

Rapport de Production

sur les causes de décès en 2022

Pierre Boulet, Zina Hebbache, Aude Robert, Elisa Zambetta,
Daniel Razakamanana, Yann Aubineau, Fanny Godet, Diane
Martin, Elise Coudin
CépiDc-Inserm

Mars 2025

Document de travail du CépiDc N°5

Ces documents de travail ne reflètent pas la position de l'Inserm et n'engagent que leurs auteurs.

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 2022. Le codage des causes de décès en CIM-10 en 2022 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 le codage obtenu lors de cette campagne à trois modes de codage et celui qui aurait été obtenu dans 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, 2022, 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 2022. The coding of causes of death in ICD-10 in 2022 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 3 coding-mode 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, 2022, statistical production of CoD data

Table des matières

1	INTRODUCTION	5
2	COLLECTE	5
2.1	EXHAUSTIVITE	5
2.2	ORIGINE DES CERTIFICATS : ELECTRONIQUE OU PAPIER	6
2.3	VERSION DE CERTIFICATS : 2017 OU 1997.....	7
2.4	VOLETS MEDICAUX COMPLEMENTAIRES (VMC).....	7
2.5	TYPE DE CERTIFICAT : GENERAL OU NEONATAL	7
3	CODAGE	8
3.1	METHODE COMBINANT TROIS MODES DE CODAGE.....	8
3.2	REPARTITION DES MODES DE CODAGE	10
3.3	ALLOCATION DES CERTIFICATS VERS LA REPRISE MANUELLE OU L'INFERENCE PAR RNP	10
3.4	CODAGE PAR SYSTEME EXPERT IRIS/MUSE.....	12
3.5	CODAGE ET REPRISE PAR L'EQUIPE DE CODAGE.....	12
3.6	CODAGE IA (DEEP LEARNING)	14
3.6.1	<i>Transformer pour coder la séquence de causes</i>	15
3.6.2	<i>Un sur-modèle pour retenir une cause initiale parmi les différentes propositions possibles</i>	17
3.7	VERIFICATIONS.....	19
3.7.1	<i>Cohérences</i>	20
3.7.2	<i>Décès spécifiques</i>	20
3.7.3	<i>Difficultés de codage connues liées au logiciel Iris</i>	21
3.7.4	<i>Vérification de la bonne application des nouvelles règles de 2016</i>	22
3.7.5	<i>Vérifications liées aux évolutions du dictionnaire (démarche des choix de code)</i>	23
3.7.6	<i>Vérifications pour raisons multiples</i>	24
4	EVALUATION DE LA STRATEGIE DE CODAGE	24
4.1	POPULATION TEST DE REFERENCE.....	24
4.2	ACCURACY, COHERENCE GLOBALE ET COMPARAISON A LA CAMPAGNE PRECEDENTE.....	26
4.3	PRECISION, RAPPEL ET ECARTS D'EFFECTIFS	29
4.4	DETAILS DE L'APPORT DES ETAPES DE REPRISE CIBLEE SUR LA PERFORMANCE GLOBALE	37
4.5	EVALUATION DE LA COHERENCE DE LA CAMPAGNE 2022 AVEC UNE CAMPAGNE A DEUX MODES SUR LE CODAGE DES CAUSES MULTIPLES.....	38
5	EVOLUTIONS DE CODAGE	38
5.1	NOUVEAUTES RELATIVES AUX RECOMMANDATIONS DE L'OMS.....	38
5.2	NOUVEAUTES INTERNES AU CEPIDC	39
5.3	MISES A JOUR DU DICTIONNAIRE DES EXPRESSIONS NOSOLOGIQUES UTILISE PAR IRIS MUSE	40
6	SPECIFICITES DE LA BASE DE DONNEES INDIVIDUELLES	40
6.1	VARIABLE SUR LE MODE DE CODAGE UTILISE POUR CODER CE CERTIFICAT.....	40
7	REFERENCES	40
8	ANNEXE	42

8.1	VERSION 2017 DU VOLET MEDICAL DU CERTIFICAT DE DECES GENERAL (28 JOURS ET PLUS).....	42
8.2	VERSION 2017 DU VOLET MEDICAL DU CERTIFICAT DE DECES NEONATAL (MOINS DE 28 JOURS)	43
8.3	CALENDRIER DE PRODUCTION DE L'ANNEE DE DECES 2022.....	44
8.4	METHODE DE CIBLAGE DES CERTIFICATS A CODER MANUELLEMENT SUR LA BASE DE PREDICTION IA	46
8.5	DESCRIPTION DES BASES D'ENTRAINEMENT ET DES BASES DE TEST DES MODELES.....	50
8.6	DEFINITIONS DETAILLEES DES VERIFICATIONS ET EVOLUTIONS PAR RAPPORT A 2021.....	51
8.6.1	<i>Evolution des vérifications</i>	51
8.6.2	<i>Certificats des décès entre 28 jours et 15 ans identifiés dans les décès sensibles ou dans les décès à vérifier</i> 59	
8.6.3	<i>Certificats avec mention de Covid-19 vérifiés</i>	61
8.7	PRECISIONS SUR LES MODELES TRANSFORMERS	61
8.8	PRECISIONS SUR LE MODELE QUI SELECTIONNE LA CAUSE INITIALE ENTRE DIFFERENTES PROPOSITIONS	66
8.8.1	<i>Base d'apprentissage</i>	67
8.8.2	<i>Problème de classification / variable cible</i>	67
8.8.3	<i>Modèle BiLSTM</i>	68
8.8.4	<i>Résultats et analyse de performance</i>	72
8.9	DICIONNAIRE DE VARIABLES DANS LE SNDS.....	73
8.9.1	<i>Table Ident</i>	73
8.9.2	<i>Table des causes</i>	79
8.10	EVALUATION DE LA COHERENCE DES CAUSES MULTIPLES AVEC UNE CAMPAGNE DE CODAGE A TROIS MODES	81
8.10.1	<i>BLEU Score</i>	81
8.10.2	<i>Population Test de référence</i>	85
8.10.3	<i>Performance globale</i>	85
8.11	CONSTITUTION DE LA POPULATION TEST DE REFERENCE	91
8.12	LISTE DES TABLEAUX	99
8.13	LISTE DES FIGURES	100

1 Introduction

Ce rapport présente les spécificités du traitement des certificats de décès en 2022. Il complète ainsi le document [‘Statistiques sur les causes de décès de A à Z’](#) plus général sur le processus de production de ces données et l’article - [Les statistiques sur les causes de décès – Courrier des statistiques N12 - 2024 | Insee](#) plus général et historique sur la statistique des causes de décès. Il poursuit les rapports de production précédents sur les décès des années 2021 et 2018/2019 [Hebbache et al, 2024, 2023].

La base de données pour les décès 2022 a été produite dans un contexte marqué par l’industrialisation d’une articulation de la production autour de trois modes de codage : automatique, deep learning et manuel. Ce rapport donne des indicateurs liés à la collecte des données et décrit le processus de codage mobilisé en 2022. Les données sur les causes de décès en 2022 ont été diffusées en septembre 2024 à Eurostat, puis à l’OMS. La base de données individuelles a été diffusée dans le SNDS fin novembre 2024. Celle-ci est également documentée ici. Ces diffusions ont été accompagnées de deux analyses des grandes causes de décès en 2022 [Fouillet et al 2022, Cadillac et al 2022].

2 Collecte

2.1 Exhaustivité

En 2022, 675 513 décès survenant sur le territoire français ont été comptabilisés par l’Insee.

Le CépiDc a reçu 670 099 volets médicaux et volets médicaux complémentaires et en a conservé 661 558 (soit 98,7%). Pour 2,1% des décès, aucun volet médical n’a été reçu, et donc aucune cause n’est déclarée.

Tableau 1. Exhaustivité de la collecte des volets médicaux (VM) pour les décès 2022 et 2021

	2022	2021
Décès Insee	675513	662149
VM reçus	670099	659381
VM supprimés	-8541	-10681
VM non reçus	13980	13449
% de VM non reçus	2,07%	2,03%

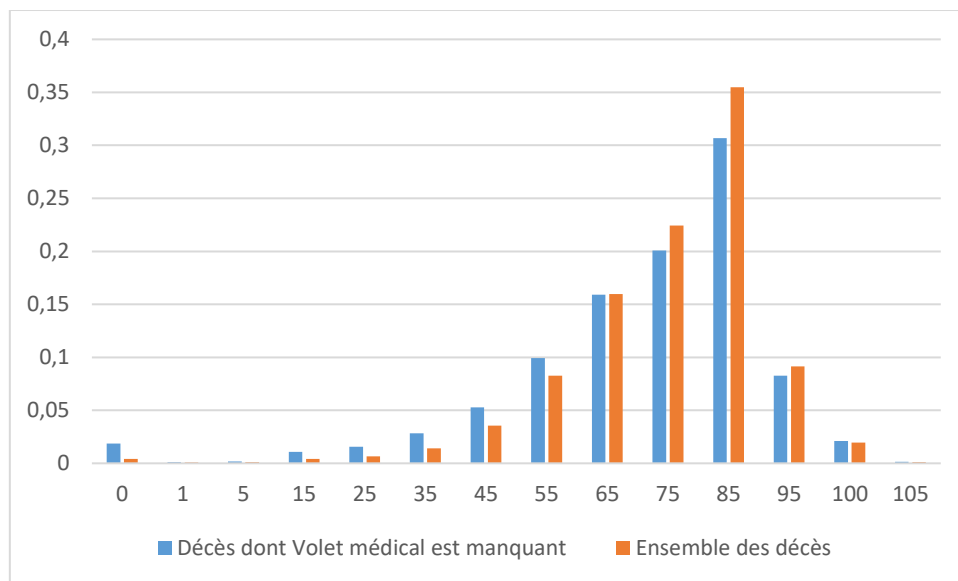
Le taux de collecte est donc de 97,93% en 2022, il était de 97,96% en 2021 et 97,79% en 2020.

Les 8500 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 13980 décès pour lesquels aucun volet médical n'a été reçu à l'Inserm sont présents dans la base de données du CépiDc avec une cause initiale codée à R99 (cause indéterminée) et la mention 'pas de certificat' dans le texte des causes. Il s'agit dans 53,1% des cas d'hommes alors qu'ils représentent 49,8% des cas dans l'ensemble des décès en 2022. Il s'agit aussi de défunts généralement plus jeunes que dans l'ensemble des décès.

Figure 1. Répartitions des décès manquants et ensemble par groupe d'âge



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. En 2022, 42% des causes initiales à R99 étaient des certificats non reçus.

2.2 Origine des certificats : électronique ou papier

Pour l'année 2022, sur l'ensemble des volets médicaux reçus et conservés au CépiDc, 423601 (64%) étaient au format papier, 237957 (36%) étaient d'origine électronique. La part de certificats électroniques est en hausse par rapport à 2021 (32% ou 209 888 certificats)

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 : 2017 ou 1997

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 1^{er} 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>, Annexe 8.1) 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 (Annexe 8.2), le certificateur peut renseigner l'enchaînement causal ayant conduit au décès.

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 2022, 608 868 (92%) correspondaient à la version de 2017 et 52 690 (8%) à la version 1997, contre respectivement 580029 (89%) et 68671 (11%) pour 2021. Si l'on se concentre sur les certificats papiers, la version 1997 représente 12% de ces certificats en 2022, contre 16% en 2021.

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 1^{er} janvier 2018.

En 2022, le CépiDc a pu prendre en compte 3171 VMC, c'est 67 VMC de plus qu'en 2021. Cette légère augmentation est portée par l'envoi des données de l'IML de Paris. La majorité des VMC (2573) étaient issus cet IML en 2022, c'est 200 de plus qu'en 2021 (2373). En effet, cet IML transmet l'ensemble des données issues de son système d'information directement au CépiDc. Les données de l'IML de Paris n'ont pas le même format que le volet médical habituel car l'IML renseigne directement le code CIM de la cause du décès telle que le médecin légiste la définit. Les autres 598 VMC pris en compte ont été collectés via l'application de certification électronique, ces VMC certifiés sur CertDc sont en baisse par rapport 2021 (731).

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

2.5 Type de certificat : général ou néonatal

En 2022, sur l'ensemble des volets médicaux reçus et conservés au CépiDc, 1863 était des volets médicaux néonataux, dont 59% rédigés électroniquement. Par ailleurs, 48 de ces volets médicaux ont été utilisés pour des décès de plus de 28 jours. Inversement, 12 décès chez des bébés de moins de 28 jours ont été certifiés en utilisant un modèle général.

3 Codage

Le codage de l'année 2022 s'est déroulé de septembre 2023 à juillet 2024 en parallèle de la finalisation du codage de l'année 2021. L'annexe 8.3 détaille l'ensemble du calendrier de production 2022.

Le processus de codage combine trois modes de codage : le codage automatique par le système expert de règles IRIS/MUSE, le codage manuel en interactif de certificats ciblés, le codage automatique (IA) impliquant des algorithmes de *deep learning* [voir Zambetta et al, 2023, 2024 ; Coudin et Robert 2024 ainsi que les rapports de production précédents]. Ce dernier mode de codage et l'évaluation globale du système mixte sont présentés dans la suite.

Dans cette section, nous décrivons comment les modes de codage interagissent et donnons les spécificités de chaque mode de codage pour l'année 2022. 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 2022.

La stratégie de codage du CépiDc intègre des algorithmes de *deep learning* capables de prédire les causes associées et la cause initiale de décès. 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.

3.1 Méthode combinant trois modes de codage

Les trois modes de codage sont combinés et suivent le schéma suivant (cf Figure 2. Répartition des modes de codage dans la campagne à trois modes, voir l'[article du courrier des stats](#)).

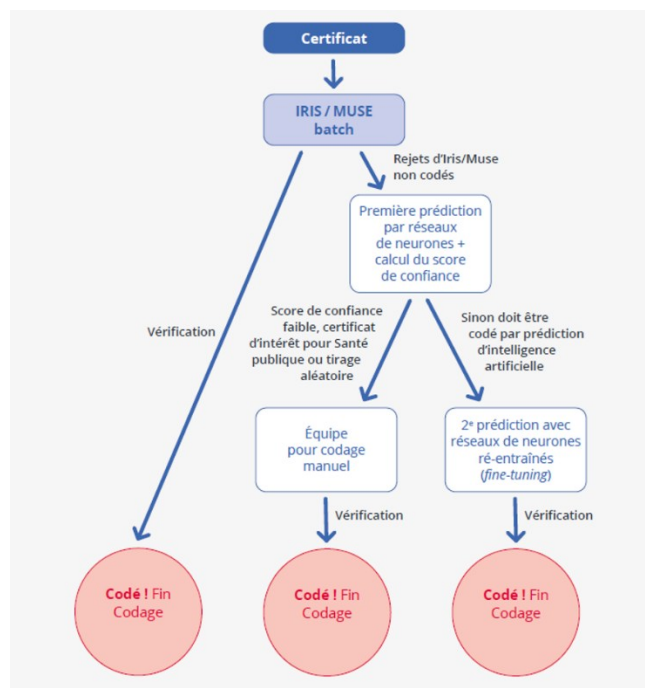


Figure 2. Répartition des modes de codage dans la campagne à trois modes

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 dont le volume est fixé en fonction des ressources disponibles pour assurer la production dans les délais. La reprise manuelle 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. Les autres certificats sont codés par inférence d'un réseau de neurones profond (RNP). L'inférence par RNP 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 2022, les modèles ayant été utilisés pour cibler des certificats sont ceux entraînés en septembre 2023, aussi utilisés pour les prédictions finales de 2021. Pour la prédiction des causes de décès des certificats restants à coder après codage manuel et automatique, les modèles ont été réentraînés en mai 2024 en prenant en compte les certificats codés jusqu'alors.

L'étape de vérifications consiste à valider ou corriger certaines données dont le codage semble a priori improbable et identifiées selon des règles déterministes 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.

3.2 Répartition des modes de codage

Au total, pour les décès 2022, 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%
 - o Version du logiciel Iris : Iris 5.8.1
 - o Version du moteur Muse utilisé dans Iris : Muse 2.9 SPECV2021SR30
- Le codage manuel assisté par Iris/Muse (13%)
- Une prédiction par *deep learning*, possiblement combinée au système expert Iris/Muse (Codage IA) : 24%.

Tableau 2. Répartition des modes de codage des données 2022 hors volets médicaux non reçus

	2022		2021	
	Effectif	Part	Effectif	Part
Codage batch système de règles	415000	61%	406000	61%
Codage inférence RNP	160000	24%	149000	23%
Codage manuel	87000	13%	93000	14%
<i>tirages aléatoires</i>	44000	7%	52000	8%
<i>reprise ciblée IA</i>	35000	5%	21000	3%
<i>décès sensibles</i>	3000	0%	3000	0%
<i>vérifications batch conduisant à modification, vérifications IA ou autres</i>	5000	1%	10000	2%
<i>codage manuel en routine avant stratégie d'allocation</i>			7000	1%
Trous de collecte	14000	2%	13000	2%
Total	676000	100%	661000	100%

Parmi les certificats de décès 2022, 87 000 certificats ont été codés manuellement, 415 000 codés par batch automatique Iris/Muse et 160 000 par IA.

Parmi les 87 000 certificats codés manuellement, 44 000 proviennent de tirages uniformes dans la population des certificats de décès de l'année et environ 43 000 de ciblage ou de vérifications (3 000 décès sensibles, 35 000 ciblage liés à l'IA, 5 000 certificats en fin de campagne repris ou codés, ainsi que les vérifications conduisant à des modifications manuelles).

3.3 Allocation des certificats vers la reprise manuelle ou l'inférence par RNP

Environ 35 000 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.

Tableau 3. Nombre de certificats repris manuellement après un ciblage IA lors de la campagne de codage 2022 (Chiffres avant synchro)

Niveau de Précision à atteindre pour chaque catégorie de la <i>shortlist</i> européenne	Nbre de certificats en reprise IA 2022
94	10428
95	4652
96	8219
96.5	5117
97	6749
Total	35165

Lecture : pour atteindre 94% de précision dans chacune des catégories de la *shortlist* européenne, on estime qu'il faut coder manuellement les 10428 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.

La méthode de ciblage est la même que celle retenue pour 2021. En première étape, on utilise le modèle utilisé pour le codage final IA de la production de 2021 pour obtenir une première prédiction de causes pour l'ensemble des certificats 2022 non codés automatiquement. Puis on calcule pour chacun de ces certificats un indicateur de confiance dans cette prédiction conditionnellement aux caractéristiques de ce certificat. Cet indicateur 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. Cet indicateur dérive d'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). (cf [Codage des causes de décès 2018-2019, approche combinant deep learning, système expert et codage manuel ciblé](#)). Ce modèle n'a pas été réestimé pour la campagne 2022, c'est le même que celui utilisé dans le cadre de la campagne 2021.

3.4 Codage par système expert Iris/Muse

La majorité des certificats sont codés automatiquement en appliquant le système expert de règles international Iris/Muse. Pour les décès 2022, 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].

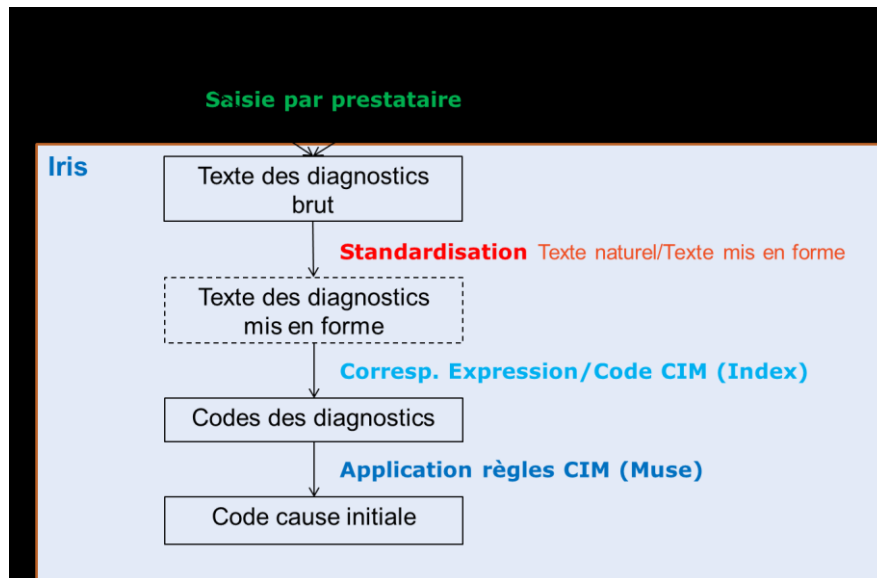


Figure 3. Codage par système expert Iris/Muse

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 2022, 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 et reprise par l'équipe de codage

L'équipe de codage code en interactif avec le logiciel Iris/Muse. Le travail de codage s'organise en trois niveaux :

Les codeurs : ils assurent le codage de premier niveau non pris en charge automatiquement par le logiciel de codage automatique Iris. Ils maîtrisent les règles de codage et connaissent les expressions médicales courantes.

Les nosologistes : Les codeurs expérimentés sont appelés "nosologistes". Ils assurent le codage expert des cas les plus complexes, et enrichissent le dictionnaire des termes rencontrés sur les certificats

Les experts : ils ont, en plus des nosologistes, une expérience médicale et maîtrisent les règles de codage. Ils sont les derniers à passer sur les échantillons, lorsque le problème concerne des cas cliniques rares et complexes et que des connaissances en physiopathologie sont nécessaires, on parle d'expertise médicale.

L'équipe de codage code les certificats ciblés par score de confiance comme présentés au-dessus, des certificats tirés aléatoirement dans l'ensemble des rejets du batch automatique (lesquels vont servir à l'entraînement ou monitoring des modèles IA, ECH1, ECH8, ECH12) ainsi que d'autres d'intérêt pour la santé publique (décès sensibles, ECH7) et la recherche (EDP, ECH2). Au total, l'équipe de codage a codé ou modifié lors des vérifications environ 87 000 certificats (Tableau 4) sur les 95 000 passés entre leurs mains.

Tableau 4. Répartition du codage manuel lors de la campagne 2022 (Effectifs après synchro)

Echantillons		Effectif codé manuellement
Tirage aléatoire 10% des lots recus - rejets de batch	ech2022- ECH001	25512
Tirage aléatoire décès EDP(1) - rejets de batch	ech2022- ECH002 - edp	10916
Tirage aléatoire dans reste après tirage de 123457	ech2022- ECH008 - alea apres 123457	3911
Tirage aléatoire dans reste après tirage de 123457910	ech2022- ECH012 - alea apres 123457910	3636
Décès sensibles	ech2022- ECH007 - DCsens	3012
Ciblage IA	ech2022- ECH003 - IA94	10331
	ech2022- ECH004 - IA95	4591
	ech2022- ECH005 - IA96	8016
	ech2022- ECH009 - IA96.5	5054
	ech2022- ECH010 - IA97	6642
Autres - hors échantillon - vérifications IA menant à codage...	ech2022- hors ech ech2022- Codage batch auto	3372
Vérifications batch menant à recodage	(iris Muse)	1913
Total		86906

(1) Echantillon démographique permanent

Les décès "sensibles" concernent des décès pour lesquels le CépiDc doit garantir l'exactitude du codage pour des raisons de santé publique (VIH, morts maternelles, décès d'enfants). Les décès sensibles codés par batch par Iris/Muse, ne sont pas reportés ici. Ils sont quant à eux vérifiés par l'équipe de codage (cf 3.7 phase de vérification présentée ensuite). Par rapport à 2021, il y a plus de décès sensibles (en particulier pour les enfants) codés par batch automatique en 2022. Mécaniquement, l'effectif de l'échantillon « décès sensibles non codés par batch Iris » diminue (3163

en 2022 pour 4639 en 2021), et celui des certificats vérifiés car relevant de décès sensibles augmente (voir après).

Tableau 5. Description et effectifs des décès identifiés comme sensibles en 2022 hors volets médicaux non reçus et avant synchro Insee

Catégorie de décès sensibles en 2022 (non codés par batch Iris)	Nbre de certificats
Moins de 28 jours (0-27 jours inclus)	1299
Au moins 28 jours – Moins de 15 ans	1191
Morts Maternelles	103
SIDA/VIH	570
Total	3163

- Comme pour 2021, Les décès sensibles correspondent en 2022 à toutes les mentions de SIDA/VIH sur le certificat, les morts maternelles ainsi que certains décès de personnes de moins de 15 ans (cf 8.6.2).

Par ailleurs à l'issue de la prédiction finale ont été repris :

- o en lien avec la tuberculose : dès que le 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, 200 certificats codés manuellement.
- o en lien avec la pharmacodépendance et la toxicomanie : le modèle de *deep learning* remonte une cause de ce type en cause initiale et la cause initiale retenue ne relève pas de la catégorie, soit 196 certificats.

Les volets médicaux complémentaires non codés automatiquement sont aussi codés manuellement, soit 295 volets médicaux.

3.6 Codage IA (Deep learning)

Les principales évolutions méthodologiques de la campagne 2022 par rapport à celle de 2021 autour des réseaux de neurones utilisés pour coder les certificats sont les suivantes :

- Simplification : abandon après étude d'impact de l'un des deux réseaux de neurones *transformers* mobilisés précédemment pour prédire une séquence de cause. Tenir en production deux modèles *transformers* (très proches finalement en termes de structure, de variables en entrée et de base d'entraînement) était trop lourd (en temps de calcul et de maintenance) et le gain observé très faible. L'étape de choix de la cause initiale (classification) a été adaptée pour conserver un support de choix assez large. On retient les trois propositions de causes initiales les plus probables (top 1, 2, 3) issues du modèle *transformer* et non la seule plus probable (voir après).
- Amélioration du modèle de classification final pour la cause initiale (surmodèle) : ce modèle prédit la cause initiale à retenir parmi les trois propositions les plus probables issues du modèle *transformer* et celle proposée par le système de règles appliqué à la séquence de codes prédite

par le modèle transformer (variables en entrée enrichies et test de performance d'autres architectures de réseau).

3.6.1 *Transformer* pour coder la séquence de causes

Le modèle Transformer de *deep learning* mobilisé dans la production 2022 (k5) pour coder la séquence de causes suit la même architecture que celui mobilisé en 2021. Pour une présentation détaillée de ces types de modèle, voir le document de travail n°2 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) et Zambetta et al (2024). On ne reprend ici que des éléments de synthèse concernant la production 2022.

Feature engineering/data pipeline.

Les séquences en entrée du modèle 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, 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 2.3, 8.1) qui permettent de mieux repérer certaines causes externes.

Ainsi, 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 206 852 *tokens* et en sortie 6 359 *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. Le modèle k5 a 142 millions de paramètres à estimer. L'annexe 8.7 illustre l'architecture de ce type de réseau, et précise les hyperparamètres choisis et les codes du modèle k5 2022.

Bases d'entraînement

Le modèle est entraîné 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. L'annexe 8.5 décrit les bases d'entraînement et de test. Les certificats codés par IA ne font pas partie des bases d'entraînement ou de test.

Une première base d'entraînement de 5 738 699 certificats a été mobilisée en début avril 2024 pour les travaux d'amélioration de la production. 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 certificats codés manuellement pour ces mêmes années au 8 juin 2023 (en respectant cette proportion quel que soit l'échantillon codé)
- 78% du codage automatique (système expert) et 56% du codage manuel de 2020, toujours tirés aléatoirement
- 96% du codage automatique et 40% du codage manuel de 2021 en date du 8 juin 2023 (hors EDP, laissé en test)
- L'ensemble des certificats 2022 codés automatiquement

La deuxième base d'entraînement qui comporte 5 764 233 certificats (+25 535) a été utilisée à la mi-avril 2024 pour produire des résultats provisoires. En plus de la première base d'entraînement, on ajoute parmi les décès 2022 :

- ECH1 (Tirage aléatoire 10%)
- ECH2 (Rejets EDP) : uniquement intégré dans le test
- ECH3 (Reprises IA - P1 - Précision CI 94%)
- ECH4 (Reprises IA - P2 - Précision CI 95%)

La troisième base d'entraînement constituée en mai 2024 est utilisée après la reprise manuelle quasiment complète. Elle comprend en plus des certificats de la deuxième base d'entraînement 50% des certificats codés manuellement lors des reprises de l'IA soit environ 16 908 certificats de plus que la première base, et au total 5 781 141 certificats. Cette base d'entraînement permet l'inférence finale des causes de décès des certificats non codés ni manuellement ni par le système expert.

Les certificats codés par les méthodes IA ne sont évidemment pas réintroduits dans la base d'entraînement.

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 532 872 observations, dont 418 824 codées manuellement pour des décès de 2016 à 2022.

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

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

Un modèle de classification en 5 classes est mobilisé pour optimalement sélectionner la cause initiale entre la prédiction directe de cause initiale par k5, proposition la plus probable (k5top 1) mais aussi proposition la deuxième plus probable (k5top 2) et proposition la troisième plus probable (k5top 3), ainsi que l'application du système expert Iris/Muse sur la séquence de causes multiples prédites par k5 (iris5).

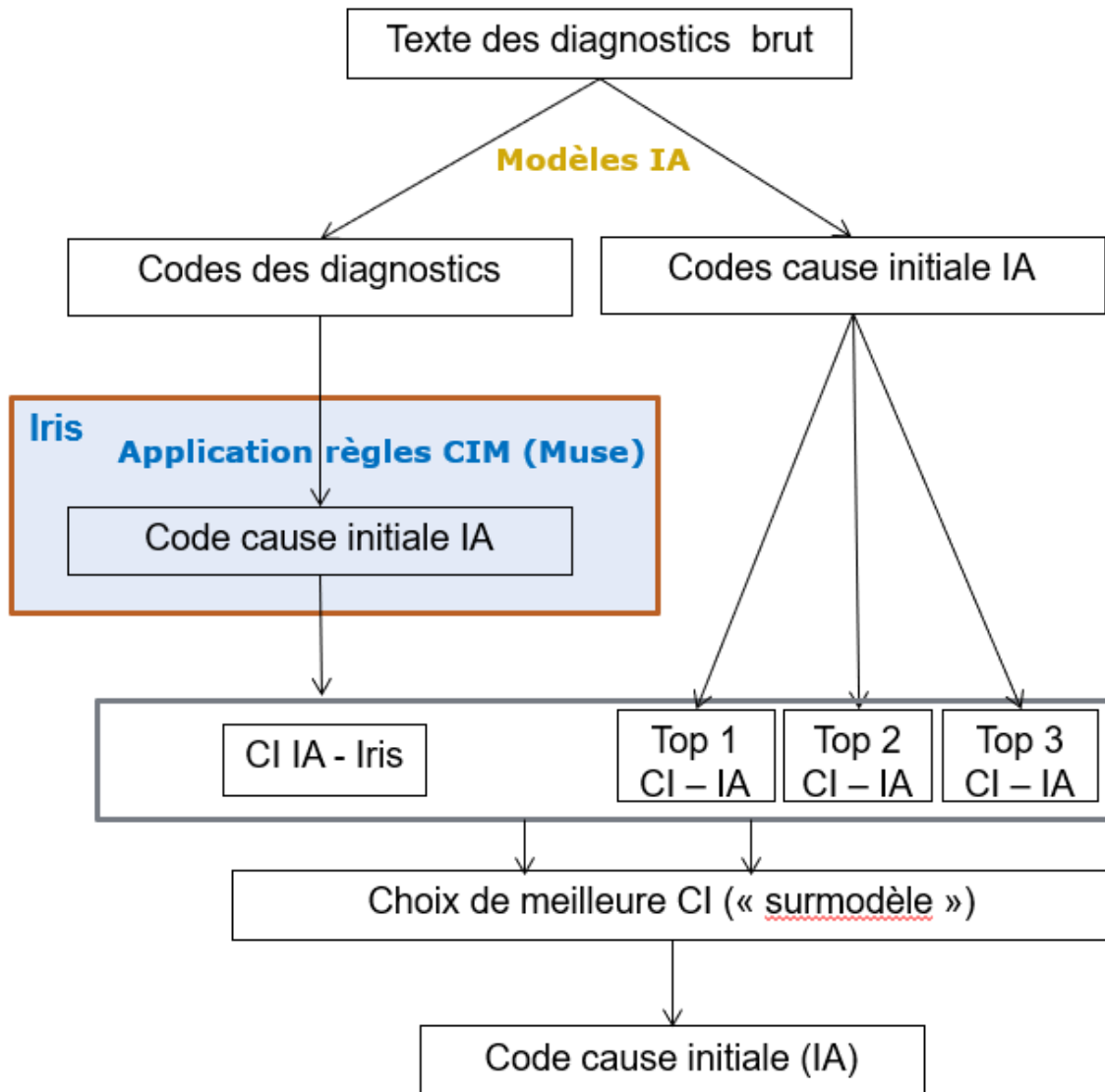


Figure 4. Description des propositions possibles de causes initiales impliquant du codage par IA

Ce sur-modèle désigne parmi les types de propositions précédentes celui dont on retiendra la proposition de cause initiale, ou si aucun des modèles n’aboutit à une bonne prédiction. Dans ce dernier cas, on retiendra la prédiction de K5Iris.

L’algorithme retenu est un BiLSTM (bidirectionnel long-term short-term memory model, voir Grave et al. 2005, Bladi et al. 1999). Il est détaillé dans l’annexe 8.8.

En résumé, les variables explicatives du modèle correspondent aux codes CIM de causes initiales proposées par les différentes classes, les causes multiples, la séquence d’entrée complète du modèle k5 (laquelle concatène les textes médicaux présents sur les certificats à des variables additionnelles d’année, sexe et âge du défunt, type et version du certificat, circonstances apparentes, etc.), ainsi que des variables numériques captant la complexité du certificat (nb de mots, de causes codées...), les probabilités des propositions et leurs écarts (pouvoir discriminant de k5), des indicateurs d’égalité

entre le résultat de l'application d'Iris et une des propositions top3 de k5 ou encore sa présence dans les causes multiples (fiabilité de la proposition).

Les variables explicatives n'entrent pas au même moment dans le BiLSTM selon qu'elles sont textuelles ou numériques. Les variables textuelles sont concaténées dans une séquence d'entrée subissant un plongement lexical sur les premières couches du modèle alors que les variables numériques gardent leur caractère numérique.

En termes d'évaluation, le sur-modèle sélectionne le bon « type » de proposition, c'est-à-dire « modèle dont on devra retenir la proposition de CI » dans 82,4% des cas dans le test (contre 81,6% pour la production 2021, 85,1% pour celle de 2018-2019, ces deux productions reposant sur deux modèles transformers lesquels n'avaient pas été entraînés sur la même base d'entraînement pour 2018/2019), et 87,1% sur le train. Au niveau du code CIM de la cause initiale (c'est-à-dire le code CIM de la cause initiale proposé par la classe/origine retenue), il est correct dans 81,7% des cas (sur la population test de référence, voir partie évaluation) pour 81,5% en 2021, et 81,9% sur 2018-2019. Ce ne serait le cas que dans 80% des cas si l'on utilisait seulement le modèle K5Iris, soit un gain de 1,7 point. Le score du train est généralement plus élevé que celui du test, car le modèle a vu et appris à partir de ces données. La différence de performance entre le train et le test pourrait indiquer un léger surapprentissage du modèle.

3.7 Vérifications

Dans le processus de codage, la phase de vérification consiste à s'assurer de la qualité du codage de certains certificats lorsque celui-ci apparaît comme « potentiellement incohérent » ou quand ces certificats concernent des décès à fort enjeu en termes de santé publique. Différents types de vérifications ont été effectuées pour l'année 2022, quel que soit le type de codage (manuel, automatique par système expert, IA). Il peut s'agir de vérifications liées à des incohérences manifestes (par exemple entre le sexe ou l'âge et la cause de décès, 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 2022 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, un certificat qui n'est pas modifié gardera son mode de codage initial, ce premier codage étant ainsi confirmé.

En 2022, environ 11 000 certificats ont ainsi été vérifiés, 2579 initialement codés par l'IA, 4883 par Iris/muse, 3592 manuellement (chiffres avant synchronisation avec l'Insee). Suite aux vérifications, 31% des certificats vérifiés changent de cause initiale : 34% des certificats codés par IrisMuse ; 30% des certificats codés manuellement ; 27% 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, cela a été le cas pour 4 625 certificats soit 42% du total. Les évolutions des vérifications par rapport à 2021 sont précisées dans l'annexe 8.6.1.

3.7.1 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 161 certificats et 95% des causes initiales et 51 % des causes associées ont été modifiées à l'issue de la vérification (pour 127 certificats et 96% de causes initiales modifiées en 2021)

3.7.2 Décès spécifiques

L'OMS et les enjeux de santé demandent de garantir un codage optimal pour certains types de décès jugés « sensibles ou spécifiques » : ie enfants, VIH, morts maternelles. Une majorité de ces certificats (3163 avant synchro) était dans l'échantillon des décès sensibles tel que décrit précédemment : ils étaient déjà vus par un expert du codage. On s'intéresse ci-dessous aux autres, ceux codés par batch automatique ou pas encore identifiés lors de la constitution de l'échantillon des décès sensibles. On détaille dans le Tableau 6 les certificats concernés ainsi que la part d'entre eux pour lesquels la vérification a conduit à un changement de codage de cause initiale au niveau le plus détaillé de la CIM. Dans la stratégie de codage de l'année 2021, le codage par batch était plus restreint, conduisant à plus de décès sensibles identifiés directement pour être codés manuellement. Si l'on somme finalement l'ensemble des cas relevant de décès sensibles qu'ils soient codés manuellement ou codés par batch et vérifiés, on retrouve le même ordre de grandeurs pour 2022 que pour 2021 (5037 pour 4721, hors intentions indéterminées). A noter que les proportions de causes initiales modifiées pour les cas DS3a et DS4 sont en dessous de 8% indiquant que le codage automatique est plutôt performant sur ces cas.

Une autre modification de l'étape de vérification a concerné l'identification des Covid long, Post Covid et des vaccinations. Plus de cas sont passés dans l'étape de vérification (1447) qu'en 2021 (719).

Tableau 6. Description et nombre de certificats vérifiés pour les décès spécifiques et part de cause initiale modifiées (Chiffres avant Synchro)

Réf	Description	Nbre de certificats vérifiés	% de CI modifiées	% de CA modifiées
DS11	Certains certificats spécifiques avec mention de Covid-19 (cf 8.6.3) (vérification de l'application de règles complexes de codage, encore récentes)	1895	20%	35%
DS10	Les certificats ayant une cause initiale de la catégorie intention indéterminée (commençant par Y1 ou Y2 ou Y3[0-4])	1825	23%	40%
CovidLong_ PostCovid	Les certificats ayant une mention de covid long ou post covid	1106	6%	34%
Vaccination	Les certificats ayant une mention de vaccination de Covid-19	341	8%	47%

DS1	Les certificats avec une mention de suicide mais non retenu en cause initiale	135	38%	33%
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 8.6.2)	610	6%	7%
DS5	Certificats avec mention de SIDA/VIH non codés dans l'échantillon des décès sensibles	209	13%	36%
DS3a	Décès d'enfants entre 0 et 27 jours (dont les enfants nés sans vie) non codés dans l'échantillon des décès sensibles	560	8%	10%
DS6	Certificats de morts maternelles non codés dans l'échantillon des décès sensibles	12	42%	67%
DS2	Les certificats avec une mention d'homicide mais non retenu en cause initiale	20	55%	65%
DS3b	Age entre 28 jours et 6 mois inclus et mention de code P non codés dans l'échantillon des décès sensibles	6	17%	17%
DS3c	Age > 6 mois et mention de code P non codés dans l'échantillons des décès sensibles	6	50%	100%
Total		6725	16%	32%

3.7.3 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 peut être considéré comme une relation causale. On détaille dans le Tableau 7 les certificats concernés.

Tableau 7. 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	% de CA 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	934	91%	94%

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	582	67%	9%
M1a	Vérification des certificats avec un code D00- D45 ou D48 en CI alors qu'il y a mention de cancer	287	52%	71%
M1b	Vérification des certificats avec un code D46 ou D47 en CI alors qu'il y a mention de cancer	144	28%	35%
M11	Anémies en cause initiale (D619, D649 et D539) avec un code informatif (R%%, Z%%, I46%, I269, I10%, I776, I50) présent en partie 1 autre qu'un code d'anémie	273	26%	31%
M7	Traitement du "ou" si mention de mort violente à côté du "ou"	304	56%	40%
M5	Notion de surdosage AVK en partie 1	137	45%	38%
M4	Code d'affection hématologique (D50-D89) en cause initiale alors qu'il existe un code C sur le certificat	148	74%	66%
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. Abandon en 2022 des cas où ces codes barrières étaient sur la dernière ligne de la P1 ou en P2	31	45%	35%
M8	D611 en CI et absence de cancer dans le certificat	23	87%	48%
M9	Cause initiale à coder en C80.9 si C34.9 ou un C41.1 en cause associée	8	100%	63%
Total		2871	66%	55%

3.7.4 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 2022, cela concernait, les hépatites virales et certaines chutes. Celles sur le diabète et les sepsis qui avaient été menées en 2021 n'ont pas été reconduites car le codage était soit automatisable soit correct.

Tableau 8. 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	% de CA 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	257	12%	49%
NR9	Les accidents (V01-X59) ou chutes W00-W19 doivent être dû à une autre cause	6	17%	50%
Total		263	13%	49%

3.7.5 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, la présence du libellé hémorragie intracérébrale pouvait être codé soit en S068 (hémorragie intracérébrale traumatique) soit en I6199 (hémorragie intracérébrale non précisée). Il a été décidé de supprimer le choix de code pour ce libellé et de permettre le codage automatique en I6199. Cependant il a été décidé de vérifier les certificats qui mentionne une hémorragie intracérébrale lorsque celle-ci est en présence d'une cause externe. En effet sur ces certificats il est possible que l'hémorragie intracérébrale soit alors codée en S068. Le contexte du certificat permet au codeur, lors de sa vérification de statuer et d'éventuellement recoder l'hémorragie intracérébrale correctement.

Les vérifications liées aux choix de code pour 2022 sont de bien moindre ampleur qu'en 2021 (total de 727 pour 2958) car certaines qui ne s'avéraient pas pertinentes ont été supprimées (elles conduisaient à modifier trop peu de causes), ou mieux ciblées.

Tableau 9. 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	% de CA modifiées
CC_Hemorragie_intracerebrale	S'il y a une notion de causes externes dans un certificat avec mention d'hémorragie intracérébrale, le codage sera S068 et non I6199	103	17%	72%
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	256	16%	60%
CC_choc_hemorragique_hypovolemique	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). En 2022 n'est plus réalisée sur le codage manuel	111	22%	87%
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.	175	14%	82%

CC_Rhabdomyolyse	S'il y a une notion de traumatisme ou cause externe (2022 : codes T ou S et Cause initiale différente de X) dans un certificat avec mention de rhabdomyolyse, on vérifie que le code M62.8 est bien approprié.	35	14%	69%
CC_noyage	Libellé « noyade » et une information de dépression (F30, F32 ou F33) en partie 1 si les circonstances apparentes du décès sont inconnues.	24	67%	17%
CC_paraplegie	On vérifie que le code G822 est approprié pour coder le libellé « paraplégi e » s'il y a aussi une mention d'AVC sur le certificat. En 2022, n'est plus réalisée sur le codage réalisé par les méthodes d'IA	23	22%	87%
total		727	18%	71%

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 2022.

3.7.6 Vérifications pour raisons multiples

Certains certificats sont vérifiés pour plusieurs raisons citées ci-dessus. Ces cas de vérifications multiples ont concerné 307 certificats en 2022 et pour 59%, la cause initiale a été modifiée à l'issue de la vérification.

4 Evaluation de la stratégie 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 2022, au sens de sa 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, de défunts EDP, observées en général dans la population des décès, et la proportion de reprise manuelle ciblée telle qu'elle a été menée en 2022. Elle est aussi représentative dans les « bonnes proportions » des décès codés automatiquement par batch. En pratique il s'agit d'une mise à jour de la population test de référence qui avait été utilisée pour évaluer la campagne 2021, en y intégrant certaines données 2022. Les données 2022 intégrées à la population test de référence n'ont pas été mobilisées dans l'entraînement des modèles. Elles sont tirées des seuls échantillons 1 et 2 dont le tirage est aléatoire et de leurs 'contreparties' codées par batch automatique, tirées pour obtenir la même proportion de codage automatique dans l'ajout à la population test de référence que dans la campagne 2022 (62,7%).

La population test de référence¹ est composée de 865 484 observations, représentative de la répartition des causes de décès sur les années 2016, 2017, 2020, 2021 et 2022. La proportion de codage automatique par *batch* dans cette population est de 58,7%, ce qui est un peu moins que le taux de codage automatique par batch effectif en 2022 (62,7%, hors trous de collecte). La conséquence sera donc une légère sous-estimation de la cohérence du codage dans les tableaux présentés ci-dessous.

On simule ensuite les apports de la reprise manuelle aléatoire et ciblée telle qu'elle a été menée en 2022, en supposant que la cause initiale codée est correcte pour les certificats relevant d'échantillons codés manuellement. Le raisonnement est le suivant : parmi les 244 949 certificats non codés automatiquement par batch en 2022², on en a tiré aléatoirement et codé manuellement 25 512 (ECH1), 10 916 (ECH2 EDP), 3 911 (ECH8), 3 636 (ECH12), soit 17,9% de l'ensemble des certificats non codés automatiquement par batch (cf Tableau 4 et Tableau 10). On va donc simuler, via un tirage uniforme, que 17,9% des certificats à coder manuellement dans le test vont passer en reprise manuelle. On va aussi considérer que les certificats relevant, 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 2022. Pour les décès sensibles, on applique les règles de repérage³ utilisées pour les identifier. Pour simuler l'impact de la reprise ciblée IA, on s'appuie sur les scores de confiance issues du modèle « 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 2022. Il y en a 31 060 dans la population test de référence. L'apport de la phase de vérifications n'est pas simulé ici. De même, 3 373 certificats ont été codés manuellement alors qu'ils ne relevaient pas d'un échantillon spécifique. Cette étape n'est pas non plus simulée ici.

Tableau 10. Description de la distribution des allocations de codage dans la base 2022 et sa simulation dans la population test de référence

	Base 2022	Répartition	Population test de référence	Répartition
Codage auto batch	416 609	63,0%	507 956	58,7%
Codages manuels aléatoires (y compris EDP)	43 975	6,6%	61 990	7,2%
Manuel EDP	10 916	1,7%	15 755	1,8%
Décès sensibles	3 012	0,5%	5 634	0,7%
Reprise manuelle ciblée IA	34 634	5,2%	46 682	5,4%
Hors échantillon	3 373	0,5%		0,0%

¹ Voir l'annexe 8.11 pour la constitution de la population test de référence.

² La phase de vérifications sur les codages batch n'est simulée ici, les certificats codés par batch mais modifiés manuellement lors des vérifications entrent donc dans les effectifs des codages batch.

³ 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 campagne 2022. En résulte de nouveau, une légère sous-estimation de la performance de la campagne 2022 dans les tableaux présentés.

Total reprise manuelle	84 994	12,8%	130 061	15,0%
Prédictions IA	159 955	24,2%	227 467	26,3%
Total	661 558	100,0%	865 484	100,0%

Source : Inserm – CépiDc

Données : Population test de référence et BCMD 2022.

Lecture : la population test de référence compte 15 755 certificats de défunts relevant de l'échantillon démographique permanent et qui n'ont pas été codés par batch automatique.

Le *Tableau 10* décrit la répartition des types de codage dans la campagne 2022 et dans sa simulation sur la population test de référence. Le taux de codage automatique est légèrement inférieur dans la simulation, comme attendu.

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 90,0% des cas (pour 89,7% sur la campagne de rattrapage 2021). Elle relève de la même catégorie de la *shortlist* européenne dans 93,7% des cas (pour 93,4% en 2021). Le *Tableau 11* **Erreur ! Source du renvoi introuvable.** et le *Tableau 12* détaillent les performances des prédictions par les modèles d'IA, combinés ou non avec Iris/Muse et avec la reprise manuelle ciblée telle qu'elle a été réalisée sur 2022.

Le *Tableau 11* se concentre sur les certificats qui auraient été codés manuellement dans une campagne traditionnelle ne combinant que batch automatique et codage manuel. Le *Tableau 12* 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 11. Cohérence (accuracy) entre causes initiales prédites par deep learning (k5), combinaison de deep learning et Iris Muse (k5 iris), proposition du surmodèle, avec ou sans prise en compte de la reprise manuelle 2022 pour les données qui auraient été codées manuellement

Année	K5	K5-IRIS/Muse	Surmodèle	Surmodèle + reprise manuelle	Nombre d'observations dans la base simulée	Surmodèle + reprise manuelle (2021)
Accuracy au niveau du code CIM-10						
Ensemble	0,790	0,800	0,817	0,900	357 528	0,897
2016	0,778	0,786	0,800	0,892	93 144	0,891
2017	0,775	0,781	0,799	0,891	93 912	0,888
2020	0,804	0,817	0,835	0,911	121 461	0,906
2021	0,803	0,818	0,837	0,909	23 520	0,901
2022	0,804	0,818	0,838	0,904	25 491	-
Accuracy au niveau de la catégorie de la shortlist européenne						
Ensemble	0,858	0,863	0,877	0,937	357 528	0,934
2016	0,85	0,853	0,866	0,931	93 144	0,931
2017	0,847	0,850	0,863	0,93	93 912	0,927
2020	0,869	0,877	0,891	0,945	121 461	0,941
2021	0,865	0,875	0,889	0,941	23 520	0,937
2022	0,866	0,875	0,893	0,941	25 491	-

Source : Inserm – CépiDc.

Données : Population test de référence - uniquement les données codées manuellement dans le cadre d'une campagne classique.

Lecture : Au niveau le plus fin de la Classification Internationale des Maladies, la première prédiction du modèle K5 est correcte dans 79,0% des cas qui auraient été codés manuellement dans une campagne traditionnelle. Si IRIS/Muse propose une cause initiale à partir des causes associées proposées par K5, on passe à 80% de cohérence. Si on utilise le surmodèle pour départager les trois prédictions proposées par K5 et celle proposée par Iris/Muse, on obtient 81,7% de certificats codés avec le même code CIM-10. En prenant en compte toutes les étapes jusqu'à la reprise manuelle, on arrive à une *accuracy* de 90%.

Au niveau le plus fin de la CIM, la prédiction du modèle k5 est correcte dans 79,0% des cas (79,1% en 2021). Appliquer Iris/Muse sur la séquence des causes prédites par k5 lorsqu'il obtient une réponse

non ambiguë fait gagner un point de cohérence. Le surmodèle qui choisit entre les trois meilleures propositions de cause initiale par k5 et k5Iris (k5 combiné avec Iris/Muse) permet de gagner 1,7 point supplémentaire. La prise en compte de la reprise aléatoire et ciblée permet de gagner 8,3 points supplémentaires (contre 8,2 points en 2021) et d’atteindre les 90%. L’évaluation de chaque étape de la reprise ciblée sera détaillée par la suite. Au niveau *shortlist* européenne, le sur-modèle permet de gagner 1,3 point d’*accuracy* par rapport à K5Iris (pour 1,3 point en 2021) et la reprise aléatoire et ciblée 6,0 points (pour 5,8 points en 2021). Au total, on atteint 93,7% de cohérence (93,4% en 2021). La cohérence est relativement homogène sur les cinq années de validité des données. Ce sont les données les plus anciennes de 2016 et 2017 qui ont le taux de cohérence le plus faible, les données récentes sont donc mieux traitées par la campagne actuelle.

Tableau 12. Cohérence entre causes initiales prédites par deep learning (k5), combinaison de deep learning et Iris Muse (k5 iris), proposition du surmodèle, avec ou sans prise en compte de la reprise manuelle 2022 pour l’ensemble de la population (et 2021 pour référence)

Année	K5 automatique +	K5-IRIS/Muse + automatique	Surmodèle + automatique	Surmodèle + reprise manuelle + automatique	Nombre d'observations	Surmodèle + reprise manuelle + automatique (2021)
Accuracy au niveau du code CIM-10						
Ensemble	0,913	0,917	0,924	0,959	865 484	0,957
2016	0,907	0,91	0,916	0,955	221 807	0,954
2017	0,907	0,909	0,917	0,955	226 856	0,954
2020	0,917	0,922	0,93	0,962	285 784	0,960
2021	0,926	0,932	0,939	0,966	63 058	0,963
2022	0,926	0,932	0,939	0,964	67 979	-
Accuracy au niveau de la catégorie de la shortlist européenne						
Ensemble	0,941	0,944	0,949	0,974	865 484	0,973
2016	0,937	0,938	0,944	0,971	221 807	0,971
2017	0,937	0,938	0,943	0,971	226 856	0,970
2020	0,944	0,948	0,954	0,977	285 784	0,975
2021	0,950	0,953	0,959	0,978	63 058	0,977
2022	0,950	0,953	0,960	0,978	67 979	-

Source : Inserm – CépiDc.

Données : Population test de référence.

Lecture : Au niveau de la catégorie agrégée de la *shortlist* d'Eurostat, le code CIM donné par le codage automatique ou, à défaut, le modèle K5 est correcte dans 94,1% des cas qui auraient été codés automatiquement ou manuellement dans une campagne traditionnelle. Si plutôt que K5, IRIS/Muse propose une cause initiale à partir des causes associées proposées par K5, on passe à 94,4% de cohérence. Si on utilise le surmodèle pour départager les trois prédictions proposées par K5 et celle proposée par Iris/Muse, on obtient 94,9% de certificats codés avec le même code CIM-10. Si finalement, on récupère la prédiction du codage automatique, ou à défaut celle du codage manuel, ou à défaut celle du surmodèle, on arrive à une *accuracy* de 97,4%.

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 95,9% des cas au niveau le plus fin de la CIM (pour 95,7% en 2021) et dans 97,4% des cas au niveau de la *shortlist* européenne (pour 97,3% en 2021), voir Tableau 12.

4.3 Précision, rappel et écarts d'effectifs

Le Tableau 13 et le Tableau 14 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 rapporté à l'ensemble des observations réellement dans la catégorie ; la F-mesure est la moyenne harmonique des deux.

Tableau 13. Performances et effectifs prédits du sur-modèle et du sur-modèle combiné avec la reprise manuelle évalués sur les observations de la population test qui auraient été codées manuellement.

Shortlist EUROSTAT	Effectifs réels (campagne traditionnelle)	Surmodèle						Surmodèle + reprise manuelle					
		Effectifs prédits	Précision	Rappel	F-mesure	pred/réels - 1	Signif.	Effectifs prédits	Précision	Rappel	F-mesure	pred/réels - 1	Signif.
01.1 Tuberculose	445	384	0,888	0,766	0,823	-0,137	****	426	0,96	0,919	0,939	-0,043	
01.2 SIDA (maladie VIH)	288	291	0,739	0,747	0,743	0,01		288	1	1	1	-	
01.3 Hépatites virales	352	348	0,713	0,705	0,709	-0,011		334	0,907	0,861	0,883	-0,051	
01.4 Autres maladies infectieuses et parasitaires	6 194	6 160	0,789	0,784	0,786	-0,005		5 975	0,934	0,901	0,917	-0,035	****
02.1.01 Tumeur lèvre, bouche, pharynx	2 914	2 727	0,935	0,875	0,904	-0,064	****	2 836	0,956	0,93	0,943	-0,027	*
02.1.02 Tumeur de l'oesophage	2 601	2 573	0,956	0,945	0,951	-0,011		2 594	0,969	0,967	0,968	-0,003	
02.1.03 Tumeur de l'estomac	2 507	2 446	0,951	0,928	0,939	-0,024		2 494	0,961	0,956	0,959	-0,005	
02.1.04 Tumeur Colon, rectum, anus	10 440	10 338	0,955	0,946	0,95	-0,01		10 468	0,966	0,968	0,967	0,003	
02.1.05 Tumeur du foie et voies biliaires	5 120	4 996	0,944	0,921	0,933	-0,024	**	5 077	0,958	0,95	0,954	-0,008	
02.1.06 Tumeur du Pancréas	5 763	5 707	0,965	0,955	0,96	-0,01		5 787	0,973	0,977	0,975	0,004	
02.1.07 Tumeur du Larynx	676	617	0,922	0,842	0,88	-0,087	***	652	0,949	0,916	0,932	-0,036	
02.1.08 Tumeur du Poumon	16 837	16 962	0,947	0,954	0,95	0,007		17 087	0,96	0,974	0,967	0,015	**
02.1.09 Mélanome malin de la peau	1 261	1 268	0,917	0,922	0,92	0,006		1 277	0,937	0,949	0,943	0,013	
02.1.10 Tumeur du sein	7 323	7 326	0,95	0,951	0,951	0		7 407	0,961	0,972	0,967	0,011	
02.1.11 Tumeur du col de l'utérus	565	560	0,934	0,926	0,93	-0,009		561	0,955	0,949	0,952	-0,007	
02.1.12 Tumeur autre partie de l'utérus	1 775	1 687	0,95	0,903	0,925	-0,05	***	1 736	0,967	0,946	0,956	-0,022	
02.1.13 Tumeur de l'ovaire	1 882	1 904	0,931	0,942	0,937	0,012		1 932	0,947	0,972	0,96	0,027	
02.1.14 Tumeur de la prostate	5 124	5 108	0,937	0,934	0,936	-0,003		5 155	0,953	0,959	0,956	0,006	
02.1.15 Tumeur du rein	2 278	2 204	0,936	0,905	0,92	-0,032	*	2 266	0,949	0,944	0,946	-0,005	
02.1.16 Tumeur de la vessie	3 152	3 097	0,945	0,929	0,937	-0,017		3 140	0,959	0,955	0,957	-0,004	
02.1.17 Tumeur cerveau et système nerveux central	2 378	2 281	0,948	0,91	0,929	-0,041	***	2 348	0,965	0,953	0,959	-0,013	
02.1.18 Tumeur de la thyroïde	278	264	0,92	0,874	0,897	-0,05		268	0,951	0,917	0,934	-0,036	
02.1.19 Maladie de Hodgkin et lymphomes	3 472	3 379	0,948	0,923	0,935	-0,027	*	3 434	0,968	0,957	0,963	-0,011	
02.1.20 Leucémie	3 848	3 848	0,937	0,937	0,937	-		3 893	0,954	0,965	0,959	0,012	
02.1.21 Autres Tumeur des tissus lymphoïdes et hématopoïétiques	2 091	2 062	0,926	0,913	0,919	-0,014		2 094	0,95	0,952	0,951	0,001	
02.1.22 Autres Tumeurs malignes	16 055	16 532	0,85	0,876	0,863	0,03	****	15 885	0,928	0,919	0,923	-0,011	*
02.2 Tumeurs non malignes	5 593	5 633	0,838	0,844	0,841	0,007		5 509	0,93	0,916	0,923	-0,015	
03 Maladies du sang et hématopoïétiques	2 193	2 068	0,693	0,653	0,673	-0,057	****	2 047	0,908	0,848	0,877	-0,067	****
04.1 Diabète sucre	7 778	7 388	0,909	0,864	0,886	-0,05	****	7 608	0,948	0,927	0,937	-0,022	**
04.2 Autres maladies endocriniennes, nutritionnelles et métaboliques	6 464	6 346	0,81	0,795	0,802	-0,018	*	6 288	0,914	0,889	0,902	-0,027	***
05.1 Démences	8 896	9 597	0,848	0,915	0,881	0,079	****	9 520	0,887	0,949	0,917	0,07	****
05.2 Abus d'alcool	1 654	1 799	0,781	0,849	0,814	0,088	****	1 659	0,926	0,929	0,927	0,003	
05.3 Pharmacodépendance, toxicomanie	223	193	0,72	0,623	0,668	-0,135	***	220	0,941	0,928	0,935	-0,013	
05.4 Autres troubles mentaux et du comportement	2 724	2 723	0,782	0,782	0,782	0		2 641	0,908	0,881	0,894	-0,03	*
06.1 Maladie de Parkinson	3 108	3 113	0,921	0,922	0,921	0,002		3 137	0,941	0,95	0,945	0,009	
06.2 Maladie d'Alzheimer	8 473	8 618	0,93	0,946	0,938	0,017	*	8 658	0,949	0,97	0,959	0,022	***
06.3 Autres maladies du système nerveux et des organes des sens	7 882	7 813	0,827	0,82	0,824	-0,009		7 816	0,925	0,918	0,921	-0,008	
07.1.1 Infarctus aigu du myocarde	6 796	7 172	0,868	0,916	0,891	0,055	****	7 079	0,911	0,949	0,93	0,042	****
07.1.2 Autres Cardiopathies ischémiques	11 762	11 654	0,886	0,878	0,882	-0,009		11 670	0,932	0,925	0,929	-0,008	
07.2 Autres maladies du cœur	25 179	25 263	0,863	0,866	0,865	0,003		25 393	0,922	0,93	0,926	0,008	*
07.3 Maladies cérébro-vasculaires	19 996	19 971	0,891	0,89	0,89	-0,001		20 112	0,938	0,943	0,941	0,006	
07.4 Autres maladies de l'appareil circulatoire	15 922	15 515	0,854	0,832	0,843	-0,026	****	15 728	0,931	0,919	0,925	-0,012	*

suite	Effectifs réels (campagne traditionnelle)	Surmodèle						Surmodèle + reprise manuelle					
		Effectifs prédits	Précision	Rappel	F-measure	pred/réels - 1	Signif.	Effectifs prédits	Précision	Rappel	F-measure	pred/réels - 1	Signif.
08.1 Grippe	833	835	0,91	0,912	0,911	0,002		838	0,955	0,96	0,958	0,006	
08.2 Pneumonie	4 916	4 842	0,848	0,836	0,842	-0,015		4 900	0,913	0,91	0,912	-0,003	
08.3.1 Asthme	458	442	0,857	0,828	0,842	-0,035		440	0,939	0,902	0,92	-0,039	
08.3.2 Autres maladies chroniques des voies respiratoires inférieures	6 095	6 250	0,88	0,903	0,891	0,025	***	6 185	0,935	0,949	0,942	0,015	
08.4 Autres maladies de l'appareil respiratoire	7 144	7 254	0,769	0,781	0,775	0,015	*	7 167	0,886	0,888	0,887	0,003	
09.1 Ulcère gastro duodénal	644	614	0,873	0,832	0,852	-0,047		627	0,954	0,929	0,941	-0,026	
09.2 Cirrhoses, fibroses et hépatites chroniques	4 381	4 496	0,894	0,917	0,905	0,026	**	4 481	0,94	0,962	0,951	0,023	*
09.3 Autres maladies de l'appareil digestif	12 139	12 121	0,85	0,849	0,85	-0,001		12 121	0,931	0,93	0,931	-0,001	
10 Maladies de la peau et du tissu sous cutané	1 285	1 266	0,776	0,764	0,77	-0,015		1 230	0,933	0,893	0,913	-0,043	*
11.1 Arthrite rhumatoïde et ostéoartrite	475	455	0,807	0,773	0,789	-0,042		452	0,942	0,897	0,919	-0,048	
11.2 Autres maladies du système ostéoarticulaire et des muscles	3 411	3 300	0,744	0,719	0,731	-0,033	**	3 277	0,933	0,897	0,914	-0,039	***
12.1 Maladies du rein et de l'urètre	4 785	4 404	0,837	0,77	0,802	-0,08	****	4 603	0,925	0,89	0,907	-0,038	****
12.2 Autres maladies génito-urinaires	2 552	2 373	0,838	0,779	0,807	-0,07	****	2 480	0,925	0,899	0,912	-0,028	*
13 Complications de grossesse et d'accouchement	58	12	0,917	0,19	0,314	-0,793	****	58	1	1	1	-	
14 Certains affections dont l'origine est périnatale	1 877	1 979	0,912	0,961	0,936	0,054	***	1 892	0,992	0,999	0,995	0,008	
15 Malformations congénitales et anomalies chromosomiques	1 481	1 150	0,858	0,666	0,75	-0,223	****	1 385	0,965	0,902	0,932	-0,065	***
16.1 Syndrome de mort subite du nourrisson	190	194	0,928	0,947	0,938	0,021		194	0,974	0,995	0,984	0,021	
16.2 Causes inconnues ou non précisées	5 377	5 466	0,821	0,834	0,827	0,017		5 287	0,897	0,882	0,89	-0,017	
16.3 Autres symptômes et états morbides non définis	7 127	7 673	0,813	0,876	0,843	0,077	****	7 735	0,852	0,925	0,887	0,085	****
17.1.1 Accidents de transport	2 456	2 322	0,942	0,89	0,915	-0,055	****	2 442	0,964	0,958	0,961	-0,006	
17.1.2 Chutes accidentelles	9 531	9 830	0,91	0,938	0,924	0,031	****	9 528	0,971	0,97	0,97	0	
17.1.3 Noyade et submersion accidentelle	431	426	0,852	0,842	0,847	-0,012		432	0,938	0,94	0,939	0,002	
17.1.4 Intoxications accidentelles	1 760	1 837	0,718	0,749	0,733	0,044	**	1 700	0,952	0,919	0,935	-0,034	*
17.1.5 Autres accidents	14 308	14 354	0,847	0,85	0,849	0,003		13 914	0,95	0,924	0,937	-0,028	****
17.2 Suicide	5 416	5 451	0,93	0,936	0,933	0,006		5 532	0,952	0,972	0,962	0,021	*
17.3 Homicides	430	378	0,894	0,786	0,837	-0,121	***	405	0,97	0,914	0,941	-0,058	
17.4 Evènement dont l'intention n'est pas déterminée	1 540	1 168	0,73	0,554	0,63	-0,242	****	1 350	0,951	0,834	0,889	-0,123	****
17.5 Autre cause externe	1 683	1 881	0,313	0,35	0,331	0,118	****	1 603	0,799	0,761	0,779	-0,048	**
18 COVID-19	14 483	14 820	0,951	0,973	0,962	0,023	****	14 771	0,965	0,984	0,974	0,02	***
Total	357 528	357 528						357 528					

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

Shortlist EUROSTAT	Effectifs réels (campagne traditionnelle)	Codage batch Iris + Surmodèle						Codage batch Iris + Surmodèle + reprise manuelle					
		Effectifs prédits	Précision	Rappel	F-mesure	pred/réels - 1	Signif.	Effectifs prédits	Précision	Rappel	F-mesure	pred/réels - 1	Signif.
01.1 Tuberculose	498	437	0,902	0,791	0,843	-0,122	****	479	0,965	0,928	0,946	-0,038	
01.2 SIDA (maladie VIH)	363	366	0,792	0,799	0,796	0,008		363	1	1	1	0	
01.3 Hépatites virales	592	588	0,83	0,824	0,827	-0,007		574	0,946	0,917	0,931	-0,03	
01.4 Autres maladies infectieuses et parasitaires	14 123	14 089	0,908	0,905	0,906	-0,002		13 904	0,972	0,957	0,964	-0,016	**
02.1.01 Tumeur lèvre, bouche, pharynx	5 339	5 152	0,965	0,932	0,948	-0,035	***	5 261	0,976	0,962	0,969	-0,015	
02.1.02 Tumeur de l'œsophage	5 199	5 171	0,978	0,973	0,975	-0,005		5 192	0,985	0,983	0,984	-0,001	
02.1.03 Tumeur de l'estomac	6 216	6 155	0,981	0,971	0,976	-0,01		6 203	0,984	0,982	0,983	-0,002	
02.1.04 Tumeur Colon, rectum, anus	24 839	24 737	0,981	0,977	0,979	-0,004		24 867	0,986	0,987	0,986	0,001	
02.1.05 Tumeur du foie et voies biliaires	12 363	12 239	0,977	0,967	0,972	-0,01		12 320	0,983	0,979	0,981	-0,003	
02.1.06 Tumeur du Pancréas	16 745	16 689	0,988	0,985	0,986	-0,003		16 769	0,991	0,992	0,991	0,001	
02.1.07 Tumeur du Larynx	1 336	1 277	0,962	0,92	0,941	-0,044	*	1 312	0,975	0,957	0,966	-0,018	
02.1.08 Tumeur du Poumon	43 688	43 813	0,979	0,982	0,981	0,003		43 938	0,984	0,99	0,987	0,006	
02.1.09 Mélanome malin de la peau	2 428	2 435	0,957	0,96	0,958	0,003		2 444	0,967	0,974	0,97	0,007	
02.1.10 Tumeur du sein	17 979	17 982	0,98	0,98	0,98	0		18 063	0,984	0,989	0,986	0,005	
02.1.11 Tumeur du col de l'utérus	1 146	1 141	0,968	0,963	0,965	-0,004		1 142	0,978	0,975	0,976	-0,003	
02.1.12 Tumeur autre partie de l'utérus	3 960	3 872	0,978	0,956	0,967	-0,022	*	3 921	0,985	0,976	0,981	-0,01	
02.1.13 Tumeur de l'ovaire	4 785	4 807	0,973	0,977	0,975	0,005		4 835	0,979	0,989	0,984	0,01	
02.1.14 Tumeur de la prostate	12 822	12 806	0,975	0,974	0,974	-0,001		12 853	0,981	0,983	0,982	0,002	
02.1.15 Tumeur du rein	4 939	4 865	0,971	0,956	0,963	-0,015		4 927	0,976	0,974	0,975	-0,002	
02.1.16 Tumeur de la vessie	7 409	7 354	0,977	0,97	0,973	-0,007		7 397	0,983	0,981	0,982	-0,002	
02.1.17 Tumeur cerveau et système nerveux central	5 703	5 606	0,979	0,962	0,971	-0,017		5 673	0,986	0,98	0,983	-0,005	
02.1.18 Tumeur de la thyroïde	525	511	0,959	0,933	0,946	-0,027		515	0,975	0,956	0,965	-0,019	
02.1.19 Maladie de Hodgkin et lymphomes	6 893	6 800	0,974	0,961	0,968	-0,013		6 855	0,984	0,979	0,981	-0,006	
02.1.20 Leucémie	8 481	8 481	0,971	0,971	0,971	0		8 526	0,979	0,984	0,981	0,005	
02.1.21 Autres Tumeur des tissus lymphoïdes et hématopoïétiques	4 631	4 602	0,967	0,961	0,964	-0,006		4 634	0,978	0,978	0,978	0,001	
02.1.22 Autres Tumeurs malignes	31 648	32 125	0,923	0,937	0,93	0,015	****	31 478	0,964	0,959	0,961	-0,005	
02.2 Tumeurs non malignes	10 981	11 021	0,917	0,92	0,919	0,004		10 897	0,965	0,957	0,961	-0,008	
03 Maladies du sang et hématopoïétiques	3 807	3 682	0,828	0,8	0,814	-0,033	***	3 661	0,949	0,912	0,93	-0,038	***
04.1 Diabète sucre	17 187	16 797	0,96	0,938	0,949	-0,023	****	17 017	0,977	0,967	0,972	-0,01	*
04.2 Autres maladies endocriniennes, nutritionnelles et métaboliques	15 071	14 953	0,919	0,912	0,916	-0,008		14 895	0,964	0,953	0,958	-0,012	*
05.1 Démences	27 202	27 903	0,948	0,972	0,96	0,026	****	27 826	0,961	0,983	0,972	0,023	****
05.2 Abus d'alcool	3 539	3 684	0,893	0,93	0,911	0,041	***	3 544	0,965	0,967	0,966	0,001	
05.3 Pharmacodépendance, toxicomanie	340	310	0,826	0,753	0,788	-0,088	*	337	0,961	0,953	0,957	-0,009	
05.4 Autres troubles mentaux et du comportement	5 375	5 374	0,89	0,889	0,89	0		5 292	0,954	0,94	0,947	-0,015	
06.1 Maladie de Parkinson	9 588	9 593	0,974	0,975	0,975	0,001		9 617	0,981	0,984	0,982	0,003	
06.2 Maladie d'Alzheimer	27 403	27 548	0,978	0,983	0,981	0,005		27 588	0,984	0,991	0,987	0,007	
06.3 Autres maladies du système nerveux et des organes des sens	16 946	16 877	0,92	0,916	0,918	-0,004		16 880	0,965	0,962	0,964	-0,004	
07.1.1 Infarctus aigu du myocarde	19 355	19 731	0,952	0,97	0,961	0,019	****	19 638	0,968	0,982	0,975	0,015	***
07.1.2 Autres Cardiopathies ischémiques	26 324	26 216	0,949	0,945	0,947	-0,004		26 232	0,97	0,966	0,968	-0,003	
07.2 Autres maladies du cœur	72 672	72 756	0,953	0,954	0,953	0,001		72 886	0,973	0,976	0,974	0,003	
07.3 Maladies cérébro-vasculaires	44 586	44 561	0,951	0,951	0,951	-0,001		44 702	0,972	0,975	0,973	0,003	
07.4 Autres maladies de l'appareil circulatoire	35 605	35 198	0,936	0,925	0,93	-0,011	***	35 411	0,969	0,964	0,967	-0,005	

suite	Shortlist EUROSTAT	Effectifs réels (campagne traditionnelle)	Codage batch Iris + Surmodèle					Codage batch Iris + Surmodèle + reprise manuelle						
			Effectifs prédits	Précision	Rappel	F-mesure	pred/réels - 1	Signif.	Effectifs prédits	Précision	Rappel	F-mesure	pred/réels - 1	Signif.
	08.1 Grippe	1 867	1 869	0,96	0,961	0,96	0,001		1 872	0,98	0,982	0,981	0,003	
	08.2 Pneumonie	17 677	17 603	0,958	0,954	0,956	-0,004		17 661	0,976	0,975	0,976	-0,001	
	08.3.1 Asthme	1 154	1 138	0,945	0,932	0,938	-0,014		1 136	0,976	0,961	0,969	-0,016	
	08.3.2 Autres maladies chroniques des voies respiratoires inférieures	14 160	14 315	0,948	0,958	0,953	0,011	*	14 250	0,972	0,978	0,975	0,006	
	08.4 Autres maladies de l'appareil respiratoire	22 888	22 998	0,927	0,932	0,929	0,005		22 911	0,964	0,965	0,965	0,001	
	09.1 Ulcère gastro duodénal	1 178	1 148	0,932	0,908	0,92	-0,025		1 161	0,975	0,961	0,968	-0,014	
	09.2 Cirrhoses, fibroses et hépatites chroniques	9 762	9 877	0,952	0,963	0,957	0,012		9 862	0,973	0,983	0,978	0,01	
	09.3 Autres maladies de l'appareil digestif	24 209	24 191	0,925	0,924	0,925	-0,001		24 191	0,966	0,965	0,965	-0,001	
	10 Maladies de la peau et du tissu sous cutané	2 260	2 241	0,873	0,866	0,87	-0,008		2 205	0,963	0,939	0,951	-0,024	
	11.1 Arthrite rhumatoïde et ostéoartrite	796	776	0,887	0,864	0,875	-0,025		773	0,966	0,938	0,952	-0,029	
	11.2 Autres maladies du système ostéoarticulaire et des muscles	4 972	4 861	0,826	0,808	0,817	-0,022	*	4 838	0,955	0,929	0,942	-0,027	**
	12.1 Maladies du rein et de l'urètre	11 562	11 181	0,936	0,905	0,92	-0,033	****	11 380	0,97	0,954	0,962	-0,016	**
	12.2 Autres maladies génito-urinaires	4 454	4 275	0,91	0,873	0,891	-0,04	****	4 382	0,958	0,942	0,95	-0,016	
	13 Complications de grossesse et d'accouchement	61	15	0,933	0,23	0,368	-0,754	****	61	1	1	1	0	
	14 Certains affections dont l'origine est périnatale	2 194	2 296	0,924	0,967	0,945	0,046	***	2 209	0,993	1	0,996	0,007	
	15 Malformations congénitales et anomalies chromosomiques	2 263	1 932	0,916	0,782	0,843	-0,146	****	2 167	0,977	0,936	0,956	-0,042	***
	16.1 Syndrome de mort subite du nourrisson	195	199	0,93	0,949	0,939	0,021		199	0,975	0,995	0,985	0,021	
	16.2 Causes inconnues ou non précisées	22 568	22 657	0,957	0,961	0,959	0,004		22 478	0,976	0,972	0,974	-0,004	
	16.3 Autres symptômes et états morbides non définis	44 384	44 930	0,968	0,98	0,974	0,012	****	44 992	0,975	0,988	0,981	0,014	****
	17.1.1 Accidents de transport	3 961	3 827	0,965	0,932	0,948	-0,034	***	3 947	0,977	0,974	0,976	-0,004	
	17.1.2 Chutes accidentelles	12 259	12 558	0,929	0,952	0,941	0,024	****	12 256	0,977	0,977	0,977	0	
	17.1.3 Noyade et submersion accidentelle	1 175	1 170	0,946	0,942	0,944	-0,004		1 176	0,977	0,978	0,977	0,001	
	17.1.4 Intoxications accidentelles	2 327	2 404	0,785	0,81	0,797	0,033	*	2 267	0,964	0,939	0,951	-0,026	
	17.1.5 Autres accidents	19 994	20 040	0,891	0,893	0,892	0,002		19 600	0,965	0,946	0,955	-0,02	****
	17.2 Suicide	12 146	12 181	0,968	0,971	0,97	0,003		12 262	0,978	0,988	0,983	0,01	
	17.3 Homicides	559	507	0,921	0,835	0,876	-0,093	***	534	0,978	0,934	0,955	-0,045	
	17.4 Evénement dont l'intention n'est pas déterminée	1 894	1 522	0,793	0,637	0,707	-0,196	****	1 704	0,961	0,865	0,911	-0,1	****
	17.5 Autre cause externe	1 980	2 178	0,407	0,447	0,426	0,1	****	1 900	0,83	0,796	0,813	-0,04	**
	18 COVID	39 954	40 291	0,982	0,99	0,986	0,008	**	40 242	0,987	0,994	0,991	0,007	*
	Total	865 484	865 484						865 484					

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 < 0,2, ** pval < 0,1, *** pval < 0,05, **** pval < 0,01 : voir <https://www.cepidc.inserm.fr/documentation/tests-statistiques-relatifs-aux-indicateurs-de-mortalite-en-population>

En rouge : les causes pour lesquelles au moins un indicateur de précision, rappel ou de F-mesures est inférieur à 0,95.

Source : Inserm – CépiDc.

Données : Base de données de test et base 2022.

Lecture : la campagne classique à deux modes batch et manuel conduit à 39954 causes initiales Covid, la campagne à trois modes de codage, à 40242, soit un écart de 0,007.

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 0,968 (niveaux très proches des niveaux simulés pour la campagne 2021). La seule catégorie pour laquelle la F-mesure est inférieure à 0,9 est « 17.5 Autres causes externes »⁴ (comme pour la campagne 2021), ce qui invite à considérer les codages dans cette catégorie avec prudence.

La F-mesure reste inférieure à 0,95 pour 8 catégories (contre 12 pour la campagne 2021) :

- « 01.1 Tuberculose » (supérieure à 0,95 pour 2021)
- « 01.3 Hépatites virales »
- « 03 Maladies du sang et hématopoïétiques »

⁴ 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)

- « 05.4 Autres troubles mentaux et du comportement »
- « 10 Maladies de la peau et du tissu sous cutané »
- « 11.1 Arthrite rhumatoïde et ostéoarthrite »
- « 12.2 Autres maladies génito-urinaires »
- « 17.4 Evènement dont l'intention n'est pas déterminée »
- « 17.5 Autre cause externe »

Des écarts conséquents et significatifs ($p < 0,05$ – au moins ***) en termes d'effectifs (test de Poisson), par catégorie, se retrouvent pour les catégories :

- « 03 Maladies du sang et hématopoïétiques » - sous-estimation de 3,8% de l'effectif attendu (contre 6,2% en 2021).
- « 05.1 Démences » - surestimation de 2,3% de l'effectif attendu (2,2% en 2021).
- « 07.1.1 Infarctus aigu du myocarde » - surestimation de 1,5% de l'effectif attendu (contre 1% non significatif en 2021).
- « 15 Malformations congénitales et anomalies chromosomiques » - sous-estimation de 4,2% (contre 3,5% non significatif en 2021).
- « 16.3 Autres symptômes et états morbides non définis » - surestimation de 1,4% (contre 0,6% non significatif en 2021).
- « 17.1.5 Autres accidents » - sous-estimation de 2% (contre 0,4% non significatif en 2021)
- « 17.4 Evènement dont l'intention n'est pas déterminée » - sous-estimation de 1% (contre 5,7% en 2021).

Certaines catégories dont les écarts étaient significativement différents en 2021 ne se retrouvent pas dans la campagne de 2022 : « 16.2 Causes inconnues ou non précisées » et « 17.5 Autres causes externes » (sous-estimation de 16% en 2021, 37% en 2018/2019, et 4% non significatif en 2022)⁵.

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

⁵ La cause de décès « 17.5 Autres causes externes » possède à la fois une F-mesure inférieure à 0,9 et un écart entre effectifs réels et prédit non significatif : le modèle n'est pas très bon pour identifier correctement les certificats appartenant effectivement à cette catégorie mais les effectifs de ceux qu'il ajoute (qui ne font pas partie de la catégorie) et ceux qu'il oublie (et qu'il devrait identifier) se compensent autour du « vrai » chiffre.

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, non réalisée aujourd'hui, des écart-types associés à ces estimateurs.

Tableau 15. Performances en termes de cohérence de chapitre de la CIM de l'ensemble de la campagne 2022 et effectifs prédits sur l'ensemble de la population test de référence.

Chapitre CIM10	Batch Iris + surmodèle + reprise manuelle						
	Effectifs réels	Effectifs prédits	Précision	Rappel	F-mesure	pred/réel - 1	Signif.
I : Certaines maladies infectieuses et parasitaires	15 576	15 320	0,971	0,955	0,963	-0,016	***
II : Tumeurs	240 055	240 022	0,993	0,993	0,993	0,000	
III : Maladies du sang et des organes hématopoïétiques et certains troubles du système immunitaire	3 807	3 661	0,949	0,912	0,930	-0,038	***
IV : Maladies endocriniennes, nutritionnelles et métaboliques	32 258	31 912	0,973	0,962	0,967	-0,011	**
IX : Maladies de l'appareil circulatoire	198 542	198 869	0,982	0,984	0,983	0,002	
V : Troubles mentaux et du comportement	36 456	36 999	0,964	0,978	0,971	0,015	****
VI : Maladies du système nerveux	53 793	53 948	0,979	0,982	0,981	0,003	
X : Maladies de l'appareil respiratoire	57 746	57 830	0,977	0,978	0,978	0,001	

XI : Maladies de l'appareil digestif	35 149	35 214	0,971	0,972	0,971	0,002	
XII : Maladies de la peau et du tissu cellulaire sous-cutané	2 260	2 205	0,963	0,939	0,951	-0,024	
XIII : Maladies du système ostéo-articulaire, des muscles et du tissu conjonctif	5 768	5 611	0,957	0,931	0,944	-0,027	***
XIV : Maladies de l'appareil génito-urinaire	16 016	15 762	0,968	0,952	0,960	-0,016	***
XV : Grossesse, accouchement et puerpéralité	61	61	1,000	1,000	1,000	0,000	
XVI : Certaines affections dont l'origine se situe dans la période périnatale	2 194	2 209	0,993	1,000	0,996	0,007	
XVII : Malformations congénitales et anomalies chromosomiques	2 263	2 167	0,977	0,936	0,956	-0,042	***
XVIII : Symptômes, signes et résultats anormaux d'examens cliniques et de laboratoire, non classés ailleurs	67 146	67 669	0,983	0,990	0,986	0,008	***
XX : Causes externes de morbidité et de mortalité	56 295	55 646	0,978	0,967	0,973	-0,012	****
XXII : Covid	39 955	40 242	0,987	0,994	0,991	0,007	*

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 15 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 16. 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 2022

	Test qui aurait été codés manuellement			Ensemble de la population Test de référence		
	Effectifs concernés	Code CIM10	Shortlist européenne	Effectifs concernés	Code CIM10	Shortlist européenne
Surmodèle	244 484	0,817	0,877	728 937	0,924	0,949
+ décès sensibles (manuel)	4 511	0,822	0,879	5 634	0,926	0,95
+ EDP et échantillons aléatoire	61 216	0,853	0,900	83 596	0,939	0,959
+ échantillons ciblées IA	46 682	0,900	0,937	46 682	0,959	0,974
+ dernières vérifications IA	635	0,900	0,937	635	0,959	0,974
Total	357 528	0,900	0,937	865 484	0,959	0,974

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

Source : Inserm – CépiDc.

Données : Population test de référence.

Lecture : Sur la population test qui aurait été codée manuellement dans une campagne traditionnelle, pour 244 484 certificats le codage provient de uniquement de l'IA. Si les 357 528 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 81,7% des cas. Coder manuellement les 4 511 décès sensibles permet d'atteindre 82,2% de cohérence (*accuracy*).

Le Tableau 16 détaille l'apport en performance de chaque étape de la reprise manuelle. Au total, c'est 8,3 points de cohérence (Code CIM-10) qui sont gagnés sur la population du test qui aurait été codée manuellement dans une campagne classique, entre une campagne entièrement codée en IA et une campagne avec l'ensemble des étapes de la stratégie manuelle (niveau équivalent en 2021). 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.

4.5 *Evaluation de la cohérence de la campagne 2022 avec une campagne à deux modes sur le codage des causes multiples*

Pour évaluer la cohérence du codage de l'ensemble du processus morbide (causes multiples), on introduit un indicateur de cohérence de séquence de codes (bleu score pondéré) qui vaut 1 lorsque deux séquences de codes sont exactement les mêmes et s'en éloigne lorsque la dissimilarité augmente (voir Annexe 8.10 sa construction). On s'intéresse à la proportion de certificats pour lesquels la séquence de codes est exactement la même, dans ce cas le bleu score sera égal à 1, et la valeur moyenne de ce bleu score. Sur les données codées manuellement dans une campagne batch iris + manuel, dans 56,2% des cas la séquence de codes prédite par le modèle k5 est exactement la même que celle codée manuellement. Et même lorsque les séquences s'éloignent c'est de peu car le score bleu pondéré moyen est de 85,6 %, ces résultats montent respectivement à 72,1% et 91,2% lorsque la reprise manuelle est ajoutée comme lors de la campagne 2022. Sur l'ensemble de la population de test de référence, incluant les certificats codés par batch, la cohérence des traductions des causes multiples en codes en tenant compte de tous les modes de codage avec celles résultant d'une campagne batch+manuel, est exactement la même dans 88,9% des cas, et le bleu score est en moyenne de 96,5%.

Ces résultats mettent en évidence une bonne adéquation des codes des causes multiples à celles résultant d'un codage manuel et une amélioration notable grâce à la reprise manuelle. Par ailleurs, les performances globales de l'algorithme montrent une stabilité relative d'une année à l'autre, renforçant la fiabilité du système dans le cadre des campagnes de codage

L'analyse peut se décliner par shortlist européenne de la cause initiale. En conclusion, Sur l'ensemble de la population test de référence (IA + batch + reprise manuelle), les niveaux de cohérence sont très élevés pour la plupart des catégories : 0,95 pour « BS » (moyenne des scores BLEU) et 0,85 pour « BS1 » (moyenne des scores BLEU lorsque celui-ci est égal à 1).

Le Bleu Score est inférieur à 95% pour 16 catégories (1.1-Tuberculose, 2.1.01-TM lèvre, bouche, pharynx, 2.1.07-TM du Larynx, 2.1.09- Mélanome malin de la peau, 2.1.11-TM du col de l'utérus, 2.1.17-TM cerveau et système nerveux central, 2.1.18-TM de la thyroïde, 2.1.19-Maladie de Hodgkin et lymphomes, 2.1.20- Leucémie, 2.1.22-Autres Tumeurs malignes, 12.2- Autres maladies génito-urinaires, 17.1.1- Accidents de transport, 17.1.5 - Autres accidents, 17.2- Suicide, 17.3- Homicides, 17.4- Evènement dont l'intention n'est pas déterminée, 17.5- Autre cause externe).

La cohérence entre les causes multiples est parfaite dans moins de 85% des cas pour 11 catégories (1.1-Tuberculose, 2.1.01-TM lèvre, bouche, pharynx, 2.1.07-TM du Larynx, 2.1.09- Mélanome malin de la peau, 2.1.11-TM du col de l'utérus, 2.1.18-TM de la thyroïde, 2.1.19-Maladie de Hodgkin et lymphomes, 2.1.20- Leucémie, 2.1.22-Autres Tumeurs malignes, 12.2- Autres maladies génito-urinaires, 17.1.1- Accidents de transport, 17.5- Autre cause externe).

5 Evolutions de codage

5.1 *Nouveautés relatives aux recommandations de l'OMS*

Rien de particulier ou de notable à signaler sur les règles de codage 2022 par rapport à 2021. La **CIM-10** n'est plus maintenue ou discutée au sein des différents groupes d'experts, maintenant que

la CIM-11 a été adoptée. Aucune mise à jour n'est désormais disponible pour la CIM10 sauf sur la Covid-19.

Pour mémoire, 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

Les codes U07 sont utilisés depuis l'année 2020 et les autres à partir de l'année 2021.

5.2 Nouveautés internes au CépiDc

Il s'agit d'évolutions concernant les implémentations en pratique des règles de la CIM pour le cas spécifique de la France. Elles ne constituent pas des ajouts à proprement parler dans la CIM-10, mais plutôt des précisions ou des points d'attention sur des éléments déjà existants, sur lesquels une mise en avant particulière a été effectuée.

Indéterminés : Les certificats avec une circonstance apparente de décès indiquant une intention « indéterminée » et avec au plus une cause mentionnant un arrêt cardiorespiratoire (ACR – code CIM commençant par R9) sont codés par la cause initiale « événement non spécifié, intention indéterminée » (code CIM Y34). Cette règle concerne 3 certificats en 2020, 11 certificats en 2021, et 317 certificats en 2022. Cela a un impact sur le nombre de certificats codés en Y34 avec une circonstance apparente de décès indiquant une intention « indéterminée », de 293 certificats concernés en 2020, ils sont passés à 366 en 2021 et 726 en 2022.

Auto-médication : Les décès imputables à l'auto-médication doivent être codés en X44, qui correspond à une intoxication accidentelle par des substances médicamenteuses, non médicamenteuses ou biologiques, lorsqu'aucune intention suicidaire ou malveillante n'est identifiée.

Tabac : Il est confirmé que le code F17.9 (troubles mentaux et comportementaux liés à l'usage du tabac, sans précision) peut être utilisé comme cause initiale dans les situations où aucune conséquence spécifique liée à la consommation de tabac n'est identifiée. Cependant, dans les cas où la consommation de tabac entraîne une issue identifiable, comme une maladie ou une complication particulière, les règles de codage peuvent imposer l'utilisation d'un code correspondant à cette conséquence en tant que cause initiale. Cette approche garantit une cohérence dans le codage, tout en reflétant fidèlement la chaîne causale de l'événement

Mise à jour des codes U82 à U85 - Résistance aux médicaments antimicrobiens et antinéoplasiques

Les codes U82 à U85, qui concernent la résistance aux médicaments antimicrobiens et antinéoplasiques, ont été modifiés par l'OMS et enrichis en 2013 avant d'être intégrés dans leur forme actuelle en 2014 dans la CIM-10. Toutefois, leur utilisation est restée limitée avant 2022, marquée par une application peu systématique, et leur absence fréquente du dictionnaire de codage.

À partir de l'année de codage 2022, une structuration plus rigoureuse de leur utilisation a été mise en place, accompagnée d'une meilleure intégration dans les pratiques de codage.

Il est essentiel de noter que ces catégories ne doivent jamais être utilisées pour coder une cause initiale de décès. Elles sont spécifiquement conçues pour être utilisées en tant que codes de causes associées, permettant d'identifier et de documenter des résistances spécifiques en complément d'autres causes identifiées dans la chaîne causale. Cette distinction garantit une précision accrue dans le codage des résistances médicamenteuses tout en respectant les règles internationales de codification.

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

Aucune mise à jour ou correction majeure du dictionnaire n'a été réalisée pour 2022.

6 Spécificités de la base de données individuelles

6.1 Variable sur le mode de codage utilisé pour coder ce certificat

Dans le cadre des évolutions du processus de codage impliquant de nouveaux outils, le CépiDc a, comme pour les années 2018, 2019 et 2021 mis en place des informations complémentaires disponibles en base de données :

- Le type de codage qui correspond à la méthode utilisée pour coder le certificat : codage manuel, automatique par système expert Iris/Muse, ou autre méthode de codage (par IA). Les modalités relatives au codage par IA sont modifiées par rapport à 2021 (voir annexe 8.9).

L'annexe 8.9 détaille le dictionnaire des codes des envois au SNDS. On rappelle que pour les certificats codés avec IA le texte diffusé correspond au libellé brut présent sur le certificat sans découpage par rang (contrairement au cas des certificats codés manuellement ou par batch automatique Iris/muse, voir annexe de production sur les données 2021).

7 Références

[Cadillac M, Fouillet A, Rivera C, Coudin É. « Les causes de décès en France en 2022 : recul du Covid-19 et hausse des maladies respiratoires », *Etudes et Résultats*, n°1312](#)

[Fouillet A, Cadillac M., Rivera C Coudin É. Grandes causes de mortalité en France en 2022 et tendances récentes. Bull Épidémiol Hebd. 2024;2024;\(18\):388-404.](#)

[Hebbache, Z. et al \(2024\) Rapport de production - Année de décès 2021 - Document de travail du CépiDc n4/2024](#)

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

Coudin, E et Robert A, « Les statistiques sur les causes de décès : classer et coder dans la classification internationale des maladies », Courrier des statistiques N12 2024- [Les statistiques sur les causes de décès – Courrier des statistiques N12 - 2024 | Insee](#)

Zambetta E, et al. (2023) 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, Elisa, Razakamanana, Nirintsoa, Robert, Aude, Clanché, François, Rivera, Cecilia, Martin, Diane, et al. Combining deep neural networks, a rule-based expert system and targeted manual coding for ICD-10 coding causes of death of French death certificates from 2018 to 2019. *International Journal of Medical Informatics.* 2024;188(105462). 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_simsd_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 : <input type="checkbox"/> date réelle OU <input type="checkbox"/> constatée		Sexe :	
Date de naissance :		<input type="checkbox"/> masculin	
		<input type="checkbox"/> féminin	
CAUSES DU DÉCÈS			
PARTIE I Maladie(s) ou affection(s) morbide(s) ayant directement provoqué le décès. <i>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...).</i>			Intervalle entre le début du processus morbide et le décès <i>En heures, jours, mois ou ans</i>
a) _____			_____
due à ou consécutive à : b) _____			_____
due à ou consécutive à : c) _____			_____
due à ou consécutive à : d) _____			_____
<small>La dernière ligne remplie doit correspondre à la cause initiale</small>			
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		GROSSESSE La femme décédée était-elle enceinte ?	
<input type="checkbox"/> Domicile (du défunt ou autre)		<input type="checkbox"/> non, pas au cours de l'année précédant le décès	
<input type="checkbox"/> EHPAD, maison de retraite		<input type="checkbox"/> pas au moment du décès, mais grossesse terminée depuis 42 jours ou moins	
<input type="checkbox"/> Voie publique		<input type="checkbox"/> pas au moment du décès, mais grossesse terminée depuis plus de 42 jours et moins d'1 an	
<input type="checkbox"/> Établissement de santé public		<input type="checkbox"/> oui, au moment du décès	
<input type="checkbox"/> Établissement de santé privé		<input type="checkbox"/> ne sait pas	
<input type="checkbox"/> Établissement pénitentiaire		La grossesse a-t-elle contribué au décès ? <input type="checkbox"/> oui <input type="checkbox"/> non <input type="checkbox"/> ne sait pas	
<input type="checkbox"/> Autre lieu ou indéterminé			
MORT SUBITE S'agit-il d'un décès brutal et inattendu, évocateur de mort subite* ?		ACTIVITÉ PROFESSIONNELLE Le décès est-il survenu lors d'une activité professionnelle* ?	
<input type="checkbox"/> oui <input type="checkbox"/> non <input type="checkbox"/> ne sait pas		<input type="checkbox"/> oui <input type="checkbox"/> non <input type="checkbox"/> ne sait pas	
<small>* décès non traumatique (adulte, enfant, nourrisson) avec mode de survenue brutal (en moins d'une heure ou probablement) et mortel (exclusion des maladies chroniques au stade terminal)</small>			
<small>* toute activité soumise de revers (y compris au domicile), au trajet domicile-travail, les déplacements professionnels, etc.</small>			
CIRCONSTANCES APPARENTES DU DÉCÈS		RECHERCHE DE LA CAUSE DU DÉCÈS	
<input type="checkbox"/> Mort naturelle		Une recherche de la cause du décès a-t-elle été demandée ?	
<input type="checkbox"/> Accidents		<input type="checkbox"/> oui, recherche médicale <input type="checkbox"/> oui, recherche médico-légale <input type="checkbox"/> non	
<input type="checkbox"/> Complications de soins médicaux, chirurgicaux		<small>Si oui, un volet médical complémentaire sera établi séparément par le médecin ayant réalisé le diagnostic des causes de décès</small>	
<input type="checkbox"/> Suicide		SIGNATURE <i>Nom lisible et cachet obligatoire du médecin</i>	
<input type="checkbox"/> Investigations en cours		_____	
<input type="checkbox"/> Atteinte à la vie d'autrui		_____	
<input type="checkbox"/> Indéterminées		_____	
EN CAS DE MORT VIOLENTE (accidentelle, délictuelle, suicidaire, criminelle) Précisez le lieu de survenue de l'évènement déclencheur :			
<input type="checkbox"/> Domicile <input type="checkbox"/> Lieu de sport <input type="checkbox"/> Voie publique			
<input type="checkbox"/> Commerce <input type="checkbox"/> Local industriel, chantier <input type="checkbox"/> Exploitation agricole			
<input type="checkbox"/> Établissement accueillant du public <input type="checkbox"/> Autre lieu ou indéterminé			
<small>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).</small>			
<small>Le certificat peut être saisi électroniquement à l'adresse suivante https://sic.certdc.inserm.fr</small>			



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

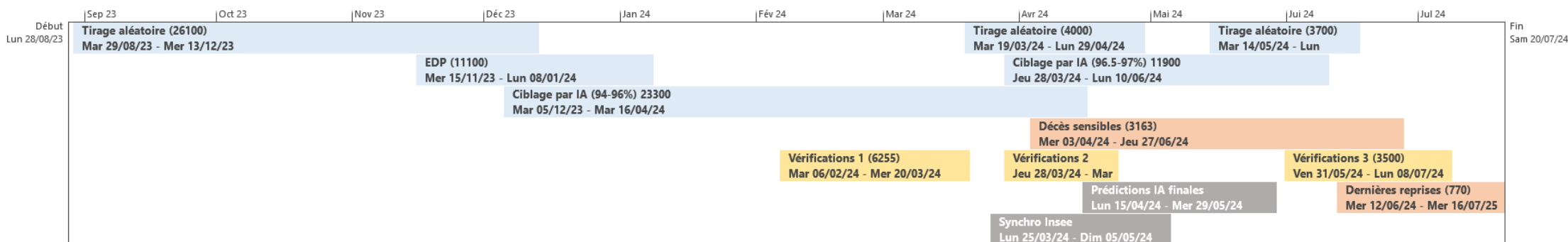
VOLET MÉDICAL À remplir et à clore par le médecin ayant constaté le décès – Renseignements confidentiels et anonymes (* instructions en annexe)			
INFORMATIONS RELATIVES À L'ENFANT			
Commune de décès : _____ Code postal : _____		Date et heure de décès : _____ à _____ h _____	Sexe : <input type="checkbox"/> masculin <input type="checkbox"/> féminin <input type="checkbox"/> indéterminé
Commune de domicile : _____ Code postal : _____		Date et heure de naissance* : _____ à _____ h _____	
Appar à 1 minute : _____		Âge gestationnel en semaines révolues d'aménorrhée : _____ Poids de naissance en grammes : _____	
INFORMATIONS RELATIVES À L'ACCOUCHEMENT		INFORMATIONS RELATIVES AUX PARENTS <i>(inscrire le code approprié)</i>	
Naissance : 1. unique 2. gémellaire 3. triple 4. quadruple 5. quintuple <input type="checkbox"/>		MÈRE Année de naissance : _____	
Numéro d'ordre de l'enfant si grossesse multiple : _____		Nationalité (en clair) : _____	
Lieu d'accouchement : 1. établissement de santé 2. domicile 3. autre <input type="checkbox"/>		Profession* (en clair) : _____	
Présentation : 1. sommet 2. autre céphalique 3. siège 4. autre <input type="checkbox"/>		exercée pendant la grossesse : 1. oui 2. non 3. chômage 4. autre situation	
Début du travail : 1. spontané 2. déclenché 3. césarienne avant travail <input type="checkbox"/>		État matrimonial : 1. célibataire 2. mariée 3. veuve 4. divorcée	
Mode d'accouchement* : 1. voie basse non instrumentale <input type="checkbox"/>		La mère vit-elle en couple ? 1. oui 2. non	
2. extraction instrumentale par voie basse 3. césarienne <input type="checkbox"/>		Nombre total de grossesses, y compris grossesse pour cet enfant : _____	
Transfert ou hospitalisation particulière *de l'enfant : 1. oui 2. non <input type="checkbox"/>		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 <i>(*Lire les instructions de remplissage en annexe)</i>			
CAUSE FŒTALE OU NÉONATALE* déterminante de la mort – Affection(s) morbide(s) ayant directement provoqué le décès. <i>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...).</i>			
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 <i>(cocher la case appropriée pour chaque point – *Lire les instructions de remplissage en annexe)</i>			
LIEU DU DÉCÈS <input type="checkbox"/> Domicile (du défunt ou autre)		RECHERCHE DE LA CAUSE DU DÉCÈS*	
<input type="checkbox"/> Établissement de santé public	<input type="checkbox"/> Voie publique	Une recherche de la cause du décès a-t-elle été demandée ?	
<input type="checkbox"/> Établissement de santé privé	<input type="checkbox"/> Autre lieu ou indéterminé	<input type="checkbox"/> oui, recherche médicale <input type="checkbox"/> oui, recherche médico-légale <input type="checkbox"/> non	
MORT INATTENDUE DU NOURRISSON S'agit-il d'un décès brutal et inattendu* ?		Si oui, un volet médical complémentaire sera établi ultérieurement par le médecin ayant réalisé le diagnostic des causes de décès.	
<input type="checkbox"/> oui <input type="checkbox"/> non <input type="checkbox"/> ne sait pas			
<i>* décès non traumatique du nourrisson avec mode de naissance brutal (en moins d'une heure ou probablement) et inattendu.</i>		SIGNATURE Nom lisible et cachet obligatoire du médecin	
CIRCONSTANCES APPARENTES DU DÉCÈS			
<input type="checkbox"/> Mort naturelle	<input type="checkbox"/> Faits de guerre		
<input type="checkbox"/> Accident	<input type="checkbox"/> Complications de soins médicaux, chirurgicaux		
<input type="checkbox"/> Atteinte à la vie de l'enfant	<input type="checkbox"/> Investigations en cours <input type="checkbox"/> Indéterminées		

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 2022

Le calendrier ci-dessous synthétise les périodes de codage pour les décès 2022. Il affiche la période de synchronisation des données avec celles de l'Insee et la période d'entraînement et prédiction IA des données finalement utilisées. Il n'affiche pas la préparation des données pour la mettre à disposition du codage (entraînement initial des modèles, prédictions pour le ciblage du codage les moins bien codés par l'IA, tirages aléatoires...) qui sont fait en amont, ni les phases de traitements statistiques et informatiques aval conduisant à la constitution des bases. Les données individuelles 2022 ont été fournies à Eurostat fin septembre 2024 et les tables individuelles complètes et détaillées au SNDS mi-novembre 2024.



Nom de la tâche	Description	Début	Fin
Tirage aléatoire (26100)	Codage manuel	Mar 29/08/23	Mer 13/12/23
EDP (11100)	Codage manuel	Mer 15/11/23	Lun 08/01/24
Ciblage par IA (94-96%) 23300	Codage manuel	Mar 05/12/23	Mar 16/04/24
Vérifications 1 (6300)	Vérifications manuelles de certificats déjà codés (en batch ou manuel)	Mar 06/02/24	Mer 20/03/24
Tirage aléatoire (4000)	Codage manuel	Mar 19/03/24	Lun 29/04/24
Synchro Insee	Alignement des bases de données CépiDc/Insee	Lun 25/03/24	Dim 05/05/24
Ciblage par IA (96.5-97%) 11900	Codage manuel	Jeu 28/03/24	Lun 10/06/24
Vérifications 2 (1400)	Vérifications manuelles de certificats déjà codés (en manuel)	Jeu 28/03/24	Mar 23/04/24
Décès sensibles (3200)	Codage manuel spécifique	Mer 03/04/24	Jeu 27/06/24
Prédictions IA finales	Entraînement et prédiction des modèles IA	Lun 15/04/24	Mer 29/05/24
Tirage aléatoire (3700)	Codage manuel	Mar 14/05/24	Lun 17/06/24
Vérifications 3 (3500)	Vérifications manuelles de certificats déjà codés (en manuel ou par IA)	Ven 31/05/24	Lun 08/07/24

Dernières reprises (770)	Codage manuel spécifique	Mer 12/06/24	Mer 16/07/24
Traitements statistiques	Constitutions des autres variables et des tables pour envoi Eurostat	Juillet	Aout
Diffusion des données causes initiales à Eurostat			20 Septembre
Traitements statistiques aval	Constitutions des autres variables et des tables pour alimentation SNDS	Septembre	Octobre
Alimentation du SNDS			18 novembre

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

Le ciblage des certificats à envoyer en reprise manuelle sur la base d'un faible score de confiance dans la prédiction de l'IA s'appuie sur le même modèle qu'utilisé pour la campagne 2021. Il n'y a pas eu de réestimation du modèle calculant le score de confiance d'un certificat donné. Il est seulement appliqué à de nouvelles données.

Pour rappel, ce score de confiance correspond à un modèle de probabilité linéaire conditionnel à certaines variables explicatives décrites ci-dessous et estimé sur 50 000 observations relevant des années 2021 et 2020, codées manuellement et issues de la base de test.

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'égalité des codes proposés par le modèle de deep learning et Iris/muse
- Ainsi que deux scores prédits par l'algorithme de deep learning k5 (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 R^2 ajusté du modèle est de l'ordre de 20%. Les accuracy de l'ordre de 80% dans le train et dans le test.

Le reste du test permet ensuite d'estimer la proportion de certificats à reprendre manuellement pour atteindre un niveau donné de précision pour chaque catégorie de diffusion (shortlist européenne. A partir des données restantes du test (uniquement les données 2020 et 2021 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 (échantillons aléatoires, EDP, décès sensibles...) sont supposés alors être correctement codés.

Ainsi, on définit :

- *eff codes* : le nombre de certificats déjà codés avec une cause initiale dans la catégorie d'intérêt (ce codage étant obtenu automatiquement ou bien codé manuellement car relevant d'un échantillon aléatoire, de l'EDP, de décès sensibles..)
- *eff_noncodes* : 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 est $a = \frac{eff\ codes}{eff\ tot}$
- La précision totale est alors $P_t = (1 - a)P_{ia} + a$ où P_{ia} est la précision pour les non-codés de la catégorie
- Si on fixe un seuil minimal P_t^* sur la précision totale, cela conduit à devoir atteindre une précision de P_{ia}^* sur les non-codés telle que

$$P_{ia}^* = \frac{(P_t^* - a)}{1 - a} = \frac{P_t^* - \frac{eff\ codes}{eff\ tot}}{\frac{eff\ non\ codes}{eff\ tot}} = \frac{P_t^* * eff\ tot - eff\ codes}{eff\ non\ codes}$$

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 P_{ia}^* 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 : $comp\ cat = P_{ia}^{-1} * P_{ia}^*$ à affecter à l'effectif des non-codés prédits dans la catégorie pour avoir les effectifs à reprendre par catégorie.

$$Eff\ a\ reprendre = P_{ia}^{-1} * Eff\ non\ code$$

Effectifs à reprendre manuellement pour atteindre XX% de précision dans la catégorie	94%	95%	96%	96,50 %	97%
01.1- Tuberculose	20	2	33	0	12
01.2- SIDA (maladie VIH)	104	31	1	10	1
01.3- Hépatites virales	58	17	22	2	6
01.4- Autres maladies infectieuses et parasitaires	1280	237	427	237	332
02.1.01-TM lèvre, bouche, pharynx	0	0	0	0	14
02.1.02-TM de l'œsophage	0	0	0	0	0
02.1.03-TM de l'estomac	0	0	0	0	0
02.1.04-TM Colon, rectum, anus	0	0	0	0	0
02.1.05-TM du foie et voies biliaires	0	0	0	0	0
02.1.06-TM du Pancréas	0	0	0	0	0
02.1.07-TM du larynx	0	0	21	0	46
02.1.08-TM du Poumon	0	0	0	0	0
02.1.09- Mélanome malin de la peau	0	0	0	0	0
02.1.10-TM du sein	0	0	0	0	0
02.1.11-TM du col de l'utérus	0	0	0	0	9
02.1.12-TM autre partie de l'utérus	0	0	0	0	0
02.1.13-TM de l'ovaire	0	0	0	0	0
02.1.14-TM de la prostate	0	0	0	0	0
02.1.15-TM du rein	0	0	0	0	0
02.1.16-TM de la vessie	0	0	0	0	0
02.1.17-TM cerveau et système nerveux central	0	0	0	0	0
02.1.18-TM de la thyroïde	0	0	0	0	2
02.1.19-Maladie de Hodgkin et lymphomes	0	0	0	17	34
02.1.20- Leucémie	0	0	0	0	0
02.1.21-Autres TM des tissus lymphoïdes et hématopoïétiques	0	0	0	0	0
02.1.22-Autres Tumeurs malignes	192	480	961	672	384
02.2 - Tumeurs non malignes	193	192	193	288	193
03 Maladies du sang et hématopoïétiques	661	96	81	80	65
04.1- Diabète sucre	0	0	89	135	89
04.2- Autres maladies endocriniennes, nutritionnelles et métaboliques	356	267	312	178	222
05.1- Démences	0	0	0	0	138
05.2- Abus d'alcool	205	144	120	48	48
05.3 - Pharmacodépendance, toxicomanie	103	5	27	2	0
05.4 - Autres troubles mentaux et du comportement	285	101	143	102	142
06.1- Maladie de Parkinson	0	0	0	0	0
06.2 - Maladie d'Alzheimer	0	0	0	0	0
06.3- Autres maladies du système nerveux et des organes des sens	379	270	379	162	270
07.1.1-Infarctus aigu du myocarde	0	0	0	110	110
07.1.2-Autres Cardiopathies ischémiques	0	0	189	190	252
07.2-Autres maladies du cœur	0	0	736	442	883
07.3-Maladies cérébraux vasculaires	0	0	341	341	454
07.4- Autres maladies de l'appareil circulatoire	280	468	654	374	373
08.1 - Grippe	0	25	19	7	32
08.2 - Pneumonie	0	0	51	101	101
08.3.1 - Asthme	16	11	14	16	3
08.3.2-Autres maladies chroniques des voies respiratoires inférieures	0	0	196	158	117

08.4- Autres maladies de l'appareil respiratoire	177	264	353	133	220
09.1 - Ulcère gastro duodénal	0	0	37	53	29
09.2 - Cirrhoses, fibroses et hépatites chroniques	0	0	76	75	50
09.3- Autres maladies de l'appareil digestif	0	453	454	377	454
10- Maladies de la peau et du tissu sous-cutané	234	18	45	36	99
11.1- Arthrite rhumatoïde et ostéoarthrite	44	39	17	7	7
11.2- Autres maladies du système ostéoarticulaire et des muscles	721	225	180	91	90
12.1-Maladies du rein et de l'urètre	299	180	209	120	120
12.2- Autres maladies génito-urinaires	49	132	99	82	82
13 - Complications de grossesse et d'accouchement	14	0	0	0	1
14- Certains affections dont l'origine est périnatale	0	0	107	24	12
15- Malformations congénitales et anomalies chromosomiques	50	83	75	58	99
16.1- Syndrome de mort subite nourrisson	0	0	0	0	0
16.2- Causes inconnues ou non précisées	0	0	362	180	543
16.3- Autres symptômes et état morbides non définis	0	0	0	0	0
17.1.1 - Accidents de transport	0	0	0	0	14
17.1.2 - Chutes accidentelles	628	539	107	360	808
17.1.3 - Noyade et submersion accidentelle	54	8	7	3	25
17.1.4 - Intoxications accidentelles	708	58	34	24	69
17.1.5 - Autres accidents	2467	740	864	370	493
17.2- Suicide	0	0	0	0	0
17.3 Homicides	213	15	4	3	19
17.4- Evènement dont l'intention n'est pas déterminée	972	36	24	24	12
17.5- Autre cause externe	867	31	22	10	32
18- Covid	0	0	0	0	0
Total	1162	516	905		761
	9	7	7	5702	0

Tableau 17. Effectifs de certificats 2022 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 58 certificats de 2022 dont la cause initiale prédite est Hépatites virales (01.3) et qui ont les 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 17 suivants. La précision globale s'obtient en supposant que les certificats codés automatiquement par Iris /Muse et codés manuellement sont correctement codés. Au final 35 924 certificats sur les 39 165 ne relevaient pas déjà de certificats déjà codés via les autres types de ciblage au moment du tirage.

Le Tableau 17 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 entraîné en septembre 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...

Au total, on simule donc qu'il faut reprendre en 2022 manuellement 31 555 certificats ciblés ainsi pour atteindre au moins 97% de précision dans chacune des catégories de la shortlist européenne. Ce nombre est beaucoup plus élevé que son équivalent en 2021, où il a suffi de reprendre manuellement 19 000 certificats ciblés pour atteindre ce même objectif. L'écart peut être dû au fait que le modèle de score de confiance n'a pas été réestimé lors de la campagne 2021 ou bien à d'autres aspects spécifiques liés à la collecte ou la mortalité de 2022 mais il faudra surveiller cet écart sur les prochaines campagnes.

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

Echantillon	Train			Test		
	Batch	Manuel	Total	Batch	Manuel	Total
ech2011_2015	0	0	2767165			0
ech2016_2017 - codage manuelle	0	298898	298898	0	187056	187056
ech2016_2017 - codage batch	670483	0	670483	11	0	11
ech2018 - codage batch	368604	0	368604			0
ech2018 - ECH001 - EDP	0	4857	4857	0	5010	5010
ech2018 - ECH002 - Deces.sensibles	0	1586	1586	0	1707	1707
ech2018 - ECH003 - IA PR1 - 94%	0	1543	1543	0	1573	1573
ech2018 - ECH004 - IA PR2 - 95%	0	557	557	0	487	487
ech2019 - codage batch	361748	0	361748			0
ech2019 - ECH001 - EDP	0	4834	4834	0	4853	4853
ech2019 - ECH002 - Deces.sensibles	0	1624	1624	0	1622	1622
ech2019 - ECH003 - IA PR1 - 94%	0	1591	1591	0	1611	1611
ech2019 - ECH004 - IA PR2 - 95%	0	890	890	0	877	877
ech2020 - codage manuelle	0	155337	155337	0	121461	121461
ech2020 - codage batch	281657	0	281657	84026	0	84026
ech2021 - codage batch	322077	0	322077	96	0	96
ech2021 - ECH1 aléa	43739	15356	59095	2131	12921	15052
ech2021 - ECH10 IA PR5 - 97%	0	1597	1597	0	1569	1569
ech2021 - ECH2 Faux EDP	14471	4323	18794	662	4706	5368
ech2021 - ECH3 - Deces.sensibles	0	965	965	0	2904	2904
ech2021 - ECH4 - aléa 5% rejets restants	0	1325	1325	0	6031	6031

ech2021 - ECH5 - EDP	0	167	167	15403	8611	24014
ech2021 - ECH6 - IA PR1 – 94%	0	3191	3191	0	3158	3158
ech2021 - ECH7 - IA PR2 – 95%	0	1541	1541	0	1590	1590
ech2021 - ECH8 - IA PR3 – 96%	0	2411	2411	0	2464	2464
ech2021 - ECH9 - IA PR4 – 96.5%	0	1960	1960	0	1932	1932
ech2021 - hors_ech	0	3360	3360	0	3544	3544
ech2021 - hors_ech_a_date	0	7958	7958			0
ech2022 - codage batch	351228	0	351228			0
Total k5_2022 from scratch		5697043			478016	
ech2022 - ECH001 aléa	41649	11616	53265	0	14376	14376
ech2022 - ECH002 - EDP			0	18719	11115	29834
ech2022 - ECH003 - IA PR1 – 94%	0	9659	9659	0	1074	1074
ech2022 - ECH004 - IA PR2 – 95%	0	4266	4266	0	474	474
Total finetuning intermédiaire		5764233			523774	
ech2022 - ECH005 - IA PR3 – 96%	0	7055	7055	0	836	836
ech2022 - ECH007 - Deces.sensibles	0	1809	1809	0	316	316
ech2022 - ECH008 - aléa rejets restants	0	955	955	0	3000	3000
ech2022 - ECH009 - IA PR4 – 96.5%	0	4222	4222	0	519	519
ech2022 - ECH010 - IA PR5 – 97%	0	2867	2867	0	690	690
ech2022 - ECH012 - aléa rejets restants				0	3737	3737
Total finetuning Final		5781141			532872	

	Codage IA en production finale
Campagne 2018	201419
Campagne 2019	211142
Campagne 2021	149260
Campagne 2022	159955

8.6 Définitions détaillées des vérifications et évolutions par rapport à 2021

8.6.1 Evolution des vérifications

8.6.1.1 Cohérence

Réf	Description 2021	Description 2022	Changement pour 2022
Cohérences	Une première série de vérifications consiste à vérifier les cohérences entre les codes	Une première série de vérifications consiste à vérifier les cohérences entre les codes	Aucun changement

de causes initiales et le sexe ou l'âge du défunt

8.6.1.2 Décès spécifiques

Réf	Description 2021	Description 2022	Changement pour 2022
DS1	Les certificats avec une mention de suicide mais non retenu en cause initiale	Les certificats avec une mention de suicide mais non retenu en cause initiale	
DS2	Les certificats avec une mention d'homicide mais non retenu en cause initiale	Les certificats avec une mention d'homicide mais non retenu en cause initiale	
DS3a	Décès d'enfants entre 0 et 27 jours (dont les enfants nés sans vie) non codés dans l'échantillon des décès sensibles	Décès d'enfants entre 0 et 27 jours (dont les enfants nés sans vie) non codés dans l'échantillon des décès sensibles	
DS3b	Age entre 28 jours et 6 mois inclus et mention de code P non codés dans l'échantillon des décès sensibles	Age entre 28 jours et 6 mois inclus et mention de code P non codés dans l'échantillon des décès sensibles	
DS3c		Age > 6 mois et mention de code P non codés dans l'échantillon des décès sensibles	Création pour 2022 : Age > 6 mois et mention de code P non codés dans l'échantillon des décès sensibles
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)	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)	
DS5	Certificats avec mention de SIDA/VIH non codés dans l'échantillon des décès sensibles	Certificats avec mention de SIDA/VIH non codés dans l'échantillon des décès sensibles	

DS6	Certificats de morts maternelles non codés dans l'échantillon des décès sensibles	Certificats de morts maternelles non codés dans l'échantillon des décès sensibles	
DS10	Les certificats ayant une cause initiale de la catégorie intention indéterminée (commençant par Y1 ou Y2 ou Y3[0-4])	Les certificats ayant une cause initiale de la catégorie intention indéterminée (commençant par Y1 ou Y2 ou Y3[0-4])	
DS11		Certains certificats spécifiques avec mention de Covid-19 (cf 8.6.3) (vérification de l'application de règles complexes de codage, encore récentes)	Création 2022 : Regroupement de DS11a et DS11 b avec les changements. Abandon définitif CI U071/U072 avec cancer mentionné en Partie I (comme réalisé sur le codage IA pour les données 2021) / Abandon pour le codage manuel (seul concerné) : CI a été modifiée par un codeur au détriment de U071/U072 / Abandon Démence Alzheimer + U071/U072 avec CI U071/U072.
DS11a	Certains certificats spécifiques avec mention de Covid-19 (cf 8.7) (vérification de l'application de règles complexes de codage, encore récentes)	Certains certificats spécifiques avec mention de Covid-19 (cf 8.7) (vérification de l'application de règles complexes de codage, encore récentes)	abandon pour 2022 au profit de DS11
DS11b	Les certificats ayant une mention de vaccination de Covid-19, post Covid-19, Covid-19 guéri ou de Covid-19 long		abandon pour 2022 au profit de CovidLong_PostCovid et Vaccination
CovidLong_ PostCovid		Les certificats ayant une mention de covid long ou post covid	Création pour 2022 : Division de DS11 Les certificats ayant une mention de covid long ou post covid

		Contenue pour les données 2021 dans Vaccination_PostCovid.
Vaccination	Les certificats ayant une mention de vaccination de Covid-19	Création pour 2022 : Division de DS11. Les certificats ayant une mention de vaccination de Covid-19 Contenue pour les données 2021 dans Vaccination_PostCovid ; non modifiée elle reste prioritaire si un certificat mentionne à la fois un post covid et une vaccination.

8.6.1.3 Difficultés de codage connues liées au logiciel Iris

Réf	Description 2021	Description 2022	Changement pour 2022
M1	Vérification des certificats avec un code D00-D48 en CI alors qu'il y a mention de cancer		Abandon en 2022 au profit de M1a et M1b
M1a		Vérification des certificats avec un code D00- D45 ou D48 en CI alors qu'il y a mention de cancer	D00-D48 -> D00- D45 ou D48 et création des verifs M1b
M1b		Vérification des certificats avec un code D46 ou D47 en CI alors qu'il y a mention de cancer	Tumeur D46 ou D47 alors qu'il existe un code C
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	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	
M3	Causes associées T81-T82 donnant X59 en CI		Abandon en 2022

M4	Code d'affection hématologique (D50-D89) en cause initiale alors qu'il existe un code C sur le certificat	Code d'affection hématologique (D50-D89) en cause initiale alors qu'il existe un code C sur le certificat	
M5	Notion de surdosage AVK en partie 1	Notion de surdosage AVK en partie 1	
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	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. Abandon en 2022 des cas où ces codes barrières étaient sur la dernière ligne de la P1 ou en P2	Abandon en 2022 des cas où ces codes barrières étaient sur la dernière ligne de la P1 ou en P2
M7	Traitement du "ou" si mention de mort violente à côté du "ou"	Traitement du "ou" si mention de mort violente à côté du "ou"	
M8	D611 en CI et absence de cancer dans le certificat	D611 en CI et absence de cancer dans le certificat	
M9	Cause initiale à coder en C80.9 si C34.9 ou un C41.1 en cause associée	Cause initiale à coder en C80.9 si C34.9 ou un C41.1 en cause associée	
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	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	
M11	Anémies en cause initiale (D619, D649 et D539) avec un code informatif présent en partie 1 autre qu'un code d'anémie	Anémies en cause initiale (D619, D649 et D539) avec un code informatif (R%%, Z%%, I46%, I269, I10%, I776, I50) présent en partie 1 autre qu'un code d'anémie	ajout à la liste des codes informatifs (R et Z) : I46%, I269, I10%, I776, I50. La vérification recherche désormais la présence de codes informatifs et anémie uniquement en P1 en excluant les cas où

l'anémie est sur la dernière ligne de la partie 1

8.6.1.4 Vérification de la bonne application des nouvelles règles de 2016

Réf	Description 2021	Description 2022	Changement pour 2022
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	Les hépatites virales ne peuvent plus être dues à d'autres pathologies : les certificats avec mention d'hépatites virales non retenues en cause initiale	
NR4-NR8	Le diabète peut désormais être dû à une liste de pathologies restreintes		Abandon en 2022
NR9	Les accidents (V01-X59) ou chutes W00-W19 doivent être dû à une autre cause	Les accidents (V01-X59) ou chutes W00-W19 doivent être dû à une autre cause	
NR12	Les sepsis et SIRS (A40-A41 et R65) peuvent avoir été dû à (D80-D84 si mentionné)		Abandon en 2022
NR13	Les sepsis et SIRS (A40-A41 et R65) peuvent avoir été dû à (E40-E46 si mentionné)		Abandon en 2022

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

Réf	Description 2021	Description 2022	Changement pour 2022
CC_Hemorragie_inracerebrale	S'il y a une notion de causes externes dans un certificat avec mention d'hémorragie intracérébrale, le codage sera S068 et non I6199	S'il y a une notion de causes externes dans un certificat avec mention d'hémorragie intracérébrale, le codage sera S068 et non I6199	Allègement appliqué au codage IA sur les données 2021 à savoir : Ne doit pas mentionner de chute en Partie I (Plage de codes W00-W19), ni

		un traumatisme en S ou T en partie 1 ni de Y44.2 et Y44.4 en partie 1 et partie 2
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.	Abandon en 2022
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	Abandon en 2022
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)	Abandon en 2022
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	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
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)	Abandon en 2022

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.	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.	Exclusion du libellé IRA anurique
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		Abandon en 2022
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.		Abandon en 2022
CC_noyade	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	Libellé « noyade » et une information de dépression (F30, F32 ou F33) en partie 1 si les circonstances apparentes du décès sont inconnues.	On ne vérifie plus que le code W74 : dépression et noyade en P1 ou P2 sans mention de suicide. Exclusion des cas suivants également à la lecture du tableau puisque la CI n'est pas à modifier : Noyade + dépression MOD=4 et CI= X71 Noyade + dépression MOD=5 et CI=Y21
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	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. En 2022, n'est	pour le codage par l'IA : exclusion des codes I6%

	plus réalisée sur le codage réalisé par les méthodes d'IA	
CC_choc_hemorragique_hypovolemique	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). En 2022 n'est plus réalisée sur le codage manuel	Création 2022 : unique sur les prédictions IA
CC_Rhabdomyolyse	S'il y a une notion de traumatisme ou cause externe (2022 : codes T ou S et Cause initiale différente de X) dans un certificat avec mention de rhabdomyolyse, on vérifie que le code M62.8 est bien approprié.	Création 2022 : Exclusion de Rhabdomyolyse + chutes (W00 à W019) et Inclusion de Rhabdomyolyse + code T ou code S et CI différente de X

8.6.2 Certificats des décès entre 28 jours et 15 ans identifiés dans les décès sensibles ou dans les décès à vérifier

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%

8.6.3 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
-
- F019 + U071/U072 avec Cause initiale U071/U072
- G122 + U071/U072 avec Cause Initiale U071/U072
- G20 + U071/U072 avec Cause initiale U071/U072
- G231 + U071/U072 avec Cause Initiale U071/U072
- G232 + U071/U072 avec Cause Initiale U071/U072
- G318 + U071/U072 avec Cause initiale U071/U072
- G319 + U071/U072 avec Cause initiale U071/U072
- K703 + U071/U072 avec Cause initiale U071/U072
- K704 + U071/U072 avec Cause Initiale U071/U072U049 en Cause initiale
- « 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

Les trois catégories suivantes ont été abandonnées depuis 2021 :

- Cause initiale U071/U072 avec cancer mentionné en Partie I
- Cause initiale a été modifiée par un codeur au détriment de U071/U072
- Démence Alzheimer + U071/U072 avec Cause initiale U071/U072.

8.7 Précisions sur les modèles Transformers

8.7.1 **Entraînement, validation et test**

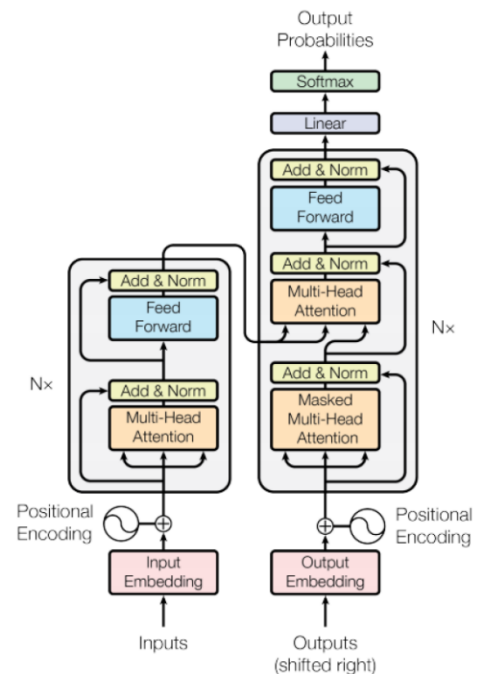
La base de données permettant l'entraînement et l'évaluation des modèles Transformers k4 et k5 se compose des textes des certificats des années passées, mis en entrée des modèles et des codes CIM 10 auxquels ils correspondent (à la fois pour les causes multiples et la cause initiale). Dans la base il y a deux types de textes, les textes bruts, tels qu'ils sont collectés et les textes "nettoyés" après passage d'étapes de standardisation et après traitement manuel. Les modèles sont exclusivement entraînés 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.

La base de test sert à évaluer la performance des modèles après entraînement. Seuls les certificats codés manuellement et provenant de tirages aléatoires sont pris en compte dans la population test de référence (357223 certificats en 2022).

8.7.2 Modèles

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

La couche initiale se compose d'un plongement lexical (embedding) permettant une représentation 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, fonction 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 fonction 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 :: 0.2
epoch :: 100
```

8.7.3 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])
```

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

Un modèle de classification en 5 classes est développé pour optimalement sélectionner la cause initiale entre la prédiction directe de cause initiale par k5, proposition la plus probable (k5top 1) mais aussi proposition la deuxième plus probable (k5top 2) et proposition la troisième plus probable (l5top 3), ainsi que l'application du système expert Iris/Muse sur la séquence de causes multiples prédite par k5 (iris5). Ces 4 propositions diffèrent de ce qui a été réalisé en 2021 où l'on choisissait entre prédictions de deux modèles *transformers* (k5 et k4) partiellement différents (voir rapport de production 2021). L'abandon du modèle k4 a conduit à devoir élargir le spectre de propositions possibles à partir du seul

modèle k5, d'où l'utilisation du top3 des propositions plutôt qu'uniquement la proposition la plus probable.

Ce modèle répond à un problème de classification en 5 classes, désignant parmi les types de proposition précédentes (k5top1, k5top2, k5top3, k5iris ou aucun) celle dont on retiendra la prédiction de cause initiale, et ceci en fonction des caractéristiques des certificats. L'algorithme retenu est un BiLSTM (bidirectionnel long-term short-term memory model, voir Grave et al. 2005, Bladi et al. 1999). D'autres architectures ont été testées (xgboost, fnet) mais n'ont pas conduit à de meilleures performances⁶.

8.8.1 Base d'apprentissage

Le modèle est entraîné sur le même jeu d'entraînement que k5, 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 541 412 certificats, voir tableau ci-dessous. Le jeu de test est le même que celui pour évaluer k5 en 2022.

Tableau 18. Effectifs de la base d'apprentissage

	Train	Test
2016	149 360	93 144
2017	149 538	93 912
2018	8 543	0
2019	8 939	0
2020	155 337	121 461
2021	44 154	23 520
2022	25 541	25 186
Total	541 412	357 223

8.8.2 Problème de classification / variable cible

L'objectif du modèle de classification est de choisir le type de proposition menant à la bonne cause initiale entre k5top1 (proposition directe de cause initiale la plus probable donnée par k5), k5top2 (deuxième plus probable), k5top3 (troisième plus probable), k5Iris (proposition provenant de

⁶ Accuracy sur test BiLSTM .8171, Fnet .8174 (au prix d'une plus grande complexité), xgboost .789

l'application d'Iris/muse sur la séquence de causes prédite par k5) et pas origine, indiquant qu'aucun des types de propositions ne mène à la bonne cause initiale.

Tableau 19. Proportion des types de proposition en base

Origine	Train : 541 412 obs	Test : 357 223 obs
iris	434 271 (80.21%)	281 259 (78.73%)
K5	42 096 (7.78%)	21 821 (6.11%)
K5_top2	24 072 (4.45%)	16 946 (4.74%)
K5_top3	8 439 (1.56%)	6 666 (1.87%)
Pas_orig	32 534 (6.01%)	30 531 (8.55%)

8.8.3 Modèle BiLSTM

On présente tout d'abord l'architecture générale du modèle utilisé. On reviendra ensuite sur ce qui compose les variables explicatives (entrées textuelles et entrées numériques).

8.8.3.1 Architecture du modèle

Il s'agit d'une architecture BiLSTM, spécifiée pour traiter séparément les entrées textuelles et les entrées numériques. La figure suivante montre l'architecture du modèle avec les différentes couches et le nombres de paramètres.

Layer (type)	Output Shape	Param #	Connected to
input_ids (InputLayer)	[(None, None)]	0	[]
embedding (Embedding)	(None, None, 100)	10298100	['input_ids[0][0]']
bidirectional (Bidirectional)	(None, 256)	234496	['embedding[0][0]']
numerical_features (InputLayer)	[(None, 19)]	0	[]
concatenate (Concatenate)	(None, 275)	0	['bidirectional[0][0]', 'numerical_features[0][0]']
dense (Dense)	(None, 32)	8832	['concatenate[0][0]']
dropout (Dropout)	(None, 32)	0	['dense[0][0]']
dense_1 (Dense)	(None, 5)	165	['dropout[0][0]']

=====
 Total params: 10,541,593
 Trainable params: 10,541,593
 Non-trainable params: 0

- ✓ ****Entrée des données textuelles**** : Le modèle prend en entrée une séquence de mots représentée par des identifiants numériques (`input_ids`). Cette séquence est transformée en vecteurs à l'aide d'une couche ****Embedding****, qui convertit chaque mot en une représentation dense de taille 100.

- ✓ ****Traitement de la séquence avec Bi-LSTM**** : La séquence transformée passe ensuite dans une couche ****Bidirectional LSTM****, qui analyse le texte dans les deux directions (de gauche à droite et de droite à gauche). Cela permet de mieux comprendre le contexte global de chaque mot. Cette couche LSTM utilise 256 unités.

- ✓ ****Entrée des variables numériques**** : En parallèle, une autre entrée est dédiée aux variables numériques (`numerical_features`), qui représentent des informations supplémentaires non textuelles comme le nombre de mots dans le certificat ou encore les probabilités de prédiction. Au total il y a 19 variables numériques ce qui explique également la taille de cette couche.

- ✓ ****Fusion des informations**** : Les sorties des deux parties (texte et variables numériques) sont combinées à l'aide d'une couche de ****Concatéation****. Cela permet au modèle d'intégrer et de traiter les informations textuelles et numériques ensemble.

- ✓ ****Régularisation**** : La combinaison passe par une couche dense de 32 unités activée par une fonction ****ReLU**** (Rectified Linear Unit), qui aide à capturer des motifs complexes. Une couche ****Dropout**** avec un taux de 40% est utilisée pour ajouter de la robustesse au modèle.
- ✓ ****Sortie du modèle**** : Enfin, une dernière couche dense avec une activation ****softmax**** est utilisée pour produire la sortie. Cette couche a autant de neurones que de classes (``nb_classes``).

8.8.3.2 *H-params*

Batch_size = 128

Embedding dimension: 100

sequence length: 222

Vocab size: 102 973

nb_classes : 5

nb_num: 19

Dense dim: 32

Dropout: 0.4

Epoch max: 20

8.8.3.3 *Stratégie d'entraînement*

ModelCheckpoint (monitor='val_loss', mode='min', save_best_only=True)

Early_stopping (patience 3, val_loss)

8.8.3.4 *Variables explicatives*

Comme vu au-dessus, les variables textuelles et les variables numériques n'entrent pas au même moment dans le modèle. Cette distinction permet d'exploiter pleinement les variables numériques⁷.

Les **variables textuelles** prises en compte correspondent aux codes CIM de causes initiales proposées par les différentes classes, les causes multiples ainsi que la séquence d'entrée complète du modèle k5 (laquelle concatène les textes médicaux présents sur les certificats à des variables additionnelles d'année, sexe et âge du défunt, type et version du certificat, circonstances apparentes, etc.).

Plus précisément,

keras5_top3 : Cause initiale (CI) prédite parmi les trois meilleures par K5.

⁷ En effet, celles-ci ne passent notamment pas par la phase d'embedding dédiée au texte. Il s'agit d'une amélioration de méthode par rapport à ce qui avait été réalisé sur les données 2021 où tous les features quelle que soit leur nature numérique ou textuelle étaient concaténées en une unique séquence d'entrée.

keras5_top1 : CI prédite en première position par K5.
keras5_top2 : CI prédite en deuxième position par K5.
keras5_iris : CI prédite par le modèle K5 utilisant l'approche Iris.
keras5top3_86postes : CI du top3, au niveau de la shortlist européenne (86 postes).
keras5top1_86postes : CI du top1, au niveau de la shortlist européenne (86 postes).
keras5iris_86postes : CI prédite par K5 Iris au niveau de la shortlist européenne (86 postes).
keras5top2_86postes : CI du top2, au niveau de la shortlist européenne (86 postes).
output_pred_k5 : Causes associées selon la classification CIM10.
Status2 : Statut du codage (final ou rejet).
LieuDecesRetenu2 : Lieu du décès retenu.
inputk5 : séquence d'entrée complète du modèle k5

Les **entrées numériques** captent la complexité du certificat (nb de mots, de causes codées...), les probabilités des propositions et leurs écarts (pouvoir discriminant de k5), des indicateurs d'égalité entre le résultat de l'application d'Iris et une des propositions top3 de k5 ou encore sa présence dans les causes multiples (fiabilité de la proposition).

nb_textes : Nombre de mots présents dans le certificat.
nb_ca : Nombre de codes CIM dans les causes associées.
diff_top12 : Différence entre les probabilités de la CI en top 1 et en top 2.
diff_top13 : Différence entre les probabilités de la CI en top 1 et en top 3.
diff_top23 : Différence entre les probabilités de la CI en top 2 et en top 3.
max_top1 : Probabilité de la CI en top 1.
max_top2 : Probabilité de la CI en top 2.
max_top3 : Probabilité de la CI en top 3.
nb_egal86 : Nombre de CI identiques dans les prédictions sur la shortlist Eurostat (86 catégories).
egal_top1 : Indicateur si la CI prédite en top 1 est identique à la CI Iris.
egal_top2 : Indicateur si la CI prédite en top 2 est identique à la CI Iris.
egal_top3 : Indicateur si la CI prédite en top 3 est identique à la CI Iris.
nb_egal1 : Indicateur si la CI prédite en top 1 est identique à la CI K5 Iris.
nb_egal2 : Indicateur si la CI prédite en top 2 est identique à la CI K5 Iris.
nb_egal3 : Indicateur si la CI prédite en top 3 est identique à la CI K5 Iris.
iris_in_ca : Présence de la CI Iris parmi les causes associées.
top1_in_ca : Présence de la CI prédite en top 1 parmi les causes associées.
top2_in_ca : Présence de la CI prédite en top 2 parmi les causes associées.
top3_in_ca : Présence de la CI prédite en top 3 parmi les causes associées.

8.8.4 Résultats et analyse de performance

8.8.4.1 Performance sur l'origine

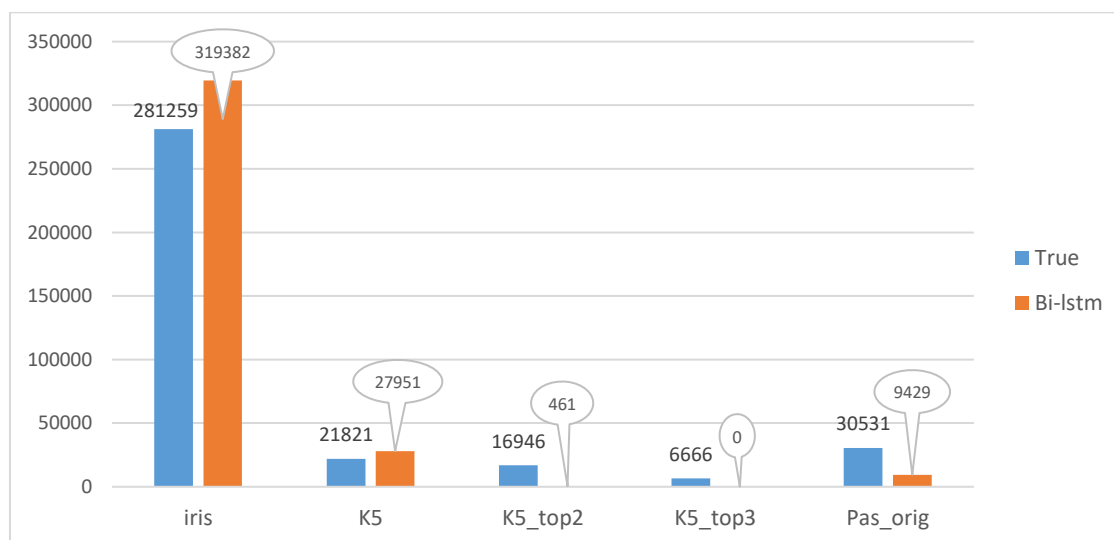
Dans 82,4% des cas (sur le test) le modèle prédit la bonne « origine », c'est-à-dire le type de proposition à retenir pour arriver à la bonne cause initiale prédite. Ce niveau est très proche de celui atteint sur la campagne 2021 (81,6%)

Train	Validation	Test
0.8705	0.8543	0.8244

Cette proportion est évidemment plus élevée sur le *train*, car le modèle a appris (a été estimé) sur ces données. L'écart de performance entre *train* et *test* ou entre *train* et *validation*, d'ailleurs, suggère une présence de léger sur-apprentissage (que l'on avait aussi un peu sur les données 2021 : train 85,9%, validation : 85,2%, test : 81,6%) à surveiller.

En pratique, le surmodèle a tendance à plus souvent prédire la classe origine la plus fréquente (Iris), voire k5. Ceci est certainement lié au caractère très déséquilibré de la répartition en des classes. Notons en particulier que sur la base de test, le surmodèle n'a jamais prédit k5-top3 (il l'a fait en production cependant).

Figure 5. Répartition selon Origine - test 2022 (357 223 obs)



8.8.4.2 Performance sur la cause initiale

Au niveau du code CIM de la cause initiale (c'est-à-dire le code CIM de la cause initiale proposé par la classe/origine retenue), il est correct dans 81,7% des cas (sur le test) pour 81,5% en 2021, et 81,9% sur 2018-2019.

8.9 Dictionnaire de variables dans le SNDS

8.9.1 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 = néonate	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 7=IAIris 8=IA	Uniquement à partir de 2018 Décrit le mode de codage utilisé. Les types de codage 3 à 8 impliquent des algorithmes de <i>deep learning</i> (IA) combinés ou non avec le système expert IRIS/Muse. Les modalités 7 et 8 sont disponibles depuis 2022 et les modalités 3 à 6 sont abandonnées sur 2022
52	Score de confiance	num		Uniquement à partir de 2018 Indicateur de confiance issu de la prédiction impliquant des algorithmes de <i>deep learning</i> permettant de cibler le codage manuel. Il est compris entre 0 et 1.

8.9.2 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 <i>deep learning</i> (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.

8.10 Evaluation de la cohérence des causes multiples avec une campagne de codage à trois modes

L'évaluation des campagnes de codage des causes de décès a jusqu'ici mis l'accent sur la cohérence de la détermination de la cause initiale de décès. Toutefois, avant d'atteindre cet objectif, il est indispensable de transcrire et de coder en CIM-10 chacune des causes mentionnées dans les certificats de décès. Dans ce cadre, l'attention se porte désormais sur une évaluation globale de la transcription intégrale du processus morbide en codes CIM-10, dépassant ainsi la simple analyse de la cause initiale.

Pour répondre à cette exigence, un score de confiance a été élaboré afin de comparer la qualité (ie cohérence) entre codage par IA, et codes réels (issus de la campagne à deux modes). Ce score, conçu pour évaluer une tâche de transformation séquentielle texte-à-texte, complète l'évaluation précédemment réalisée sur la cause initiale et renforce l'évaluation de la campagne à trois modes en y ajoutant une analyse de la traduction complète du processus morbide.

8.10.1 BLEU Score

L'évaluation de la cohérence de la transcription des processus morbides sur les certificats de décès s'est portée sur la métrique BLEU Score, largement utilisée pour évaluer la qualité des prédictions dans des contextes de traitement automatique du langage. Ici, le BLEU Score est calculé pour chaque ligne du certificat en tenant compte des codes CIM-10 associés.

Pour obtenir une mesure à l'échelle du certificat en particulier pour être en mesure de comparer les codages des processus morbides de certificats présentant des nombres de causes très différents ou des structures par ligne différente, on utilise une pondération dynamique des scores en fonction de la position des lignes et des parties du certificat. En effet, le score attribué à la dernière ligne de la Partie I est pondéré plus fortement, car cette ligne contient fréquemment la cause initiale de décès. À l'inverse, les lignes de la Partie II reçoivent un poids plus faible, étant donné que la cause initiale y apparaît rarement.

Le paragraphe suivant détaille la définition et la méthodologie mise en œuvre.

Définition du Score BLEU

Le score BLEU a été introduit par Papineni et al. en 2002 dans l'article "*BLEU: a Method for Automatic Evaluation of Machine Translation*". Il permet de mesurer la précision des n-grams générés par un système de traduction automatique en les comparant à des n-grams de référence, dits "gold".

Le calcul du BLEU se fait en deux étapes :

1 - Précision des n-grams : La précision p_n pour un ordre n des n-grams est donnée par la formule suivante :

$$\text{Précision} = \frac{\text{Nombre total de ngrams correctement prédits}}{\text{Nombre total ngrams (gold)}}$$

2 - Moyenne géométrique des précisions : La moyenne géométrique des précisions p_n est pondérée par des poids w_n , un pour chaque ordre de n-grams. Cette moyenne est donnée par :

$$\text{Moyenne géométrique des précisions} = \prod_{n=1}^N p_n^{w_n}$$

Où N est l'ordre maximal des n-grams, et w_n , est le poids associé à chaque ordre de n-grams.

3 - Facteur de pénalité : Le score BLEU intègre également un facteur de pénalité BP pour corriger les biais induits par des traductions trop courtes. Ce facteur est défini comme suit :

$$BP (\text{Facteur de pénalité}) = \begin{cases} 1 & \text{si } c > r \\ e^{(1-\frac{r}{c})} & \text{si } c \leq r \end{cases}$$

Où c est la longueur de la séquence générée et r est la longueur de la séquence de référence (gold).

Le score BLEU final est donné par la formule :

$$BLEU = \text{Facteur de pénalité} * \text{Moyenne géométrique des précisions}$$

Choix pour l'implémentation du BLEU Score

Dans cette implémentation, nous avons opté pour les paramètres suivants :

- Utilisation de n-grams de taille 2 (utilisation d'uni-grammes et de bi-grammes de mots).
- Poids égaux pour les n-grams de taille 1 et 2, soit $w_1 = w_2 = 0.5$
- Calcul du score BLEU pour chaque ligne d'un certificat, avec une pondération dynamique selon les sections du certificat détaillée par la suite.

Calcul du Score BLEU avec Pondération Dynamique

La pondération dynamique a pour objectif de rendre comparables les bleus scores de certificats ayant des tailles différentes (nombre de causes) ou des structures différentes (partie 1 et partie 2). Elle surpondère la ligne où l'on s'attend à retrouver la cause initiale du décès et au contraire affecte un poids relatif plus faible à la partie 2.

Les lignes d'un certificat sont organisées en deux parties distinctes, chacune dotée d'une pondération spécifique. Un coefficient total de **10** est assigné à l'ensemble du certificat. On choisit une

répartition pondérée et équilibrée des coefficients entre les différentes sections et lignes du certificat, tout en mettant en avant les lignes clés.

- **Première partie (Partie I) :**
- La première partie du certificat comprend jusqu'à 4 lignes. Le poids est réparti proportionnellement entre les lignes, selon leur nombre. La dernière ligne de cette partie se voit attribuer un poids **trois fois supérieur** à celui des autres lignes de la Partie I.
- **Deuxième partie (Partie II) :**
- La deuxième partie, si elle est présente, reçoit une pondération qui correspond à **un tiers** de celle attribuée à la Partie I. Cette pondération réduit donc l'importance relative de cette section par rapport à la première.

Cas 1 : Pas de Partie II

Pour un certificat contenant uniquement la Partie I, le poids des lignes est calculé en fonction du nombre de lignes a de la Partie I. L'équation pour déterminer les poids des lignes est la suivante :

$$10 = (a - 1)p_1 + 3p_1$$

Cela donne la solution pour p_1 (poids de la Partie I) et p_{d1} (poids de la dernière ligne de la Partie I) :

$$p_1 = \frac{10}{a+2}; p_{d1} = \frac{30}{a+2}$$

Cas 2 : Avec Partie II

Dans le cas où une Partie II est présente, l'équation à résoudre devient :

$$10 = (a - 1)p_1 + 3p_1 + \frac{p_1}{3}$$

Les poids sont calculés comme suit avec p_1 : poids de la Partie I, p_{d1} : poids de la dernière ligne de la Partie I et p_2 : poids de la Partie II

$$p_1 = \frac{10}{a + \frac{7}{3}}; p_{d1} = \frac{30}{a + \frac{7}{3}}; p_2 = \frac{10}{3a + 7}$$

Cela permet de pondérer les scores BLEU en fonction du nombre de lignes dans le certificat.

Exemples en pratique

Les exemples suivants illustrent les choix méthodologiques adoptés pour l'évaluation et la pondération des lignes du certificat de décès, tels qu'ils ont été décrits précédemment.

Exemple 1 :

Partie I :

Ligne 1 : traumatisme crânien

Ligne 2 : pneumopathie hypoxiémisante

Ligne 3 : arrêt cardiaque hypoxique

Partie II : cirrhose child b

		Codage du processus morbide dans une campagne traditionnelle	Codage du processus morbide par algorithme d'IA par K5
Partie I	Ligne 1	S069	S069
	Ligne 2	J189	J189
	Ligne 3	I469	I469
Partie II		K746	K746

Ce premier exemple illustre un codage parfait du processus morbide par l'intelligence artificielle. Dans ce cas, les poids attribués aux parties et aux lignes du certificat n'ont aucun impact sur le score final, qui atteint la valeur maximale de 1.

Exemple 2 :

Partie I :

Ligne 1 : dissociation électromécanique

Ligne 2 : hyperkaliémie

Ligne 3 : ischémie aiguë membre inférieur

Ligne 4 : canule cec sur choc cardiogénique

Partie II : cardiopathie ischémique terrain vasculaire insuffisance rénale chronique

		Codage du processus morbide dans une campagne traditionnelle	Codage du processus morbide par algorithme d'IA par K5	Poids attribué dynamiquement	score
Partie I	Ligne 1	I500	I500	1.6	1
	Ligne 2	E875	E875	1.6	1
	Ligne 3	I743	I743	1.6	1
	Ligne 4	Y848	Z998, R570	4.7	0
Partie II		R570, I259, I99, N189	I259, I99, N189	0.5	0.6

Dans ce deuxième exemple, la transcription du processus morbide en codes CIM-10 par l'intelligence artificielle présente deux erreurs :

1 - Première erreur (ligne 4) :

L'algorithme a mal retranscrit le texte en code CIM-10, ce qui a entraîné un BLEU Score de 0 pour cette ligne.

2 - Deuxième erreur (Partie II) :

Le code "R570" a été omis dans la Partie II du certificat, ce qui a donné un BLEU Score de 0,6 pour cette section.

En calculant une moyenne pondérée des scores, le certificat obtient un score global de 0,5. L'erreur de transcription sur la ligne 4 impacte fortement ce résultat, rendant ce certificat éligible à une reprise manuelle.

Cependant, l'algorithme a correctement codé la cause initiale, identifiée comme "I259". Cette métrique, conçue pour compléter l'analyse de cohérence liée à la cause initiale, offre une évaluation plus approfondie. En effet, elle permet d'examiner des certificats qui ne seraient pas forcément détectés comme problématiques par l'évaluation de la cause initiale seule. Cela enrichit le processus global d'analyse.

8.10.2 Population Test de référence

La base de test utilisée pour évaluer la cohérence de la traduction du processus morbide est identique à celle utilisée précédemment pour analyser la cohérence de la cause initiale dans la campagne à trois modes. Cependant, certains certificats ont été exclus parmi les certificats codés manuellement dans la population test de référence, afin de garantir une analyse fondée sur des données cohérentes et comparables. Plus précisément, les observations où le nombre de lignes du "gold standard" ne correspond pas à celui du texte brut ont été exclues. Ces écarts proviennent d'ajouts de lignes par les codeurs pour compléter des informations manquantes (entre le texte saisi par le prestataire et celui rédigé par le médecin) ou de suppressions de lignes pour des mentions non codables, comme "cf. médecin traitant" en première ligne. Après ce filtrage, le nombre d'observations passe de 357 528 à 335 335, garantissant ainsi une base fiable pour l'analyse.

8.10.3 Performance globale

Dans la population test de référence, le Tableau 20 analyse spécifiquement les certificats qui auraient été codés manuellement dans une campagne traditionnelle combinant codage par batch et codage manuel. L'algorithme de deep learning prédit exactement la même séquence de causes que celle codée manuellement dans 56,2% des cas (cf Tableau 3 « BS1 »). Et même lorsqu'il s'en éloigne c'est de peu car le score bleu pondéré moyen est de 85,6 %, ces résultats montent respectivement à 72,1% (cf Tableau 22 « BS1 ») et 91,2% lorsque la reprise manuelle est ajoutée comme lors de la campagne

2022. Ces résultats mettent en évidence une bonne adéquation des codes des causes multiples à celles résultant d'un codage manuel et une amélioration notable grâce à l'intervention humaine. Par ailleurs, les performances globales de l'algorithme montrent une stabilité relative d'une année à l'autre, renforçant la fiabilité du système dans le cadre des campagnes de codage (voir Tableau 20/ tableau 1)

Tableau 20. BLEU score pondéré des causes multiples prédites par deep learning (k5), ainsi que de celles obtenues en combinant les prédictions du modèle avec la reprise manuelle sur les données qui auraient été codées manuellement dans une campagne traditionnelle

Année	Transcription du processus morbide par IA	Transcription du processus morbide par IA + reprise manuelle	Nb obs
ensemble	0,856	0,912	335335
2016	0,85	0,909	81978
2017	0,854	0,912	87007
2020	0,863	0,916	118299
2021	0,852	0,905	23008
2022	0,858	0,907	25043

Sur l'ensemble de la population de test de référence, incluant les certificats codés par batch, la cohérence des traductions des causes multiples en tenant compte de tous les modes de codage, est exacte dans 88,9% des cas (cf Tableau 22 « BS1 »), et le bleu score est en moyenne de 96,5%.

Tableau 21. BLEU score pondéré des causes multiples si on combine les deux modes de codages par deep learning (k5) et batch, ainsi que sur la campagne complète de codage (trois modes de codage)

Année	Transcription du processus morbide par IA + batch	Transcription du processus morbide par IA + batch + reprise manuelle	Nb obs
ensemble	0,943	0,965	843291
2016	0,941	0,964	210641
2017	0,942	0,965	219951
2020	0,943	0,965	282622

2021	0,945	0,965	62546
2022	0,947	0,966	67531

Le Tableau 22 détaille la cohérence de la transcription du processus morbide au sein de la shortlist Eurostat. Cette cohérence est d'abord évaluée sur la proportion des cas qui auraient été codés manuellement dans une campagne classique sur la population test de référence, puis sur l'ensemble de la campagne. Elle est mesurée en moyennant les scores BLEU selon le chapitre Eurostat correspondant à la cause initiale réelle, ce qui peut être interprété comme une mesure de rappel. Deux moyennes sont calculées : la première (notée BS dans le Tableau 22) correspond au score BLEU défini dans la section 1.3 de cette annexe, tandis que la seconde (notée BS1) est calculée uniquement lorsque le score BLEU est égal à 1. Cette distinction est introduite pour tenir compte de notre définition subjective de l'évaluation, qui repose sur des impératifs pratiques liés à la production des données. En ajoutant la moyenne BS1, nous identifions les cas de transcription parfaite par l'algorithme de deep learning, ce qui permet d'évaluer objectivement la séquence des causes associées tout en écartant les choix pratiques liés au processus de production.

Sur l'ensemble de la population test de référence (IA + batch + reprise manuelle), les niveaux de cohérence sont très élevés pour la plupart des catégories : 0,95 pour « BS » (moyenne des scores BLEU) et 0,85 pour « BS1 » (moyenne des scores BLEU lorsque celui-ci est égal à 1).

Le Bleu Score est inférieur à 95% pour 16 catégories (1.1-Tuberculose, 2.1.01-TM lèvre, bouche, pharynx, 2.1.07-TM du Larynx, 2.1.09- Mélanome malin de la peau, 2.1.11-TM du col de l'utérus, 2.1.17-TM cerveau et système nerveux central, 2.1.18-TM de la thyroïde, 2.1.19-Maladie de Hodgkin et lymphomes, 2.1.20- Leucémie, 2.1.22-Autres Tumeurs malignes, 12.2- Autres maladies génito-urinaires, 17.1.1- Accidents de transport, 17.1.5 - Autres accidents, 17.2- Suicide, 17.3- Homicides, 17.4- Evènement dont l'intention n'est pas déterminée, 17.5- Autre cause externe).

La cohérence entre les causes multiples n'est parfaite que dans moins de 85% des cas pour 11 catégories (1.1-Tuberculose, 2.1.01-TM lèvre, bouche, pharynx, 2.1.07-TM du Larynx, 2.1.09- Mélanome malin de la peau, 2.1.11-TM du col de l'utérus, 2.1.18-TM de la thyroïde, 2.1.19-Maladie de Hodgkin et lymphomes, 2.1.20- Leucémie, 2.1.22-Autres Tumeurs malignes, 12.2- Autres maladies génito-urinaires, 17.1.1- Accidents de transport, 17.5- Autre cause externe).

Tableau 22. Cohérence des causes multiples prédites au niveau de la shortlist, sur la population de test de référence qui aurait été codée manuellement puis sur la campagne complète

Libellé	Part sur la population de test de référence qui aurait été codée manuellement dans une campagne classique de codage				
	Nb Obs	Transcription du processus morbide par IA		Transcription du processus morbide par IA + reprise manuelle	
		BS	BS1	BS	BS1
01.1- Tuberculose	421	0,802	0,470	0,900	0,774
01.2- SIDA (maladie VIH)	276	0,832	0,478	1,000	1,000
01.3- Hépatites virales	324	0,798	0,420	0,923	0,806
01.4- Autres maladies infectieuses et parasitaires	5786	0,844	0,539	0,923	0,791
02.1.01-TM lèvre, bouche, pharynx	2765	0,806	0,492	0,858	0,616
02.1.02-TM de l'œsophage	2478	0,884	0,629	0,910	0,708
02.1.03-TM de l'estomac	2366	0,865	0,581	0,897	0,677
02.1.04-TM Colon, rectum, anus	9877	0,871	0,583	0,901	0,668
02.1.05-TM du foie et voies biliaires	4816	0,876	0,626	0,907	0,707
02.1.06-TM du Pancréas	5488	0,885	0,634	0,917	0,722
02.1.07-TM du Larynx	633	0,820	0,491	0,887	0,684
02.1.08-TM du Poumon	15987	0,884	0,607	0,912	0,693
02.1.09- Mélanome malin de la peau	1199	0,871	0,615	0,893	0,684
02.1.10-TM du sein	6952	0,889	0,603	0,912	0,682
02.1.11-TM du col de l'utérus	529	0,851	0,558	0,895	0,669
02.1.12-TM autre partie de l'utérus	1677	0,855	0,574	0,900	0,682
02.1.13-TM de l'ovaire	1771	0,874	0,608	0,904	0,689
02.1.14-TM de la prostate	4859	0,885	0,574	0,911	0,664
02.1.15-TM du rein	2146	0,862	0,568	0,897	0,671
02.1.16-TM de la vessie	2983	0,862	0,550	0,893	0,644
02.1.17-TM cerveau et système nerveux central	2238	0,801	0,562	0,860	0,681
02.1.18-TM de la thyroïde	268	0,838	0,549	0,882	0,675
02.1.19-Maladie de Hodgkin et lymphomes	3294	0,813	0,490	0,863	0,619
02.1.20- Leucémie	3630	0,844	0,540	0,884	0,651

Population test de référence complète, campagne qui combine les trois modes de codage				
nb Obs	Transcription du processus morbide par IA + batch		Transcription du processus morbide par IA + batch + reprise manuelle	
	BS	BS1	BS	BS1
474	0,824	0,530	0,911	0,800
351	0,868	0,590	1,000	1,000
564	0,884	0,667	0,956	0,888
13715	0,934	0,805	0,967	0,912
5190	0,897	0,729	0,925	0,795
5076	0,943	0,819	0,956	0,858
6075	0,947	0,837	0,960	0,874
24276	0,948	0,830	0,960	0,865
12059	0,951	0,851	0,963	0,883
16470	0,962	0,878	0,972	0,907
1293	0,912	0,751	0,945	0,845
42838	0,957	0,853	0,967	0,885
2366	0,935	0,805	0,946	0,840
17608	0,956	0,843	0,965	0,874
1110	0,929	0,789	0,950	0,842
3862	0,937	0,815	0,957	0,862
4674	0,952	0,852	0,964	0,882
12557	0,956	0,835	0,965	0,870
4807	0,939	0,807	0,954	0,853
7240	0,943	0,815	0,956	0,853
5563	0,920	0,824	0,944	0,872
515	0,916	0,765	0,939	0,831
6715	0,908	0,750	0,933	0,813
8263	0,931	0,798	0,949	0,847

02.1.21-Autres TM des tissus lymphoïdes et hématopoïétiques	1990	0,868	0,567	0,901	0,670	4530	0,942	0,810	0,957	0,855
02.1.22-Autres Tumeurs malignes	15065	0,795	0,509	0,874	0,705	30658	0,899	0,759	0,938	0,855
02.2- Tumeurs non malignes	5288	0,820	0,498	0,907	0,740	10676	0,911	0,751	0,954	0,871
03- Maladies du sang et hématopoïétiques	2081	0,827	0,504	0,935	0,827	3695	0,903	0,720	0,964	0,903
04.1- Diabète sucre	7324	0,878	0,521	0,916	0,662	16733	0,946	0,790	0,963	0,852
04.2- Autres maladies endocriniennes, nutritionnelles et métaboliques	6089	0,886	0,578	0,939	0,767	14696	0,953	0,825	0,975	0,904
05.1- Démences	8337	0,913	0,635	0,935	0,723	26643	0,973	0,886	0,980	0,913
05.2- Abus d'alcool	1550	0,871	0,575	0,945	0,823	3435	0,942	0,808	0,975	0,920
05.3- Pharmacodépendance, toxicomanie	203	0,793	0,453	0,963	0,901	320	0,869	0,653	0,976	0,938
05.4- Autres troubles mentaux et du comportement	2542	0,854	0,526	0,921	0,753	5193	0,929	0,768	0,961	0,879
06.1- Maladie de Parkinson	2935	0,911	0,615	0,934	0,706	9415	0,972	0,880	0,979	0,908
06.2 - Maladie d'Alzheimer	7964	0,903	0,630	0,926	0,713	26894	0,971	0,890	0,978	0,915
06.3- Autres maladies du système nerveux et des organes des sens	7484	0,849	0,541	0,915	0,739	16548	0,932	0,792	0,962	0,882
07.1.1-Infarctus aigu du myocarde	6360	0,889	0,621	0,918	0,715	18919	0,963	0,872	0,973	0,904
07.1.2-Autres Cardiopathies ischémiques	11157	0,880	0,539	0,919	0,677	25719	0,948	0,800	0,965	0,860
07.2-Autres maladies du cœur	23560	0,871	0,574	0,919	0,724	71053	0,957	0,859	0,973	0,909
07.3-Maladies cérébraux vasculaires	19058	0,877	0,597	0,918	0,722	43648	0,946	0,824	0,964	0,879
07.4- Autres maladies de l'appareil circulatoire	14905	0,849	0,521	0,917	0,724	34588	0,935	0,793	0,964	0,881
08.1- Grippe	768	0,860	0,529	0,909	0,689	1802	0,940	0,799	0,961	0,867
08.2- Pneumonie	4566	0,890	0,653	0,931	0,775	17327	0,971	0,909	0,982	0,941
08.3.1- Asthme	430	0,858	0,509	0,911	0,705	1126	0,946	0,813	0,966	0,887
08.3.2-Autres maladies chroniques des voies respiratoires inférieures	5719	0,877	0,525	0,914	0,666	13784	0,949	0,803	0,964	0,861
08.4- Autres maladies de l'appareil respiratoire	6658	0,852	0,572	0,914	0,760	22402	0,956	0,873	0,975	0,929
09.1 - Ulcère gastro duodéal	602	0,830	0,472	0,912	0,736	1136	0,910	0,720	0,953	0,860

09.2 - Cirrhoses, fibroses et hépatites chroniques	4135	0,883	0,586	0,919	0,710
09.3- Autres maladies de l'appareil digestif	11301	0,835	0,501	0,907	0,710
10- Maladies de la peau et du tissu sous-cutané	1209	0,826	0,515	0,927	0,798
11.1- Arthrite rhumatoïde et ostéoarthrite	454	0,848	0,498	0,928	0,784
11.2- Autres maladies du système ostéoarticulaire et des muscles	3016	0,761	0,366	0,928	0,797
12.1-Maladies du rein et de l'urètre	4487	0,833	0,524	0,911	0,747
12.2- Autres maladies génito-urinaires	2300	0,787	0,456	0,896	0,717
13- Complications de grossesse et d'accouchement	54	0,446	0,111	1,000	1,000
14- Certains affections dont l'origine est périnatale	1704	0,676	0,322	1,000	1,000
15- Malformations congénitales et anomalies chromosomiques	1372	0,693	0,319	0,935	0,832
16.1- Syndrome de mort subite nourrisson	173	0,973	0,902	0,997	0,983
16.2- Causes inconnues ou non précisées	4845	0,870	0,710	0,907	0,792
16.3- Autres symptômes et état morbides non définis	6464	0,917	0,788	0,939	0,840
17.1.1- Accidents de transport	2108	0,742	0,452	0,819	0,610
17.1.2 - Chutes accidentelles	8945	0,878	0,555	0,947	0,809
17.1.3 - Noyade et submersion accidentelle	375	0,741	0,496	0,875	0,771
17.1.4 - Intoxications accidentelles	1670	0,771	0,451	0,947	0,884
17.1.5 - Autres accidents	13431	0,827	0,560	0,925	0,810
17.2- Suicide	4439	0,740	0,499	0,807	0,624
17.3- Homicides	338	0,582	0,287	0,910	0,834
17.4- Evènement dont l'intention n'est pas déterminée	1366	0,636	0,366	0,915	0,859
17.5- Autre cause externe	1540	0,731	0,406	0,909	0,799
18- Covid	14211	0,877	0,548	0,903	0,642
Total	335335	0,856	0,562	0,912	0,721

9516	0,949	0,820	0,965	0,874
23371	0,920	0,759	0,955	0,860
2184	0,903	0,732	0,959	0,888
775	0,911	0,706	0,958	0,874
4577	0,843	0,582	0,953	0,866
11264	0,934	0,810	0,964	0,899
4202	0,883	0,702	0,943	0,845
57	0,475	0,158	1,000	1,000
2021	0,727	0,428	1,000	1,000
2154	0,805	0,566	0,959	0,893
178	0,974	0,904	0,997	0,983
22036	0,971	0,936	0,980	0,954
43721	0,988	0,969	0,991	0,976
3613	0,850	0,680	0,894	0,772
11673	0,906	0,659	0,959	0,854
1119	0,913	0,831	0,958	0,923
2237	0,829	0,590	0,961	0,914
19117	0,878	0,691	0,947	0,866
11169	0,897	0,801	0,923	0,851
467	0,697	0,484	0,935	0,880
1720	0,711	0,497	0,933	0,888
1837	0,775	0,502	0,924	0,831
39682	0,956	0,838	0,965	0,872
843291	0,943	0,826	0,965	0,889

8.11 Constitution de la population test de référence

Cette annexe décrit la méthodologie et la logique de la création de la population test de référence utilisée dans la partie 4 Evaluation de la stratégie de codage.

Rappel de la formation de la population test de référence

Pour évaluer la stratégie de reprise manuelle ciblée constitutive d'une campagne à 3 modes, il nous faut une base de données :

- (1) Constituée de certificats de décès codés manuellement et non-utilisés pour entraîner les modèles d'intelligence artificielle (IA) (principe méthodologique).
- (2) Ou ayant été codé par système automatique IRIS/Muse (qu'importe s'ils ont été utilisés pour entraîner l'IA, car les codes batch ne sont pas comparés aux codes IA).
- (3) Respectant la répartition entre codage automatique et manuel dans le cadre d'une campagne traditionnelle pour chaque année (pour conserver la répartition des causes de décès de chaque année, en prenant l'hypothèse que la distribution des causes de décès est relativement stable sur les années utilisées).
- (4) Respectant d'autres distributions : la part des décès sensibles⁸ et des défunts EDP⁹ observées dans la population des décès, et la proportion de reprise manuelle ciblée ou aléatoire utilisée lors de la production.

Pour l'évaluation de la campagne de 2021, une population de référence avait déjà été constituée sur les données 2016-2017-2020-2021¹⁰. Pour l'évaluation de la campagne de 2022, cette population de référence de 2021 a été reprise et complétée avec des certificats de 2022. La population

⁸ 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 campagne 2022. En résulte de nouveau, une légère sous-estimation de la performance de la campagne 2022 dans les tableaux présentés.

⁹ Le champ « EDP » fait référence à la base de données Echantillon Démographique Permanent (EDP) produite par l'Insee, dont le champ est défini tel que suit : Toutes les personnes nées « un des 16 jours suivants : du 2 et au 5 janvier, du 1er au 4 avril, du 1er au 4 juillet, ou du 1er au 4 octobre » - « L'échantillon démographique permanent : en 50 ans, l'EDP a bien grandi ! », Isabelle Robert-Bobée, Natacha Gualbert, [Courrier des statistiques](#) (6, 2021)

¹⁰ Voir « 4. Evaluation de la méthode de codage », [Rapport de Production 2021](#) et « Annexe 5 - Compléments sur l'évaluation de la stratégie de reprise manuelle ciblée », [Codage des causes de décès de 2018 et 2019 en CIM10 - Approche combinant deep learning, système expert et codage manuel ciblé](#)

test de référence n'est donc pas recalculée pour chaque nouvelle campagne mais on y ajoute des certificats supplémentaires pour chaque nouvelle année¹¹.

(1) (2) Il est très souvent requis par la méthodologie des modèles d'IA supervisés de laisser de côté une partie importante des données à disposition (annotées, c'est-à-dire déjà codées) pour permettre l'évaluation finale d'un modèle « en situation réelle », face à des certificats inconnus et pour lesquels nous disposons d'un codage jugé exact à partir duquel nous pouvons évaluer la correspondance avec le codage imputé par IA. Le Tableau 23 offre à voir l'ensemble et les caractéristiques des certificats disponibles pour évaluer le modèle de l'année 2022.

Les années 2016-2017-2020 sont des années de campagne « classique », c'est-à-dire que seul le codage automatique et le codage manuel ont été utilisés. Leur segmentation entre *train/test* a été réalisé pendant la campagne de 2018-2019 où, en raison d'un impératif de rattrapage du calendrier de la production des données sur les causes de décès après l'épidémie de COVID-19, la quasi-totalité des certificats ont été codés via le système automatique IRIS/Muse ou par un modèle d'intelligence artificielle. Il n'y a donc pas de certificats de 2018-2019 disponibles pour composer la population de référence.

A partir de 2021, la stratégie de production à trois modes (IRIS, manuel, IA) implique de fine-tuner et de tester le modèle d'IA sur des certificats de l'année en cours de production. Des échantillons de certificats, aléatoires ou centrés sur des codes ou situations particulières, sont prélevés parmi les certificats de l'année en cours de production pour être rajoutés à la base d'entraînement et à celle de *test*. Pour les besoins d'évaluation, seuls les certificats issus d'échantillons aléatoires ou représentatifs de 2021-2022 seront ajoutés à la population test de référence, car ce sont les seuls qui nous assurent respecter la distribution des causes de décès pour ces deux années, lorsque complétés par des certificats codés par batch automatique dans la proportion adéquate.

¹¹ Et les certificats des années précédentes seront éventuellement retirés pour que la distribution globale des causes de décès de la population test de référence soit aussi proche que possible de celle de la campagne évaluée

Tableau 23. Données disponibles pour composer la population de référence - incluant segmentation test/train

Année	Echantillon	Test		Train	
		Batch Automatique Iris	Manuel	Batch Automatique Iris	Manuel
2016	ech2016_2017 - auto	-	-	335 644	-
2016	ech2016_2017	-	93 144	-	149 841
2017	ech2016_2017 - auto	11	-	345 478	-
2017	ech2016_2017	-	93 912	-	150 143
2020	ech2020 - auto	84 026	-	291 795	-
2020	ech2020	-	121 461	-	156 331
2021	ech2021 - auto	96	-	329 925	-
2021	ech2021 - ECH1 - aléa	2 131	12 921	44 858	15 529
2021	ech2021 - ECH2 - décès jour edp	662	4 706	14 785	4 374
2021	ech2021 - ECH4 - aléa manuel hors ECH1/ECH5	-	6 031	-	1 335
2022	ech2022 - auto	-	-	358 941	-
2022	ech2022- ECH001 - aléa	-	14 376	42 541	11 762
2022	ech2022- ECH002 - edp	18 719	11 115	-	-
Total		105 645	357 666	1 763 967	489 315

Source : CépiDc – Inserm.

Données : Base de données sur les causes médicales de décès en France - certificats codés automatiquement par Iris et échantillons à tirages aléatoires ou représentatifs

Lecture : Lors de l'année 2022, deux échantillons aléatoires ou représentatifs ont été tirés pendant la production. L'échantillon (1), tiré aléatoirement parmi tous les certificats de décès, dont 14 376 certificats codés manuellement et non-utilisés pour entraîner l'IA, et l'échantillon (2) composé de la totalité des certificats relevant du champ EDP.

(3) Nous avons donc accès au total à 357 696 certificats codés manuellement, auxquels nous allons associer la part de certificats codés automatiquement égale à celle observée pour chacune des

campagnes utilisées. Le Tableau 24 rappelle la part de certificats codés par batch observés chaque année en début de campagne et la part de codage manuel.

Tableau 24. Répartition des modes de codages souhaitée pour la population test de référence

Année	% codage automatique par batch Iris observé pendant la campagne	% codage manuel recherché pour la population test de référence
2016	58,0	42,0
2017	58,6	41,4
2020	57,5	42,5
2021	62,4	37,6
2022	62,3	37,7

Lecture : En début de campagne de 2016, la part de codage automatique était de 58%. Donc, à l'intérieur de la population test de référence, pour les certificats de 2016, on va donc chercher à obtenir 42% de codés manuellement et 58% de codés automatiquement.

Dit autrement, pour 2016, les 93 144 certificats codés manuellement devront représenter 42% des certificats. On peut donc déduire (par règle de trois) que le nombre de certificats codés automatiquement sera autour de $\frac{9\,314\,400}{42} - 93\,144 \approx 128\,627$ certificats.

La répartition des modes de codage de la population test de référence est donc égale à la réalité de 2016-2017-2020 (campagnes dites « classiques », sans IA) mais est différente pour 2021-2022 (le mode « IA » prenant une partie du codage). On pioche donc nos effectifs de certificats codés manuellement ou automatiquement dans les données du Tableau 23 de telle manière à obtenir la distribution du Tableau 24.

Méthodologie de tirage des données pour chaque année

Cette partie détaille de façon fine l'origine des données de la population test de référence, le Tableau 25 offre plus bas les effectifs de la base de test finale, détaillés pour les années 2021-2022.

Années sans codage IA : 2016-2017-2020¹²

¹² Rappel : la quasi-totale absence de codage manuel pour les années 2018-2019 (rattrapage) rend impossible l'intégration de données de ces années dans la population test de référence.

Les données de base à disposition pour intégrer la population test de référence sont les données codées manuellement présentes dans la base de *test* des modèles d'IA. Ces certificats vont ensuite être complétés par des certificats codés par batch automatique, sans qu'ils aient besoin d'être issus d'une base de test¹³.

2016 : On a 93 144 certificats codés manuellement, tirés aléatoirement pour faire partie de la base de test. Ils composent donc 38,33% des données manuelles à disposition. Nous les récupérons tous et tirons dans la même proportion les données codées automatiquement (335 644 x 38,33% ≈ 128 663 certificats codés automatiquement). Nous obtenons bien, pour l'année 2016, la répartition escomptée (entre codage manuel et codage automatique) dans la population test de référence.

2017 : On a 93 912 certificats codés manuellement, tirés aléatoirement pour faire partie de la base de test. Ils composent donc 38,48% des données manuelles à disposition. Nous les récupérons tous et tirons dans la même proportion les données codées automatiquement (345 489 x 38,48% ≈ 132 944 certificats codés automatiquement). Nous obtenons bien, pour l'année 2017, la répartition escomptée (entre codage manuel et codage automatique) dans la population test de référence.

2020 : On a 121 461 certificats codés manuellement, tirés aléatoirement pour faire partie de la base de test. Ils composent donc 43,72% des données manuelles à disposition. Nous les récupérons tous et tirons dans la même proportion les données codées automatiquement (335 644 x 43,72% ≈ 164 323 certificats codés automatiquement). Nous obtenons bien, pour l'année 2020, la répartition escomptée (entre codage manuel et codage automatique) dans la population test de référence.

Années avec codage IA : 2021-2022

Les années 2021-2022 n'ont pas eu un codage manuel exhaustif mais ont fait l'objet d'une stratégie de ciblage du codage manuel. L'idée est d'accélérer la production de la base des causes médicales de décès en mobilisant un modèle d'intelligence artificielle pour coder une partie des données qui était auparavant codées manuellement. Désormais, les données soumises au codage manuel (appelés « échantillons de reprise manuelle ») sont sélectionnées soit aléatoirement soit dans un but d'optimisation de l'efficacité de l'ensemble du codage. C'est à partir des tirages aléatoires (plus précisément de leur partie *manuel ET test*) que nous pouvons compléter la population test de référence pour les années 2021-2022.

2021 : Trois¹⁴ tirages aléatoires ont été réalisés pendant la campagne 2021. Les données de ces tirages pouvaient être déjà codées automatiquement ou ont été codées manuellement. Parmi ces tirages, une partie *test* a été isolée à des fins d'évaluation .

¹³ Les certificats codés par système expert IRIS/Muse ne sont jamais comparés à des certificats codés par l'IA, il n'y a donc aucune contrainte méthodologique particulière sur les certificats codés par batch automatique (si ce n'est d'avoir été tirés aléatoirement dans la bonne année).

¹⁴ Plus précisément 4 tirages aléatoires ont été réalisés pendant la campagne 2021 : le quatrième est constitué des certificats de décès des personnes *nées* un jour EDP – au moment de la constitution de la

Echantillon 1 (tirage aléatoire de 100 lots) : 14 376 certificats sont codés manuellement et appartiennent à la base de test. Ils composent donc 45,41% des données manuelles à disposition. Nous les récupérons tous et tirons dans la même proportion les données codées automatiquement du même échantillon ($46\,989 \times 45,41\% \approx 21\,341$ certificats codés automatiques).

Echantillon 2 (décès un jour EDP¹⁵) : 4 706 certificats sont codés manuellement et appartiennent à la base de test. Ils composent donc 51,83% des données manuelles à disposition. Nous les récupérons tous et tirons dans la même proportion les données codées automatiquement du même échantillon ($15\,447 \times 45,41\% \approx 8\,006$ certificats codés automatiques).

Echantillon 4 (tirage aléatoire de 7500 certificats non-codés-automatiquement hors échantillon 1 et hors échantillon 2) : parmi eux 6 139 certificats appartiennent à la base de test. Ils composent donc 82,31% des données manuelles à disposition. Nous les récupérons tous. Cependant nous ne pouvons pas tirer dans le même échantillon les données codées automatiquement car il n'y en a pas. Il nous faut donc constituer la part de données codées automatiquement à partir des certificats codés par batch hors échantillons 1 et 2 (pour éviter les doublons). Si le tirage avait été fait en tenant en compte les données codées automatiquement, on aurait pioché la proportion de codage automatique du début de campagne, soit $\frac{\text{effectifs codage automatique}}{\text{effectifs hors codage automatique}} \times 6\,139 = \frac{407\,857}{245\,689} \times 6\,139 \approx 10\,191$ certificats. Ces 10 191 certificats sont donc tirés parmi le batch automatique de 2021 hors les échantillons 1 et 2.

En sommant nous obtenons bien, pour l'année 2021, la répartition escomptée (entre codage manuel et codage automatique) dans la population test de référence.

2022 : Deux tirages aléatoires ont été réalisés pendant la campagne 2022. Les données de ces tirages pouvaient être déjà codées automatiquement ou ont été codées manuellement. Parmi ces tirages, une partie *test* a été isolée pour contribuer à l'amélioration de la qualité de l'IA.

Echantillon 1 (tirage aléatoire de 100 lots) : Parmi 26 138 certificats codés manuellement, 14 376 appartiennent à la base de test. Comme pour l'échantillon 4 de 2021, on va venir compléter l'échantillon avec dans les certificats codés par batch¹⁶ hors échantillon 2 dans la bonne proportion (c'est-à-dire la proportion de codage automatique du début de campagne), soit $\frac{\text{effectifs codage automatique}}{\text{effectifs hors codage automatique}} \times 14\,376 = \frac{420\,201}{254\,147} \times 14\,376 \approx 23\,769$ certificats. Ces 23 769

population test de référence, cet échantillon n'était pas disponible. Il n'avait donc pas été ajouté alors même qu'il répond aux critères pour y appartenir.

¹⁵ Cet échantillon avait été décrit par erreur dans le rapport *Codage des causes de décès de 2018 et 2019 en CIM10 Approche combinant deep learning, système expert et codage manuel ciblé* comme « Il s'agit des décès ayant eu lieu en début de trimestre ».

¹⁶ On aurait pu tirer spécifiquement dans la partie automatique de l'échantillon 1 (comme pour les échantillons 1 et 2 de 2021) : cela revient à la même chose c

certificats sont donc tirés parmi le batch automatique de 2022 en dehors de l'échantillon 2¹⁷. A la suite de ce tirage aléatoire, 2 580 appartiennent à la partie automatique de l'échantillon 1.

Echantillon 2 (né un jour EDP) : Parmi les 11 115 certificats codés manuellement, l'entièreté appartiennent à la base de test. De même, les 18 719 certificats codés automatiquement de l'échantillon forment déjà le rapport adapté à l'année 2022 (37,3% manuel / 62,7% automatique). L'ensemble des certificats de cet échantillon est donc ajouté à la population test de référence.

En sommant nous obtenons bien, pour l'année 2022, la répartition escomptée (entre codage manuel et codage automatique) dans la population test de référence.

Au total, la population test de référence est composée de 865 484 certificats, dont 58,6% sont codés par batch automatique Iris.

¹⁷ Les certificats de l'échantillon 2 appartenant déjà à un échantillon, on risquerait d'introduire des doublons dans la population test de référence. On préfère donc les exclure.

Tableau 25. Répartition des types de codage dans la population test de référence

Année	Echantillon	Test		Train	Total manuel annuel	Total automatique annuel	Total
		Batch Automatique Iris	Manuel	Batch Automatique Iris			
2016	ech2016_2017 - auto	-	-	128 663	93 144	128 663	221 807
	ech2016_2017	-	93 144	-			
2017	ech2016_2017 - auto	3	-	132 941	93 912	132 944	226 856
	ech2016_2017	-	93 912	-			
2020	ech2020 - auto	36 452	-	127 871	121 461	164 323	285 784
	ech2020	-	121 461	-			
2021*	ech2021 - auto	2	-	9 760	23 520	39 538	63 058
	ech2021 - ECH1 - aléa	991	12 866	20 350			
	ech2021 - ECH2 - décès jour edp	348	4 685	7 658			
	ech2021 - ECH4 - aléa manuel hors ECH1/ECH5	-	5 969	-			
	ech2021 - ECH5 - edp	429	-	-			
2022	ech2022- auto	-	-	21 308	25 491	42 488	67 979
	ech2022- ECH001 - aléa	-	14 376	2 461			
	ech2022- ECH002 - edp	18 719	11 115	-			
Total		56 944	357 528	451 012	357 528	507 956	865 484

* Les chiffres présentés pour 2021 ont une différence 146 observations par rapport au rapport de production de 2021 : ces certificats ont été supprimés lors de la synchronisation des certificats avec l'Insee, on les retire donc également des bases de test.
 Source : CépiDc – Population test de référence 2022
 Données : Population test de référence 2022
 Lecture : Dans la base finale de test, on trouve 25 491 certificats codés manuellement issus de l'année 2022. 14 376 d'entre eux sont issus d'un échantillon tiré entièrement aléatoirement en tout début de production et 11 115 d'un échantillon représentatif de la population (naissance un jour EDP).

8.12 Liste des tableaux

Tableau 1. Exhaustivité de la collecte des volets médicaux (VM) pour les décès 2022 et 2021.....	5
Tableau 2. Répartition des modes de codage des données 2022 hors volets médicaux non reçus.....	10
Tableau 3. Nombre de certificats repris manuellement après un ciblage IA lors de la campagne de codage 2022 (Chiffres avant synchro)	11
Tableau 4. Répartition du codage manuel lors de la campagne 2022 (Effectifs après synchro).....	13
Tableau 5. Description et effectifs des décès identifiés comme sensibles en 2022 hors volets médicaux non reçus et avant synchro Insee	14
Tableau 6. Description et nombre de certificats vérifiés pour les décès spécifiques et part de cause initiale modifiées (Chiffres avant Synchro).....	20
Tableau 7. Description des vérifications liées au système expert et nombre total de certificats associés et part de cause initiale modifiées	21
Tableau 8. Description des vérifications liées aux nouvelles règles et nombre total de certificats associés et part de cause initiale modifiées.	22
Tableau 9. Description des vérifications liées à la démarche des choix de code et nombre total de certificats associés et part des causes initiales modifiées.....	23
Tableau 10. Description de la distribution des allocations de codage dans la base 2022 et sa simulation dans la population test de référence	25
Tableau 11. Cohérence (accuracy) entre causes initiales prédites par deep learning (k5), combinaison de deep learning et Iris Muse (k5 iris), proposition du surmodèle, avec ou sans prise en compte de la reprise manuelle 2022 pour les données qui auraient été codées manuellement.....	27
Tableau 12. Cohérence entre causes initiales prédites par deep learning (k5), combinaison de deep learning et Iris Muse (k5 iris), proposition du surmodèle, avec ou sans prise en compte de la reprise manuelle 2022 pour l'ensemble de la population (et 2021 pour référence)	28
Tableau 13. Performances et effectifs prédits du sur-modèle et du sur-modèle combiné avec la reprise manuelle évalués sur les observations de la population test qui auraient été codées manuellement.	30
Tableau 14. Performances et effectifs prédits du batch combiné au sur-modèle et du batch combiné au sur-modèle combiné et à la reprise manuelle évaluées sur l'ensemble de la population test de référence.....	32
Tableau 15. Performances en termes de cohérence de chapitre de la CIM de l'ensemble de la campagne 2022 et effectifs prédits sur l'ensemble de la population test de référence.....	35
Tableau 16. 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 2022	37
Tableau 17. Effectifs de certificats 2022 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).	49
Tableau 18. Effectifs de la base d'apprentissage	67
Tableau 19. Proportion des types de proposition en base	68
Tableau 20. BLEU score pondéré des causes multiples prédites par deep learning (k5), ainsi que de celles obtenues en combinant les prédictions du modèle avec la reprise manuelle sur les données qui auraient été codées manuellement dans une campagne traditionnelle	86
Tableau 21. BLEU score pondéré des causes multiples si on combine les deux modes de codages par deep learning (k5) et batch, ainsi que sur la campagne complète de codage (trois modes de codage).....	86
Tableau 22. Cohérence des causes multiples prédites au niveau de la shortlist, sur la population de test de référence qui aurait été codée manuellement puis sur la campagne complète	87
Tableau 23. Données disponibles pour composer la population de référence - incluant segmentation test/train	93
Tableau 24. Répartition des modes de codages souhaitée pour la population test de référence.....	94
Tableau 25. Répartition des types de codage dans la population test de référence.....	98

8.13 Liste des figures

<i>Figure 1. Répartitions des décès manquants et ensemble par groupe d'âge.....</i>	<i>6</i>
<i>Figure 2. Répartition des modes de codage dans la campagne à trois modes</i>	<i>9</i>
<i>Figure 3. Codage par système expert Iris/Muse.....</i>	<i>12</i>
<i>Figure 4. Description des propositions possibles de causes initiales impliquant du codage par IA.....</i>	<i>18</i>
<i>Figure 5. Répartition selon Origine - test 2022 (357 223 obs).....</i>	<i>72</i>