %**
17.1.2 Chutes accidentelles	93 %	91 %	**92 %**	95 %	94 %	**95 %**
17.1.3 Noyade et submersion accidentelles	92 %	91 %	**91 %**	97 %	97 %	**97 %**
17.1.4 Intoxications accidentelles	74 %	79 %	**76 %**	82 %	86 %	**84 %**
17.1.5 Autres accidents	83 %	81 %	**82 %**	90 %	89 %	**89 %**
17.2 Suicides et lésions auto-infligées	96 %	94 %	**95 %**	98 %	97 %	**98 %**
17.3 Homicides	83 %	83 %	**83 %**	87 %	87 %	**87 %**
17.4 Événements dont l'intention n'est pas déterminée	78 %	70 %	**74 %**	80 %	72 %	**76 %**
17.5 Autres causes externes de morbidité et mortalité	34 %	32 %	**33 %**	43 %	41 %	**42 %**

Lecture > Précision : Quand le modèle prédit un décès par tuberculose, c'est vrai 92 % des cas parmi les certificats codés par l'intelligence artificielle, et 93 % des cas sur l'ensemble de décès. Rappel : Parmi les décès codés par l'IA 85 % des décès effectivement dus à la tuberculose ont bien été repérés comme tel par le modèle Si on prend en compte l'ensemble des certificats, cette proportion monte à 86 %. La « F-Mesure » est la moyenne harmonique des deux taux précédents : 2* (P*R) / (P+ R)
Champ > Ensemble des certificats de décès de 2016 et 2017.

Annexe 10. Série des effectifs de causes de décès entre 2015 et 2020

Les données portent sur l'ensemble des décès en France de personne résident en France.

Les chiffres de 2015, 2016, 2017 et 2020 sont les chiffres définitifs du CépiDc.

Ceux relatifs à 2018 et 2019 sont issus partiellement de la modélisation par IA.

	2015	2016	2017	2018 prov.	2018 prov.	2020
1.1 Tuberculose	434	403	402	372	401	295
1.2 SIDA (maladie VIH)	390	334	237	267	244	201
1.3 Hépatites virales	600	587	773	456	445	351
1.4 Autres maladies infectieuses et parasitaires	9 797	9 180	10 193	10 242	10 694	10 208
2.1.1 Tumeurs malignes de la lèvre, de la cavité buccale et du pharynx	3 924	3 936	3 809	3 665	3 474	3 636
2.1.2 Tumeur maligne de l'œsophage	3 893	3 902	3 865	3 766	3 775	3 630
2.1.3 Tumeur maligne de l'estomac	4 559	4 602	4 612	4 541	4 420	4 258
2.1.4 Tumeur maligne du côlon, rectum et anus	17 658	18 029	17 996	17 324	17 337	17 197
2.1.5 Tumeur maligne du foie et des voies biliaires intrahépatiques	8 518	8 776	8 551	8 510	8 566	8 727
2.1.6 Tumeur maligne du pancréas	10 921	11 300	11 467	11 745	12 182	12 476
2.1.7 Tumeur maligne du larynx	1 091	1 069	1 000	942	882	827
2.1.8 Tumeur maligne de la trachée, des bronches et du poumon	32 150	31 877	31 402	31 172	31 105	30 935
2.1.9 Mélanome malin de la peau	1 850	1 748	1 767	1 755	1 796	1 756
2.1.10 Tumeur maligne du sein	12 580	12 936	13 013	13 081	12 967	13 008
2.1.11 Tumeur maligne du col de l'utérus	763	801	817	857	775	770
2.1.12 Tumeur maligne d'autres parties de l'utérus	2 755	2 838	2 903	2 882	2 884	2 845
2.1.13 Tumeur maligne de l'ovaire	3 491	3 495	3 545	3 315	3 466	3 341
2.1.14 Tumeur maligne de la prostate	8 919	9 022	9 212	9 310	9 408	9 178
2.1.15 Tumeur maligne du rein	3 640	3 597	3 612	3 455	3 332	3 483
2.1.16 Tumeur maligne de la vessie	5 230	5 349	5 146	5 287	5 209	5 345
2.1.17 Tumeur maligne du cerveau et du système nerveux central	3 885	3 964	4 087	3 721	3 944	4 035
2.1.18 Tumeur maligne de la thyroïde	412	378	420	431	388	362

2.1.19 Maladie de Hodgkin et lymphomes	4 843	4 869	4 936	4 649	4 745	4 875
2.1.20 Leucémie	5 936	6 016	6 134	5 905	5 923	6 165
2.1.21 Autres tumeurs malignes des tissus lymphoïde et hématopoïétique	3 385	3 433	3 230	3 234	3 319	3 283
2.1.22 Autres tumeurs malignes	21 315	21 738	22 106	23 436	24 033	23 018
2.2 Tumeurs non-malignes (bénignes et incertaines)	7 441	7 527	7 587	7 781	7 832	7 656
3. Maladies du sang et des organes hématopoïétiques et certains troubles du système immunitaire	2 207	2 291	2 570	3 219	3 188	2 801
4.1 Diabète sucré	12 268	11 848	11 927	11 756	12 063	12 264
4.2 Autres maladies endocriniennes, nutritionnelles et métaboliques	9 357	9 407	10 189	10 496	11 116	11 334
5.1 Démence	19 309	19 755	19 661	21 058	20 744	18 596
5.2 Abus d'alcool (y compris psychose alcoolique)	2 594	2 577	2 460	2 698	2 739	2 472
5.3 Pharmacodépendance, toxicomanie	160	230	189	211	241	229
5.4 Autres troubles mentaux et du comportement	3 344	3 452	3 608	3 948	4 127	4 091
6.1 Maladie de Parkinson	6 192	6 642	6 826	6 897	6 865	7 012
6.2 Maladie d'Alzheimer	20 872	21 111	20 962	20 396	19 194	18 244
6.3 Autres maladies du système nerveux et des organes des sens	10 944	11 128	11 782	12 170	12 573	12 360
7.1.1 Infarctus aigu du myocarde	14 659	14 031	13 976	13 438	13 258	12 922
7.1.2 Autres cardiopathies ischémiques	19 310	18 985	19 053	19 032	18 482	18 170
7.2 Autres maladies du cœur	53 623	53 184	53 652	53 935	50 940	48 061
7.3 Maladies cérébrovasculaires	32 176	32 213	31 776	31 834	31 763	31 112
7.4 Autres maladies de l'appareil circulatoire	25 019	25 117	25 165	24 612	24 450	24 498
8.1 Grippe	1 915	961	2 501	2 297	2 776	871
8.2 Pneumonie	13 371	13 305	13 920	14 162	14 264	11 559
8.3.1 Asthme	891	929	914	863	855	721
8.3.2 Autres maladies chroniques des voies respiratoires inférieures	10 746	10 416	10 747	11 058	10 899	9 372
8.4 Autres maladies de l'appareil respiratoire	15 811	15 722	16 675	16 492	16 407	16 188
9.1 Ulcère gastro-duodénal	853	867	862	782	724	837

9.2 Cirrhoses, fibroses et hépatites chroniques	7 056	6 914	6 775	6 671	6 630	6 776
9.3 Autres maladies de l'appareil digestif	16 081	16 396	16 533	16 021	16 368	17 362
10. Maladies de la peau et du tissu cellulaire sous-cutané	1 379	1 489	1 623	1 628	1 724	1 639
11.1 Arthrite rhumatoïde et ostéoarthrite	555	565	578	565	534	583
11.2 Autres maladies du système ostéo-articulaire, des muscles et du tissu conjonctif	3 651	3 589	3 424	2 942	3 236	3 440
12.1 Maladies du rein et de l'uretère	7 637	7 572	8 105	8 072	8 714	8 580
12.2 Autres maladies de l'appareil génito-urinaire	2 461	2 550	2 752	2 941	3 098	3 512
13. Complications de grossesse, accouchement et puerpéralité	40	40	41	30	18	41
14. Certaines affections dont l'origine se situe dans la période périnatale	1 571	1 501	1 685	1 616	1 607	1 443
15. Malformations congénitales et anomalies chromosomiques	1 694	1 675	1 624	1 477	1 539	1 502
16.1 Syndrome de la mort subite du nourrisson	165	176	139	165	109	114
16.2 Causes inconnues ou non précisées	25 361	27 198	29 680	30 797	34 612	34 657
16.3 Autres symptômes et états morbides mal définis	29 163	28 069	29 700	31 899	32 437	33 001
17.1.1 Accidents de transport	3 199	3 186	3 054	2 664	2 564	2 144
17.1.2 Chutes accidentelles	7 684	7 781	8 262	8 996	9 082	9 073
17.1.3 Noyade et submersion accidentelles	904	920	884	879	770	668
17.1.4 Intoxications accidentelles	2 042	1 800	1 725	1 175	1 316	1 505
17.1.5 Autres accidents	13 991	13 694	14 202	13 019	13 982	14 272
17.2 Suicides et lésions auto-infligées	9 118	8 592	8 367	8 882	8 622	8 986
17.3 Homicides	336	312	281	435	494	472
17.4 Événements dont l'intention n'est pas déterminée	873	785	1 102	1 062	1 187	1 552
17.5 Autres causes externes de morbidité et mortalité	844	1 391	1 525	2 583	2 560	1 361
18 - Covid 19	-	-	-	-	-	69 238
Total	**591 806**	**592 072**	**604 298**	**607 974**	**612 417**	**667 496**

Champ > Décès de personnes résidant et décédées en France entre 2015 et 2020.
Source > CépiDc, 2020, Statistique sur les causes de décès, 2018 et 2019 : données provisoires.

Annexe 11. Série des taux standardisés de mortalité par causes de 2015 à 2020

Dans la colonne « R » (Risque) on fait apparaître un + quand la simulation pour 2016 et 2017 a fait apparaitre une tendance significative à la surestimation de la catégorie, qui a pu impacter les chiffres de 2018 et 2019. Le signe – signale une potentielle sous-estimation.

	2015	2016	2017	2018 prov.	2018 prov.	R.	2020
1.1 Tuberculose	0,6	0,6	0,6	0,5	0,6	-	0,4
1.2 SIDA (maladie VIH)	0,6	0,5	0,4	0,4	0,4		0,3
1.3 Hépatites virales	0,9	0,9	1,2	0,7	0,7		0,5
1.4 Autres maladies infectieuses et parasitaires	14,5	13,2	14,1	14,0	14,4	+	13,5
2.1.1 Tumeurs malignes de la lèvre, de la cavité buccale et du pharynx	6,4	6,3	6	5,7	5,3	-	5,5
2.1.2 Tumeur maligne de l'œsophage	6,5	6,3	6,2	5,9	5,9		5,5
2.1.3 Tumeur maligne de l'estomac	7,3	7,3	7,2	6,9	6,6		6,3
2.1.4 Tumeur maligne du côlon, rectum et anus	27,5	27,4	26,8	25,4	24,9		24,3
2.1.5 Tumeur maligne du foie et des voies biliaires intrahépatiques	14,1	14,2	13,6	13,4	13,1		13,1
2.1.6 Tumeur maligne du pancréas	17	17,3	17,2	17,4	17,6		17,8
2.1.7 Tumeur maligne du larynx	1,9	1,8	1,7	1,5	1,4		1,3
2.1.8 Tumeur maligne de la trachée, des bronches et du poumon	53,1	51,7	50,1	48,9	47,9		46,9
2.1.9 Mélanome malin de la peau	2,9	2,7	2,7	2,7	2,7		2,6
2.1.10 Tumeur maligne du sein	16,8	16,9	16,8	16,6	16,1		16
2.1.11 Tumeur maligne du col de l'utérus	1,1	1,1	1,1	1,2	1,1		1,1
2.1.12 Tumeur maligne d'autres parties de l'utérus	3,6	3,7	3,7	3,6	3,6		3,4
2.1.13 Tumeur maligne de l'ovaire	4,7	4,7	4,7	4,3	4,4		4,2
2.1.14 Tumeur maligne de la prostate	17,5	17,2	17,1	16,9	16,6		15,9
2.1.15 Tumeur maligne du rein	5,9	5,7	5,7	5,3	5,0		5,1
2.1.16 Tumeur maligne de la vessie	9	9	8,5	8,4	8,2		8,3
2.1.17 Tumeur maligne du cerveau et du système nerveux central	6,1	6,2	6,3	5,7	6,0	-	6
2.1.18 Tumeur maligne de la thyroïde	0,6	0,6	0,6	0,6	0,5		0,5

2.1.19 Maladie de Hodgkin et lymphomes	7,6	7,5	7,4	7,0	6,9	-	7
2.1.20 Leucémie	9,4	9,3	9,3	8,8	8,7		8,8
2.1.21 Autres tumeurs malignes des tissus lymphoïde et hématopoïétique	5,3	5,3	4,9	4,7	4,8		4,7
2.1.22 Autres tumeurs malignes	33,6	33,5	33,4	34,8	35,0		32,9
2.2 Tumeurs non-malignes (bénignes et incertaines)	11,4	11,2	11	11,0	10,9		10,4
3. Maladies du sang et des organes hématopoïétiques et certains troubles du système immunitaire	3,2	3,3	3,6	4,5	4,4	+	3,7
4.1 Diabète sucré	18,4	17,4	17	16,4	16,5		16,5
4.2 Autres maladies endocriniennes, nutritionnelles et métaboliques	13,4	13	13,4	13,7	14,2		14,3
5.1 Démence	26,3	25,6	24,4	25,6	24,4		21,2
5.2 Abus d'alcool (y compris psychose alcoolique)	4,2	4,1	3,9	4,2	4,3		3,8
5.3 Pharmacodépendance, toxicomanie	0,3	0,4	0,3	0,3	0,4		0,4
5.4 Autres troubles mentaux et du comportement	5	4,9	5,1	5,4	5,6	+	5,4
6.1 Maladie de Parkinson	9,7	10,2	10,2	10,1	9,9		9,9
6.2 Maladie d'Alzheimer	26,8	26,2	25,2	24,0	22,0		20,5
6.3 Autres maladies du système nerveux et des organes des sens	16,9	16,7	17,4	17,7	18,0		17,4
7.1.1 Infarctus aigu du myocarde	22,7	21,3	20,8	19,5	19,0	+	18,3
7.1.2 Autres cardiopathies ischémiques	30,5	29,1	28,3	27,7	26,5		25,6
7.2 Autres maladies du cœur	77,6	74,1	72,5	70,8	65,0		60,5
7.3 Maladies cérébrovasculaires	46	44,9	42,9	42,2	41,3		39,7
7.4 Autres maladies de l'appareil circulatoire	36,1	34,9	33,9	32,5	31,3		31
8.1 Grippe	2,7	1,4	3,4	3,1	3,7		1,2
8.2 Pneumonie	20	19,2	19,3	19,3	18,7		15,3
8.3.1 Asthme	1,2	1,2	1,2	1,1	1,1		0,9
8.3.2 Autres maladies chroniques des voies respiratoires inférieures	17,4	16,4	16,4	16,4	15,8		13,5
8.4 Autres maladies de l'appareil respiratoire	24	23	23,6	23,0	22,2	+	21,8
9.1 Ulcère gastro-duodénal	1,3	1,3	1,2	1,1	1,0		1,1

9.2 Cirrhoses, fibroses et hépatites chroniques	11,4	11,1	10,7	10,4	10,2		10,3
9.3 Autres maladies de l'appareil digestif	23,7	23,5	23,1	21,9	22,0	-	22,9
10. Maladies de la peau et du tissu cellulaire sous-cutané	1,9	2	2,1	2,1	2,1		2
11.1 Arthrite rhumatoïde et ostéoarthrite	0,7	0,7	0,7	0,7	0,6		0,7
11.2 Autres maladies du système ostéo-articulaire, des muscles et du tissu conjonctif	5,3	5,1	4,8	4,0	4,3		4,5
12.1 Maladies du rein et de l'uretère	11,5	11	11,4	11,2	11,8		11,3
12.2 Autres maladies de l'appareil génito-urinaire	4	3,9	4,1	4,2	4,3		4,8
13. Complications de grossesse, accouchement et puerpéralité	0,1	0,1	0,1	0,0	0,0	-	0,1
14. Certaines affections dont l'origine se situe dans la période périnatale	1	1	1,1	1,1	1,1		1
15. Malformations congénitales et anomalies chromosomiques	2,1	2,1	2	1,8	1,9	+	1,8
16.1 Syndrome de la mort subite du nourrisson	0,1	0,1	0,1	0,1	0,1		0,1
16.2 Causes inconnues ou non précisées	38,2	39,8	42,2	43,1	47,1	+	46,3
16.3 Autres symptômes et états morbides mal définis	41,3	38,3	38,9	40,7	40,3		39,8
17.1.1 Accidents de transport	5	5	4,8	4,1	4,0	-	3,3
17.1.2 Chutes accidentelles	11,4	11,2	11,5	12,2	12,0	-	11,7
17.1.3 Noyade et submersion accidentelles	1,4	1,4	1,4	1,3	1,2		1
17.1.4 Intoxications accidentelles	3,1	2,7	2,6	1,7	1,9	-	2,2
17.1.5 Autres accidents	20,9	19,9	20	17,8	18,8		18,9
17.2 Suicides et lésions auto-infligées	14,8	13,9	13,4	14,1	13,6		14,1
17.3 Homicides	0,5	0,5	0,4	0,7	0,8		0,7
17.4 Événements dont l'intention n'est pas déterminée	1,4	1,3	1,8	1,7	1,9		2,4
17.5 Autres causes externes de morbidité et mortalité	1,2	2,1	2,2	3,8	3,6	-	1,9
18 - Covid 19	0	0	0	0,0	0,0		92,9
Total	890,6	867,4	861,7	850,0	837,7		899,0

Champ > Décès de personnes résidant et décédées en France entre 2015 et 2020.
Source > CépiDc, 2020, Statistique sur les causes de décès, 2018 et 2019 : données provisoires.

DREES MÉTHODES
N° 8 • mars 2023

Les statistiques provisoires sur les causes de décès en 2018 et 2019

Directeur de la publication
Fabrice Lenglart

Responsable d'édition
Valérie Bauer-Eubriet

ISSN
2495-120X

Ministère des Solidarités et de la Santé
Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)

14 avenue Duquesne - 75 350 paris 07 SP
Retrouvez toutes nos publications sur drees.solidarites-sante.gouv.fr et nos données sur www.data.drees.sante.fr