Centre d'épidémiologie sur les causes médicales de décès – Institut national de la santé et de la recherche médicale – Documents de travail du CépiDc

Rapport de Production

Année de décès : 2021

Zina Hebbache, Pierre Boulet, Aude Robert, Elisa Zambetta, Daniel Razakamana, Elise Coudin, Diane Martin
CépiDc-Inserm

Mars 2024

Document de travail du CépiDc N°4

Résumé

Ce rapport de production revient sur l'ensemble des traitements réalisés par le CépiDc pour produire la base de données sur les causes médicales de décès en 2021. Le codage des causes de décès en CIM-10 en 2021 en France combine un codage automatique par le système expert de codage IRIS/MUSE, des prédictions par des algorithmes d'apprentissage profond, et un codage manuel en particulier ciblé sur les certificats d'intérêt spécifique pour la recherche et la santé publique. Une attention particulière porte sur l'évaluation de la cohérence entre ce nouveau type de campagne de codage et une campagne traditionnelle qui ne mobiliserait pas les réseaux de neurones profonds. Les spécificités de la base de données individuelles accessibles via le SNDS sont aussi précisées.

Mots-clés : causes de décès, mortalité, CIM10, 2021, production statistique des données sur les causes de décès

Abstract

This production report details the treatments carried out by CépiDc to produce the database on medical causes of death in 2021. The coding of causes of death in ICD-10 in 2021 in France combines automatic coding by the IRIS/MUSE expert coding system, predictions by *deep learning* algorithms, and manual coding targeted in particular at certificates of specific interest for research and public health. Particular attention is paid to assessing the coherence between this new type of coding campaign and a traditional campaign that would not mobilize deep neural networks. The database of individual data accessible via the SNDS are also presented.

Keywords: causes of death, mortality, ICD10, 2021, statistical production of CoD data

Table des matières

1 Introduction

Le processus de production des données du CépiDc est décrit en détail dans le document 'Statistiques sur les causes de décès de A à Z' accessible sur le site internet du CépiDc.

Ce rapport présente les spécificités du traitement des certificats de décès d'une année donnée et apporte des précisions sur le processus de production. Ce rapport concerne l'année 2021.

La base de données pour les décès 2021 a été produite dans un contexte marqué par le développement de nouvelles méthodes de codage automatique impliquant des algorithmes de *deep learning* développées au CépiDc. Ces avancées avaient déjà été mises en œuvre avec succès pour la production des données de 2018 et 2019 dans un contexte de rattrapage. Fort des résultats obtenus, une nouvelle stratégie de codage a émergé avec pour objectif de générer des échantillons de codage manuels ciblés en volume significativement plus importants que ce qui était précédemment envisageable pour les données de 2018 et 2019. Ainsi, le processus classique de codage a été une nouvelle fois adapté et les nouvelles méthodes de codage automatique ont été utilisées pour permettre de produire la base plus rapidement que les autres années. Ce processus de codage spécifique se rapproche du processus pérenne envisagé au CépiDc mais sera encore adapté pour l'année 2022. Ce rapport donne des indicateurs liés à la collecte des données pour l'année 2021 et décrit le processus de codage mis en place et ses évolutions. Enfin, la base de données finalisée et diffusée dans le SNDS est documentée également.

2 Collecte

2.1 Exhaustivité

En 2021, 662 149 décès ont été comptabilisés par l'Insee sur le territoire français. Le CépiDc a reçu 659 381 volets médicaux et en a conservé 648 700 (soit 98,3%). Pour 2% des décès, aucun volet médical n'a été reçu, et donc aucune cause n'est déclarée.

193 volets médicaux de décès de moins de 28 jours n'ont pas été reçus.

Tableau 1. Exhaustivité de la collecte des volets médicaux (VM) pour les décès 2021

Décès Insee	662149
VM reçus	659381
VM supprimés	-10681
VM non reçus	13449
% de VM non reçus	2,03%

Les volets médicaux supprimés correspondent majoritairement à des doublons (le décès a fait l'objet de plusieurs volets médicaux). Il y a aussi quelques cas où des certificats n'ont pas été appariés à un

décès comptabilisé par l'Insee. Il peut s'agir de volets médicaux pour lesquels les données individuelles étaient incomplètes ou erronées (le certificateur a pu avoir fait un autre volet médical, c'est donc similaire à un doublon) ou encore il peut s'agir de volets médicaux tests (des tests sur l'application de certification électronique par exemple), mais ce dernier cas reste très anecdotique.

Les décès pour lesquels aucun volet médical n'a été reçu seront présents dans la base de données du CépiDc et leur cause initiale sera R99 (cause indéterminée). Tous les décès dont la cause initiale est R99 ne sont pas forcément des certificats non reçus au CépiDc. Il peut simplement s'agir de certificats dont la partie cause de décès n'a pas été complétée par le médecin ou insuffisamment pour identifier une cause initiale de décès informative. Pour repérer la différence de situation entre les certificats non reçus et ceux dont la cause était trop peu informative pour être codée autrement qu'en R99, il faut regarder s'il y a la mention « pas de certificat » dans les causes. En 2021, 43% des R99 étaient des certificats non reçus.

2.2 Origine des certificats

Pour l'année 2021, sur l'ensemble des volets médicaux reçus et conservés au CépiDc, 438846 (68%) étaient au format papier, 209888 (32%) étaient d'origine électronique.

Les causes rédigées par le médecin sur les certificats papiers sont saisies par un prestataire qui applique des règles de saisie précises permettant de faciliter leur codage automatique (corrections de fautes d'orthographe, suppression des articles…). A partir des décès 2018, ces règles de standardisation sont également appliquées sur les causes des certificats électroniques rejetés par le système expert de codage. Dans ce cas, le prestataire de saisie a appliqué les mêmes règles de standardisation que celles appliquées sur les formulaires papier.

2.3 Version de certificats

En 2016, lors de la parution du volume 2 de la classification internationale des maladies (CIM), qui est le volume comprenant toutes les nouvelles règles de codage, l'organisation mondiale de la santé (OMS) proposait un nouveau modèle de certificat. La France a mis en place ce certificat en 2018.

A partir du 1er janvier 2018, une nouvelle version du certificat ("certificat 2017", conforme à l'arrêté du 17 juillet 2017 https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000035388290) a été mise en circulation. Celle-ci contient des informations complémentaires concernant notamment les circonstances apparentes de décès. Le médecin peut désormais renseigner s'il semble s'agir d'un suicide, un accident, une atteinte à la vie d'autrui, une mort naturelle (maladie) en cochant une case. De plus, dans la nouvelle version du certificat néonatal, le certificateur peut renseigner l'enchaînement causal ayant conduit au décès. Sur la version de 2017, il ne pouvait qu'indiquer la cause ayant directement provoqué le décès et d'autres causes indirectement impliquées dans le décès. Il y a également la notion de mort subite du nourrisson.

L'application de certification électronique a mis en place la nouvelle version de certificat le 02/01/2018 au matin. En papier, il persiste dans les faits une cohabitation des deux versions (1997/2017) pour des raisons pratiques et logistiques (les médecins utilisent les stocks de certificats ancien modèle encore en leur possession) pendant plusieurs années.

Sur l'ensemble des volets médicaux reçus et conservés au CépiDc pour l'année 2021, 580029 (89%) correspondaient à la nouvelle version du certificat, 68671 (11%) l'ancienne. 16% des certificats papiers correspondaient à l'ancienne version.

2.4 Volets médicaux complémentaires (VMC)

Un volet médical complémentaire doit être rédigé en cas d'investigations médicales ou médico-légales post-mortem. Il est identique au volet médical initial et est complété sur la plateforme de certification électronique. Il se substitue au volet médical initial s'il apporte davantage d'informations concernant les causes du décès. Les VMC sont collectés à partir du 1er janvier 2018.

En 2021, le CépiDc a pu prendre en compte 3104 VMC dont une grande partie (2373) étaient issus de l'IML de Paris. En effet, cet IML transmet l'ensemble de ces données issues de son système d'information directement au CépiDc. Les VMC de l'IML de Paris n'ont pas le même format que le volet médical initial pour la description des causes de décès car l'IML renseigne directement le code CIM de la cause initiale du décès telle que le médecin légiste la définit. Les autres 731 VMC pris en compte ont été collectés via l'application de certification électronique.

En conclusion, il existe une disparité géographique sur la collecte des VMC car leur complétion reste non exhaustive dans les IML de province.

2.5 Type de certificat

En 2021, sur l'ensemble des volets médicaux reçus et conservés au CépiDc, 1874 était des volets médicaux néonatals, dont 52% rédigé électroniquement. 45 de ces volets médicaux ont été utilisés pour des décès de plus de 28 jours. Inversement, 5 décès chez des bébés de moins de 28 jours ont été certifiés en utilisant un modèle général.

3 Codage

Le codage de l'année 2021 s'est déroulé de Septembre 2022 à Novembre 2023 en parallèle du codage des années 2018 et 2019 dans une période de rattrapage du retard de production. L'annexe 8.3 détaille le calendrier du codage 2021.

La stratégie de codage du CépiDc intègre des algorithmes de *deep learning* capables de prédire les causes associées et la cause initiale de décès [voir Zambetta et al. (2023a, 2023b), Hebbache et al. (2023)]. Elle se différencie d'une campagne de codage dite « traditionnelle » correspondant à du codage automatique par le système expert (63%) et du codage manuel uniquement pour le reste. La dernière année produite avec cette stratégie est l'année 2020, elle-même produite avant les années 2018-2019 du fait de la crise sanitaire.

Cette nouvelle stratégie fait suite aux travaux pour la production des données de décès 2018 et 2019. Le processus de codage s'appuie sur plusieurs outils qui interagissent entre eux et constituent une méthode combinant trois modes de codage : le codage automatique par le système expert, le codage manuel de certificats ciblés, le codage automatique (IA) impliquant des algorithmes de *deep learning*. Ce dernier mode de codage et l'évaluation globale du système mixte sont présentés dans la suite.

Dans ce paragraphe, nous décrivons comment les modes de codage interagissent, donnons les spécificités de chaque mode de codage pour l'année 2021. Nous décrivons enfin l'étape de

« vérifications » qui permet de valider ou de corriger le codage de certaines données ciblées - notamment celles faisant intervenir de nouvelles règles de codage, ou des cas où des imperfections du codage automatique par batch ont été relevées, ou encore des décès à fort enjeu de santé publique – et indiquons comment elle a été mise en place pour l'année 2021.

3.1 Méthode combinant trois modes de codage

Les trois modes de codage sont combinés et s'organisent de la façon suivante (cf Figure 1).

L'ensemble des certificats de décès est traité par le système expert Iris/Muse pour une première tentative de codage automatique par batch. En cas de rejet, le codage manuel va se concentrer sur une partie des certificats. Ce ciblage concerne des tirages aléatoires, des décès sensibles et des catégories de causes de décès pour lesquelles le troisième mode de codage (IA) est estimé comme moins performant. Enfin, pour le reste des certificats c'est également ce troisième mode (codage IA) qui sera utilisé. Ce troisième mode peut être combiné avec le système expert IRIS/MUSE pour la détermination de la cause initiale. Les modèles de *deep learning* sont donc utilisés à deux moments du traitement des certificats :

- le ciblage des certificats pour le codage manuel (on utilise alors les modèles utilisés lors de la production annuelle précédente) : les certificats sont ciblés selon les catégories les moins bien prédites par l'IA pour la cause initiale et par ordre de priorité selon leur score de confiance.
- puis la prédiction du codage des causes de décès pour les certificats restant à coder après la phase de codage manuel, soit en toute fin de production afin d'enrichir les modèles avec le codage manuel de l'année en cours de codage.

En 2021, les modèles ayant été utilisés pour cibler des certificats sont ceux entrainés en avril 2023 (pour rappel l'année 2020 a été produite avant 2018-2019 et sans l'usage d'algorithmes de *deep learning*). Pour la prédiction des causes de décès des certificats restants à coder après codage manuel et automatique, les modèles ont été réentraînés en septembre 2023 en prenant en compte les certificats codés jusqu'alors.

L'étape de vérifications consiste à valider ou corriger certaines données selon des règles préétablies. Ces vérifications portent sur certaines catégories de certificats et sont réalisées par l'équipe de codage. Elles peuvent porter sur du codage automatique Iris/Muse, du codage manuel ou encore sur le codage IA. Pour le codage IA, ces vérifications sont en pratique du codage manuel : pour des raisons technico-fonctionnelles, à ce jour, les agents ne peuvent pas avoir accès aux prédictions détaillées de l'IA via l'interface de codage. Ils ont accès à la cause initiale prédite et le modèle utilisé mais pas aux causes associées prédites.

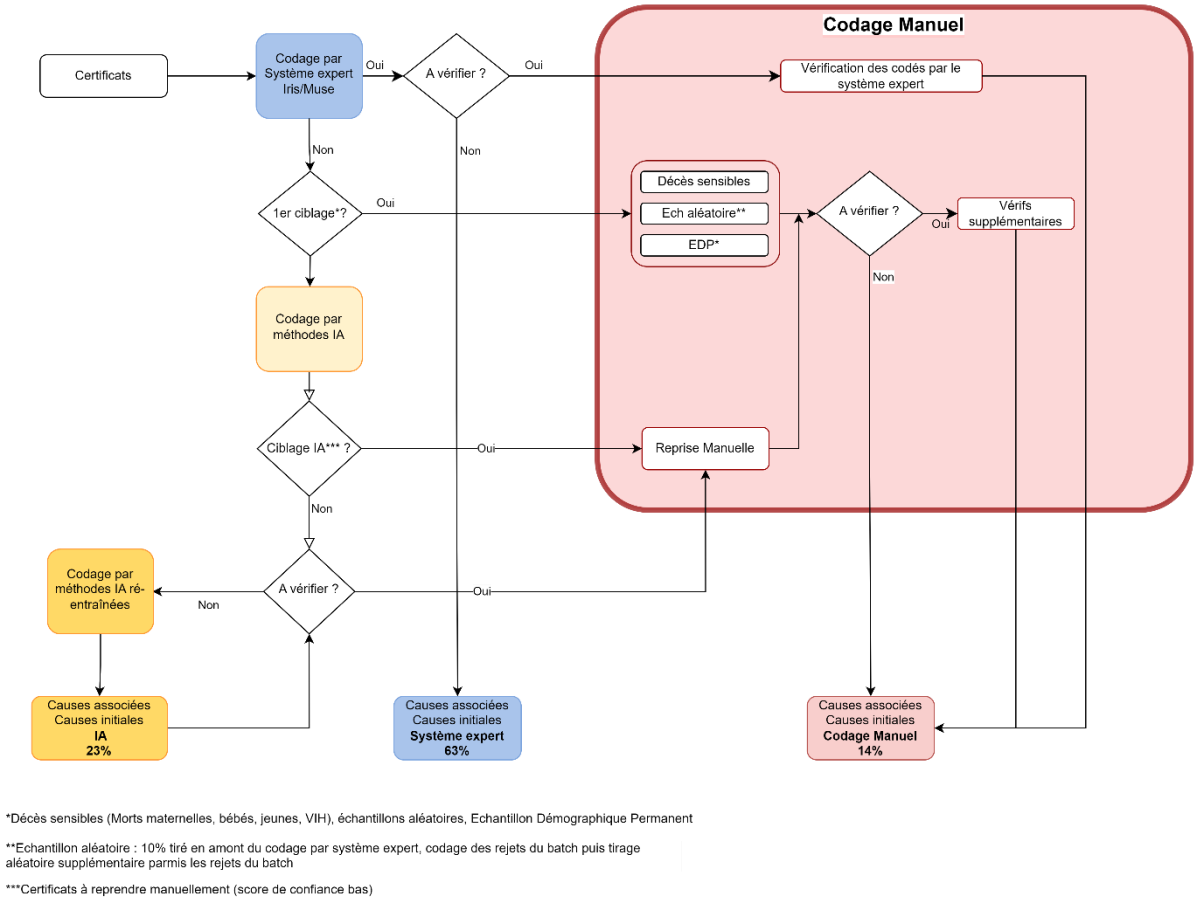


Figure 1. Circuit de codage combinant les trois modes et répartition pour l'année 2021

3.2 Répartition des modes de codage

Au total, pour les décès 2021, le CépiDc a eu recours aux trois modes de codage pour coder les causes multiples et déterminer la cause initiale, selon la répartition suivante :

- Le codage automatique à l'aide du système expert Iris/Muse (Batch) : 63%
  - Version du logiciel Iris : Iris 5.8.1
  - Version du moteur Muse utilisé dans Iris : Muse 2.9 SPECV2021SR30
- Le codage manuel assisté par Iris/Muse (14%)
- Une prédiction par *deep learning*, possiblement combinée au système expert Iris/Muse (Codage IA) : 23%.

Année / Type de codage	Manuel	Codage prédit par algo IA	Codage automatique IRIS/MUSE	Total
2021 (Nbre)	92930	149275	406495	648700
2021 (%)	14%	23%	63%	100%

Tableau 2. Répartition des modes de codage des données 2021 hors volets médicaux non reçus

Parmi les certificats de décès 2021, 93 000 certificats ont été codés manuellement, 406 000 codés par batch automatique Iris/Muse et 149 200 par IA. Parmi les 93 000 certificats codés manuellement, 54 000 proviennent de tirages uniformes dans la population des certificats de décès de l'année et

environ 39 000 de ciblages ou de vérifications (4 600 décès sensibles, 24 300 ciblages liés à l'IA, 2 300 certificats en fin de campagne repris ou codés, ainsi que les vérifications conduisant à des modifications manuelles).

3.3 Codage manuel

Le codage manuel des certificats non codés par batch a été ciblé selon différentes catégories de certificats et des échantillons ont été définis.

Pour la production 2021, le codage manuel (hors vérifications) couvre :

- Tirage aléatoire de 10% des certificats reçus, codage manuel des certificats non codé en batch par Iris/Muse : 29 086 certificats
- Tirage aléatoire parmi l'ensemble des certificats non codés en batch par Iris/muse : 16 073 certificats
- L'échantillon démographique permanent (EDP[1]) non codé en batch par Iris/Muse: 8447 certificats
- Une partie des décès "sensibles". Il s'agit de types de décès pour lesquels le CépiDc doit garantir l'exactitude du codage pour des raisons de santé publique (VIH, morts maternelles, décès d'enfants) : 4 639 certificats. Les décès sensibles codés par batch par Iris/Muse sont quant à eux vérifiés par l'équipe de codage (cf phase de vérification 3.6). Certains de ces certificats ont été codés manuellement dans le cadre de l'EDP ou du tirage aléatoire.

Tableau 3 Description et effectifs des décès identifiés comme sensibles en 2021 hors volets médicaux non reçus

Catégorie de décès sensibles en 2021	**Nbre de certificats**
Moins de 28 jours (0-27 jours inclus)	1859
Au moins 28 jours – Moins de 15 ans	1931
Morts Maternelles	122
SIDA/VIH	727

Les décès sensibles correspondent en 2021 à toutes les mentions de SIDA/VIH sur le certificat, les morts maternelles ainsi que presque tous les décès de moins de 15 ans. Une partie des certificats de décès entre 2 et 15 ans codables automatiquement par le système expert n'ont pas été repris manuellement.

- Les certificats issus des catégories les moins bien prédites par l'IA :

24 310 certificats choisis de façon à atteindre au moins 97% de cohérence de codage pour chaque catégorie de la *shortlist* Eurostat)

[1] Panel sociodémographique longitudinal de 4% de la population de l'insee : https://www.insee.fr/fr/metadonnees/source/serie/s1166

- Des lots de certificats ont été spécifiquement générés à l'issue de la prédiction finale car la performance de la stratégie de codage menée jusqu'alors restait insuffisante pour certaines catégories de décès (sous-estimation) :
    - en lien avec la tuberculose : dès lors qu'un modèle de *deep learning* remonte une cause de type tuberculose en cause initiale ou associée et la tuberculose n'est pas retenue en cause initiale, 256 certificats codés manuellement.
    - en lien avec la pharmacodépendance et la toxicomanie : un des modèles de *deep learning* remonte une cause de ce type en cause initiale et la cause initiale retenue ne relève pas de la catégorie, soit 227 certificats.
- Les certificats ayant une mention de vaccination de Covid-19, post Covid, Covid guéri ou de Covid long non codés par Iris/Muse : 1537
- Les volets médicaux complémentaires non codés automatiquement : 241

3.4 Codage par système expert Iris/Muse

La majorité des certificats sont codés automatiquement en appliquant le logiciel international Iris/Muse. Pour les décès 2021, il s'agit de la version d'Iris 5.8.1 et de la version de Muse 2.9 [voir les *metadata Causes of Death* sur le site d'Eurostat pour l'historique des versions utilisées du logiciel].

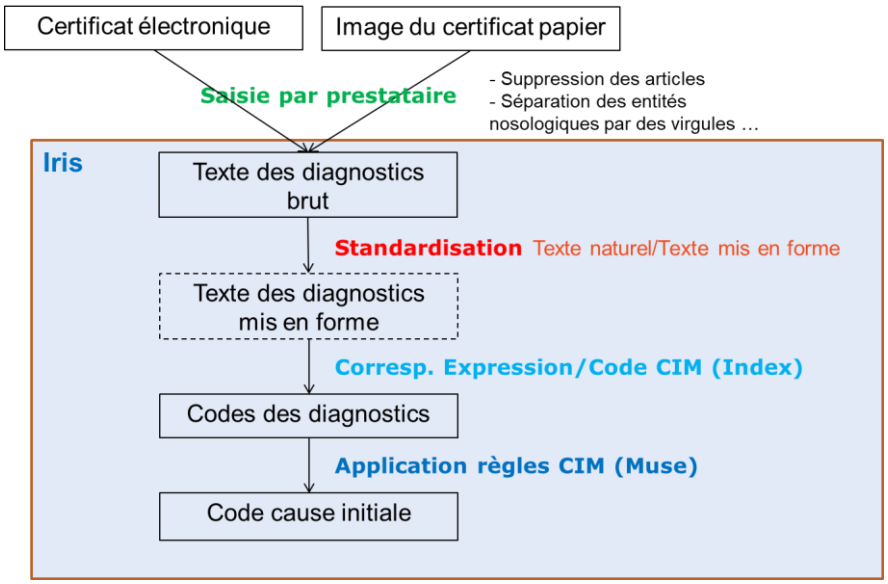


Figure 2. Codage par système expert Iris/Muse

Les textes des diagnostics issus des certificats électroniques et papier bruts, saisis ou corrigés par le prestataire de saisie et numérisation sont utilisés en entrée du logiciel Iris. Une étape de standardisation faisant intervenir des règles d'expressions régulières est réalisée sur ce texte brut afin de tenter de trouver une correspondance entre le libellé ainsi standardisé (« mis en forme ») et un code CIM à l'aide de l'index (ou « dictionnaire) implémenté dans Iris et maintenu par le CépiDc. En cas de réussite, les codes de diagnostics sont interprétés par le moteur Muse qui va appliquer les règles internationales d'identification de la cause initiale pour fournir le code qui sera retenu en cause initiale. En 2021, le système Iris/Muse conclut à une cause initiale pour 63% des certificats de décès. La majorité des certificats non codés sont rejetés à la première étape (le texte ne correspond pas à un libellé du dictionnaire). Pour en savoir plus : 'Statistiques sur les causes de décès de A à Z'.

3.5 Codage IA (Deep learning)

3.5.1 Ciblage IA des échantillons en reprise

Un peu plus de 24 300 certificats ont été ciblés par une méthode d'IA pour être envoyés en reprise manuelle. La reprise a été ordonnée par priorité de façon à assurer une précision de 94, 95 … puis finalement 97% pour les causes initiales de chaque catégorie de la *shortlist* européenne. Ceci a été réalisé en codant manuellement dans ces catégories les certificats pour lesquels un indicateur de confiance dans la prédiction IA de ce certificat était le plus bas. Au total, en prenant en compte l'ensemble de la population des décès de l'année quel que soit le mode de codage retenu, on estime qu'après cette reprise ciblée on atteint ou dépasse 97% de précision en comparant à une campagne traditionnelle (n'utilisant que du batch automatique et du codage manuel assisté) pour chacune des catégories de la *shortlist*. Pour au moins 97% des certificats codés dans une catégorie donnée de la *shortlist* européenne, la catégorie de la *shortlist* européenne dans laquelle va le code CIM de la cause initiale obtenu est la même que celle qui aurait résulté d'une campagne de codage traditionnelle.

Niveau de Précision à atteindre pour chaque catégorie de la *shortlist* européenne	Nbre de certificats en reprise IA 2021
94	6496
95	3180
96	4994
96.5	4250
97	5390
Total	24310

Tableau 4. Nombre de certificats repris manuellement après un ciblage IA lors de la campagne de codage 2021

Lecture : pour atteindre 94% de précision dans chacune des catégories de la *shortlist* européenne, on estime qu'il faut coder manuellement les 6496 certificats ciblés par l'IA : ceux pour qui les catégories en dessous du seuil de 94% de précision correspondent aux scores de confiance les plus bas.

Cette stratégie de ciblage est la même que celle retenue pour 2018 et 2019 mais son envergure en 2021 est beaucoup plus étendue.

En première étape, on utilise le modèle K5Iris (cf Codage des causes de décès 2018-2019, approche combinant *deep learning*, système expert et codage manuel ciblé) pour obtenir une prédiction de cause initiale pour chacun des certificats 2021 non codés automatiquement par batch. Ce modèle a été entrainé en avril 2023 et prend en compte dans sa base de *train* les certificats de 2021 codés à cette date (hors ceux qui ont été réservés pour alimenter la base de test).

Puis on estime l'indicateur de confiance dans la prédiction conditionnellement aux caractéristiques des certificats. Cet indicateur sera ensuite calculé pour chaque certificat. Il mesure la probabilité estimée de cohérence entre la cause initiale prédite par le modèle de *deep learning* et la cause initiale qu'aurait codée l'équipe de codage : plus il est élevé (proche de 1) plus on considère que la cause

initiale prédite a de fortes chances d'être la bonne. Pour ce faire, on estime un modèle de probabilité linéaire qui explique l'égalité entre les codes CIM de la cause initiale codée par l'équipe de codage et celle prédite par le modèle IA (cf 8.4).

3.5.2 Modèles de *deep learning* mobilisés et bases d'entrainement

Les modèles de *deep learning* mobilisés dans la production 2021 sont comme pour la production des données finales 2018 et 2019 des réseaux de neurones de type encoder-decoder, précisément des transformers. Une présentation détaillée de ces modèles est disponible dans le document de travail n2 du CépiDc « Codage des causes de décès de 2018 et 2019 en CIM10 Approche combinant *deep learning*, système expert et codage manuel ciblé », Zambetta et al. (2023). On ne reprend ici que des éléments de synthèse et les précisions concernant la production 2021.

3.5.2.1 les Transformers

Feature engineering/data pipeline.

Les séquences en entrée des modèles sont les concaténations des textes inscrits sur chaque ligne du certificat séparés par un token indiquant le numéro de la ligne. D'autres variables sont également ajoutées à la séquence sous forme de *tokens* spéciaux. Ces variables additionnelles comprennent systématiquement le sexe, le groupe d'âge, l'année de décès. Elles diffèrent ensuite selon le modèle. Le premier modèle (k4) ne contient pas de variable additionnelle supplémentaire. Le deuxième modèle (k5) contient en plus des variables précédentes, le type de certificat (électronique ou papier) la version du certificat (modèle 1997 ou 2017), ainsi que les circonstances apparentes de décès, nouvelle variable introduite dans les modèles de certificats 2017 (cf 8.1) qui permettent de mieux repérer certaines causes externes.

Ainsi pour le modèle k5, la phrase d'entrée est :

Paper-back/elec_certificate CertificateVersion sex agegroup yearofdeath sepLine1 text_ written_on_line_1 sepLine2 text_written_on_line2 … … sepLine7 death circumstances sepUC

La phrase de sortie / l'*output* a la même structure que la phrase d'entrée, simplement les codes en CIM remplacent les textes bruts, les circonstances apparentes de décès ne sont pas répétées. Le code de la cause initiale termine la phrase.

Paper-back /elec_certificate certificateVersion sex agegroup yearofdeath sepLine1 ICDcod11 ICDcod12 sepLine2 ICDcod2 … … sepLine7 sepUC ICDcodeUC

Exemple de séquence d'entrée et de séquence de sortie du modèle k5:

input sequence : certificatpapier versioncertificat1997 femme age55ans annee2017 lignecause1 arrêt cardio respiratoire lignecause2 épanchement pleural lignecause3 métastases pulmonaires lignecause4 cancer sein lignecause7 mort naturelle causeinitiale

output sequence : [*start*] certificatpapier versioncertificat1997 femme age55ans annee2017 lignecause1 r092 lignecause2 j90 lignecause3 c780 lignecause4 c509 lignecause7 causeinitiale c509 [*end*]

Pour être utilisées par le modèle ces séquences sont découpées en éléments de base ou « *token* ». L'algorithme qui réalise ce découpage est "*Tokenizer*". En entrée, il découpe les séquences en mots après une étape de normalisation simple (passage en minuscule, suppression des accents et des caractères spéciaux). En sortie il découpe les séquences en codes CIM unitaires et *tokens* spéciaux. Le dictionnaire en entrée de k5 comprend 150 037 *tokens* et en sortie 6 333 *tokens*. Le dictionnaire en entrée de k4 comprend 149 775 et en sortie 6 328 *tokens*.

Architecture du modèle

L'architecture de *Transformers* utilisée ici est similaire à publication initiale (Vaswani 2017). Elle est de type encodeur/décodeur. Les entrées sont représentées par leur plongement dans un espace vectoriel de taille finie (512) (*embedding*) et la position des mots dans la phrase (*positional encoding*). L'encodeur des modèles *Transformers* applique à la séquence d'entrée plusieurs fois les mêmes couches combinant une modélisation du mécanisme d'attention à plusieurs têtes (qui permet de tenir compte des liens entre les mots) et une couche *feed-forward* complètement connectée, tout cela suivi d'une normalisation. Le décodeur répète aussi ces mêmes couches sur la séquence de sortie en intercalant une modélisation du mécanisme d'attention sur *l'output* de l'encodeur. Chaque groupe de couches se termine aussi par une couche *feed-forward* complètement connectée et une étape de normalisation. La sortie du décodeur passe ensuite par une transformation linéaire et une fonction *softmax* permettant de convertir *l'output* du décodeur en probabilités prédites du mot suivant. Les modèles k4 et k5 ont chacun 151 millions de paramètres à estimer. L'annexe 8.8 illustre l'architecture de ce type de réseau, et précise les hyperparamètres choisis et les codes du modèle k5 2021 et du modèle k4 2021.

Bases d'entrainement

Les modèles sont entraînés sur des certificats de décès "annotés", c'est-à-dire pour lesquels on dispose de la séquence des causes multiples codées en CIM et de la cause initiale choisie. Dans le cas de la production 2021 les deux modèles k4 et k5 sont entrainés sur les mêmes bases d'entrainement.

Une première base d'entraînement de 5 330 657 certificats a été mobilisée en avril / mai 2023 pour l'étape de ciblage via l'utilisation de k5 Iris.

La deuxième base d'entraînement qui comporte 5 317 845 certificats a été utilisée en juin 2023 pour produire des résultats provisoires. Elle se compose de :

- l'ensemble des données annotées des années 2011 à 2015 (codage automatique et codage manuel),
- l'ensemble des certificats codés automatiquement (batch) pour 2016 et 2017 ainsi que 300 000 observations tirées aléatoirement parmi celles codées manuellement pour 2016 et 2017
- l'ensemble des certificats codés automatiquement (système expert) pour les années 2018 et 2019 ainsi que la moitié des observations codées manuellement au 8 juin 2023 (en respectant cette proportion quel que soit l'échantillon codé)
- 78% du codage automatique (système expert) de 2020 et 56% du codage manuel, toujours tirés aléatoirement

- 96% du codage automatique 2021 et 40% du codage manuel en date du 8 juin 2023 (hors EDP, laissé en test)

La troisième base d'entrainement est utilisée après la reprise manuelle complète. Elle comprend en plus des certificats de la deuxième base d'entrainement 50% des certificats codés manuellement lors des reprises de l'IA soit environ 10 713 certificats de plus que la première base, et au total 5 382 558 certificats. Cette base d'entraînement permet l'identification finale des causes de décès des certificats non codés ni manuellement ni par le système expert.

La base d'entraînement est toujours séparée en échantillon d'entraînement proprement dit et échantillon de validation comprenant 20% du *train* (tiré aléatoirement une fois, avant l'entraînement).

Base de test

Le test, qui inclut uniquement des observations annotées qui ne figurent pas dans la base d'entraînement, comprend 478 129 observations, dont 365 087 codées manuellement.

La table en 8.5 récapitule tous les échantillons et leurs inclusions dans les bases de test et de *train*.

Stratégie d'entraînement

Le modèle k5 a d'abord été entraîné sur la première base de *train*/validation comprenant près de 5,3 millions d'observations au printemps 2023. Ce sont ces résultats après application d'Iris/Muse qui ont été mobilisés pour réaliser le ciblage des certificats à reprendre (cf plus haut).

Les modèles k4 et k5 ont été réentraînés « *from scratch* » sur la troisième base d'entrainement en septembre 2023 (5,38 millions d'observations).

Les certificats de 2018 et 2019 codés par les méthodes IA ne sont évidemment pas réintroduits dans la base d'entrainement.

3.5.2.2 Un sur-modèle pour retenir une cause initiale parmi les différentes propositions possibles

L'output prédit par chaque réseau mobilisé peut conduire à deux propositions pour la cause initiale. En effet, il est possible de s'appuyer directement sur la cause initiale prédite par le modèle, située en dernière position de la phrase. Il est aussi possible d'appliquer le système expert de codage Iris/Muse sur la séquence de causes multiples prédites, et retenir le choix de cause initiale auquel il aboutit, lorsqu'il y en a un. De plus, les deux modèles k4 et k5 peuvent proposer deux causes initiales différentes, et deux séquences de causes pouvant conduire lorsqu'on applique Iris/Muse à des propositions de causes initiales différentes aussi. Au total, il y a donc potentiellement 4 propositions de cause initiale - celles provenant directement des algorithmes k4 et k5, et celles après passage d'Iris/Muse sur les séquences de causes prédites par les algorithmes, K4Iris et K5Iris; ainsi que deux séquences de causes prédites. A noter que lorsque Iris ne conclut pas, c'est la cause initiale directement prédite par l'algorithme qui est reprise. On ne dispose alors que d'une seule proposition par algorithme.

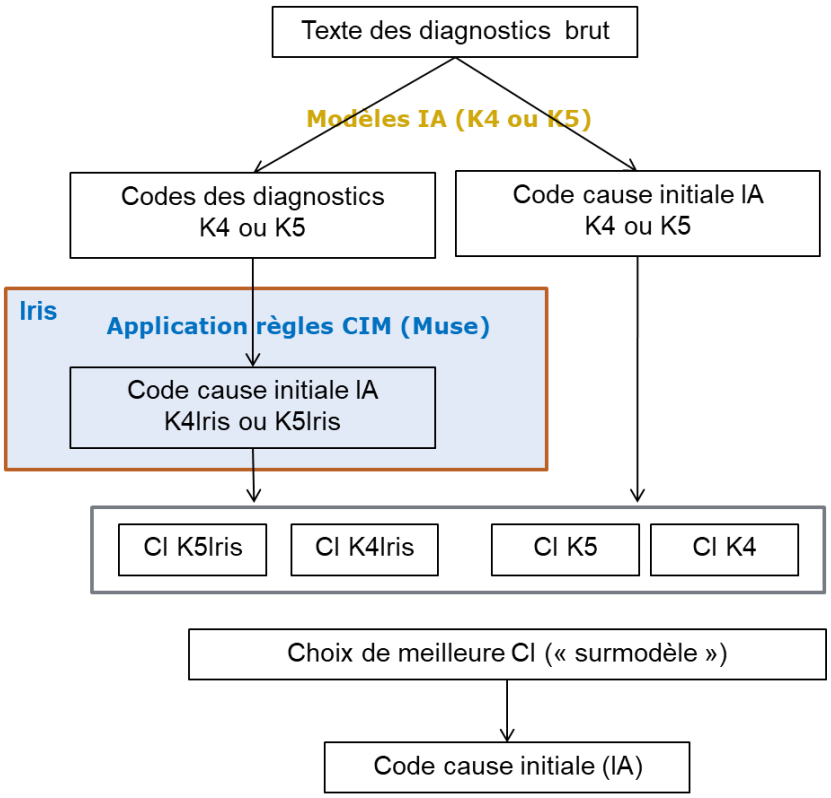


Figure 3. Description des propositions possibles de causes initiales impliquant du codage par IA

On a recours à un "sur-modèle" relevant aussi de l'apprentissage supervisé pour choisir entre ces propositions. Ce sur-modèle répond à un problème de classification en 5 classes, désignant parmi les modèles précédents celui dont on retiendra la proposition de cause initiale et par extension de causes associées, ou si aucun des modèles n'aboutit à une bonne prédiction (10% des cas dans le *train*). Dans ce dernier cas, on retiendra la prédiction de K5Iris.

Les séquences en entrée du sur-modèle concatènent les codes en CIM-10 des causes initiales et des causes associées prédites par k4 et k5, et leurs regroupements au niveau de la shortlist européenne (86 postes). On ajoute comme autres variables explicatives dans la séquence la valeur de la probabilité associée à la sortie de k4 et celle de k5 ainsi que l'écart entre cette probabilité et la probabilité de la deuxième cause la plus probable selon le modèle. Cette dernière variable capte le pouvoir discriminant de l'algorithme. La séquence comprend enfin aussi le type de certificat (électronique ou papier), les circonstances apparentes de décès, le nombre de causes associées (indicateur de complexité du certificat), et le nombre de fois où les modèles prédisent le même code pour la cause initiale (indicateur de fiabilité de cette proposition). Ainsi la séquence d'entrée est :

"keras4_ci keras5_ci keras4iris_ci keras5iris_ci keras4_86postes keras5_86postes keras5iris_86postes keras4iris_86postes keras4_list_causes_associees keras5_list_causes_associees certificat_type age CircApparDeces proba_max4 proba_diff4 proba_max5 proba_diff5 nb_causes_k4 nb_causes_k5 nb_egal"

Le modèle retenu est un *long-term short-term memory* bidirectionnel (BiLSTM). L'entraînement se fait sur la base d'apprentissage de k4 et de k5, uniquement sur les données codées manuellement depuis 2016 (509 933 observations). Les prétraitements, l'architecture du modèle et les codes sont reportés en annexe 8.9. Les variables ayant le plus fort pouvoir explicatif, selon les valeurs de Shapley sont le

nombre de fois où le même code est proposé, ainsi que les regroupements au niveau de la *shortlist* européenne des causes initiales prédites (valeurs de Shapley, voir 8.9). En termes d'évaluation, le sur-modèle sélectionne la bonne « classe », c'est-à-dire « modèle dont on devra retenir la proposition de CI » dans 81,6% des cas dans le test (contre 85,6% pour la production 2018/2019 à l'époque pour les deux modèles k4 et k5 n'avaient pas été entrainés sur la même base d'apprentissage), et 85,9% sur le *train*. La valeur de la CI sélectionnée par application du sur-modèle est égale dans 81,5% des cas à la CI que l'on aura obtenue dans le cas d'un codage manuel (81,9% en 2018/2019) contre 80,1% si l'on utilise seulement le modèle K5Iris, soit un gain de 1,4 point (sur le test). C'est moins que ce que l'on obtenait pour la production 2018-2019 où le gain était de 2,4 points. La réduction du gain du sur-modèle provient du fait que pour la production 2021 les deux modèles k4 et k5 ayant été entrainés sur la même base d'apprentissage sont beaucoup plus proches dans leurs prédictions et l'apport du sur-modèle est plus faible.

3.6 Vérifications

Dans le processus de codage, la phase de vérification consiste à s'assurer de la qualité de codage de certaines catégories de certificats identifiées comme étant à risque de mauvais codage. Différents types de vérification ont été effectuées pour l'année 2021, quel que soit le type de codage (manuel, automatique par système expert, IA). Il peut s'agir de vérifications liées à des incohérences manifestes (vérifications dites de « cohérence »), des vérifications liées à des anomalies identifiées du système expert, ou encore liées à des évolutions de règles de codage pour lesquels il est nécessaire de vérifier leur bonne application. Cette phase de vérifications pour l'année 2021 est décrite dans les paragraphes ci-dessous. Un certificat qui a été modifié dans le cadre des vérifications correspond finalement à un certificat codé manuellement.

En 2021, environ 11 000 certificats ont ainsi été vérifiés, 3527 initialement codés par l'IA, 4511 par Iris/muse, 3143 manuellement. 33% des certificats vérifiés changent de cause initiale après vérification : 33% des certificats codés par IrisMuse ; 28% des certificats codés manuellement ; 35% des certificats codés par l'IA. Il peut aussi y avoir eu des modifications concernant les causes associées qui n'ont pas conduit à une modification de cause initiale.

3.6.1 Cohérences

Une première série de vérifications consiste à vérifier les cohérences entre les codes de causes initiales et le sexe ou l'âge du défunt (cancer de la prostate chez une femme, pathologie hormonale à un âge trop ou pas assez avancé…). De plus, certains codes qui ne peuvent pas être utilisés en cause initiale de décès selon les recommandations doivent être modifiés. Ces vérifications ont concerné 127 certificats et 96% des causes initiales ont été modifiées à l'issue de la vérification.

3.6.2 Décès spécifiques

L'OMS et les enjeux de santé publique mettent en évidence des décès spécifiques pour lesquels il faut garantir un codage optimal selon les informations à disposition sur le certificat. Une majorité de ces

certificats étaient dans l'échantillon des décès sensibles tel que décrit en 3.3 et ont été codés par des experts. Pour les autres (non identifiés au moment de la constitution de l'échantillon des décès sensibles), ils ont été vérifiés par des experts. On détaille dans le Tableau 5 les certificats concernés.

Tableau 5. Description et nombre de certificats vérifiés pour les décès spécifiques et part de cause initiale modifiées

Réf	Description	Nbre de certificats vérifiés	% de CI modifiées
DS11 a	Certains certificats spécifiques avec mention de Covid-19 (cf 8.7) (vérification de l'application de règles complexes de codage, encore récentes)	1929	17%
DS10	Les certificats ayant une cause initiale de la catégorie intention indéterminée (commençant par Y1 ou Y2 ou Y3[0-4])	1280	25%
DS11 b	Les certificats ayant une mention de vaccination de Covid-19, post Covid-19, Covid-19 guéri ou de Covid-19 long	719	5%
DS1	Les certificats avec une mention de suicide mais non retenu en cause initiale	162	52%
DS4	Décès d'enfants entre 28 jours et 15 ans avec une mort violente ou pathologie spécifique (dont codes P non codés dans l'échantillon des décès sensibles (cf 8.6)	77	10%
DS5	Certificats avec mention de SIDA/VIH non codés dans l'échantillon des décès sensibles	68	28%
DS3a	Décès d'enfants entre 0 et 27 jours (dont les enfants nés sans vie) non codés dans l'échantillon des décès sensibles	62	23%*
DS6	Certificats de morts maternelles non codés dans l'échantillon des décès sensibles	11	27%
DS2	Les certificats avec une mention d'homicide mais non retenu en cause initiale	16	70%
DS3b	Age entre 28 jours et 6 mois inclus et mention de code P non codés dans l'échantillon des décès sensibles	2	50%
Total		4326	

*Hors 43 données manquantes avant vérifications

3.6.3 Difficultés de codage connues liées au logiciel Iris

Le logiciel (système expert IRIS/MUSE) qui permet le codage par *batch*, le codage manuel semi-assisté et une partie du codage réalisé avec des algorithmes de *deep learning* a des lacunes identifiées. Certains certificats ont donc été vérifiés pour palier à ces lacunes.

De plus, suite à un choix de privilégier l'amélioration du taux de codage automatique, des vérifications ont été mises en place sur des certificats dont on sait que le codage est erroné. Par exemple : la standardisation du terme « ou » en « , » permet de coder automatiquement et de façon correcte la majorité des certificats, mais une partie, notamment en lien avec les morts violentes doit être vérifiée pour s'assurer que l'interprétation du « ou » est correcte selon les règles de codage. De même avec le

terme « sur » qui est automatiquement considéré comme une relation causale. On détaille dans le Tableau 6 les certificats concernés.

Tableau 6. Description des vérifications liées au système expert et nombre total de certificats associés et part de cause initiale modifiées

Réf	Description	Nbre de certificats vérifiés	Part de CI modifiées
M10	Mention de "sur" suivi de certains codes laissant à penser que le « sur » ne doit pas être considéré comme un lien de causalité, avec impact potentiel sur la CI	735	96%
M2	Cause initiale en Y avec un code T sur le certificat : dans ces cas-là, le plus souvent une autre pathologie doit être en CI	505	70%
M1	Vérification des certificats avec un code D00-D48 en CI alors qu'il y a mention de cancer	428	43%
M11	Anémies en cause initiale (D619, D649 et D539) avec un code informatif présent en partie 1 autre qu'un code d'anémie	376	33%
M7	Traitement du "ou" si mention de mort violente à côté du "ou"	247	62%
M5	Notion de surdosage AVK en partie 1	162	45%
M4	Code d'affection hématologique (D50-D89) en cause initiale alors qu'il existe un code C sur le certificat	124	82%
M6	Codes en CI qui commencent par :'M60'ou 'M79' ou 'N94' ou 'N93'' ou 'N92' : Muse s'arrête sur ces codes même s'il y a d'autres codes sur les lignes suivantes	100	22%
M3	Causes associées T81-T82 donnant X59 en CI	70	7%
M8	D611 en CI et absence de cancer dans le certificat	17	77%
M9	Cause initiale à coder en C80.9 si C34.9 ou un C41.1 en cause associée	5	100%
Total		2769	

3.6.4 Vérification de la bonne application des nouvelles règles de 2016

L'OMS a introduit des changements dans les règles de sélection de la cause initiale depuis 2016. Selon la complexité de ces changements, des vérifications sont encore réalisées pour s'assurer de la bonne compréhension et application de ces règles. En 2021, cela concernait les hépatites virales, le diabète, certaines chutes, et les sepsis.

Tableau 7. Description des vérifications liées aux nouvelles règles et nombre total de certificats associés et part de cause initiale modifiées.

Réf	Description	Nbre de certificats vérifiés	Part de CI modifiées
NR1	Les hépatites virales ne peuvent plus être dues à d'autres pathologies : les certificats avec mention d'hépatites virales non retenues en cause initiale	255	16%

NR4-NR8	Le diabète peut désormais être dû à une liste de pathologies restreintes	161	6%
NR13	Les sepsis et SIRS (A40-A41 et R65) peuvent avoir être dû à (E40-E46 si mentionné)	20	35%
NR9	Les accidents (V01-X59) ou chutes W00-W19 doivent être dû à une autre cause	6	83%
NR12	Les sepsis et SIRS (A40-A41 et R65) peuvent avoir être dû à (D80-D84 si mentionné)	2	100%
Total		444	

3.6.5 Vérifications liées aux évolutions du dictionnaire (démarche des choix de code)

A partir de 2018, le dictionnaire a été mis à jour, notamment les libellés faisant l'objet de "choix de code". On définit un "choix de code" comme étant un libellé qui peut être codé de deux façons différentes selon le contexte du certificat. Cela nécessitait systématiquement un regard humain alors que dans certains cas, le choix était toujours le même ou les autres n'étaient pas pertinents. Des suppressions de choix de codes au profit d'un code "préféré" ont été réalisées et ont permis de diminuer les certificats inutilement codés manuellement. Cependant, certaines situations peuvent laisser le doute et des vérifications ont donc été définies.

Par exemple, le libellé "épuisement" pouvait être codé soit en R53 (malaise, fatigue) soit en T73.3 (surmenage lié à effort intensif) soit en X50 (surmenage lié à effort répété). Il a été décidé de supprimer le choix de code pour ce libellé et de permettre le codage automatique en R53. Cependant, il a été décidé de vérifier les certificats qui mentionnent un épuisement chez les moins de 75 ans car c'est atypique et pourrait correspondre à un effort intensif ou sportif. Le contexte du certificat permet au codeur, lors de sa vérification de statuer et d'éventuellement recoder l'épuisement correctement.

Tableau 8. Description des vérifications liées à la démarche des choix de code et nombre total de certificats associés et part des causes initiales modifiées.

Réf	Description	Nbre de certificats vérifiés	Part de CI modifiées
CC_Hemorragie_innracerebrale	S'il y a une notion de causes externes dans un certificat avec mention d'hémorragie intracérébrale, le codage sera S068 et non I6199	763	16%
CC_IRC	On vérifie que le code J96.1 est approprié pour les certificats avec un libellé « IRC » et une notion de diabète ou certaines pathologies rénales/urinaires.	729	20%
CC_Hemorragie_interne	S'il y a une notion de traumatisme dans un certificat avec mention d'hémorragie interne, le code R5809 ne sera pas forcément approprié	277	24%
CC_OAP	On vérifie que le code I50.1 est approprié pour le libellé "OAP" dans les certificats avec mention d'"IR" et "IC" (potentiellement à reprendre en J81)	252	14%

CC_Alcool	S'il y a une notion de cause externe dans un certificat avec mention d'alcool, on vérifie que le code F10.2 est approprié	231	15%
CC_choc_hemorragique_intravolemique	S'il y a une notion d'acte diagnostique et thérapeutique invasif dans un certificat avec mention de choc hémorragique ou hypovolémique, on vérifie que le code R57.1 est approprié (potentiellement à reprendre en T179)	230	30%
CC_IRA	On vérifie que le code J96.0 est approprié pour les certificats avec un libellé « IRA » et une notion de diabète ou certaines pathologies rénales/urinaires.	158	16%
CC_epuisement	Libellé « épuisement » chez une personne de moins de 75 ans est souvent dû à un effort sportif. On vérifie que le code R53 est approprié chez cette catégorie	133	2%
CC_Thabdomyolyse	S'il y a une notion de traumatisme ou cause externe dans un certificat avec mention de rhabdomyolyse, on vérifie que le code M62.8 est approprié.	81	5%
CC_noyage	Libellé « noyade » et une information de dépression (F30, F32 ou F33) en partie 1 si les circonstances apparentes du décès sont inconnues. On vérifie que le code W74	74	23%
CC_paraplegie	On vérifie que le code G822 est approprié pour coder le libellé « paraplégie » s'il y a aussi une mention d'AVC sur le certificat	30	30%
total		2958	

Le tableau ci-dessus décrit les vérifications définies suite à cette démarche de suppression de choix de code dans le dictionnaire et réalisées pour la production 2021.

3.6.6 Vérifications pour raisons multiples

Certains certificats sont vérifiés pour plusieurs raisons citées ci-dessus. Ces cas de vérifications multiples ont concerné 558 certificats en 2021 et pour 56%, la cause initiale a été modifiée à l'issue de la vérification.

4 Evaluation de la méthode de codage

4.1 Population test de référence

La base de test, constituée d'observations annotées qui ont été exclues de l'apprentissage, va permettre d'évaluer la performance de la campagne de codage 2021, au sens de la cohérence entre le codage qui aurait été obtenu dans une campagne classique de codage combinant codage par *batch* et codage manuel assisté et celui obtenu ici. Pour l'évaluation, on construit une population de test représentative de la distribution des causes de décès dans la population. Celle-ci respecte les proportions de décès sensibles[2], de défunts EDP, observées en général dans la population des décès,

[2] La définition des décès sensibles dans la base de test est celle qui a été utilisée lors de la production des données finales 2018 et 2019. Elle est un peu plus restrictive que celle qui a effectivement eu lieu en

et la proportion de reprise manuelle ciblée telle qu'elle a été menée en 2021. Elle est aussi représentative dans les bonnes proportions des décès codés automatiquement par *batch*. Ainsi, en première étape, on se restreint aux seuls échantillons tirés aléatoirement (tests 2016, 2017, 2020 codés manuellement, et les sélections aléatoires de lots ou de certificats pour 2021), soit 332 183 observations. La deuxième étape consiste à compléter cette base avec des tirages proportionnels dans le codage automatique par batch pour chaque sous-échantillon. On obtient alors une population test de référence, composée de 797 651 observations, représentative de la répartition des causes de décès sur les années 2016, 2017, 2020 et 2021. La proportion de codage automatique par *batch* dans cette population est de 58% ce qui est un peu moins que le taux de codage automatique par batch effectif en 2021 (63%). La conséquence sera donc une légère sous-estimation de la cohérence du codage dans les tableaux présentés ci-dessous. Il s'agit de la même base que celle utilisée pour évaluer la production 2018 et 2019 – pour les détails de sa construction voir Zambetta et al. 2023.

On simule ensuite les apports de la reprise manuelle aléatoire et ciblée telle qu'elle a été menée en 2021, en supposant que la cause initiale codée est correcte pour les certificats relevant d'échantillons codés manuellement. Le raisonnement est le suivant : parmi les 245 689 certificats non codés automatiquement par batch en 2021, on en a codé manuellement 28 063 (ECH1), 7 453 (ECH4), 8 573 (ECH2), soit 18,6% de l'ensemble des certificats non codés automatiquement par batch (hors EDP). On va donc simuler, via un tirage uniforme, que 18,6% des certificats à coder manuellement dans le test (hors EDP) vont passer en reprise manuelle (soit 59 017 dans le test). On va aussi considérer que les certificats relevant de l'EDP (échantillon démographique permanent), des décès sensibles, ainsi que ceux relevant de la reprise manuelle ciblée par IA sont correctement codés car ils ont été codés manuellement dans la campagne 2021. Pour l'EDP, on repère les défunts nés les 2,3,4,5 janvier ou les 1,2,3,4 avril, juillet ou octobre et dont les certificats n'ont pas déjà été codés par *batch* (soit 14 715 observations dans le test) ce qui correspond à la définition de l'EDP. Pour les décès sensibles, on applique les règles de repérage utilisées pour les identifier sur les données de 2018/2019, soit 5 213 certificats dans le test) (cf Rapport de Production 2018-2019). Pour simuler l'impact de la reprise ciblée IA, on s'appuie sur les scores de confiance issues de k5 et les prédictions de cause de K5Iris. On applique la même part de reprise dans les catégories sur lesquelles la reprise manuelle a été ciblée que ce qui a été fait sur 2021. Il y en a 31 060 dans la population test de référence. L'apport de la phase de vérifications n'est pas simulé ici.

	codage auto batch	EDP manuel	décès sensibles	reprise aléatoire	reprise ciblée IA	total reprise manuelle	prédictions	total
nb de certificats	465468	14715	5213	59017	31060	108062	224121	797651
%	58,4%	1,8%	0,7%	7,4%	3,9%	13,5%	28,1%	100%

Tableau 9. description de la population test de référence

Note : les effectifs ne somment pas en ligne car certains certificats relèvent de plusieurs catégories.

Lecture : la population test de référence compte 14 715 certificats de défunts relevant de l'échantillon démographique permanent et qui n'ont pas été codés par batch automatique.

campagne 2021. En résulte de nouveau, une légère sous-estimation de la performance de la campagne 2021 dans les tableaux présentés.

4.2 Accuracy, cohérence globale et comparaison à la campagne précédente

Sur la partie de la population test de référence qui aurait été codée manuellement dans une campagne classique de codage, la cause initiale obtenue en combinant la prédiction du sur-modèle et la reprise manuelle ciblée égale la cause initiale codée par l'équipe de codage au niveau le plus fin de la CIM dans 89,7% des cas (pour 84,1% sur la campagne de rattrapage 2018/2019 ; pour rappel, 2020 a été produite avant 2018/2019 et sans algorithmes de *deep learning*). Elle relève de la même catégorie de la *shortlist* européenne dans 93,4% des cas (pour 89,3% en 2018/2019). Les **Erreur ! Source du renvoi introuvable.** et Tableau 11 détaillent les performances des prédictions par les différents modèles, combinés ou non avec Iris/Muse et avec la reprise manuelle ciblée telle qu'elle a été réalisée sur 2021.

Le Tableau 10 se concentre sur les certificats qui auraient été codés manuellement dans une campagne traditionnelle ne combinant que batch automatique et codage manuel. Le Tableau 11 reporte les mêmes indicateurs calculés sur l'ensemble de la population test de référence, c'est-à-dire y compris les certificats codés par batch pour lesquels il n'y a pas eu de changement entre cette campagne et une campagne traditionnelle (cohérence complète), de façon à fournir un niveau global de cohérence entre campagnes prenant en compte tous les modes de codage.

Tableau 10. Cohérence (accuracy) entre les causes initiales prédites par *deep learning* (k4 ou k5), combinaison de *deep learning* et Iris Muse, sur-modèle combiné ou non à la reprise manuelle et la cause initiale codée sur la population test de référence, pour la partie qui aurait été codée manuellement dans le cadre d'une campagne ne combinant que batch et codage manuel

Année	K5	IRIS5	K4	IRIS4	Surmodèle	Surmodèle+ reprise manuelle	Nb obs
	Accuracy au niveau du code CIM-10						
ensemble	0,791	0,801	0,786	0,799	0,815	0,897	332 183
2016	0,785	0,791	0,781	0,790	0,805	0,891	93 144
2017	0,780	0,785	0,776	0,784	0,801	0,888	93 912
2020	0,803	0,819	0,797	0,815	0,831	0,906	121 461
2021	0,798	0,817	0,792	0,815	0,827	0,901	23 666
	Accuracy au niveau de la catégorie de la shortlist européenne						
ensemble	0,860	0,865	0,856	0,863	0,876	0,934	332 183
2016	0,856	0,859	0,853	0,857	0,870	0,931	93 144
2017	0,851	0,852	0,848	0,852	0,865	0,927	93 912
2020	0,869	0,878	0,862	0,874	0,887	0,941	121 461
2021	0,865	0,877	0,858	0,875	0,884	0,937	23 666

Lecture : au niveau le plus fin de la CIM, la prédiction du modèle k5 est correcte dans 79,1% des cas qui auraient été codés manuellement dans une campagne traditionnelle, en prenant en compte toutes les étapes jusqu'à la reprise manuelle on arrive à une accuracy de 89,7%.

Au niveau le plus fin de la CIM, la prédiction du modèle k5 est correcte dans 79,1% des cas (78,5% en 2018/2019). Appliquer Iris/Muse sur la séquence des causes prédites par k5 lorsqu'il obtient une réponse non ambiguë fait gagner un point de cohérence. Les performances du modèle k4 sont très proches de k5 mais un peu moins bonnes. Pour autant, les deux modèles se complètent puisqu'en les

combinant via le sur-modèle on atteint 81,5% d'accuracy (contre 81,9% en 2018/2019). La prise en compte de la reprise aléatoire et ciblée permet de gagner 8,2 points supplémentaires (contre 2 points en 2018/2019) et d'atteindre les 89,7%. L'évaluation de chaque étape de la reprise ciblée sera détaillée par la suite. Au niveau *shortlist* européenne, le sur-modèle permet de gagner 1,4 point d'accuracy par rapport à K5Iris (k5 combiné avec Iris/Muse) (pour 1,7 point en 2018/2019) et la reprise aléatoire et ciblée 5,8 points (pour 1,6 point en 2018/2019). Au total, on atteint 93,4% de cohérence (89,4% en 2018/2019). Enfin, les performances sont stables au fil des années et ont même tendance à s'améliorer dans les années les plus récentes.

Tableau 11. Cohérence (accuracy) entre les causes initiales prédites par batch, *deep learning* (k4 ou k5), combinaison de *deep learning* et Iris Muse, sur-modèle combiné ou non à la reprise manuelle et la cause initiale codée sur la population test de référence, sur l'ensemble de la population test de référence

Année	K5 + batch	IRIS5 + batch	K4 + batch	IRIS4 + batch	Surmodèle + batch	Surmodèle+ reprise manuelle + batch	Nb obs
	Accuracy au niveau du code CIM-10						
ensemble	0,913	0,917	0,911	0,916	0,923	0,957	797 651
2016	0,910	0,912	0,908	0,912	0,918	0,954	221 807
2017	0,909	0,911	0,907	0,911	0,918	0,954	226 856
2020	0,916	0,923	0,914	0,922	0,928	0,960	285 784
2021	0,924	0,932	0,922	0,931	0,935	0,963	63 204
	Accuracy au niveau de la catégorie de la shortlist européenne						
ensemble	0,942	0,944	0,940	0,943	0,948	0,973	797 651
2016	0,940	0,941	0,938	0,940	0,945	0,971	221 807
2017	0,938	0,939	0,937	0,939	0,944	0,970	226 856
2020	0,944	0,948	0,941	0,946	0,952	0,975	285 784
2021	0,949	0,954	0,947	0,953	0,957	0,977	63 204

Lecture : dans 91,7% des cas la CI à 4 positions obtenue par codage batch si possible ou bien par prédiction de k5 combiné à Iris/muse (K5Iris) est la même que celle que l'on aurait obtenue en procédant à une campagne traditionnelle de codage combinant batch et codage manuel assisté. On monte à 94,4% de cohérence au niveau de la *shortlist* européenne

Si maintenant on prend en compte le fait qu'une très grande partie des certificats est codée par batch et que pour ces certificats le codage par rapport à une campagne classique ne change pas (la cohérence est donc parfaite) on obtient une cohérence parfaite dans 95,7% des cas au niveau le plus fin de la CIM (pour 93,4% en 2018/2019) et dans 97,3% des cas au niveau de la *shortlist* européenne (pour 95,6% en 2018/2019), voir Tableau 11.

4.3 Précision, rappel et écarts d'effectifs

Les Tableau 12 et Tableau 13 présentent les précisions, rappels, F-mesures et les effectifs prédits par catégorie de la *shortlist* européenne, pour le sur-modèle combiné au batch et lorsque l'on tient aussi compte de la reprise manuelle. La précision est la part de prédictions correctes rapportée à l'ensemble des prédictions dans la catégorie, le rappel est la part des observations correctement prédites par le modèle rapportée à l'ensemble des observations réellement dans la catégorie ; la F-mesure est la moyenne harmonique des deux.

Tableau 12. Performances et effectifs prédits du sur-modèle et du sur-modèle combiné avec la reprise manuelle évaluées sur les observations de la population test qui auraient été codées manuellement.

		Surmodèle						Surmodèle + reprise manuelle					
Test qui aurait été codé manuellement dans une campagne traditionnelle	Effectifs réels	Précision	Rappel	F-measure	Effectifs prédits	Pred/Réels - 1	sign de diff	Précision	Rappel	F-measure	Effectifs prédits	Pred/Réels - 1	sign de diff
01.1	424	0,887	0,816	0,85	390	-0,08	*	0,97	0,92	0,944	402	-0,052	
01.2	260	0,792	0,658	0,718	216	-0,169	****	0,985	1	0,992	264	0,015	
01.3	334	0,757	0,689	0,721	304	-0,09	*	0,916	0,853	0,884	311	-0,069	
01.4	5737	0,798	0,767	0,782	5515	-0,039	****	0,926	0,895	0,91	5545	-0,033	***
02.1.01	2761	0,934	0,88	0,906	2604	-0,057	****	0,957	0,929	0,943	2680	-0,029	*
02.1.02	2447	0,95	0,955	0,952	2458	0,004		0,965	0,972	0,968	2466	0,008	
02.1.03	2359	0,934	0,936	0,935	2365	0,003		0,955	0,967	0,961	2388	0,012	
02.1.04	9820	0,955	0,947	0,951	9739	-0,008		0,969	0,969	0,969	9812	-0,001	
02.1.05	4782	0,94	0,933	0,936	4746	-0,008		0,957	0,96	0,959	4798	0,003	
02.1.06	5411	0,969	0,963	0,966	5374	-0,007		0,978	0,979	0,979	5416	0,001	
02.1.07	654	0,898	0,875	0,886	637	-0,026		0,933	0,931	0,932	653	-0,002	
02.1.08	15882	0,943	0,959	0,951	16151	0,017	***	0,965	0,978	0,972	16103	0,014	**
02.1.09	1168	0,913	0,929	0,921	1188	0,017		0,936	0,953	0,944	1189	0,018	
02.1.10	6828	0,944	0,957	0,95	6923	0,014		0,963	0,976	0,97	6920	0,013	
02.1.11	524	0,922	0,929	0,926	528	0,008		0,96	0,96	0,96	524	-	
02.1.12	1664	0,939	0,913	0,926	1617	-0,028		0,961	0,948	0,955	1642	-0,013	
02.1.13	1785	0,955	0,939	0,947	1756	-0,016		0,966	0,966	0,966	1785	-	
02.1.14	4825	0,94	0,942	0,941	4834	0,002		0,96	0,964	0,962	4845	0,004	
02.1.15	2147	0,929	0,915	0,922	2116	-0,014		0,957	0,948	0,953	2127	-0,009	
02.1.16	2952	0,936	0,941	0,938	2968	0,005		0,958	0,963	0,96	2968	0,005	
02.1.17	2205	0,939	0,903	0,921	2121	-0,038	**	0,963	0,943	0,953	2160	-0,02	
02.1.18	267	0,878	0,891	0,885	271	0,015		0,902	0,933	0,917	276	0,034	
02.1.19	3253	0,932	0,935	0,934	3266	0,004		0,955	0,961	0,958	3273	0,006	
02.1.20	3643	0,939	0,932	0,936	3617	-0,007		0,956	0,954	0,955	3635	-0,002	
02.1.21	1959	0,917	0,918	0,918	1960	0,001		0,944	0,95	0,947	1972	0,007	
02.1.22	15015	0,856	0,874	0,865	15331	0,021	***	0,935	0,926	0,93	14877	-0,009	
02.2	5261	0,839	0,843	0,841	5289	0,005		0,922	0,919	0,92	5245	-0,003	
3	2033	0,713	0,618	0,662	1762	-0,133	****	0,924	0,826	0,872	1818	-0,106	****
04.1	7313	0,904	0,847	0,875	6855	-0,063	****	0,943	0,907	0,925	7033	-0,038	****
04.2	5987	0,806	0,775	0,79	5754	-0,039	****	0,912	0,877	0,894	5755	-0,039	****
05.1	8407	0,844	0,924	0,882	9200	0,094	****	0,892	0,952	0,921	8976	0,068	****
05.2	1510	0,794	0,825	0,809	1570	0,04	*	0,91	0,921	0,916	1528	0,012	
05.3	199	0,676	0,628	0,651	185	-0,07		0,935	0,869	0,901	185	-0,07	
05.4	2493	0,807	0,765	0,786	2363	-0,052	****	0,906	0,868	0,887	2390	-0,041	***
06.1	2915	0,91	0,934	0,922	2991	0,026	*	0,939	0,96	0,949	2978	0,022	
06.2	7994	0,939	0,945	0,942	8045	0,006		0,957	0,963	0,96	8047	0,007	
06.3	7316	0,831	0,815	0,823	7174	-0,019	**	0,925	0,913	0,919	7221	-0,013	
07.1.1	6433	0,874	0,905	0,89	6661	0,035	****	0,917	0,942	0,929	6609	0,027	***
07.1.2	11020	0,864	0,874	0,869	11154	0,012		0,916	0,925	0,92	11134	0,01	
07.2	23508	0,859	0,864	0,862	23625	0,005		0,915	0,925	0,92	23764	0,011	**
07.3	18752	0,884	0,889	0,887	18846	0,005		0,933	0,939	0,936	18869	0,006	
07.4	14925	0,849	0,83	0,839	14601	-0,022	****	0,922	0,91	0,916	14719	-0,014	**
08.1	760	0,908	0,918	0,913	769	0,012		0,958	0,957	0,957	759	-0,001	
08.2	4640	0,836	0,842	0,839	4671	0,007		0,906	0,908	0,907	4653	0,003	
08.3.1	425	0,823	0,833	0,828	430	0,012		0,914	0,927	0,921	431	0,014	
08.3.2	5630	0,868	0,892	0,88	5787	0,028	***	0,922	0,939	0,93	5733	0,018	*
08.4	6656	0,773	0,765	0,769	6589	-0,01		0,885	0,87	0,878	6545	-0,017	*
09.1	598	0,837	0,841	0,839	601	0,005		0,925	0,931	0,928	602	0,007	
09.2	4084	0,887	0,915	0,901	4211	0,031	***	0,937	0,956	0,946	4168	0,021	*
09.3	11248	0,853	0,845	0,849	11145	-0,009		0,926	0,919	0,923	11169	-0,007	
10	1185	0,754	0,752	0,753	1182	-0,003		0,911	0,887	0,899	1154	-0,026	
11.1	435	0,788	0,72	0,752	397	-0,087	**	0,914	0,878	0,896	418	-0,039	
11.2	3136	0,716	0,724	0,72	3173	0,012		0,897	0,878	0,887	3068	-0,022	
12.1	4459	0,807	0,769	0,788	4252	-0,046	****	0,909	0,876	0,893	4296	-0,037	***
12.2	2352	0,81	0,78	0,794	2265	-0,037	**	0,899	0,878	0,888	2297	-0,023	
13	51	0,885	0,451	0,597	26	-0,49	****	1	1	1	51	-	
14	1762	0,94	0,942	0,941	1766	0,002		0,997	1	0,999	1767	0,003	
15	1384	0,863	0,721	0,786	1156	-0,165	****	0,947	0,896	0,921	1310	-0,053	***
16.1	174	0,895	0,931	0,913	181	0,04		0,983	0,994	0,989	176	0,011	
16.2	4747	0,789	0,888	0,835	5343	0,126	****	0,857	0,929	0,892	5145	0,084	****
16.3	6428	0,838	0,873	0,855	6698	0,042	****	0,886	0,918	0,902	6654	0,035	****
17.1.1	2283	0,947	0,896	0,921	2159	-0,054	****	0,964	0,938	0,951	2221	-0,027	*
17.1.2	8520	0,91	0,935	0,922	8745	0,026	***	0,962	0,968	0,965	8570	0,006	
17.1.3	395	0,842	0,861	0,851	404	0,023		0,934	0,924	0,929	391	-0,01	
17.1.4	1610	0,76	0,691	0,724	1463	-0,091	****	0,929	0,887	0,908	1537	-0,045	**
17.1.5	12758	0,845	0,847	0,846	12796	0,003		0,932	0,927	0,929	12688	-0,005	
17.2	4999	0,916	0,926	0,921	5053	0,011		0,962	0,963	0,962	5007	0,002	
17.3	382	0,819	0,662	0,732	309	-0,191	****	0,981	0,95	0,965	370	-0,031	
17.4	1404	0,684	0,624	0,653	1281	-0,088	****	0,921	0,857	0,888	1306	-0,07	****
17.5	1570	0,447	0,299	0,358	1049	-0,332	****	0,881	0,714	0,789	1272	-0,19	****
18	12936	0,95	0,971	0,96	13212	0,021	***	0,965	0,979	0,972	13123	0,014	*

Note: les degrés de significativité des écarts entre effectifs prédits et réels proviennent de tests d'égalité supposant que les effectifs réels suivent des lois de Poisson, * pval<.2, ** pval<.1, *** pval<.05, **** pval<.01

Tableau 13. Performances et effectifs prédits du batch combiné au sur-modèle et au sur-modèle combiné et à la reprise manuelle évaluées sur l'ensemble de la population test de référence.

				Batch + Surmodèle						Batch+ Surmodèle + Reprise manuelle					
Population test de référence (y compris codage batch)			Effectifs réels	Précision	Rappel	F-measure	Effectifs prédits	Pred/Réels - 1	sign de diff	Précision	Rappel	F-measur	Effectifs prédits	Pred/Réels - 1	sign de diff
01.1	Tuberculose	A15-19, B90	476	0,900	0,836	0,867	442	-0,071	*	0,974	0,929	0,951	454	-0,046	
01.2	SIDA (maladie VIH)	B20-B24	332	0,844	0,732	0,784	288	-0,133	***	0,988	1,000	0,994	336	0,012	
01.3	Hepatites virales	B15-B19,B94.2	560	0,860	0,814	0,837	530	-0,054		0,952	0,913	0,932	537	-0,041	
01.4	Autres maladies infectieuses et parasitaires	autres A et autres B	12936	0,912	0,897	0,904	12714	-0,017	**	0,968	0,953	0,961	12744	-0,015	**
02.1.01	TM levre. bouche. pharinx	C00-C14	4996	0,964	0,934	0,949	4839	-0,031	***	0,977	0,961	0,969	4915	-0,016	
02.1.02	TM de l'œsophage	C15	4797	0,975	0,977	0,976	4808	0,002		0,982	0,986	0,984	4816	0,004	
02.1.03	TM de l'estomac	C16	5790	0,973	0,974	0,974	5796	0,001		0,982	0,987	0,984	5819	0,005	
02.1.04	TM Colon.rectum.anus	C18-C21	23061	0,981	0,977	0,979	22980	-0,004		0,987	0,987	0,987	23053	-0,000	
02.1.05	TM du foie et voies billaires	C22	11426	0,975	0,972	0,973	11390	-0,003		0,982	0,983	0,983	11442	0,001	
02.1.06	TM du pancreas	C25	15433	0,989	0,987	0,988	15396	-0,002		0,992	0,993	0,993	15438	0,000	
02.1.07	TM du larynx	C32	1271	0,948	0,935	0,942	1254	-0,013		0,965	0,965	0,965	1270	-0,001	
02.1.08	TM de la trachee, bronche, poumon	C33-C34	40493	0,977	0,984	0,981	40762	0,007	*	0,986	0,992	0,989	40714	0,005	
02.1.09	Mélanome malin de la peau	C43	2241	0,954	0,963	0,959	2261	0,009		0,966	0,975	0,971	2262	0,009	
02.1.10	TM du sein	C50	16601	0,977	0,982	0,980	16696	0,006		0,985	0,990	0,987	16693	0,006	
02.1.11	TM du col de l'uterus	C53	1048	0,961	0,965	0,963	1052	0,004		0,980	0,980	0,980	1048	-	
02.1.12	TM autre partie de l'uterus	C54-C55	3630	0,973	0,960	0,966	3583	-0,013		0,982	0,976	0,979	3608	-0,006	
02.1.13	TM de l'ovaire	C56	4424	0,982	0,976	0,979	4395	-0,007		0,986	0,986	0,986	4424	-	
02.1.14	TM de la prostate	C61	11882	0,976	0,976	0,976	11891	0,001		0,984	0,985	0,984	11902	0,002	
02.1.15	TM du rein	C64	4626	0,967	0,961	0,964	4595	-0,007		0,980	0,976	0,978	4606	-0,004	
02.1.16	TM de la vessie	C67	6874	0,972	0,975	0,973	6890	0,002		0,982	0,984	0,983	6890	0,002	
02.1.17	TM cerveau et syst. nerveux central	C70-C72	5232	0,975	0,959	0,967	5148	-0,016		0,985	0,976	0,980	5187	-0,009	
02.1.18	TM de la thyroide	C73	490	0,933	0,941	0,937	494	0,008		0,946	0,963	0,954	499	0,018	
02.1.19	Maladie de Hodgkin et lymphomes	C81-C86	6393	0,965	0,967	0,966	6406	0,002		0,977	0,980	0,979	6413	0,003	
02.1.20	Leucemie	C91-C95	7856	0,972	0,969	0,970	7830	-0,003		0,979	0,978	0,979	7848	-0,001	
02.1.21	Autres TM des tissus lymphoides et hematopoietiques	C88,C90,C96	4290	0,962	0,962	0,962	4291	0,000		0,974	0,977	0,976	4303	0,003	
02.1.22	Autres Tumeurs malignes	autres C	29282	0,926	0,935	0,930	29598	0,011	**	0,967	0,962	0,964	29144	-0,005	
02.2	Tumeurs non malignes	D00-D48	10175	0,916	0,919	0,918	10203	0,003		0,960	0,958	0,959	10159	-0,002	
3	Maladies du sang et hematopoietiques	D50-D89	3491	0,843	0,777	0,809	3220	-0,078	****	0,958	0,899	0,927	3276	-0,062	****
04.1	Diabete sucre	E10-E14	16008	0,958	0,930	0,944	15550	-0,029	****	0,975	0,958	0,966	15728	-0,017	***
04.2	Autres maladies endocriniennes. nutritionnelles et metaboliques	autres E	13704	0,917	0,902	0,909	13471	-0,017	***	0,963	0,946	0,954	13472	-0,017	***
05.1	Demences	F01,F03	25311	0,945	0,975	0,960	26104	0,031	****	0,963	0,984	0,973	25880	0,022	****
05.2	Abus d'alcool	F10\.0	3230	0,902	0,918	0,910	3290	0,019		0,958	0,963	0,960	3248	0,006	
05.3	Pharmacodependance. toxicomanie	F11\.0-F16\.0 F18\.0-F19\.0	308	0,796	0,760	0,777	294	-0,045		0,959	0,916	0,937	294	-0,045	
05.4	Autres troubles mentaux et du comportement	F04-F09, F17\.0,F20-F99	4907	0,905	0,881	0,893	4777	-0,026	**	0,953	0,933	0,943	4804	-0,021	*
06.1	Maladie de Parkinson	G20	8866	0,970	0,978	0,974	8942	0,009		0,980	0,987	0,983	8929	0,007	
06.2	Maladie d'Alzheimer	G30	25747	0,981	0,983	0,982	25798	0,002		0,987	0,989	0,988	25800	0,002	
06.3	Autres maladies du systeme nerveux et des organes des sens	G00-G12, G14, G21-G25, G31-H95	15541	0,921	0,913	0,917	15399	-0,009		0,965	0,959	0,962	15446	-0,006	
07.1.1	Infarctus aigu du myocarde	I21-I22	18023	0,954	0,966	0,960	18251	0,013	**	0,970	0,979	0,974	18199	0,010	*
07.1.2	Autres Cardiopathies ischemiques	I20,I23-I25	24438	0,938	0,943	0,941	24572	0,005		0,962	0,966	0,964	24552	0,005	
07.2	Autres maladies du coeur	I30-I51	67415	0,951	0,952	0,952	67532	0,002		0,970	0,974	0,972	67671	0,004	
07.3	Maladies cerebro vasculaires	I60-I69	41319	0,947	0,950	0,949	41413	0,002		0,970	0,972	0,971	41436	0,003	
07.4	Autres maladies de l'appareil circulatoire	I00-I15, I26-I28,I70-I99	33025	0,932	0,923	0,928	32701	-0,010	**	0,965	0,959	0,962	32819	-0,006	
08.1	Grippe	J09-J11	1668	0,958	0,963	0,960	1677	0,005		0,981	0,980	0,981	1667	-0,001	
08.2	Pneumonie	J12-J18	16322	0,953	0,955	0,954	16353	0,002		0,973	0,974	0,974	16335	0,001	
08.3.1	Asthme	J45-J46	1077	0,930	0,934	0,932	1082	0,005		0,966	0,971	0,969	1083	0,006	
08.3.2	Autres maladies chroniques des voies respiratoires inferieures	J40-J44, J47	13006	0,942	0,953	0,948	13163	0,012	*	0,966	0,974	0,970	13109	0,008	
08.4	Autres maladies de l'appareil respiratoire	J00-J06, J20-J39, J60-J99	21100	0,929	0,926	0,927	21033	-0,003		0,964	0,959	0,962	20989	-0,005	
09.1	Ulcere gastro duodenal	K25-K28	1081	0,910	0,912	0,911	1084	0,003		0,959	0,962	0,960	1085	0,004	
09.2	Cirrhoses. fibroses et hepatites chroniques	K70, K73-K74	8986	0,948	0,961	0,955	9113	0,014	*	0,971	0,980	0,976	9070	0,009	
09.3	Autres maladies de l'appareil digestif	K00-K22, K29-K66, K71-K72, K75-K92	22147	0,926	0,921	0,923	22044	-0,005		0,962	0,959	0,961	22068	-0,004	
10	Maladies de la peau et du tissu sous-cutane	L00-L99	2067	0,859	0,858	0,858	2064	-0,001		0,949	0,935	0,942	2036	-0,015	
11.1	Arthrite rumathoide et osteroarthrite	M05-M06, M15-M19	726	0,878	0,832	0,854	688	-0,052	*	0,949	0,927	0,938	709	-0,023	
11.2	Autres maladies du syst. osteoarticulaire et des muscles	M00-M02, M08-M13, M20-M99	4537	0,803	0,810	0,806	4574	0,008		0,929	0,915	0,922	4469	-0,015	
12.1	Maladies du rein et de l'uretre	N00-N29	10646	0,921	0,903	0,912	10439	-0,019	***	0,963	0,948	0,955	10483	-0,015	*
12.2	Autres maladies genito-uniraires	N30-N99	4029	0,891	0,871	0,881	3942	-0,022	*	0,941	0,929	0,935	3974	-0,014	
13	Complications de grossesse et d'accouchement	O00-O99	54	0,897	0,481	0,627	29	-0,463	****	1,000	1,000	1,000	54	-	
14	Certains affections dont l'origine est périnatale	P00-P96	2048	0,948	0,950	0,949	2052	0,002		0,998	1,000	0,999	2053	0,002	
15	Malformations cong. et anomalies chrom.	Q00-Q99	2105	0,916	0,817	0,863	1877	-0,108	****	0,966	0,932	0,948	2031	-0,035	*
16.1	Mort subite nourisson	R95	179	0,898	0,933	0,915	186	0,039		0,983	0,994	0,989	181	0,011	
16.2	Causes inconnues ou non precisees	R96-R99	20174	0,946	0,974	0,959	20770	0,030	****	0,964	0,983	0,974	20572	0,020	****
16.3	Autres symptomes et etat morbides non definis	R00-R94	40404	0,973	0,980	0,977	40674	0,007	*	0,981	0,987	0,984	40630	0,006	
17.1.1	Accidents de transport	V01-V99, Y85	3678	0,968	0,935	0,951	3554	-0,034	***	0,978	0,961	0,970	3616	-0,017	
17.1.2	Chutes accidentelles	W00-W19	11146	0,931	0,950	0,940	11371	0,020	***	0,971	0,975	0,973	11196	0,004	
17.1.3	Noyade et submersion accidentelle	W65-W74	1090	0,942	0,950	0,946	1099	0,008		0,976	0,972	0,974	1086	-0,004	
17.1.4	Intoxications accidentelles	X40-X49	2163	0,826	0,770	0,797	2016	-0,068	****	0,948	0,916	0,932	2090	-0,034	*
17.1.5	Autres accidents	W20-W64, W75-X39, X50-X59, Y86	18254	0,891	0,893	0,892	18292	0,002		0,952	0,949	0,950	18184	-0,004	
17.2	Suicide	X60-X84, Y87.0	11281	0,963	0,967	0,965	11335	0,005		0,983	0,984	0,983	11289	0,001	
17.3	Homicides	X85-Y09, Y87.1	499	0,869	0,741	0,800	426	-0,146	****	0,986	0,962	0,974	487	-0,024	
17.4	Evenement dont l'intention n'est pas determinee	Y10-Y34, Y87.2	1709	0,745	0,691	0,717	1586	-0,072	****	0,936	0,882	0,908	1611	-0,057	***
17.5	Autre cause externe	Y35-Y84, Y88-Y89	1847	0,563	0,404	0,470	1326	-0,282	****	0,903	0,757	0,823	1549	-0,161	****
18	Covid	U07,U10, U12	35680	0,982	0,989	0,986	35956	0,008	*	0,987	0,993	0,990	35867	0,005	

Note: les degrés de significativité des écarts entre effectifs prédits et réels proviennent de tests d'égalité supposant que les effectifs réels suivent des lois de Poisson, * pval<.2, ** pval<.1, *** pval<.05, **** pval<.01

Au total, et sur l'ensemble de la population test de référence, la combinaison du batch, du sur-modèle et de la campagne de reprise ciblée permet d'atteindre des niveaux de cohérence avec une campagne de codage classique très élevés pour la plupart des catégories, avec une moyenne des F-mesures par catégorie à 0,966. La seule catégorie pour laquelle la F-mesure est inférieure à 0,9 est 17.5 Autres causes externes[3], ce qui invite à considérer les codages dans cette catégorie avec une certaine prudence. La F-mesure reste inférieure à 0,95 pour 12 catégories (01.3 hépatites virales, 03 les maladies du sang, 05.1 démences, 05.3 pharmacodépendance, 05.4 autres troubles mentaux et du comportement, 10 maladies de la peau, 11.1 arthrite rhumatoïde, 11.2 autres maladies du système musculosquelettique, 12.2 autres maladies génito-urinaires, 15 malformations congénitales, 17.4 intentions indéterminées, 17.5 autres causes externes)

Des écarts à la fois significatifs et conséquents en termes d'effectifs (test de Poisson), par catégorie, se retrouvent pour

- 03 les maladies du sang, sous-estimation de 6% de l'effectif attendu
- 05.1, les démences, surestimation de 2% l'effectif attendu
- 16.2, causes inconnues, surestimation de 2% de l'effectif attendu
- 17.4, les intentions indéterminées, sous-estimation de 6%
- 17.5, les autres causes externes, sous-estimation de 16% (contre 37% en 2018/2019)

Comme attendu, la reprise manuelle ciblée améliore la performance, et ceci largement dans les catégories ciblées que ce soit de décès sensibles (01.2, VIH/Sida, 13 complications de grossesse, 14 15 et 16.1 codes périnataux, malformations congénitales et morts subites du nourrisson, représentatifs des décès des jeunes enfants) ou les catégories ciblées par la reprise IA (01.3 hépatites virales, 03 maladies du sang, 05.3 pharmacologie, 11.2 autres maladies musculosquelettiques, 17.1.4 intoxications accidentelles, 17.3 homicides, 17.4 intentions indéterminées et 17.5 autres causes externes) ou encore celles faites à la finalisation spécialement ciblées sur les risques de tuberculose (01.1), homicides (17.3) et pharmacodépendance (05.3).

L'approche retenue de cibler la reprise manuelle sur la base des précisions par catégorie, améliore aussi les rappels. Et, l'intention de dépasser 97% de précision pour chaque catégorie qui a guidé le ciblage n'est pas satisfaite pour 34 catégories, 8 catégories ont des précisions inférieures à 95% : 02.1.18 (thyroïde), 10 (peau), 11.1, 11.2, 12.2, 17.1.4, 17.4 et 17.5. Ceci peut notamment s'expliquer dans des aléas de tirage et par des effets de reports entre catégories. Tout cela devra être mis en regard des aléas de tirage, via une analyse des écart-types associés à ces estimateurs non réalisée aujourd'hui.

Tableau 14. Performances en termes de cohérence de chapitre de la CIM de l'ensemble de la campagne 2021 et

[3] A titre de comparaison pour la campagne 2018 /2019 les F-mesures étaient inférieures à 0,9 pour 10 catégories sur 71 (les hépatites virales, les maladies du sang et hématopoïétiques, pharmacologie, maladies de la peau, arthrite rhumatoïde, autres maladies musculosquelettiques, les maladies génito-urinaires, les intoxications accidentelles, les intentions indéterminées et les autres causes externes)

effectifs prédits sur l'ensemble de la population test de référence.

Population test de référence (y compris codage batch)	Effectifs réels	Batch + Surmodèle + Reprise manuelle					
		Précision	Rappel	F-measure	Effectifs prédits	Pred/Réels - 1	sign de diff
I - certaines maladies infectieuses et parasitaires	14071	0,968	0,953	0,960	14304	- 0,016	**
II - tumeurs	222453	0,992	0,993	0,992	222311	0,001	
III - maladies du sang...	3276	0,958	0,899	0,927	3491	- 0,062	****
IV maladies endocriniennes, nutritionnelles...	29200	0,972	0,955	0,963	29712	- 0,017	****
V - troubles mentaux et du comportement	34226	0,964	0,977	0,970	33756	0,014	***
VI - maladies du système nerveux..	50175	0,980	0,981	0,980	50154	0,000	
IX - maladies du système circulatoire	184677	0,981	0,983	0,982	184220	0,002	
X - maladies du système respiratoires	53183	0,975	0,975	0,975	53173	0,000	
XI - maladies du système digestif	32223	0,968	0,968	0,968	32214	0,000	
XII - maladies de la peau...	2036	0,949	0,935	0,942	2067	- 0,015	
XIII - maladies du système musculosquelettique	5178	0,934	0,918	0,926	5263	- 0,016	
XIV - maladies du système génitourinaire	14457	0,959	0,945	0,952	14675	- 0,015	**
XV - complications de grossesse...	54	1,000	1,000	1,000	54	-	
XVI - ... origine périnatale	2053	0,998	1,000	0,999	2048	0,002	
XVII - malformations congénitales	2031	0,966	0,932	0,948	2105	- 0,035	*
XVIII - symptomes mal définis...	61383	0,982	0,992	0,987	60757	0,010	***
XIX et XX - causes externes	51108	0,977	0,967	0,972	51667	- 0,011	***
XXI Covid	35867	0,987	0,993	0,990	35680	0,005	

Note: les degrés de significativité des écarts entre effectifs prédits et réels proviennent de tests d'égalité supposant que les effectifs réels suivent des lois de Poisson, * pval<.2, ** pval<.1, *** pval<.05, **** pval<.01

Le Tableau 14 détaille ce même type d'indicateur au niveau du chapitre de la CIM.

4.4 *Détails de l'apport des étapes de reprise ciblée sur la performance globale*

Tableau 15. Evaluation sur la population test des gains en cohérence (accuracy) des différentes étapes de reprises manuelles telles qu'elles ont été menées pour la campagne 2021

	Test qui aurait été codé manuellement			Ensemble de la population test de référence	
	effectif concerné	code CIM 10	shortlist européenne	code CIM 10	shortlist européenne
Surmodèle	223904	0,815	0,815	0,923	0,923
+ décès sensibles (manuel)	4197	0,820	0,878	0,925	0,949
+ EDP et échantillons aléatoires	72805	0,860	0,905	0,942	0,961
+ échantillons ciblés IA	31060	0,896	0,934	0,957	0,972
+ dernières précisions	217	0,897	0,934	0,957	0,973
total	332 183	332 183	332 183	797 651	797 651

Note : pour sommer les effectifs reportés, un certificat est compté uniquement dans la première catégorie (ligne) où il apparait.

Lecture : sur la population test qui aurait été codée manuellement dans une campagne traditionnelle, pour 223 904 le codage provient de l'IA. Si les 332 183 certificats à coder manuellement l'avaient été par prédiction d'IA (sans reprise manuelle) le code CIM 10 obtenu pour la CI aurait été le même qu'en manuel dans 81,5% des cas. Coder manuellement les 4197 décès sensibles permet d'atteindre 82% de cohérence (accuracy). En prenant en compte aussi les certificats codés par batch (colonnes de droite), on atteint une cohérence de 92,3 et 92,5 %.

Le Tableau 15 détaille l'apport en performance de chaque étape de la reprise manuelle. L'ensemble de la reprise manuelle améliore de 8,2 points l'accuracy sur la population test qui aurait été codée manuellement dans le cadre d'une campagne classique de codage, faisant passer cette cohérence

/accuracy de 81,5% à 89,7% (pour 2,2 points en 2018/2019). Pour autant les apports en performance de chaque étape ne sont pas les mêmes. En rapportant le différentiel d'accuracy au % repris manuellement, on voit que reprendre un décès sensible est 2 fois plus efficace que la reprise d'un certificat tiré aléatoirement, et reprendre un certificat ciblé par l'IA 2,1 fois plus efficace. Ceci peut renseigner sur les proportions de reprise à allouer à ces différentes étapes, sans omettre cependant l'apport de ces reprises manuelles sur la base d'entraînement et tout en prenant en compte l'intérêt en termes de santé publique.

5 Evolutions de codage

5.1 Nouveautés relatives aux recommandations OMS

En 2020, dans le cadre de la pandémie Covid-19, l'OMS a édité les codes d'urgence U07.1 et U07.2. Ces codes servent à coder respectivement le Covid-19, virus identifié et le Covid-19, virus non identifié. Ils sont utilisés dans le codage de la mortalité dès l'année 2020.

En milieu d'année 2020, des mises à jour ont été publiées. Il s'agit des codes allant de U08 à U12 (cf site de l'OMS).

U08 : Antécédents personnels de Covid-19
U09 : Post Covid-19
U10 : Syndrome inflammatoire multi-systémique associé à la Covid-19
U11 : Nécessité de se faire vacciner contre la Covid-19
U12 : Vaccination contre la Covid-19 causant des effets indésirables

Ces codes ont commencé à être utilisés à partir de l'année 2021.

5.2 Précisions dans l'application des règles de l'OMS

Les règles de codage décrites dans le Volume 2 de la CIM-10 depuis 2016, sont appliquées à partir de l'année 2017 et continuent de s'appliquer.
Depuis 2017, le code D68.3 remplace le code Y44.2 en cause initiale. En effet, pendant longtemps, l'anticoagulation en cause initiale était codée en Y44.2 mais un code plus précis existe dans le Chapitre III Maladies du sang et des organes hématopoïétiques et certains troubles impliquant le mécanisme immunitaire (D50-D89).

L'OMS recommande depuis 2016, de ne pas retenir les codes peu informatifs I46.0 (arrêt cardiaque réanimé) et I46.9 (arrêt cardiaque sans précision) en causes initiales. Dans les données du CépiDc, ces dernières deviennent systématiquement des R99 à partir de l'année de décès 2019.

5.3 Mises à jour du dictionnaire des expressions nosologiques utilisé par Iris Muse

En 2021, aucune mise à jour ou correction majeure du dictionnaire n'a été réalisée. Seuls les codes mis à jour par l'OMS (cf 5.1) ont été ajoutés au dictionnaire pour faciliter le codage.

6 Spécificités de la base de données individuelles

6.1 Nouvelle version de certificats de décès

La nouvelle version du certificat de décès est en circulation depuis 2018 (cf rapport de production 2018-2019), des nouvelles informations sont collectées et des nouvelles variables ont été ajoutées à la base de données. Les détails sur les variables et leurs modalités de réponse sont dans l'annexe.

6.2 Nouveau mode de codage

Dans le cadre des évolutions du processus de codage impliquant de nouveaux outils, le CépiDc a, comme pour les années 2018 et 2019 mis en place des informations complémentaires disponibles en base de données :

- Le type de codage qui correspond à la méthode utilisée : codage manuel, automatique par système expert Iris/Muse ou autre méthode de codage (notamment impliquant des algorithmes de *deep learning*, lesquels sont détaillés).
- L'indicateur de confiance associé au codage automatique impliquant des algorithmes de *deep learning*. Cet indicateur est celui utilisé pour cibler les certificats à coder manuellement (3.5.1 ; 8.4).

Le détail sur ces deux nouvelles variables et leurs modalités de réponse sont dans l'annexe 8.108.10.

6.3 Spécificités liées à l'utilisation de l'IA dans le codage

6.3.1 Intervalles pas toujours pris en compte

Les intervalles de temps entre les entités nosologiques déclarés sur le certificat peuvent influer sur la détermination de la cause initiale. Ils peuvent aussi conduire à un codage différent selon qu'ils soient indiqués ou non. Par exemple, ils peuvent orienter vers un code de séquelle, de maladie chronique ou encore congénital. Certains codes peuvent donc être impactés par un intervalle.

Exemple 1 : Contexte chez un homme de 25 ans
Codage de causes de décès sans intervalle :

Partie 1	Texte	Intervalle	Code CIM
	a) accident voie publique		V892

Cause Initiale retenue : **V892 (Personne blessée dans un accident de la circulation avec un véhicule à moteur, sans précision)**

Codage des mêmes causes avec intervalle :

Partie 1	Texte	Intervalle	Codes CA
	a) accident voie publique	**1 AN**	V892(1An)

Cause Initiale retenue : **Y850 (Séquelles d'un accident de véhicule à moteur)**

Exemple 2 : Contexte chez un femme de 72 ans
Sans intervalle :

Partie 1	Texte	Intervalle	Code CIM

a) démence alzheimer		G309

Cause Initiale retenue : **G309 (Maladie d'Alzheimer, sans précision)**

Avec intervalle :

Partie 1	Texte	Intervalle	Code CIM
	a) démence alzheimer	**4 ANS**	G309(4ans)

Cause Initiale retenue : **G301 (Maladie d'Alzheimer à début tardif)**

Dans le codage par le système expert ou le codage manuel, les intervalles sont analysés et pris en compte. Ce n'est pas systématique pour les certificats codés en recourant à l'IA. Tout d'abord, les intervalles ne sont pas inclus comme *features* (variables explicatives) en entrée des modèles transformers. Pour autant, au vu des données annotées passées, il est tout-à-fait possible mais jamais garanti que l'algorithme soit en mesure de capter des liens non apparents lui permettant de prédire directement la cause initiale. En revanche, lorsque l'on applique Iris/Muse sur la séquence de causes prédites par un transformer (cf 3.5.2.1) pour déterminer la cause initiale, il n'a pas été possible de mobiliser les intervalles pour les certificats papiers, seulement pour les électroniques (car l'information numérisée au bon format n'est pas disponible).

6.3.2 Liens de causalité au sein d'une même ligne

Dans le cas où on combine IA + Iris/Muse (cf 3.5.2.2), : les liens de causalités indiqués au sein d'une même ligne ne sont pas pris en compte pour l'identification de la CI ce qui peut avoir un impact sur la qualité de la CI retenue, spécifiquement quand le certificateur a indiqué plusieurs causes par ligne avec un enchaînement causal.

6.3.3 Libellés diffusés des causes codées par l'IA

Pour les certificats codés avec algorithme de *deep learning* (combinés ou non au système expert Iris/Muse) le libellé diffusé de chaque ligne de cause correspondra au texte brut présent sur le certificat sans découpage par rang. Pour le codage par le système expert ou le codage manuel, le libellé diffusé correspond au texte « nettoyé » classé par rang. Dans ce dernier cas on a bien un libellé associé à un (ou plusieurs) codes. Avec du codage IA, on n'a pas le « découpage » permettant d'identifier quel code correspond à quel libellé quand le certificateur a indiqué plusieurs codes par lignes.

Exemples : Le texte brut indiqué par le certificateur en ligne 6 du certificat codé par l'IA est répété autant de fois qu'il y a de codes CIM prédits sur la ligne de causes.

Codage manuel			
Ligne	Rang sur la ligne	Texte	CodeCIM
1	1	SAM	D761
1	2	CIVD	D65
1	3	insuffisance hépatique	K729
1	4	insuffisance rénale	N19
2	1	lymphome T	C844

Codage par système expert			
1	1	anurie	R34
2	1	déshydratation	E86
2	2	hypernatrémie	E870

3	1	syndrome infectieux sévère	A419
6	1	COVID	U071
6	2	AIT	G459
6	3	diabète	E149
6	4	AOMI	I702
6	5	troubles cognitifs	R418

Codage IA			
1	1	détresse respiratoire	J960
2	1	insuffisance respiratoire chronique	J961
6	1	sclérose latérale amyotrophique, SAOS, accident ischémique vasculaire cérébral	G122
6	2	sclérose latérale amyotrophique, SAOS, accident ischémique vasculaire cérébral	G473
6	3	sclérose latérale amyotrophique, SAOS, accident ischémique vasculaire cérébral	I635

7 Références

Hebbache, Z. et al. (2023) « Rapport de production – Années de décès 2018 et 2019 – Données définitives », Document de travail du CépiDc n3/2023. https://www.cepidc.inserm.fr/sites/default/files/2023-10/DT_CEPIDC_N3_Rapport%20de%20production%202018-2019_0.pdf

Zambetta E, et al. (2023a) Codage des causes de décès de 2018 et 2019 en CIM10 - Approche combinant *deep learning*, système expert et codage manuel ciblé [Internet]. Centre d'épidémiologie sur les causes médicales de décès; 2023 Sep. (Document de travail du CépiDc). Report No.: 2 (version française). Available from: https://www.cepidc.inserm.fr/documentation/codage-des-causes-de-deces-de-2018-et-2019-en-cim10-approche-combinant-deep-learning-systeme-expert-et-codage-manuel-cible-document-de-travail-cepidc-n22023

Falissard L. et al. (2020). A Deep Artificial Neural Network-Based Model for Prediction of Underlying Cause of Death From Death Certificates: Algorithm Development and Validation. JMIR Med Inform. 2020 Apr 28;8(4):e17125.

Zambetta, E. et al (2023b), « Combining Deep Neural Networks, Rule-Based Expert System and Targeted Manual Coding for ICD-10 Cause of Death Coding of French Death Certificates in 2018 – 2019. » Available at SSRN: https://ssrn.com/abstract=4693074 or http://dx.doi.org/10.2139/ssrn.4693074

Vaswani, A et al. (2017) Attention is all you need. In Advances in Neural Information Processing Systems, pages 5998–6008, 2017

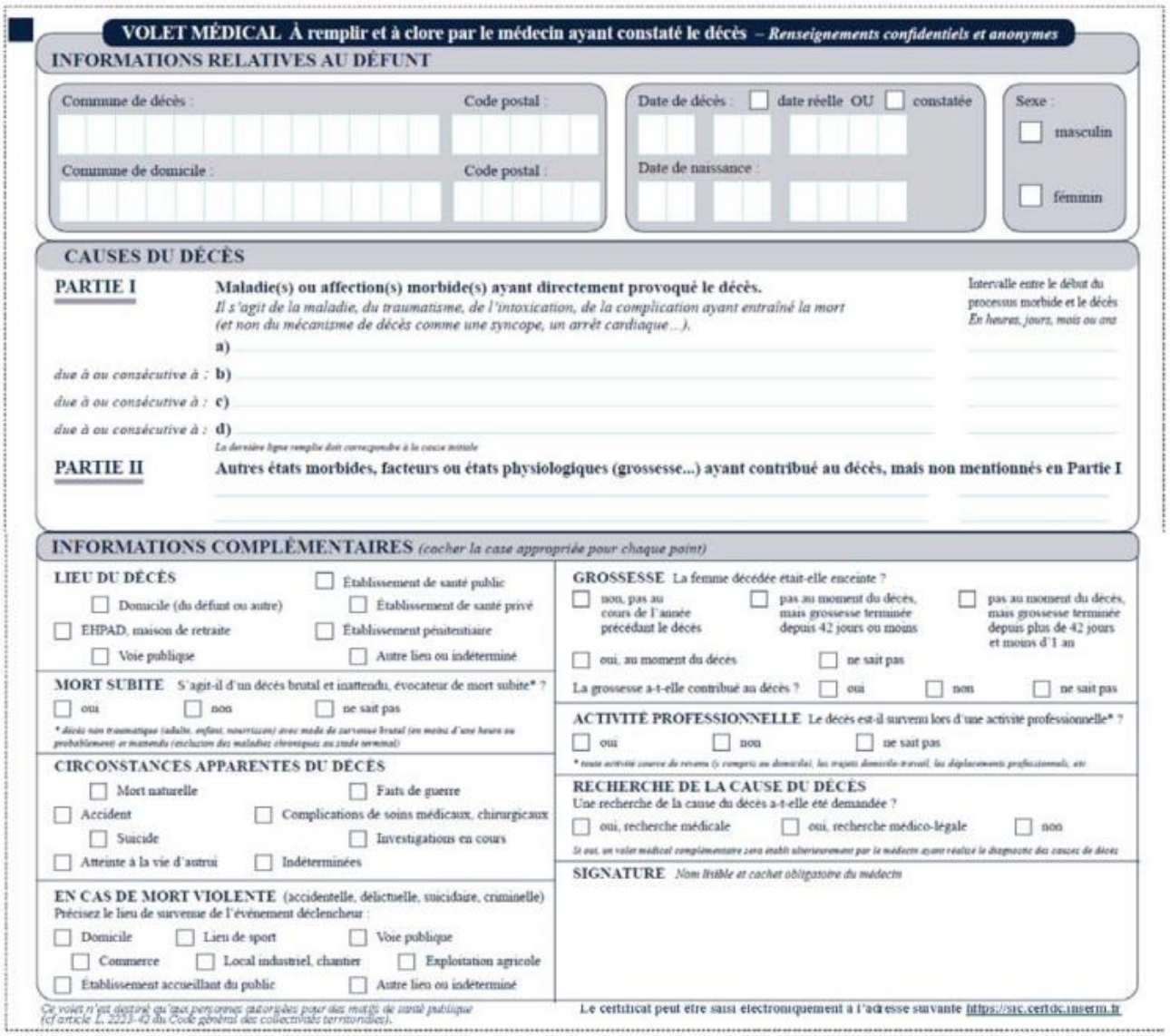
Eurostat (2012). European shortlist of causes of death, 2012

World Health Organization, International Statistical Classification of Diseases and Related Health Problems, 10th revision, 2019. URL: https://icd.who.int/browse10/Content/statichtml/ICD10Volume2_en_2019.pdf [accès 16/10/2023].

Eurostat : Causes of death metadata : France https://ec.europa.eu/eurostat/cache/metadata/EN/hlth_cdeath_simscd_fr.htm [accès 01/03/2024]

8 ANNEXE

8.1 Version 2017 du volet médical du certificat de décès général (28 jours et plus)



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 ☐ Établissement de santé privé ☐ EHPAD, maison de retraite ☐ Établissement pénitentiaire ☐ Voie publique ☐ 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

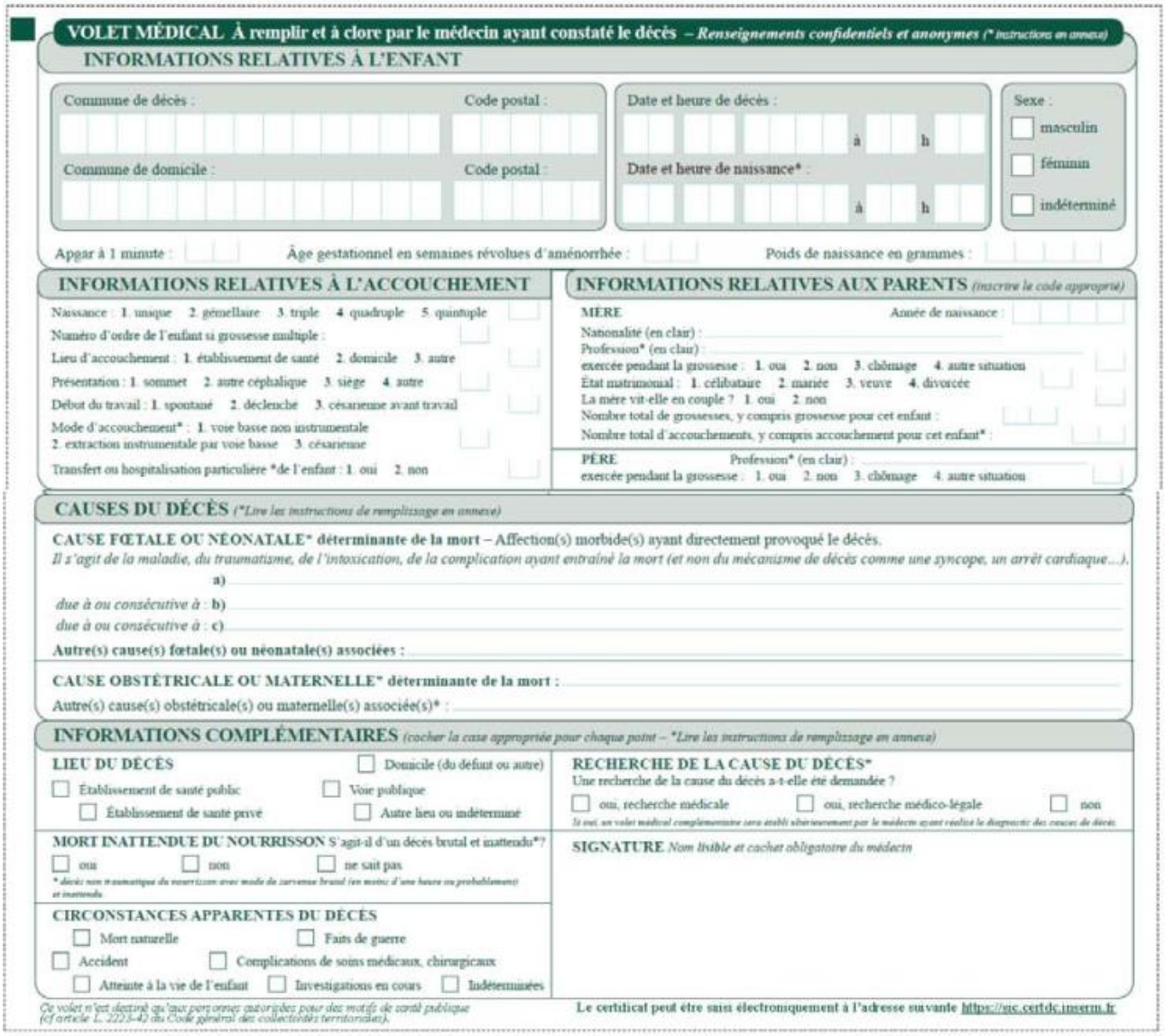
RECHERCHE DE LA CAUSE DU DÉCÈS Une recherche de la cause du décès a-t-elle été demandée ? ☐ oui, recherche médicale ☐ oui, recherche médico-légale ☐ non
Si oui, un volet médical complémentaire sera établi ultérieurement par le médecin ayant réalisé le diagnostic des causes de décès

SIGNATURE *Nom lisible et cachet obligatoire du médecin*

Ce volet n'est destiné qu'aux personnes autorisées pour des motifs de santé publique (cf article L. 2223-42 du Code général des collectivités territoriales).

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

8.2 Version 2017 du volet médical du certificat de décès néonatal (moins de 28 jours)



VOLET MÉDICAL À remplir et à clore par le médecin ayant constaté le décès – *Renseignements confidentiels et anonymes (* instructions en annexe)*

INFORMATIONS RELATIVES À L'ENFANT

Commune de décès : Code postal :

Commune de domicile : Code postal :

Date et heure de décès : à h

Date et heure de naissance* : à h

Sexe : masculin féminin indéterminé

Apgar à 1 minute : Âge gestationnel en semaines révolues d'aménorrhée : Poids de naissance en grammes :

INFORMATIONS RELATIVES À L'ACCOUCHEMENT

Naissance : 1. unique 2. gémellaire 3. triple 4. quadruple 5. quintuple

Numéro d'ordre de l'enfant si grossesse multiple :

Lieu d'accouchement : 1. établissement de santé 2. domicile 3. autre

Présentation : 1. sommet 2. autre céphalique 3. siège 4. autre

Début du travail : 1. spontané 2. déclenché 3. césarienne avant travail

Mode d'accouchement* : 1. voie basse non instrumentale 2. extraction instrumentale par voie basse 3. césarienne

Transfert ou hospitalisation particulière *de l'enfant : 1. oui 2. non

INFORMATIONS RELATIVES AUX PARENTS *(inscrire le code approprié)*

MÈRE Année de naissance :

Nationalité (en clair) :

Profession* (en clair) :

exercée pendant la grossesse : 1. oui 2. non 3. chômage 4. autre situation

État matrimonial : 1. célibataire 2. marié 3. veuve 4. divorcée

La mère vit-elle en couple ? 1. oui 2. non

Nombre total de grossesses, y compris grossesse pour cet enfant :

Nombre total d'accouchements, y compris accouchement pour cet enfant* :

PÈRE Profession* (en clair) :

exercée pendant la grossesse : 1. oui 2. non 3. chômage 4. autre situation

CAUSES DU DÉCÈS *(*Lire les instructions de remplissage en annexe)*

CAUSE FŒTALE OU NÉONATALE* déterminante de la mort – 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...).

a)

due à ou consécutive à : b)

due à ou consécutive à : c)

Autre(s) cause(s) fœtale(s) ou néonatale(s) associées :

CAUSE OBSTÉTRICALE OU MATERNELLE* déterminante de la mort :

Autre(s) cause(s) obstétricale(s) ou maternelle(s) associée(s)* :

INFORMATIONS COMPLÉMENTAIRES *(cocher la case appropriée pour chaque point – *Lire les instructions de remplissage en annexe)*

LIEU DU DÉCÈS Domicile (du défunt ou autre) Établissement de santé public Voie publique Établissement de santé privé Autre lieu ou indéterminé

MORT INATTENDUE DU NOURRISSON S'agit-il d'un décès brutal et inattendu*? oui non ne sait pas

* décès non traumatique du nourrisson avec mode de survenue brutal (en moins d'une heure ou probablement) et inattendu

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

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

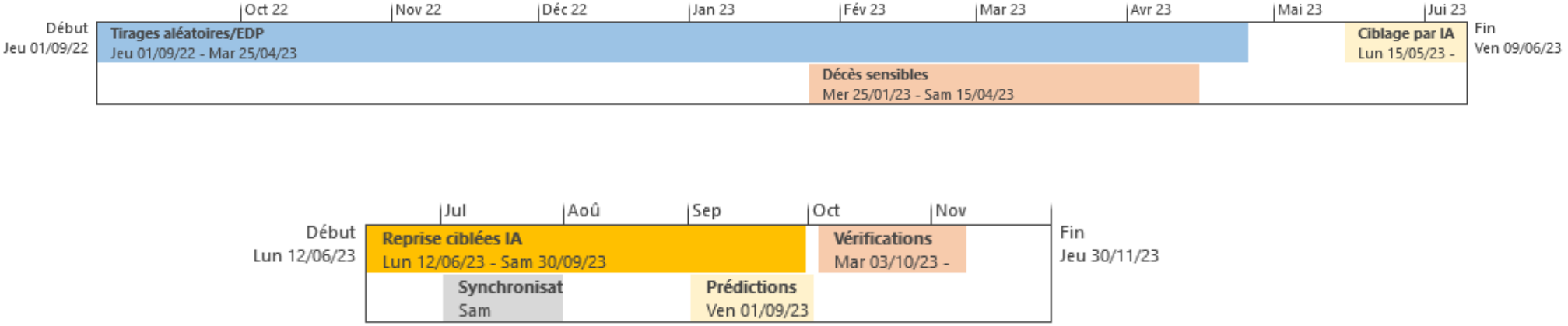
Si oui, un volet médical complémentaire sera établi ultérieurement par le médecin ayant réalisé le diagnostic des causes de décès.

SIGNATURE *Nom lisible et cachet obligatoire du médecin*

Ce volet n'est destiné qu'aux personnes autorisées pour des motifs de santé publique (cf article L. 2223-42 du Code général des collectivités territoriales).

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

8.3 Calendrier de production de l'année de décès 2021



Etape de production	Début	Fin
Codage manuel tirages aléatoires/EDP	Jeu 01/09/22	Mar 25/04/23
Codage manuel Décès sensibles	Mer 25/01/23	Sam 15/04/23
Synchro Insee	Sam 01/07/23	Dim 30/07/23
Ciblage par IA	Lun 15/05/23	Ven 09/06/23
Reprise ciblées IA	Lun 12/06/23	Sam 30/09/23
Prédictions IA finales	Ven 01/09/23	Lun 02/10/23
Vérifications	Mar 03/10/23	Jeu 09/11/23

8.4 *Méthode de ciblage des certificats à coder manuellement sur la base de prédiction IA*

On mobilise 50 000 observations codées manuellement issues de la base de test concernant les années 2020 et 2021 pour réaliser cette estimation. Les variables explicatives entrant dans le modèle sont

- la cause initiale prédite après passage par Iris/muse regroupée au niveau de la *shortlist* eurostat (de loin la plus explicative), et avec un regroupement supplémentaire pour les cas très rares (keras_iris_ci86 et keras_iris_ci2)

- des proxies de la longueur et de la complexité du texte du certificat (nombre de mots dans le certificat, polynôme jusqu'à l'ordre 3, nombre de codes dans la séquence),

- la capacité ou pas d'Iris/muse à aboutir à une cause initiale

- l'homogénéité des codes proposés par le modèle de *deep learning* et Iris/muse

- ainsi que deux scores prédits par l'algorithme de *deep learning* (la probabilité associée au code de la cause initiale prédite par le modèle et la différence entre cette probabilité et la probabilité de la cause initiale la deuxième plus probable selon l'algorithme). Cette dernière variable est censée capter le pouvoir discriminant de l'algorithme dans sa prédiction.

- Le sexe et le groupe d'âge sont aussi inclus dans le modèle.

1_ {CI prédite = CI codée} || nb_mots + nb_codes + status_final_lab + tranche_age + sexe + certif + ci_prob + ci_diff + Keras_iris_ci86 + Homog + nb_mots2 + nb_mots3 + nb_codes2 + nb_codes3

Le R2 ajusté du modèle est de l'ordre de 20%. Les accuracy de l'ordre de 80% dans le *train* et dans le test.

On prédit ensuite cet indicateur pour les données restant du test et pour les données à coder de 2021.

A partir des données restant du test (uniquement les données 2020 et 2021 et codées manuellement sont mobilisées environ 100 000 observations) on simule l'impact d'une reprise manuelle ciblée sur les $\alpha \%$ des données présentant les plus faibles scores de confiance puis on estime la proportion à reprendre par catégorie prédite.

Pour calculer la proportion de certificats à reprendre manuellement catégorie prédite par catégorie prédite, on se concentre sur les seules catégories prédites regroupées au niveau de la *shortlist* eurostat pour lesquelles le niveau de cohérence (précision) entre prédiction par *deep learning* et codage manuel est inférieur à 97%. On calcule ainsi pour chacune des catégories de la *shortlist* eurostat sur la base du test, une précision à atteindre pour dépasser une précision totale de 94%, 95%, 96%, et 97%, tout en tenant compte du fait que les certificats codés automatiquement et des certificats déjà codés (echantillon 1 aléatoire, EDP, décès sensibles…) sont supposés alors être correctement codés.

Ainsi, on définit

eff_codes (pour 2021) : le nombre de certificats déjà codés avec une cause initiale dans la catégorie (ce codage étant obtenu automatiquement ou bien codé manuellement)

eff_noncodes (pour 2021): le nombre de certificats non codés et pour lesquels l'algorithme de *deep learning* prédit une cause initiale dans la catégorie. C'est parmi eux que l'on cherche à cibler ceux à coder manuellement en priorité

eff_tot est la somme des deux

Et le taux de codage dans la catégorie a = eff codes/eff tot

La précision totale est alors Pt= (1-a)Pia + a où Pia est la précision pour les non-codés de la catégorie

Si on fixe un seuil minimal Pt* sur la précision totale, cela conduit à devoir atteindre une précision de Pia* sur les non-codés telle que

Pia*=(Pt* -a)/(1 - a) = (Pt* - eff codes/eff tot)/ (eff non codes/ eff tot) = (Pt* eff tot - eff codes)/eff non codes

Or la précision pour les non-codés peut se voir comme une fonction du taux de reprise manuelle supplémentaire dans la catégorie. Elle vaut la précision simulée dans la catégorie s'il n'y a pas de reprise supplémentaire (estimée sur le test) et va jusqu'à 1 si on considère que toute la catégorie est reprise manuellement. L'inversion de cette fonction pour la valeur Pia* donne le taux de reprise manuelle à réaliser sur la catégorie, en se concentrant sur les certificats pour lesquels l'indicateur de confiance est le plus faible :

Taux reprise comp cat =Pia^{-1}(Pia*) à affecter à l'effectif des non-codés prédits dans la catégorie pour avoir les effectifs à reprendre par catégorie.

Eff_ a reprendre = P_ia^-1 * Eff_noncode

Effectifs à reprendre manuellement pour atteindre XX% de précision dans la catégorie	94%	95%	96%	96.5%	97%
01.1- Tuberculose	6	13	1	32	0
01.2- SIDA (maladie VIH)	14	0	0	0	0
01.3- Hepatites virales	27	11	12	12	3
01.4- Autres maladies infestieuses et parasitaires	802	174	209	139	244
02.1.07-TM du marynx	0	0	6	11	0
02.1.22-Autres Tumeurs malignes	0	330	496	578	578
02.2- Tumeurs non malignes	27	133	214	81	240
03- Maladies du sang et hematopoetiques	437	47	82	36	71
04.1- Diabete sucre	0	0	0	38	153
04.2- Autres maladies endocrinienness, nutritionnelles et metaboliques	138	208	243	173	173
05.2- Abus d'alcool	150	118	108	42	65
05.3- Pharmacodependance, toxicomanie	42	28	8	14	1
05.4- Autres troubles mentaux et du comportement	156	62	94	62	62
06.3- Autres maladies du systeme nerveux et des organes des sens	84	209	251	167	168
07.1.1-Infarctus aigu du myocarde	0	0	0	0	62
07.1.2-Autres Cardiopathies ischemiques	0	0	0	106	212
07.2-Autres maladies du coeur	0	0	0	353	589
07.3-Maladies cerebro vasculaires	0	0	0	192	385
07.4- Autres maladies de l'appareil circulatoire	0	308	540	309	308
08.1- Grippe	1	1	0	1	1
08.2- Pneumonie	0	0	0	70	70
08.3.1- Asthme	9	9	12	9	9
08.3.2-Autres maladies chroniques des voies respiratoires inferieures	0	0	60	89	120
08.4- Autres maladies de l'appareil respiratoire	0	207	242	172	138
09.1 - Ulcere gastro duodenal	0	0	9	22	43
09.2 - Cirrhoses, fibroses et hepatites chroniques	0	0	0	63	63
09.3- Autres maladies de l'appareil digestif	0	59	356	238	296
10- Maladies de la peau et du tissu sous-cutane	151	27	28	20	34
11.1- Arthrite rumathoide et osteroarthrite	28	9	28	11	5
11.2- Autres amaldies du systeme osteoarticulaire et des muscles	306	102	190	87	73
12.1-Maladies du rein et de l'uretre	127	177	203	77	101
12.2- Autres maladies genito-uniraires	0	51	113	38	76
15- Malformations congenitales et anomalies chromosomiques	0	0	0	4	14
16.2- Causes inconnues ou non precisees	0	0	0	186	186
17.1.2 - Chutes accidentelles	342	137	616	479	411
17.1.3 - Noyade et submersion accidentelle	59	9	9	11	21
17.1.4 - Intoxications accidentelles	610	43	54	32	21
17.1.5 - Autres accidents	1216	655	748	281	374
17.3- Homicides	158	5	11	3	2
17.4- Evenement dont l'intention n'est pas determinee	404	42	35	12	18
17.5- Autre cause externe	1229	30	44	15	30
Total	6523	3190	5022	4265	5420

Tableau 16. effectifs de certificats 2021 à coder manuellement pour atteindre une précision de codage de 94, 95, 96 et 97% minimum dans chacune des catégories de la shortlist européenne (les effectifs des colonnes s'ajoutent).

Lecture : en reprenant manuellement les 6 certificats de 2021 dont la cause initiale prédite est hépatite virale (01.3) aux indicateurs de confiance les plus bas, selon les simulations sur le test, on atteindrait une précision globale de 94% pour cette catégorie, et de 95% si on code manuellement les 13 suivants La précision globale s'obtient en supposant que les certificats codés automatiquement par Iris /Muse et codés manuellement sont correctement codés.- Au final 24310 certificats sur les 24 420 ne relevaient pas déjà de certificats déjà codés via les autres types de ciblage au moment du tirage.

Le Tableau 16 indique le nombre de certificats envoyés en reprise manuelle (en ciblant ceux dont l'indicateur de confiance est le plus faible) selon la catégorie de la *shortlist* européenne de laquelle relevait la prédiction de la cause initiale par K5iris entrainé au printemps 2023. On voit ainsi que la reprise ciblée par l'IA a concerné principalement des certificats relevant de catégories de type

« autres » (difficiles à classer), des causes externes et certaines catégories à faible effectifs comme les maladies du sang etc.

8.5 Description des bases d'entrainement et des bases de test des modèles

	test			train		
Echantillon	manuel	batch	total	manuel	batch	total
Tout 2011-2015			0			3445333
Batch auto 2016-2021		84121	84121		1367184	1367184
Codage manuel 2016/2017	187056		187056	299984		299984
2018 - ECH001 - codage manuel EDP	5010		5010	4881		4881
2018 - ECH002 - décès sensibles	1707		1707	1589		1589
2018 - ECH003 - IA PR1	1573		1573	1543		1543
2018 - ECH004 - IA PR2	487		487	558		558
2019 - ECH001 - codage manuel EDP	4853		4853	4852		4852
2019 - ECH002 - décès sensibles	1622		1622	1627		1627
2019 - ECH003 - IA PR1	1611		1611	1591		1591
2019 - ECH004 - IA PR2	877		877	893		893
2020 - manuel	121461		121461	156331		156331
2021 - ECH1 lots aléatoires	12921	2131	15052	15529	44858	60387
2021 - ECH2 - aléatoire	4706	662	5368	4374	14785	19159
2021 - ECH3 - décès sensibles	2906	0	2906	1016		1016
2021 - ECH4	6139	0	6139	1335		1335
2021 - ECH5	8614	15403	24017	170		170
2021 - hors_ech	3544		3544	3412		3412
Total (train1)			*467404*			*5371845*
2021 - ECH06 - PR1 - 94	3158		3158	3191		3191
2021 - ECH07 - PR2 - 95	1590		1590	1541		1541
2021 - ECH08 - PR3 - 96	2464		2464	2411		2411
2021 - ECH09 - PR4 - 96.5	1932		1932	1973		1973
2021 - ECH10 - PR5 - 97	1569		1569	1597		1597
	12		12			
Total (train2)			*478129*			*5382558*
2018 - codage IA en production finale			201419			
2019 - codage IA en production finale			211142			
2021 - codage IA en production finale			149260			

8.6 Certificats des décès entre 28 jours et 15 ans identifiés dans les décès spécifiques à vérifier hors morts violentes

Descriptif	Code en CI ou en cause associée (Chapitre ou Intervalle de codes)	Exceptions (non considéré comme spécifique à vérifier)
Maladie du système nerveux à l'exception de …	G%	G12%, G40%, G41%, G70%, G71%, G72%, G80%, G93%
Troubles mentaux et du comportement	F%	
Maladie de l'appareil respiratoire à l'exception de …	J%	J09%, J10%, J11%, J12%, J21%, J35%, J45%, J46%, J840
Maladie de l'appareil digestif à l'exception de …	K%	K35%, K65%
Maladie de la peau et du tissu cellulaire sous-cutané	L%	
Tumeur à l'exception de …	C01-C98	C222, C40%, C41%, C49%, C62%, C64%, C71%, C72%, C91%, C92%, C93%, C94%, C95%

8.7 Certificats avec mention de Covid-19 vérifiés

Les certificats concernés par la vérification en lien avec l'application des règles de codage sur le Covid-19 appartiennent aux catégories suivantes :

- Cause initiale non Covid hors cancer hors mort violente, avec U071/U072 présent en cause associée
- Cause initiale U071/U072 avec cancer mentionné en Partie I
- K703 + U071/U072 avec Cause initiale U071/U072
- Démence Alzeimer + U071/U072 avec Cause initiale U071/U072
- G20 + U071/U072 avec Cause initiale U071/U072
- Cause initiale a été modifiée par un codeur au détriment de U071/U072
- G319 + U071/U072 avec Cause initiale U071/U072
- F019 + U071/U072 avec Cause initiale U071/U072
- ou + U071/U072
- Mention de ''contexte Covid-19'' codé en U071 en P1 sans mention de comorbidités liées à la Covid-19
- Mention de ''contexte Covid-19'' codé en U071 en P2 au lieu du code Z attendu
- G318 + U071/U072 avec Cause initiale U071/U072

8.8 Précisions sur les modèles Transformers

8.8.1 Entrainement / validation / test

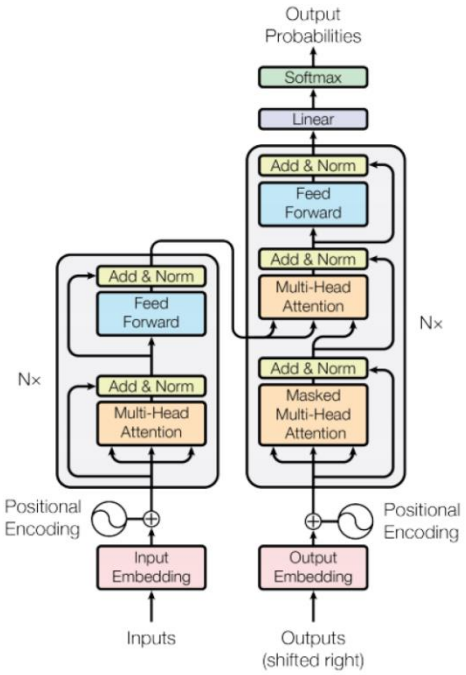
La base de données permettant l'entrainement et l'évaluation des modèles Transformers k4 et k5 se compose des textes des certificats des années passées, mis en entrée des modèles et des codes CIM 10 auxquels ils correspondent (à la fois pour les causes multiples et la cause initiale). Dans la base il y a deux types de textes, les textes bruts, tels qu'ils sont collectés et les textes "nettoyés" après passage d'étapes de standardization et après traitement manuel. Les modèles sont exclusivement entrainés sur les textes bruts, les seuls disponibles pour les certificats à coder. La base d'apprentissage contient des certificats codés manuellement et des certificats codés par batch automatique.

Le tableau 7.7 décrit la base d'apprentissage finale (celle utilisée pour l'entrainement fina et l'évaluation des modèle utilisés dans la production de l'année 2021).

La base de test sert à évaluer la performance des modèles après entrainement. Seuls les certificats codés manuellement et provenant de tirages aléatoires sont pris en compte dans la population test de reference (332183 certificats).

8.8.2 Model

Les modèles utilisés reprennent l'architecture Transformer propose par Vaswani et al. in 2017, (voir schema). Il s'agit d'un modèle d'apprentissage supervisé sequence-to-sequence (seq-2-seq) avec une structure encoder/decoder.



La couche initiale se compose d'un plongement lexical (embedding) permettant une representation vectorielle des mots/tokens ainsi qu'un positional encoding, qui capte l'information sur la position / le rang du mot/token dans la séquence. Dans chaque bloc d'encodage et de décodage, une couche d'attention (multiheaded attention) permet au modèle de tenir compte des relations entre les mots, même dans une longue séquence. De plus, une couche *feed-forward* tient compte du contexte dans son ensemble. Le décodeur contient une couche d'attention croisée, qui combine la sortie de l'encodeur avec la sortie du mécanisme d'attention à têtes multiples masquée. Le Transformer se termine par une couche softmax, dont les dimensions sont équivalentes à la taille du vocabulaire de sortie.

- *Optimizer, function de perte et hyper-paramètres*

L'optimisation se fait grâce à l'optimizer Adam pour lequel les taux d'apprentissage (learning rates) ont été choisis selon une stratégie "warm-up" de 5000. Les hyper-paramètres de l'optimizer Adam sont

beta_1 :: 0.9
beta_2 :: 0.98
epsilon :: 1e-9

La function de perte est la *sparse categorical cross-entropy*.

Les autres hyperparamètres du modèle sont

sequence_length :: 100
batch_size :: 200
buffer_size :: 5000
embed_dim :: 514
latent_dim :: 2048
num_heads :: 8
dropout :: 0.1
epoch :: 100

8.8.3 Programmes du Transformers

```
"""
## Create vocabulary : Train + Test subset
"""
tab_vocab = pd.concat([train_samples, val_samples, test])
print("Tab vocabulary :", tab_vocab.shape)

# Création du vocabulaire
inp_texts = tab_vocab['input'].to_list()
tar_texts = tab_vocab['output'].to_list()

text_vectorization_inp = Tokenizer(
    num_words=None,
    filters='-+=><!%/;.')(?°:, '',
    lower=True,
    split=' ',
)

text_vectorization_tar = Tokenizer(
    num_words=None,
    filters='-+=><!%/;.')(?°:, '',
    lower=True,
    split=' ',
)

# Input text
text_vectorization_inp.fit_on_texts(inp_texts)
voc_input = text_vectorization_inp.word_index

with open('inp_vocabulaire.json', 'w') as fp:
    json.dump(text_vectorization_inp.to_json(), fp)

pickle.dump({'config': text_vectorization_inp.get_config()}
            , open("inp_vectorization.pkl", "wb"))
# Output text
```

```
text_vectorization_tar.fit_on_texts(tar_texts)
voc_output = text_vectorization_tar.word_index

with open('tar_vocabulaire.json', 'w') as fp:
    json.dump(text_vectorization_tar.to_json(), fp)

pickle.dump({'config': text_vectorization_tar.get_config()}
            , open("tar_vectorization.pkl", "wb"))

inp_vocab_size = len(voc_input)
tar_vocab_size = len(voc_output)
print("inp_vocab_size :", inp_vocab_size) #150037
print("tar_vocab_size :", tar_vocab_size) #6333

"""
## Train and validation data
"""
inp_seq_val = text_vectorization_inp.texts_to_sequences(val_samples['input'].to_list())
inp_seq_val = pad_sequences(inp_seq_val, maxlen=sequence_length, padding="post",
truncating="post")

tar_seq_len = sequence_length + 1
tar_seq_val = text_vectorization_tar.texts_to_sequences(val_samples['output'].to_list())
tar_seq_val = pad_sequences(tar_seq_val, maxlen=tar_seq_len, padding="post",
truncating="post")

val_dataset = make_dataset(buffer_size, batch_size, inp_seq_val, tar_seq_val)

inp_seq_train = text_vectorization_inp.texts_to_sequences(train_samples['input'].to_list())
inp_seq_train = pad_sequences(inp_seq_train, maxlen=sequence_length, padding="post",
truncating="post")

tar_seq_train = text_vectorization_tar.texts_to_sequences(train_samples['output'].to_list())
tar_seq_train = pad_sequences(tar_seq_train, maxlen=tar_seq_len, padding="post",
truncating="post")

train_dataset = make_dataset(buffer_size, batch_size, inp_seq_train, tar_seq_train)

"""
## Training
"""
print("Num GPUs Available: ", len(tf.config.list_physical_devices('GPU')))
print(tf.test.is_built_with_cuda())

def transformer(sequence_length,
                inp_vocab_size,
                tar_vocab_size,
                d_model,
```

```
            latent_dim,
            num_heads,
            dropout):
    encoder_inputs = keras.Input(shape=(None,), dtype="int64", name="encoder_inputs")
    x = PositionalEmbedding(sequence_length, inp_vocab_size, d_model)(encoder_inputs)
    encoder_outputs = TransformerEncoder(d_model, latent_dim, num_heads)(x)
    encoder_outputs = layers.Dropout(dropout)(encoder_outputs)
    encoder = keras.Model(encoder_inputs, encoder_outputs)

    decoder_inputs = keras.Input(shape=(None,), dtype="int64", name="decoder_inputs")
    encoded_seq_inputs = keras.Input(shape=(None, d_model),
name="decoder_state_inputs")
    x = PositionalEmbedding(sequence_length, tar_vocab_size, d_model)(decoder_inputs)
    x = TransformerDecoder(d_model, latent_dim, num_heads)(x, encoded_seq_inputs)
    x = layers.Dropout(dropout)(x)
    decoder_outputs = layers.Dense(tar_vocab_size, activation="softmax")(x)
    decoder = keras.Model([decoder_inputs, encoded_seq_inputs], decoder_outputs)

    decoder_outputs = decoder([decoder_inputs, encoder_outputs])
    return keras.Model(
        [encoder_inputs, decoder_inputs], decoder_outputs, name="transformer"
    )

model = transformer(sequence_length,
                    inp_vocab_size,
                    tar_vocab_size,
                    embed_dim,
                    latent_dim,
                    num_heads,
                    dropout)
model.summary()

class CustomSchedule(tf.keras.optimizers.schedules.LearningRateSchedule):
    def __init__(self, d_model, warmup_steps=6000):
        super(CustomSchedule, self).__init__()

        self.d_model = d_model
        self.d_model = tf.cast(self.d_model, tf.float32)

        self.warmup_steps = warmup_steps

    def __call__(self, step):
        arg1 = tf.math.rsqrt(step)
        arg2 = step * (self.warmup_steps ** -1.5)

        return tf.math.rsqrt(self.d_model) * tf.math.minimum(arg1, arg2)

    def get_config(self):
```

```
        config = {
            'd_model': self.d_model,
            'warmup_steps': self.warmup_steps,
        }
        return config

learning_rate = CustomSchedule(embed_dim)
optimizer = tf.keras.optimizers.Adam(learning_rate,
                                     beta_1=0.9,
                                     beta_2=0.98,
                                     epsilon=1e-9)

model.compile(
    optimizer, loss="sparse_categorical_crossentropy", metrics=["accuracy"]
)

"""
## Training Model
"""
checkpoint_dir = os.path.dirname(checkpoint_filepath)

model_checkpoint_callback = tf.keras.callbacks.ModelCheckpoint(
    filepath=checkpoint_filepath,
    save_weights_only=True,
    monitor='val_loss',
    mode='min',
    save_best_only=True,
    verbose=1)

save_logs_callback = SaveLogsCallback(checkpoint_filepath + '_log.txt')

history = model.fit(train_dataset,
                    epochs=62,
                    validation_data=val_dataset,
                    callbacks=[model_checkpoint_callback, save_logs_callback])
```

8.9 *Précisions sur le modèle qui sélectionne la cause initiale entre différentes propositions*

Pour sélectionner la cause initiale qui sera finalement retenue parmi les 4 possiblement différentes sorties des modèles – prédictions directes de la cause initiale par k4 et k5, et application du le système expert Iris/Muse sur les séquences de causes multiples prédites par k4 et k5, appelés iris4 et iris5 – on a recours à un modèle de classification relevant également de l'apprentissage supervisé. Ce modèle répond à un problème de classification en 5 classes, désignant parmi les modèles précédents celui dont on retiendra la prédiction de cause initiale selon les caractéristiques des certificats. L'algorithme retenu pour ce sur modèle est un BiLSTM (*long-term short term memory bidirectionnel*, voir Graves et al 2005, Baldi et al. 1999), modèle utilisé classiquement en analyse des séquences et qui s'avère le plus

performant parmi les algorithmes testés. D'autres modèles (LSTM, FastText, XGboost ainsi qu'un Transformer dédié) ont aussi été testés mais se sont avérés moins performants.

8.9.1 Bases d'apprentissage

Le modèle est entrainé sur le même jeu d'entrainement que k5 et k4, mais uniquement sur les données codées manuellement et à partir de 2016 (année à partir de laquelle on sait distinguer le mode de codage), soit 509 933 certificats, voir tableau ci-dessous.

Manual	Train	Test
All	509 933	332 183
2016	149 841	93 144
2017	150 143	93 912
2018	8 568	0
2019	8 960	0
2020	156 330	121 461
2021	36 091	23 666

Table 1: Distribution of Certificates in the Database

Le jeu de test est le même que celui pour évaluer k5 et k4.

8.9.2 Modèle de sélection de la cause initiale

L'objectif du modèle de sélection de la cause initiale est de choisir la bonne cause initiale parmi les quatre propositions des modèles. Ce modèle prédit en réalité cinq classes : "k4", "k4_iris", "k5", "k5_iris" et "pas_orig", indiquant l'origine de la proposition à retenir. La cinquième classe "pas_orig" indique qu'aucun des modèles n'a prédit la bonne cause initiale. Dans ce cas là, on retiendra la proposition d'iris5, modèle de référence. La table 2 montre la proportion des cinq classes dans la base de données, c'est iris5 qui fournit le plus souvent la bonne cause initiale. Cela vient du fait que lorsque plusieurs modèles remontent correctement la même cause initiale, c'est la classe iris5 (modèle de référence) qui est affectée en priorité. La proportion des « Pas-Origin » autour de 10% est plus forte que pour la production 2018 /2019 pour laquelle elle est autour de 6%. Il s'agit d'un point d'attention, certainement lié à l'uniformisation des bases d'entrainement entre k4 et k5 pour la production de 2021, qui n'avait pas été réalisée pour la production 2018 /2019 mais qui doit être approfondi et surveillé.

Classes	Train	Test
Keras5_iris	82.6%	80.1%
keras4_iris	3.6%	3.7%
keras5	2.8%	2,3%
Keras4	0,1%	0.8%
Pas_orig	9.9%	13%

Table 2: Proportion of classes in the database

8.9.3 Data processing

Les sequences d'entrée concatènent les codes CIM 10 de la cause initiale prédite par k5, k4, k4_iris, k5_iris, ainsi que leur aggregation au niveau de la shortlist européenne., la liste des codes des causes multiples prédites par k4 et par k5, le fait que le certificate soit électronique ou papier, le groupe d'âge, les circonstances apparentes de décès, Les probabilités de sortie associées à la cause initiale prédite par k4 et à celle de k5, les différences de probabilités entre les deux causes initiales les plus probables estimées par k4 et par k5, pouvoir discriminant des modèles, le nombre de causes

associées, un indicateur du nombre de fois où les 4 modèles (k4, k5, iris4 et iris5) produisent des résultats similaires : c'est également un indicateur de la fiabilité de la cause initiale proposée

La sequence d'entrée du modèle est donc:

"keras4_ci keras5_ci keras4iris_ci keras5iris_ci keras4_86postes keras5_86postes keras5iris_86postes keras4iris_86postes keras4_list_causes_associees keras5_list_causes_associees certificat_type age CircApparDeces proba_max4 proba_diff4 proba_max5 proba_diff5 nb_causes_k4 nb_causes_k5 nb_egal"

La figure suivante rapporte les valeurs de Shapley des variables explicatives (grâce au package SHAP : "Shapley Additive Explanations"). La valeur de Shapley mesure pour chaque variable explicative l'importance du rôle de cette variable dans la prédiction. Les variables qui ont le plus de poids lors de la prédiction sont ici l'indicatrice de cohérence entre les modèles (nombre de fois où la bonne cause initiale est remontée par un modèle), les causes initiales agrégées au niveau shortlist européenne, les circonstances apparentes de décès, la séquence des causes multiples et les probabilités de confiance des modèles k4 et k5.

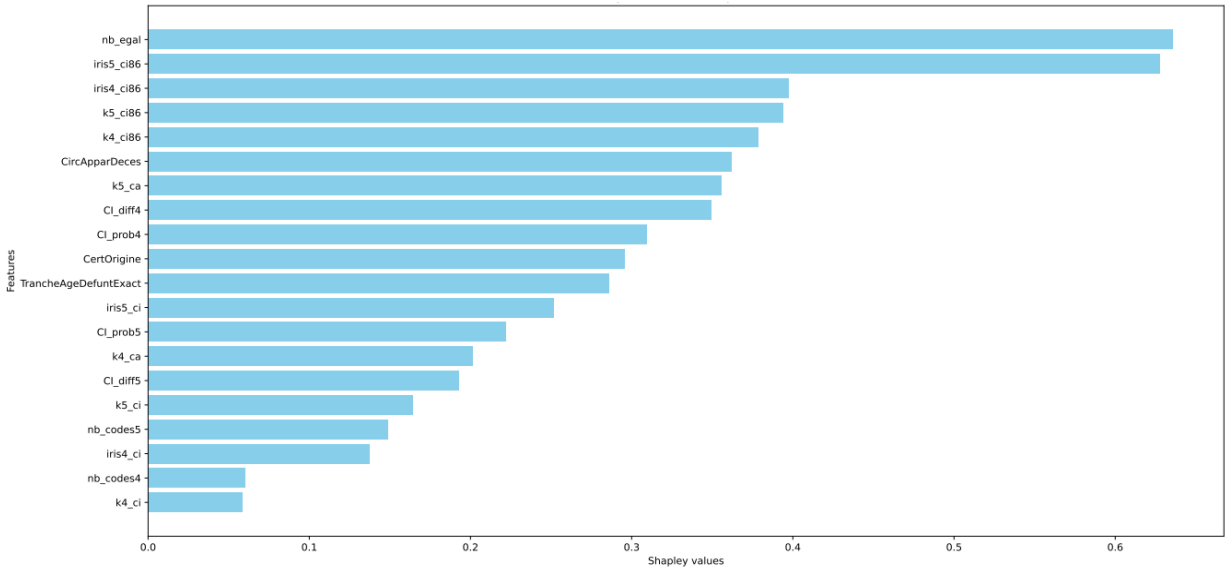


Figure A2.1 : Valeurs de Shapley des variables explicatives

La séquence d'entrée est en amont convertie en vecteurs numériques pour être utilisés comme entrée dans le modèle, les étapes sont les suivantes:

1. Tokenisation : la phrase est divisée en mots appelés *tokens*
2. Indexation : chaque token est associé à un indice unique dans un dictionnaire de mots.
3. Transformations en séquence d'indice : la phrase est ensuite représentée sous forme de séquence d'indices correspondant aux *tokens*.
4. Padding : pour assurer que toutes les séquences ont la même longueur des valeurs sont ajoutées pour remplir les séquences les plus courtes et tronquer les séquences plus longues.

Le vecteur numérique utilisé en entrée du modèle passe d'abord par une couche *d'embedding,* où chaque mot est représenté par un vecteur de dimension fixe. Lors de l'entraînement du modèle, les vecteurs de représentation des mots sont ajustés par le modèle de façon à saisir les relations sémantiques entre les mots, ce qui signifie que des mots ayant des significations similaires auront des vecteurs proches dans l'espace de projection. Ensuite, la couche BiLSTM permet d'extraire des informations séquentielles importantes dans la séquence et les représente sous forme de vecteurs caractéristiques : *features.* La couche de *Fully Connected* permet de déterminer la classe de la séquence.

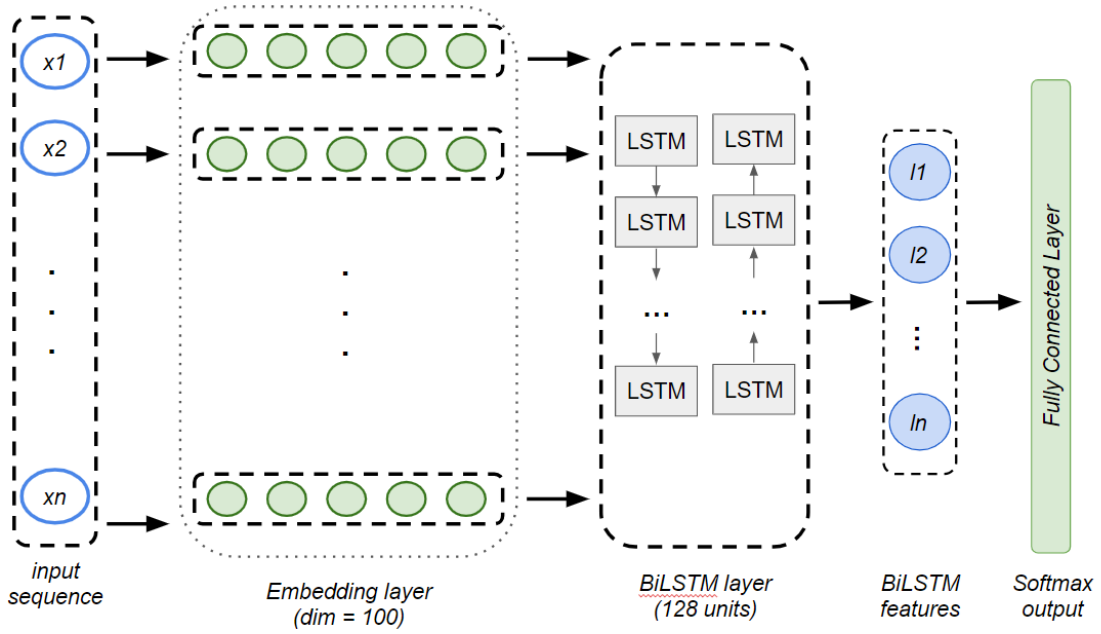


Figure A2.2 : réseau BiLSTM pour la sélection de la cause initiale

8.9.4 Hyperparamètres et fonction de perte

La fonction de perte utilisée est l'entropie croisée Le sur-modèle à été entraîné en utilisant les hyper-paramètres résumés dans le tableau ci-dessous. L'algorithme Adam a été utilisé pour minimiser la fonction de perte. Une approche d'adaptation dynamique du taux d'apprentissage au cours des epochs a été utilisée pour améliorer la convergence et optimiser les performances d'apprentissage.

Hyperparameters	Value
Sequence length	196
Optimizer	Adam
Batch size	128
Vocabulary size	3 137
Embedding dimension	100

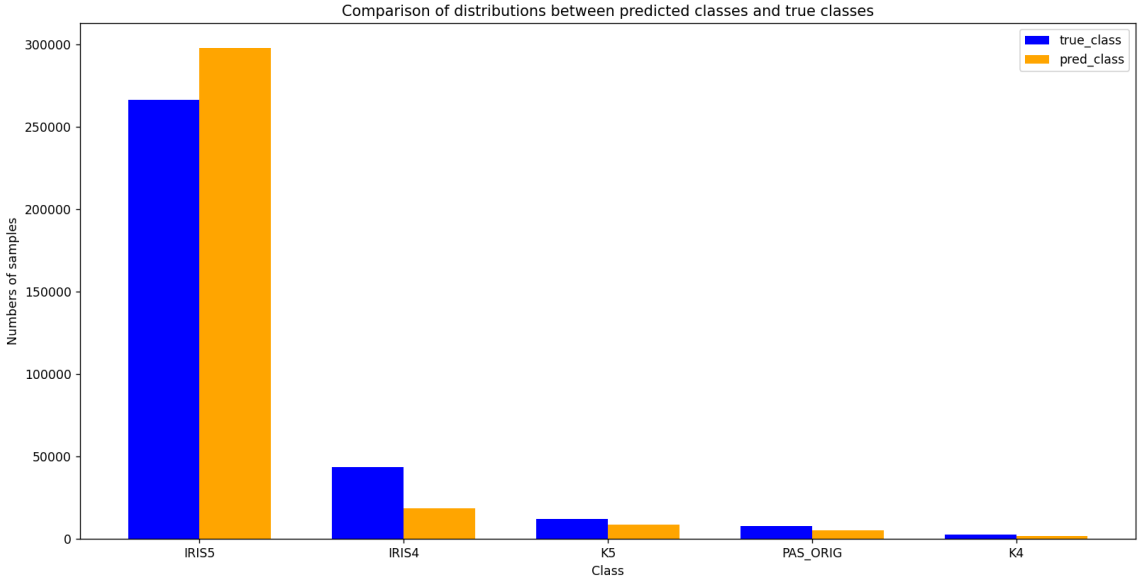
Table 3: Selected hyperparameters

8.9.5 Résultats et analyse de performance

La performance finale est de 81,6% (cf : table 4, pour 85,6% pour la production 2018 et 2019). C'est-à-dire que dans 81,6% des cas, le modèle de sélection prédit la bonne "classe"/"origine". La Figure ci-dessous illustre la distribution des prédictions par classe sur la base de test. Le modèle prédit principalement iris5. En revanche, les causes initiales de keras4 et pas_orig sont très rarement récupérées.

Train	Validation	Test
85.91%	85.15%	81,62%

Table 4: BiLSTM performance results



Au niveau du code CIM prédit maintenant (le code CIM de la cause initiale proposé par le modèle/l'origine et retenu par le modèle de sélection), le modèle permet de prédire le bon code CIM-10 pour la cause initiale dans 81,5% des cas sur la base de test (pour 81,9% en 2018-2019). Le modèle de selection permet donc d'augmenter la performance de 1,4 points, étant donné que le modèle iris 5 a une performance de 80,1%.

8.9.6 Programme BiLSTM

```
"""
## Create vocabulary : Train + Val subset
"""
# Création du vocabulaire
texts = tab['text'].to_list()

tokenizer = Tokenizer()

# Input text
tokenizer.fit_on_texts(texts)
sequences = tokenizer.texts_to_sequences(texts)
```

```
vocab_size = len(tokenizer.word_index) + 1

max_sequence_length = max([len(seq) for seq in sequences])
print("max_sequence_length : ", max_sequence_length)

with open('inp_vocabulaire.json', 'w') as fp:
    json.dump(tokenizer.to_json(), fp)

x_train_pad = tokenizer.texts_to_sequences(x_train)
x_train_pad = pad_sequences(x_train_pad, maxlen=max_sequence_length, padding="post", truncating="post")

x_val_pad = tokenizer.texts_to_sequences(x_val)
x_val_pad = pad_sequences(x_val_pad, maxlen=max_sequence_length, padding="post", truncating="post")

"""
## Encode labels
"""
# Encode the categorical labels
label_encoder = LabelEncoder()
labels = tab['origine']
label_encoder.fit(labels)
num_classes = len(label_encoder.classes_)
print("Nb classes :", num_classes)

# Save the LabelEncoder model to a file
filename = 'label_encoder_model.joblib'
joblib.dump(label_encoder, filename)

y_train_cat = label_encoder.transform(y_train)
y_train_cat = to_categorical(y_train_cat, num_classes=num_classes)

y_val_cat = label_encoder.transform(y_val)
y_val_cat = to_categorical(y_val_cat, num_classes=num_classes)


"""
## Create deep model
"""
print("Num GPUs Available: ", len(tf.config.list_physical_devices('GPU')))
print(tf.test.is_built_with_cuda())

cum_sch = 256
batch_size = 128

def bilstm1(vocab_size, num_classes, sequence_length):
    # Création du modèle
```

```
    model = Sequential()
    model.add(Embedding(input_dim=vocab_size,
output_dim=100, input_length=sequence_length))
    model.add(Bidirectional(LSTM(128)))
    model.add(Dense(num_classes, activation='softmax'))
    return model


checkpoint_dir = os.path.dirname(checkpoint_filepath)
model_checkpoint_callback = tf.keras.callbacks.ModelCheckpoint(
        filepath=checkpoint_filepath,
        save_weights_only=True,
        monitor='val_loss',
        mode='min',
        save_best_only=True,
        verbose=1)

# Assuming you have training data `X_train` and labels `y_train`
print(max_sequence_length)
model = bilstm1(vocab_size, num_classes, max_sequence_length)

class CustomSchedule(tf.keras.optimizers.schedules.LearningRateSchedule):
    def __init__(self, d_model, warmup_steps=5000):
        super(CustomSchedule, self).__init__()

        self.d_model = d_model
        self.d_model = tf.cast(self.d_model, tf.float32)

        self.warmup_steps = warmup_steps

    def __call__(self, step):
        arg1 = tf.math.rsqrt(step)
        arg2 = step * (self.warmup_steps ** -1.5)

        return tf.math.rsqrt(self.d_model) * tf.math.minimum(arg1, arg2)

    def get_config(self):
        config = {
            'd_model': self.d_model,
            'warmup_steps': self.warmup_steps,
        }
        return config

learning_rate = CustomSchedule(cum_sch)
optimizer = tf.keras.optimizers.Adam(learning_rate,
                                     beta_1=0.9,
                                     beta_2=0.98,
                                     epsilon=1e-9)
```

```
model.compile(optimizer=optimizer, loss='categorical_crossentropy', metrics=['accuracy'])

# Open a text file in write mode
file_path = 'variable.txt'
with open(file_path, 'w') as file:
    # Write the variable value to the file
    file.write(f"max_sequence_length = {max_sequence_length}\n\n")
    file.write(f"batch_size = {batch_size}\n\n")
    file.write(f"cum_sch = {cum_sch}\n\n")

    # Write the function definition to the file
    function_source = inspect.getsource(bilstm1)
    file.write(function_source)


history = model.fit(x_train_pad, y_train_cat, batch_size=batch_size, epochs=30,
          validation_data=(x_val_pad, y_val_cat),
          callbacks=model_checkpoint_callback, shuffle=True)
```

8.10 Dictionnaire de variables dans le SNDS

Table Ident

Colonne	Nom de la variable	Formats	Modalités des valeurs SNDS	Commentaires
1	Identifiant IdDécès	varchar (64)		
2	Version du certificat	Num	1 = version de 1997 2 = version de 2017	La version 2017 apparaît à partir de 2017 (expérimentation)
3	Statut de traitement	Num	1 = Codé 2 = En cours	Les données finalisées envoyées dans le SNDS sont toutes à 1
4	Type de certificat	Num	1 = adulte 2 = néonat	Variable de mauvaise qualité ; pour identifier le type de certificat, regarder l'âge du défunt et la présence ou non de valeurs dans les variables spécifiques du certificat néonatal
5	Type de support	Num	1 = électronique 2 = papier	
6	Type de volet	Num	1 = initial 2 = Complémentaire	Les volets médicaux complémentaires apparaissent à partir de 2018
7	Département de décès	Varchar (3)		2 caractères pour métropole, 3 pour DOM TOM
8	Commune de décès	Varchar (3)		3 caractères pour métropole, 2 pour DOM TOM
9	Département de domicile	Varchar (3)		2 caractères pour métropole, 3 pour DOM TOM
10	Commune de domicile	Varchar (3)		3 caractères pour métropole, 2 pour DOM TOM
11	Date du décès	Date		

12	Lieu du décès	num	0= Non renseigné 1 = Domicile 2 = Etablissement public de santé 3 = établissement privé de santé 4 = EHPAD, maison de retraite, 5 = voie publique 6 = autre lieu ou indéterminé 7=établissement pénitentiaire (certificat de 2017 uniquement	
13	Année de naissance	Car (4)		
14	Mois de naissance	Car (2)		
15	Sexe	num	1 = masculin 2 = féminin 9 = indéterminé	
16	Activité professionnelle	num	1 = Retraité 2 = Inactif autre que retraité 3 = actif	Variable de l'INSEE
17	Profession et catégorie socio-professionnelle	varchar (2)	Nomenclature PCS INSEE	Variable de l'INSEE
18	Etat matrimonial	num	1 = célibataire 2 = marié 3 = veuf 4 = divorcé	Variable de l'INSEE
19	Cause initiale du décès	varchar (4)	CIM 10	

20	Recherche de la cause de décès	num	1 = non 2 =oui, résultats disponibles 3 =oui, résultats non disponibles 4 =oui recherche médicale (certificat de 2017 uniquement), 5 = oui recherche médico-légale (certificat de 2017 uniquement)	Valeur vide de 2006 à 2009, question n'était pas posée avant 2010. Valeur 2 et 3 uniquement pour les versions 1997 Valeur 4 et 5 uniquement pour les nouveaux certificats à partir de 2017
21	Grossesse "Le décès est-il survenu pendant une grossesse (ou moins d'un an après) " (ancienne version) ? La femme décédée était-elle enceinte ?" (nouvelle version)	num	1 = non, pas au cours de l'année précédant le décès 2 = Pas au moment du décès mais grossesse terminée depuis 42 jours au moins 3 = Pas au moment du décès mais grossesse terminée depuis plus de 42 jours et moins de 1 an 4 = Oui, au moment du décès 5 = Ne sait pas	Cette variable est vide avant 2015 Libellé de la question et modalités de réponses ont changé entre la version 1997 et 2017 mais la variable reste la même
22	Délai entre fin de grossesse et décès	Varchar(4)	mois + jour, ex : 0502 = 5 mois + 2 jours après la date de fin de grossesse	Variable mal renseignée, de mauvaise qualité et uniquement sur la version 1997.
23	Ancien certif : Est-ce un Accident de travail ? Nouveau certif : Est-ce pendant une activité professionnelle ?	num	1 = oui 2 = Non 3 = sans précision (ancien) ou Ne sait pas (nouveau)	Libellé de la question et modalités de réponses ont changé mais la variable reste la même
24	Apgar à une minute	num	0 à 10 (score d'Apgar, 0 = mort apparente, 10 = état optimal)	Variable spécifique du certificat néonatal
25	Âge gestationnel en semaines révolues d'aménorrhée	num	en semaines	variable spécifique du certificat néonatal
26	Poids de naissance en grammes	num	en g	variable spécifique du certificat néonatal

27	Type de naissance	num	1 = Unique 2 = Gémellaire 3 = Triple 4 = Quadruple 5 = Quintuple	variable spécifique du certificat néonatal
28	N° d'ordre de l'enfant si grossesse multiple	num	1 à 5	variable spécifique du certificat néonatal
29	Lieu d'accouchement	num	1 = Etablissement de sante 2 = Domicile 3 = Autres	variable spécifique du certificat néonatal
30	Présentation de l'enfant	num	1 = Sommet 2 = Autres céphaliques 3 = Siège 4 = Autres	variable spécifique du certificat néonatal
31	Début du travail	num	1 = Spontané 2 = Déclenché 3 = Césarienne	variable spécifique du certificat néonatal
32	Mode d'accouchement	num	1 = Voie basse 2 = Extraction 3 = Césarienne	variable spécifique du certificat néonatal
33	Transfert ou hospitalisation particulière de l'enfant	num	1 = Oui 2 = Non	variable spécifique du certificat néonatal
34	Année de naissance de la mère	Varchar(4)		variable spécifique du certificat néonatal
35	Activité professionnelle de la mère	num	1 = En activité 2 = Non au chômage 3 = Autres 4 = Au chômage 5 = Non	variable spécifique du certificat néonatal 2006 à 2015 uniquement les 3 1ere modalités, les autres modalités arrivent avec le nouveau certificat néonatal en 2017
36	Profession de la mère exercée pendant la grossesse	Varchar (50)	libellé de la profession	variable spécifique du certificat néonatal

37	Etat matrimonial de la mère	num	1 = célibataire 2 = mariée 3 = veuve 4 = divorcée	variable spécifique du certificat néonatal
38	La mère vit elle en couple	num	1 = Oui 2 = Non	variable spécifique du certificat néonatal
39	Nombre total de grossesses	num		variable spécifique du certificat néonatal
40	Nombre total d'accouchements	num		variable spécifique du certificat néonatal
41	Activité professionnelle du père		1 = En activité 2 = Non au chômage 3 = Autres 4 = Au chômage 5 = Non	variable spécifique du certificat néonatal 2006 à 2015 uniquement les 3 1ere modalités, les autres modalités arrivent avec le nouveau certificat
42	Profession du père exercée pendant la grossesse		libellé de la profession	
43	Mort subite	num	1 = Ne sait pas 2 = Oui 3 = Non	Uniquement dans la version 2017
44	Circonstance apparente du décès	num	1 = Indéterminé 2 = Mort naturelle 3 = Accident 4 = Suicide 5 = Atteinte à la vie du défunt 6 = Fait de guerre 7 = Investigations en cours 8 = Complications de soins 9 = Atteinte à la vie de l'enfant (certificat néonatal uniquement) ;	Uniquement dans la version 2017

45	Codification du lieu si mort violente	num	1 = Autre lieu ou indéterminé 2 = Domicile 3 = Etablissement accueillant du public 4 = Exploitation agricole 5 = Lieu de sport 6 = Commerce 7 = Voie publique 8 = Local industriel - chantier	Uniquement dans la version 2017
46	Fiabilité de la date de décès	num	1 = date réelle 2 = Hypothèse	Uniquement dans la version 2017
47	Mort inattendue du nourrisson	num	1 = Ne sait pas 2 = Oui 3 = Non	variable spécifique du certificat néonatal Uniquement dans la version 2017
48	Groupe d'âge mortalité néonatale		1 = < 7 jours 2= 7 jours <= âge < 28 jours 3 = 28 jours<= âge <= 365 jours	
49	La grossesse a-t-elle contribué au décès ?	num	1 = Oui 2 = Non 3 = Ne sait pas	Uniquement dans la version 2017
50	CauseInitialeTypeCode	num	1 = CIM 9 2 = CIM 10 3 = CIM 11	
51	TypeCodage	num	1=codage Iris Automatique 2= manuel 3= K5Iris 4=K4Iris 5=K5 6=K4	Uniquement à partir de 2018 Décrit le mode de codage utilisé. Lee types de codage 3 à 6 impliquent des algorithmes de *deep learning* combinés ou non avec le système expert IRIS/Muse
52	Score de confiance	num		Uniquement à partir de 2018 Indicateur de confiance issu de la prédiction impliquant des algorithmes de *deep learning* permettant de cibler le codage manuel. Il est compris entre 0 et 1.

Table des causes

Colonne	Nom de la variable	Formats	Modalités des valeurs SNDS	Commentaires
1	identifiant IdDécès	Varchar (64)		
2	N° de ligne de la cause sur le certificat	numérique	1 à 6	Numéro de la ligne de cause
3	rang de la cause	numérique	1 à 40	rang de la cause sur la ligne si plusieurs causes par ligne
4	libellé de la cause*	Varchar (400)	texte de la cause	Pour les certificats codés avec algorithme de *deep learning* (typecodage >=3) le texte de la ligne correspondra au texte brut présent sur le certificat sans découpage par rang
5	code CIM de la cause**	Varchar (4)	Tables ir_cci_v* et ir_cim_v*	Ce champ va évoluer avec la CIM 11 et passera en varchar 6
6	TypeCodage			Variables non informatives à l'échelle de la ligne de cause, contenu vide.
7	Score de confiance			Variables non informatives à l'échelle de la ligne de cause, contenu vide.

* Certains libellés de causes peuvent avoir plusieurs codes CIM associés. Dans ces cas-là (hors codage avec de l'IA), les codes sont indiqués à la suite (implémentation de rangs) et le texte associé n'est pas répété. C'est pourquoi on peut retrouver des codes CIM sans libellé de texte associé.

** Certains libellés de causes peuvent ne pas avoir de CodeCIM correspondant : il s'agit de texte non informatif non pris en compte dans le codage. C'est pourquoi on peut retrouver des textes sans codes associés dans cette table.

8.11 Liste des tableaux