Les statistiques sur les causes de décès

Classer et coder… dans la classification internationale des maladies

Elise Coudin et Aude Robert***

La statistique sur les causes de décès repose sur l'analyse des textes écrits par les médecins sur les certificats de décès, lesquels sont ensuite classés et codés d'après la classification internationale des maladies de l'Organisation mondiale de la santé. Cette statistique et la classification dans laquelle elle est codée se sont construites de concert depuis la fin du XIXe siècle. La statistique sur les causes de décès est aujourd'hui produite par le Centre d'épidémiologie sur les causes médicales de décès (CépiDc) de l'Institut national de la santé et de la recherche médicale (Inserm). Depuis l'année de décès 2021, sa production articule trois modes de codage entre eux : codage automatique par système expert, codage interactif par équipe d'experts-nosologistes et codage par prédictions de réseaux de neurones profonds entraînés sur l'historique des certificats de décès déjà codés. Sa mise en œuvre, qui s'appuie sur les critères de qualité en matière de statistique européenne, permet notamment de réduire les calendriers de production et d'avoir une démarche d'évaluation et une supervision statistiques systématiques.

Causes-of-death (CoD) statistics are based on an analysis of the texts written by medical practitioners on death certificates, which are then classified and coded according to the World Health Organization's International Classification of Diseases. The statistics and the International statistical classification of diseases (ICD) classification have been developed together since the end of the 19th century. Today, causes-of-death (CoD) statistics are produced by the Centre for Epidemiology on Medical Causes of Deaths (CépiDc) of the National Institute of Health and Medical Research (INSERM). Since the year of death 2021, the CoD statistics production uses three different coding methods: automatic coding by an expert system, interactive coding by a team of expert nosologists, and predictive coding by deep neural networks trained on the history of previously coded death certificates. The implementation, which is based on European statistical quality criteria, makes it possible to shorten production delays and initiate systematic statistical evaluation and monitoring procedures.

* Directrice du Centre d'épidémiologie sur les causes médicales de décès, INSERM.
elise.coudin@inserm.fr

** Responsable de l'équipe automatisation, Centre d'épidémiologie sur les causes médicales de décès, INSERM.
aude.robert@inserm.fr

Pour hiérarchiser, suivre les problèmes de santé publique et analyser leurs disparités géographiques, temporelles ou sociales, les épidémiologistes, statisticiens et démographes utilisent les statistiques sur les causes de décès. Ces statistiques mobilisent peu de variables : tout au plus quelques informations démographiques et géographiques en complément du texte décrivant la ou les causes de décès. À cette apparente simplicité s'oppose une construction statistique plutôt complexe, résultat de l'histoire et de la spécificité de la source. La **codification** du texte rédigé par le médecin, en codes de nomenclature, c'est-à-dire l'attribution d'un « code » de la classification internationale des maladies (CIM) au texte médical décrivant la cause de décès pour obtenir des indicateurs statistiques pertinents, joue un rôle essentiel[1]. Cette codification s'est complexifiée au cours du temps, avec la précision grandissante de la classification internationale des maladies, le besoin d'un traitement homogène et comparable entre pays. Elle repose aujourd'hui en France sur une articulation entre système expert à base de règles, équipe d'experts-nosologistes (la nosologie étant l'étude des classifications médicales) et depuis peu, inférence d'algorithmes d'intelligence artificielle. L'articulation entre ces trois modes a pour but d'optimiser la qualité du codage dans son ensemble, pour un usage statistique, en respectant les délais de production.

Cet article décrit l'histoire de cette statistique, sa construction et ses spécificités, en particulier la place de la codification. Puis, il présente les outils de codage mobilisés dans la production aujourd'hui, du traditionnel au plus innovant.

De l'origine de la statistique sur les causes de décès au besoin de santé publique

Les prémices de la statistique sur les causes de décès sont très anciennes (Vallin et Meslé, 1988). Les grandes épidémies de peste auraient motivé des relevés et des comptages de décès par cause. Initialement « opérations de circonstance » (Bouvier-Colle et alii, 1990), ceux-ci sont mobilisés dans des études à l'origine des disciplines statistiques, épidémiologiques et démographiques. En France, les comptages par cause de décès s'établissent en parallèle du développement des registres paroissiaux, puis de l'état civil à partir de la fin du XVIII^{e}. Le XIX^{e} siècle marque le début d'une véritable construction statistique. Pour les besoins politiques de santé publique, de « salubrité publique » selon la terminologie de l'époque, le relevé des causes de décès se systématise. En France, la statistique des causes de décès naît officiellement en 1886, à la suite d'une circulaire gouvernementale demandant aux maires des villes de plus de 10 000 habitants de relever une fois par quinzaine les décès dus à sept maladies infectieuses (fièvre typhoïde, variole, rougeole, scarlatine, coqueluche, diphtérie, diarrhée infantile) élargies à 27 rubriques dès l'année suivante.

1 Pour une approche statistique historique voir le chapitre Classer et coder de *La politique des grands nombres* d'Alain Desrosières (Desrosières, 1993).

Statistique et classification internationale se construisent de concert depuis la fin du XIXe siècle

> ***Une bonne statistique est indispensable pour apprécier avec exactitude l'état sanitaire du pays, pour diriger avec efficacité la lutte contre les maladies, pour mesurer avec précision les résultats obtenus.***

Les relevés par l'État sont généralisés à l'ensemble des causes de décès puis à l'ensemble du territoire à la fin du XIXe et début du XXe siècle. Cette collecte systématique requiert une codification dans une classification statistique commune et univoque. Alors qu'au cours du XIXe siècle, une trop grande variété de listes et terminologies est utilisée (empêchant les comparaisons ou le suivi), les travaux en nosologie pour répondre aux besoins statistiques aboutissent, non sans détours[2], aux grands principes d'une classification commune. Le principe d'une classification internationale des maladies est adopté en 1893 par l'Institut international statistique à partir de celle de synthèse proposée par Jacques Bertillon[3], chef des travaux statistiques de la ville de Paris (en 14 chapitres, 161 rubriques détaillées). Sont construites ensuite puis adoptées par les pays les premières versions de la CIM. Cette dernière est gérée par l'Organisation mondiale de la santé (OMS) à partir de 1945 (***encadré 1***).

En France, on collecte l'ensemble des causes de décès sur l'ensemble du territoire, depuis 1906 (Aubenque et alii, 1978). Celles-ci alimentent les tableaux de Statistique sanitaire publiés par la Direction de l'assistance et de l'hygiène publique du ministère de l'Intérieur. Le simple comptage effectué par les mairies devient un recueil d'information organisé et systématisé quand, à partir de 1925, l'information sur la cause de décès est ajoutée au bulletin de décès de l'état civil. Ceci permet d'avoir des informations démographiques sur le défunt (sexe, groupe d'âge, département de décès puis de résidence). La statistique générale de France (SGF) est alors chargée du recueil et de l'élaboration de cette statistique. « Une bonne statistique est indispensable pour apprécier avec exactitude l'état sanitaire du pays, pour diriger avec efficacité la lutte contre les maladies, pour mesurer avec précision les résultats obtenus. » écrit Huber, directeur de la SGF. Ce verbatim résume ainsi les trois finalités de cette collecte, toujours d'actualité : veille sanitaire, politiques de santé publique et statistique.

La naissance du certificat à caractère médical

En 1937, est créé le certificat médical de décès, confidentiel, rempli par un médecin qui y déclare la cause en texte libre. Cette déclaration devient alors obligatoire. Le bulletin de décès anonymisé et le certificat médical cacheté sont transmis de la mairie à la Direction départementale de la santé à des fins de veille sanitaire. Le bulletin de décès sur lequel est reporté le texte de la cause est ensuite transmis à la SGF, laquelle codifie dans la classification des maladies et diffuse la statistique.

2 Voir aussi l'historique très complet sur le site du CépiDc *https://www.cepidc.inserm.fr/causes-medicales-de-deces/cim-9/historique*.
3 Qui propose au même congrès une première classification sur les professions.

▶ Encadré 1. Classification internationale des maladies

La codification des causes de décès dans une nomenclature assurant comparabilité dans l'espace et le temps est un enjeu majeur et international dès le début du XXe siècle. « Le but de la Classification statistique internationale des maladies et des problèmes de santé connexes est de permettre l'analyse systématique, l'interprétation et la comparaison des données de mortalité et de morbidité recueillies dans différents pays ou régions et à des époques différentes (...). » (Organisation mondiale de la santé, 2008 ; Rey, 2016). La nomenclature doit assurer par son organisation une certaine stabilité temporelle et régionale de la codification des pathologies dans un contexte où les progrès de la médecine peuvent être rapides et les variations régionales fortes. Élaborée à partir de la classification Bertillon à la fin du XIXe, adoptée internationalement, la CIM est réactualisée tous les

Évolution de la classification internationale des maladies : de la CIM 1 à la CIM 11*

CIM 1 - 1900
Première révision de la nomenclature internationale des causes de décès

179 rubriques, 14 chapitres

I	Maladies générales
II	Maladies du système nerveux et des organes des sens
III	Maladies de l'appareil circulatoire
IV	Maladies de l'appareil respiratoire
V	Maladies de l'appareil digestif
VI	Maladies de l'appareil génito-urinaire et de ses annexes
VII	Maladies puerpérales
VIII	Maladies de la peau et de ses annexes
IX	Maladies des organes de locomotion
X	Vices de conformation
XI	Maladies du 1er âge
XII	Maladies de la vieillesse
XIII	Affections produites par des causes extérieures
XIV	Maladies mal définies

Morbidité

Organisation mondiale de la santé

Règles cause initiale

CIM 6 - 1948
Modèle standard de certificat, Règles de détermination de la cause initiale

17 chapitres, 765 rubriques, près de 2 000 codes

CIM 10 - 1992

21 chapitres, 12 000 rubriques, ~ 16 000 codes

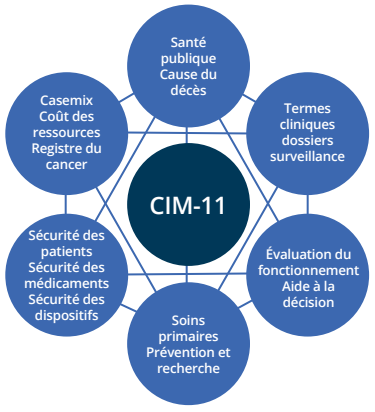
I	A00-B99	Certaines maladies infectieuses et parasitaires
II	C00-D48	Tumeurs
III	D50-D89	Maladies du sang et des organes hématopoïétiques et certains troubles du système immunitaire
IV	E00-E90	Maladies endocriniennes, nutritionnelles et métaboliques
V	F00-F99	Troubles mentaux et du comportement
VI	G00-G99	Maladies du système nerveux
VII	H00-H59	Maladies de l'œil et de ses annexes
VIII	H60-H95	Maladies de l'oreille et de l'apophyse mastoïde
IX	I00-I99	Maladies de l'appareil circulatoire
X	J00-J99	Maladies de l'appareil respiratoire
XI	K00-K93	Maladies de l'appareil digestif
XII	L00-L99	Maladies de la peau et du tissu cellulaire sous-cutané
XIII	M00-M99	Maladies du système ostéo-articulaire, des muscles et du tissu conjonctif
XIV	N00-N99	Maladies de l'appareil génito-urinaire
XV	O00-O99	Grossesse, accouchement et puerpéralité
XVI	P00-P96	Certaines affections dont l'origine se situe dans la période périnatale
XVII	Q00-Q99	Malformations congénitales et anomalies chromosomiques
XVIII	R00-R99	Symptômes, signes et résultats anormaux d'examens cliniques et de laboratoire, non classés ailleurs
XIX	S00-T98	Lésions traumatiques, empoisonnements et certaines autres conséquences de causes externes
XX	V01-Y98	Causes externes de morbidité et de mortalité
XXI	Z00-Z99	Facteurs influant sur l'état de santé et motifs de recours aux services de santé
XXII	U00-U99	Codes d'utilisation particulière

* *https://www.who.int/standards/classifications/classification-of-diseases.*

10 ans environ. Ces évolutions sont gérées par l'OMS depuis 1945. Les versions de la CIM sont de plus en plus riches et complexes. À l'origine, ce principe de classement reposait sur une distinction entre les maladies générales et celles localisées à un organe donné, plus facile à collecter que l'étiologie de la pathologie. La dimension étiologique s'est étendue au fil du temps et des progrès de la médecine, puis avec son usage pour élaborer les statistiques de morbidité hospitalière. La première révision de la CIM (en 1900) comportait 179 rubriques, la CIM 9 (utilisée pour coder les causes de décès de 1979 à 1999) 6 000 rubriques, et la CIM 10 (utilisée depuis l'an 2000) compte 12 000 rubriques et 16 000 codes. La transition à la CIM 11, adoptée en 2022 par l'OMS et dont l'objectif d'interopérabilité et de réutilisation dépasse celui de la statistique de santé, constituera un défi pour les prochaines années.

CIM 10 - 2019

I	Certaines maladies infectieuses et parasitaires					
II	Tumeurs					
	C00-C97	Tumeurs malignes				
		C00-C75	Tumeurs malignes, primitives ou présumées primitives, de siège précisé, à l'exception des tissus lymphoïde, hématopoïétique et apparentés			
			C00-C14	Tumeurs malignes de la lèvre, de la cavité buccale et du pharynx		
			...	...		
			C50-C50	Tumeur maligne du sein		
				C50	Tumeur maligne du sein	
					C50.0	Mamelon et aréole
					C50.1	Partie centrale du sein
					C50.2	Quadrant supéro-interne du sein
					C50.3	Quadrant inféro-interne du sein
					C50.4	Quadrant supéro-externe du sein
					C50.5	Quadrant inféro-externe du sein
					C50.6	Prolongement axillaire du sein
					C50.8	Lésion à localisations contiguës du sein
					C50.9	Sein, sans précision
			...	...		
			C73-C75	Tumeurs malignes de la thyroïde et d'autres glandes endocrines		
III	Maladies du sang et des organes hématopoïétiques et certains troubles du système immunitaire					
...	...					
XXII	Codes d'utilisation particulière					



CIM 11 - 2022
Post-coordination, interopérabilité, réutilisation, cas d'usages plus larges que statistiques

Ce circuit de recueil d'information statistique assure la confidentialité de l'information médicale transmise, tout en circulant par les mains des acteurs de la veille sanitaire au plus tôt. Le revers de la médaille est qu'il s'avère lourd, difficile à coordonner et peut conduire à des pertes d'informations. Ces contraintes existent toujours aujourd'hui du fait d'un circuit de transmission des certificats papiers quasi inchangé.

L'Insee remplace la SGF en 1945 et décentralise la codification des causes en direction régionale, introduisant une hétérogénéité dans les pratiques. Par ailleurs, l'information collectée devient plus précise et se standardise suivant les besoins de la classification. Dès 1955, le certificat médical permet au médecin de déclarer plusieurs causes de décès. Et dès 1958, le modèle de certificat en France suit le standard international élaboré dans la CIM et recommandé par l'OMS.

▶ Une production confiée à l'Inserm

En 1968, à l'occasion d'une nouvelle révision de la CIM, la mission de codification des causes de décès est confiée à l'Inserm afin qu'elle soit réalisée par un personnel médical qualifié et de manière centralisée. Jusqu'en 1987, les professionnels médicaux de la Section information en santé publique de l'Inserm, puis du Service commun d'information sur les causes médicales de décès (SC8) codifient des causes et reportent les codes sur les bulletins de décès compilés ensuite par l'Insee. À partir de 1988, avec le développement de l'informatisation, l'Inserm obtient la maîtrise complète du processus de production jusqu'à la diffusion : l'Insee met à disposition un extrait informatique de son fichier démographique de mortalité. Les experts-nosologistes y adossent les codes des causes et l'Inserm peut ainsi publier des statistiques provisoires, puis définitives. Depuis 1997, le Centre d'épidémiologie sur les causes médicales de décès de l'Inserm (CépiDc, remplaçant le SC8 chargé du codage et de l'exploitation et le SC25 chargé de la saisie) assure la production et la diffusion de la statistique.

▶ Une statistique officielle européenne suivant les standards de l'OMS

> ***La construction de la statistique sur les causes de décès est internationale dès sa création.***

La construction de la statistique sur les causes de décès est internationale dès sa création. À partir de 2007-2008, avec la mise en place du système statistique européen, elle devient une statistique officielle européenne soumise au règlement 223[4]. Elle entre dans le champ d'application du règlement relatif aux statistiques communautaires de la santé publique ainsi que de la santé et de la sécurité au travail, lequel découle du développement de programmes d'action communautaire dans le domaine de la santé publique. Ce règlement et son règlement d'application[5] assurent la comparabilité de la statistique au

4 Voir les références juridiques en fin d'article.
5 CE 1338/2008 et UE 328/2011.

sein de l'Europe et définissent les conditions de collecte de l'information. Celle-ci s'appuie sur des certificats de décès nationaux conformes aux recommandations de l'OMS et sur un codage des causes dans la classification internationale des maladies de l'OMS. Les normes d'évaluation de la qualité sont celles du code des bonnes pratiques en matière de statistique européenne[6] (pertinence, exactitude, actualité, ponctualité, accessibilité, comparabilité, cohérence) ; les concepts, les champs, les variables, leur ventilation, les périodes de référence et les délais de transmission sont fixés. En particulier, les États membres doivent transmettre à Eurostat les données d'une année dans un délai de 24 mois à compter de la fin de cette même année. Le CépiDc de l'Inserm devient en 2017 une « autorité statistique nationale » (ONA : *Other National Authority*) productrice de statistique officielle, hors service statistique public.

Les grandes causes de décès en France et en Europe en 2021

Concrètement, l'indicateur le plus courant est la répartition des décès selon leurs causes initiales. La cause initiale du décès est « la maladie ou le traumatisme qui a déclenché l'évolution morbide conduisant directement au décès ou les circonstances dans le cas d'un traumatisme » (Organisation mondiale de la santé, 2008). Ainsi, en 2021, les cancers sont la première cause de décès en France avec plus d'un quart des décès ; en Europe ils représentent 22 % des décès. En moyenne en Europe, ce sont les maladies de l'appareil circulatoire qui se situent à la première place avec près d'un tiers des décès, contre 21 % en France (***figure 1***). Ceci n'est qu'un aperçu des tables de données sur les causes de décès mises à jour, et diffusées[7] chaque année après collecte auprès des États membres par Eurostat[8] (Eurostat, 2024), et en France, sur le site du CépiDc (Fouillet et alii, 2023, 2024 ; Cadillac et alii 2023, 2024). La base de données individuelles avec l'ensemble des causes, des textes et autres variables alimente chaque année le Système national des données de santé.

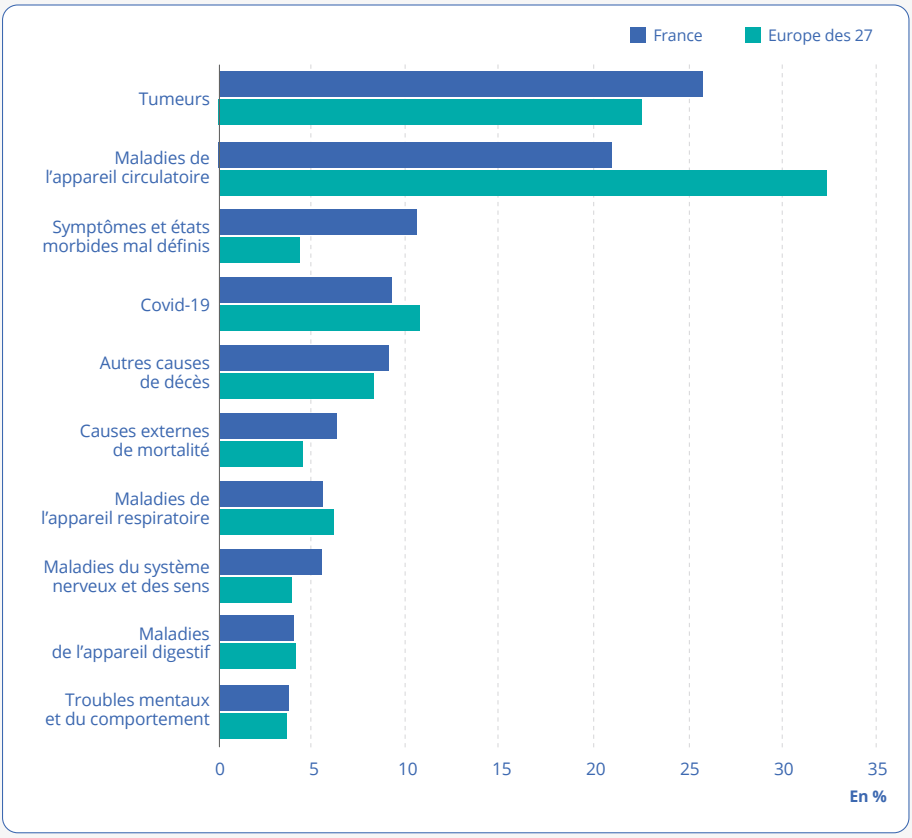
Cependant, à l'apparente simplicité de ces statistiques, ventilées par sexe, âge, localisation du décès ou lieu de résidence du défunt, s'oppose une complexité en pratique de la source et de son traitement. Et ceci, dès la collecte de l'information.

La certification des décès en pratique aujourd'hui

Aujourd'hui, chaque décès donne lieu à la rédaction d'un certificat de décès par un médecin. La validation de ce certificat est nécessaire pour fermer le cercueil et procéder à l'inhumation. La forme de ce certificat et les finalités[9] d'usage des informations qu'il

6 *https://www.insee.fr/fr/information/4137879.*
7 Voir les tables « Causes de décès » (hlth_cdeath) sur la base de données d'Eurostat : *https://ec.europa.eu/eurostat/web/health/database.*
8 Eurostat, l'Office statistique de l'Union européenne, est chargé de publier des statistiques et des indicateurs européens, permettant d'effectuer des comparaisons entre les pays et les régions.
9 Ces informations ne peuvent être utilisées que pour des motifs de santé publique, à des fins de veille et d'alerte, par l'État, les agences régionales de santé et l'Agence nationale de santé publique ; pour l'établissement de la statistique nationale des causes de décès et pour la recherche en santé publique par l'Inserm ; pour l'alimentation du Système national des données de santé et les traitements de données afférents concernant la santé ; et pour l'établissement de statistiques par l'Insee et le service statistique public dans le cadre de la loi de 1951. Voir les références juridiques en fin d'article.

Figure 1 - Les 10 causes de décès les plus fréquentes en 2021 en France et en Europe



Lecture : 32 % des décès en Europe sont causés par les maladies de l'appareil circulatoire et 21 % en France.
Champ : Tous les décès survenus en France et en Europe.
Source : Eurodatabase, causes of death, hlth_cdeath, *https://ec.europa.eu/eurostat/web/health/database*.

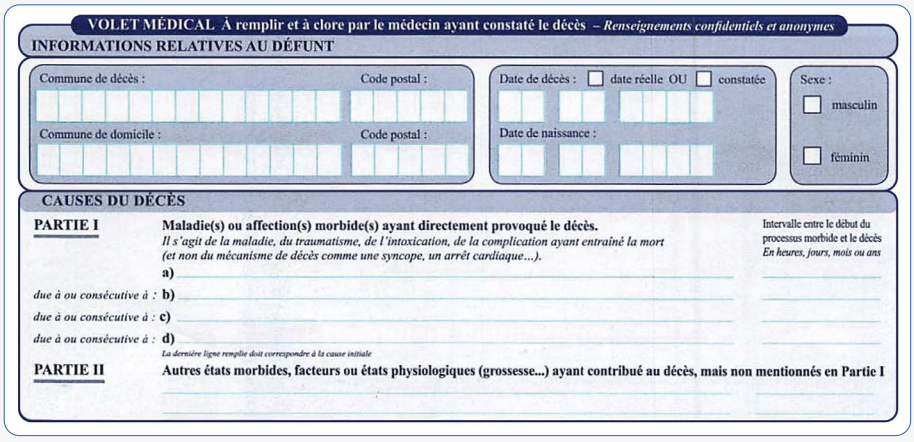
contient sont définies dans le Code général des collectivités territoriales[10]. Un volet administratif avec les noms et prénoms est destiné aux opérateurs funéraires et à la mairie qui dressera l'acte de décès, et un volet médical confidentiel sans les noms et prénoms du défunt est transmis à l'Inserm. Depuis 2022, le certificat de décès doit obligatoirement être rempli sur support électronique dans les établissements de santé, sauf exceptions[11]. Début 2024, un peu plus de 40 % des certificats étaient électroniques.

Le volet médical arrive au CépiDc au format électronique ou numérisé via un prestataire de saisie. Il suit le modèle international de l'OMS (***figure 2***). Dans une première partie (Partie 1), le médecin écrit sur quatre lignes l'enchaînement causal des maladies qui ont

10 Voir les références juridiques en fin d'article.
11 Voir les références juridiques en fin d'article.

conduit au décès (aussi appelé « processus morbide »), de la cause immédiate à inscrire en première ligne, à la cause initiale. Le médecin est invité à remplir une cause en texte libre par ligne, chaque cause indiquée sur une ligne étant « due à » celle indiquée sur la ligne suivante (relation causale attendue). Dans la Partie 2, le médecin indique les autres états morbides, facteurs ou états physiologiques qui ont pu contribuer au décès sans être directement impliqués dans l'enchaînement conduisant au décès. Il consigne aussi sur le volet médical des informations notamment sur le lieu de décès, l'état de grossesse, les circonstances apparentes du décès (mort naturelle, mort subite, accident, suicide, etc.).

▶ Figure 2 - Volet médical du certificat de décès*



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

INFORMATIONS RELATIVES AU DÉFUNT

Commune de décès : Code postal :

Commune de domicile : Code postal :

Date de décès : ☐ date réelle OU ☐ constatée

Date de naissance :

Sexe : ☐ masculin ☐ féminin

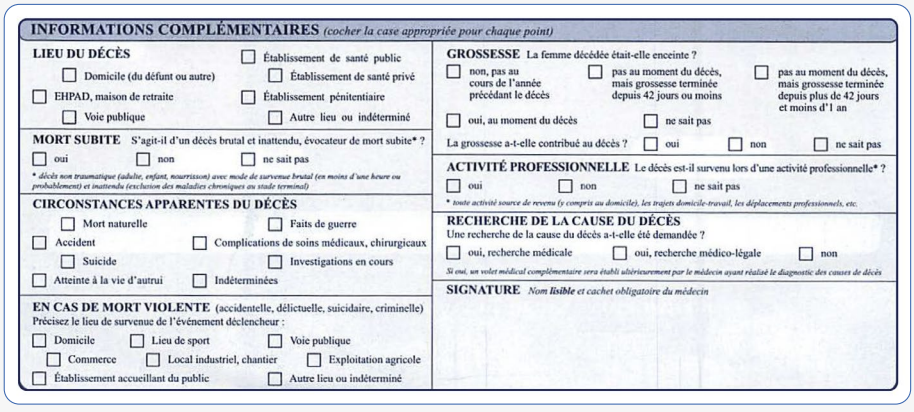
CAUSES DU DÉCÈS

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

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

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

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



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

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

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

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

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

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

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

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

SIGNATURE *Nom lisible et cachet obligatoire du médecin*

* *Voir les références juridiques en fin d'article.*

▶ Un texte médical riche, polysémique, parfois sans enchaînement causal

Le constat d'un décès relevant d'un diagnostic médical, le médecin qui en prend la responsabilité est libre dans sa rédaction, sans pré-remplissage. C'est d'ailleurs une recommandation de l'OMS. En conséquence, le texte[12] du processus morbide est tout aussi libre et contient de nombreux libellés possibles :

- toutes les maladies et pathologies, avec leurs étiologie, intensité, localisation et autres caractéristiques associées (telles que « cancer », « tabac », « génétique », « sévère », « stade 4 », « poumon », « gauche », « maladies osseuses ») ;
- des acronymes (comme « AVC », « ACR », « IRM », « BPCO »[13]), parfois polysémiques (« IRC » ou « IRA » peuvent signifier selon le contexte « insuffisance rénale » « chronique » ou « aiguë » ou « insuffisance respiratoire » « chronique » ou « aiguë », ou encore « TA » « tentative d'autolyse » ou « tension artérielle ») ;
- des abréviations (« K » pour cancer), parfois en grec (ψ, γ pour grossesse, etc.), des signes (ensemble vide pour « sans ») ;
- différentes appellations ayant le même sens (par exemple « cancer », « tumeur », « néo », « néoplasme », « carcinome », etc.).

Il peut aussi y avoir des incohérences dans l'enchaînement causal, voire entre la Partie I et la Partie II du certificat. On peut retrouver plusieurs entités sur la même ligne du formulaire avec des liens de causalité exprimés (par exemple par le terme « sur » : « métastases sur cancer pulmonaire ») ou pas (« infection sur cathéter »).

> ***La spécificité des causes de décès provient bien de cette information textuelle riche en termes et faiblement structurée qu'il faut transformer en un langage statistique commun.***

La spécificité des causes de décès provient bien de cette information textuelle riche en termes et faiblement structurée qu'il faut transformer en un langage statistique commun. Ainsi l'opération de codification des causes de décès se distingue pour plusieurs raisons de celles sur d'autres sources de la statistique publique, pour lesquelles il s'agit de classifier de simples libellés. Tout d'abord, la classification comprend plusieurs milliers de rubriques, sous forme d'une arborescence touffue. Par exemple, si l'étiologie d'une pathologie permet de la reclasser dans un chapitre définissant plus précisément son origine, on peut être amené à changer de chapitre : ainsi une démence sera classée généralement en F « troubles mentaux » mais elle peut l'être dans d'autres chapitres si elle dérive d'autres pathologies (infectieuses, du système nerveux, etc.). En complément de la codification des causes, il faut vérifier la cohérence du processus causal comme préalable à la détermination de la cause initiale, ce qui est une opération complexe.

12 Pour les certificats papier, le prestataire numérise le processus causal (parties 1 et 2) par saisie vocale.
13 « AVC » : accident vasculaire cérébral, « ACR » : arrêt cardio-respiratoire, « IRM » : imagerie par résonance magnétique, « BPCO » : bronchopneumopathie chronique obstructive.

En 2021, un certificat non vide comprend en moyenne 3,5 codes de causes, 15 % des certificats en ont 6 ou plus (et 1 %, 11 codes ou plus). De plus, 4 145 codes différents de la CIM ont été utilisés au moins une fois, dont 2 675 codes en cause initiale.

Coder les causes de décès, un exercice complexe

Coder les causes de décès dans la CIM a deux finalités distinctes : déterminer dans le texte écrit par le médecin les entités nosologiques[14] pour leur affecter un code de la nomenclature CIM et déterminer la cause initiale du décès.

Pour satisfaire ce double objectif, la classification internationale des maladies décrit en complément de la nomenclature (c'est-à-dire les codes) un ensemble de règles de codage, rassemblées dans un volume 2, le « guide de référence » (OMS, 2008). Ce manuel d'instruction au codage outille le codeur : une douzaine de règles appliquées sur la séquence causale selon un algorithme précis permet de déterminer la cause initiale du décès, d'une façon systématique, tout en corrigeant des erreurs ou des incohérences possibles dans la chaîne causale déclarée ou encore en privilégiant certaines pathologies à suivre car d'intérêt de santé publique[15] (***figure 3***).

> ***Ce besoin d'homogénéité pour assurer la comparabilité de la statistique dans le temps et l'espace a motivé l'automatisation du codage.***

Ce besoin d'homogénéité pour assurer la comparabilité de la statistique dans le temps et l'espace a motivé l'automatisation du codage. Le caractère systématique du raisonnement et la présence de règles précises décrites dans la CIM font de l'exercice un candidat idéal. Les outils de cette automatisation ont évolué dans le temps, de l'interface interactive permettant un codage assisté par système de règles au batch automatique de ce même système, jusqu'aux réseaux de neurones profonds entraînés sur les multiples données déjà codées.

Les systèmes experts nés du partage international des règles de décision

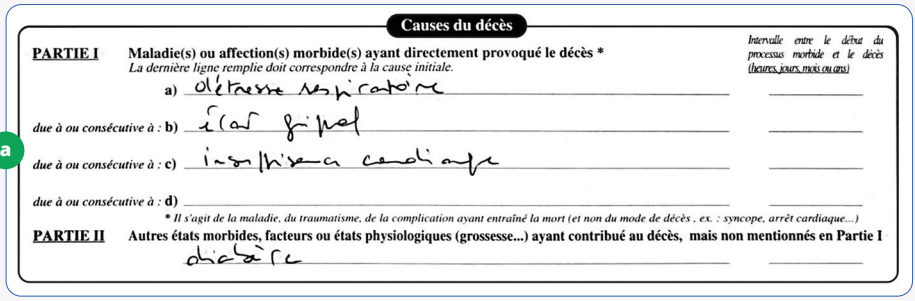
Les systèmes experts de codage des causes de décès sont utilisés en France depuis 2000, à l'occasion de la mise en œuvre de la CIM 10. Styx, développé en France, puis Iris à partir de 2012, sont des logiciels d'aide au codage avec une interface homme-machine, utilisant cette démarche (Pavillon et Laurent, 2003 ; Iris Institute, 2024). Ils s'appuient sur le modèle international du certificat de décès et intègrent les règles du volume 2 de la CIM 10 sous

14 Une entité nosologique est un terme classifiable dans la CIM.

15 Les experts du CépiDc sont membres du *Mortality reference group* (groupe de référence sur la mortalité), le groupe de l'Organisation mondiale de la santé (OMS) chargé du maintien et des évolutions des règles de codage de la CIM en matière de mortalité. Le centre collaborateur français de l'OMS sur la famille des classifications internationales présidé par l'Agence du numérique en santé regroupe l'Inserm-CépiDc (pour la mortalité), la Caisse nationale de l'assurance maladie (qui intervient sur la nomenclature de classement des remboursements, appelée l'*International Classification of Health Interventions* (ICHI), l'Agence technique de l'information sur l'hospitalisation (pour la Classification internationale des maladies et ICHI en morbidité), l'École des hautes études en santé publique (pour la Classification internationale du fonctionnement, du handicap et de la santé).

▶ Figure 3 - Causes de décès indiquées par le médecin

Une des règles de codage indique que la séquence entre la cause immédiate et la cause initiale doit être entièrement causale. Ainsi l'état grippal peut entraîner une détresse respiratoire, mais une insuffisance cardiaque ne peut pas conduire à un état grippal : ce sera donc l'état grippal qui sera retenu comme cause initiale, même si le médecin a indiqué l'insuffisance cardiaque sur la dernière ligne de la Partie 1 (figure 3a).


Causes du décès

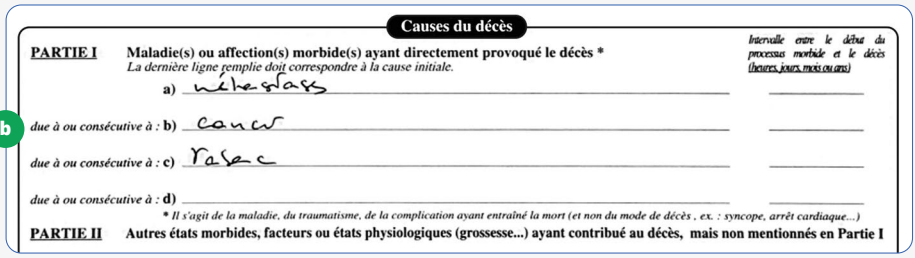
PARTIE I — Maladie(s) ou affection(s) morbide(s) ayant directement provoqué le décès *
La dernière ligne remplie doit correspondre à la cause initiale.
Intervalle entre le début du processus morbide et le décès (heures, jours, mois ou ans)

a) détresse respiratoire
due à ou consécutive à : b) état grippal
a due à ou consécutive à : c) insuffisance cardiaque
due à ou consécutive à : d)

* Il s'agit de la maladie, du traumatisme, de la complication ayant entraîné la mort (et non du mode de décès , ex. : syncope, arrêt cardiaque...)

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

Le même raisonnement permet de retenir comme cause initiale le cancer pulmonaire, même si le médecin a indiqué tabac dans la ligne suivante (figure 3b).


Causes du décès

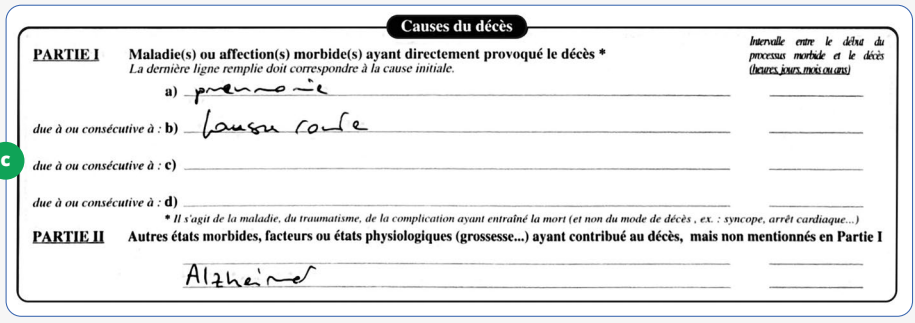
PARTIE I — Maladie(s) ou affection(s) morbide(s) ayant directement provoqué le décès *
La dernière ligne remplie doit correspondre à la cause initiale.
Intervalle entre le début du processus morbide et le décès (heures, jours, mois ou ans)

a) métastase
b due à ou consécutive à : b) cancer
due à ou consécutive à : c) tabac
due à ou consécutive à : d)

* Il s'agit de la maladie, du traumatisme, de la complication ayant entraîné la mort (et non du mode de décès , ex. : syncope, arrêt cardiaque...)

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

D'autres règles permettent d'aller rechercher la cause initiale en partie 2 si les causes en Partie 1 ne sont pas assez informatives. Ainsi, on ira chercher Alzheimer en cause initiale même s'il n'est indiqué qu'en partie 2 (figure 3c).


Causes du décès

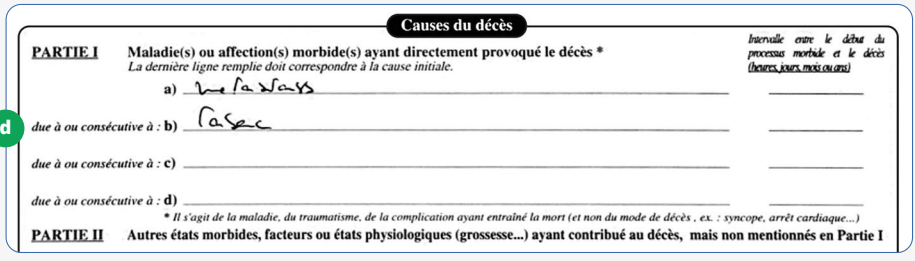
PARTIE I — Maladie(s) ou affection(s) morbide(s) ayant directement provoqué le décès *
La dernière ligne remplie doit correspondre à la cause initiale.
Intervalle entre le début du processus morbide et le décès (heures, jours, mois ou ans)

a) pneumopathie
due à ou consécutive à : b) cancer poumon
c due à ou consécutive à : c)
due à ou consécutive à : d)

* Il s'agit de la maladie, du traumatisme, de la complication ayant entraîné la mort (et non du mode de décès , ex. : syncope, arrêt cardiaque...)

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

Et on ira chercher « cancer » en cause initiale, dans l'exemple ci-dessous même si elle n'est pas précisée explicitement, car celle-ci est « évidente » (figure 3d).


Causes du décès

PARTIE I — Maladie(s) ou affection(s) morbide(s) ayant directement provoqué le décès *
La dernière ligne remplie doit correspondre à la cause initiale.
Intervalle entre le début du processus morbide et le décès (heures, jours, mois ou ans)

a) métastase
d due à ou consécutive à : b) cancer
due à ou consécutive à : c)
due à ou consécutive à : d)

* Il s'agit de la maladie, du traumatisme, de la complication ayant entraîné la mort (et non du mode de décès , ex. : syncope, arrêt cardiaque...)

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

forme de relations entre codes CIM, rassemblées dans des « tables de décisions ». Ces « tables » ont été développées dès le début des années 1980 par le *National Center for Health Statistics* (NCHS, le centre national de statistiques sur la santé) américain dans le module « *Automated Classification of Medical Entities* » (ACME, Classification automatisée des entités médicales) du système expert américain *Mortality Medical Data System* (Système de données médicales sur la mortalité) (National Center for Health Statistics, 2024 ; Lu et Wu, 2005). Elles ont été partagées internationalement puis traduites de la CIM 8 à la CIM 9, et de la CIM 9 à la CIM 10 (Navarra et alii, 2016) et suivent les mises à jour annuelles officielles de l'OMS. Les systèmes experts comme Iris comportent deux étapes.

Tout d'abord, il faut traduire le texte libre rédigé par les médecins en codes CIM. Cette première étape dépend de la langue. Après des standardisations du texte du certificat (synonymes, abréviations, termes non pertinents[16]), il est mis en correspondance avec un dictionnaire contenant 157 000 expressions, pour lesquelles chaque terme est associé à un ou plusieurs codes CIM 10.

Il s'agit ensuite de sélectionner la cause initiale selon les règles et directives de la CIM. Cette seconde étape s'applique à l'aide des tables de décision[17], maintenues depuis 2011 par un consortium international de pays au sein de l'Institut Iris. Depuis le développement du *Multicausal and Unicausal Selection Engine* (MUSE, moteur de sélection de cause initiale et de cause multiple), cette étape prend aussi en compte les circonstances de décès, l'âge et le sexe du défunt.

Une trentaine de pays utilisent le logiciel Iris (Allemagne, Canada, Italie, Pays-Bas, Grande-Bretagne, etc.).

Les systèmes experts étaient tout d'abord utilisés uniquement en interactif : tous les certificats étaient vus au moins une fois par un agent de l'équipe de codage[18]. L'organisation du travail de l'équipe de codage reste aujourd'hui encore très marquée par l'usage du système expert : les codeurs corrigent, mettent en forme, simplifient le texte sans en modifier le fond, pour permettre au système expert de coder en interactif et les nosologistes, et en dernier lieu les experts tranchent sur les cas délicats.

▶ Du batch automatique au codage interactif

Le logiciel Iris/Muse est lancé en batch de façon hebdomadaire dès réception dans la base de codage. Même s'il échoue à coder certains cas, il apporte des informations sur le type de blocage rencontré (cause manquante, plusieurs causes initiales possibles, etc.), et guide le travail de l'expert-codeur. Le taux de codage automatique complet du certificat est de 63 % pour la France. À titre de comparaison, il est de 65 % pour les Pays-Bas et de 80 % pour l'Italie et l'Angleterre[19]. En France, la première étape de codage pêche souvent : pour 83 % des rejets du batch, Iris/Muse n'a pas réussi à coder l'ensemble des codes CIM 10 attendus. Ceci provient du fait de fautes d'orthographe, de textes non pertinents pour le codage

16 Au total près de 1 000 expressions normalisées.
17 Les tables de décision comprennent près de 30 millions de relations entre paires de codes CIM.
18 Depuis 2015, l'équipe de codage vérifie seulement les cas improbables ou sensibles.
19 Hors décès pour lesquels une enquête est ouverte (*inquest*, 5 % des décès). Dans ces cas en Grande-Bretagne, l'information est transmise sous la forme d'un compte-rendu médical et non d'un certificat de décès.

> ***Maintenir les tables de décision et le dictionnaire est lourd et la capitalisation de l'expertise humaine doit se faire en parallèle de la campagne de codage.***

sur les certificats (« médecin traitant », « smur », etc.), ou encore des expressions polysémiques mentionnées plus haut, qui peuvent s'ajouter aux difficultés de lecture du texte sur les certificats papiers. Si toutes les causes sont codées, Iris-Muse identifie le code de la cause initiale dans 90 % des cas.

Les systèmes de règles connaissent d'autres limites. Maintenir les tables de décision et le dictionnaire se révèle lourd et la capitalisation de l'expertise humaine doit se faire en parallèle de la campagne de codage. En France, les performances du système expert étaient insuffisantes pour respecter les délais de diffusion des données. Aussi, un troisième mode de codage a été introduit dans la production régulière de la statistique depuis la production des causes de décès 2021. Il repose sur des algorithmes d'apprentissage profond, lesquels ont l'avantage de pouvoir proposer une codification complète de tous les certificats qui leur sont soumis en un temps record.

Une troisième technique, basée sur l'intelligence artificielle (IA)

> ***Avoir recours à l'intelligence artificielle dans le codage des causes de décès est innovant.***

Avoir recours à l'intelligence artificielle dans le codage des causes de décès est innovant[20]. Cette innovation est possible grâce au développement récent des modèles d'apprentissage profond de type réseaux de neurones profonds (RNP) pour le traitement du texte, à celui des puissances de calcul, à leurs applications à l'informatique médicale et à la présence d'un historique conséquent de certificats déjà codés. Un RNP va transformer l'enchaînement des termes médicaux décrivant le processus morbide (séquence d'entrée) en un enchaînement de codes de la CIM et proposer une cause initiale (séquence de sortie). On parle alors d'algorithme « *sequence to sequence* » (*seq-to-seq*). En traitement automatique des langues et en apprentissage statistique, la transformation de séquence s'apparente à une traduction et celle de chaque terme médical en code à un problème de « classification ». Les RNP mobilisés ici sont des « *Transformers* » (Vaswani et alii, 2017). Les « *Transformers* » ont révolutionné les tâches de traduction grâce à une prise en compte du contexte (liens entre les mots d'une phrase) tout en ne nécessitant pas de puissances de calcul trop importantes[21]. Les calculs sous-jacents sont hautement parallélisés et nécessitent moins de données d'entraînement. Enfin, ces RNP sont disponibles en librairies open source[22].

20 Cette avancée est notable au niveau international. Seuls les États-Unis et le Portugal en complément de la France utilisent ou ont utilisé des RNP pour leur production de statistiques officielles sur les causes de décès (National Center for Health Statistics, 2023 ; Pita Ferreira et alii, 2022).
21 À l'inverse des *recurrent neural networks* par exemple.
22 Ici, Keras et TensorFlow sont mobilisées (*https://keras.io/* ; *https://www.tensorflow.org/?hl=fr*).

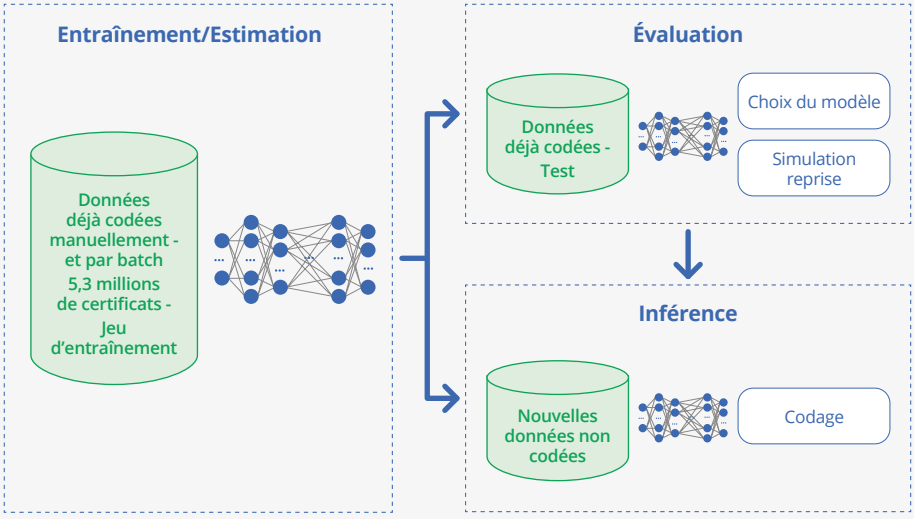
L'architecture des « *Transformers* » est de type encodeur/décodeur (Zambetta et alii, 2024 ; Hebbache et alii, 2024). Cela veut dire que la séquence d'entrée du modèle comme celle de sortie sont représentées dans un espace vectoriel de dimension beaucoup plus réduite que celles de leur corpus (tous les termes apparaissant au moins une fois dans la séquence d'entrée, respectivement celle de sortie). Cette étape de réduction de la dimension s'appelle « plongement lexical » ou *embedding* et s'appuie sur la proximité linguistique (Babet et alii, 2023). Par ailleurs, le « *Transformer* » modélise les éléments de contexte dans la phrase grâce au mécanisme d' « attention à plusieurs têtes » qui s'appuie sur des calculs parallélisés. En sortie, est calculée une distribution de probabilités qui permet de retenir la suite de codes la plus probable et composer ainsi la séquence de sortie.

Des travaux menés au CépiDc ont montré que les « *Transformers* » étaient performants dans le cas de la codification des causes de décès, dès lors que ces modèles étaient entraînés sur suffisamment de données déjà annotées (Falissard et alii, 2022). Après la preuve de concept et le potentiel démontré en recherche, l'enjeu a été de les intégrer dans la production courante, dans un cadre de qualité et en suivant les principes de la statistique européenne.

Un mécanisme sophistiqué relevant de l'apprentissage statistique

En pratique, l'étape des pré-traitements permet de transformer le texte des certificats en entrée numérique interprétable par un RNP (***encadré 2***). Puis, comme en apprentissage statistique supervisé, l'étape d'entraînement consiste à estimer les millions de paramètres (« poids ») du réseau en utilisant l'historique des certificats déjà codés. Lors de l'étape de test, la performance du ou des réseaux à prédire avec

Figure 4 - Réseaux de neurones profonds - De l'entraînement à l'inférence



exactitude la séquence de sortie est évaluée. Cette évaluation guide le choix du réseau estimé finalement retenu dans la production finale. Ce modèle, appliqué sur les textes d'un nouveau certificat, est désormais capable de prédire une séquence de codes et une cause initiale. Dans l'étape de prédiction, appelée aussi « inférence », on utilise donc ce réseau, en production, sur les certificats à coder (***figure 4***).

Ces RNP sont capables de « coder » environ 100 000 certificats par jour et par machine. Contrairement au système de règles qui rejette un certificat s'il n'a pas les règles pour le coder, un RNP prédira toujours une séquence de sortie quel que soit le certificat en entrée, pour peu que son texte suive correctement la structure spécifiée. Il associera aux termes de cette séquence prédite, des probabilités qui pourront par la suite être mobilisées pour estimer la qualité de la prédiction.

Les RNP sont utilisés pour prédire la séquence des codes et pour déterminer la cause initiale de certains certificats. Or, pour déterminer la cause initiale, il y a plusieurs stratégies : soit utiliser le code directement prédit par le modèle, soit appliquer le système expert Iris-muse sur la séquence des causes prédites par le modèle pour en déduire la cause initiale. L'approche retenue vise à retenir le meilleur des deux : elle consiste à entraîner un autre réseau de neurones capable de choisir entre les propositions possibles.

▶ Encadré 2. Le codage grâce aux réseaux de neurones profonds : des pré-traitements à l'inférence

Pour coder les causes de décès en CIM 10 en utilisant un réseau de neurones, on procède par étapes. Dans l'étape des pré-traitements, les données des certificats doivent être structurées en séquence interprétable par le modèle. On concatène les textes écrits sur chaque ligne du certificat de décès en ajoutant des variables de contexte. Pour une femme de 55 ans, décédée en 2017 et dont le certificat mentionnait « Ligne 1 : arrêt cardio respiratoire ; Ligne 2 : « dû à » épanchement pleural ; Ligne 3 : « dû à » métastases pulmonaires ; Ligne 4 : « dû à » cancer sein ; circonstances apparentes : mort naturelle, on obtient :

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

Les séquences en sortie ont quasiment la même structure sauf que les codes CIM 10 remplacent les termes. Elles se terminent par le code de la cause initiale, par exemple :

[start] certificatpapier versioncertificat1997 femme age55ans annee2017 lignecause1 r092 lignecause2 j90 lignecause3 c780 lignecause4 c509 lignecause7 causeinitiale c509 [end].

On découpe ensuite les séquences en « bouts », appelés « tokens » (*tokenizer*). Les tokens sont constitués pour la séquence d'entrée des mots qui se retrouvent au moins une fois dans le texte d'un certificat (150 000 différents environ), et pour la séquence de sortie de codes de la CIM (6 300 différents).

« Inférence » veut dire « prédiction » : il s'agit d'appliquer le réseau de neurones à la séquence de textes d'un certificat pour en prédire la séquence de codes. En un jour, on prédit environ 100 000 certificats sur une machine avec une unité de traitement graphique (GPU).

La probabilité que la cause soit correctement prédite, ainsi que l'écart de probabilité entre le code le plus probable et le second code le plus probable, sont mobilisés dans l'étape de ciblage.

Au préalable il aura fallu « entraîner » le modèle, c'est-à-dire estimer les multiples paramètres qu'il contient pour l'adapter au mieux à la tâche de codage qu'on lui demande. L'« entraînement », c'est-à-dire l'estimation du modèle, minimise une fonction de coût (*loss fonction*) : on calcule les poids ou les paramètres (de l'ordre de 100 millions) en minimisant les écarts entre les séquences prédites par le modèle et les séquences données par la base d'entraînement. La base d'entraînement utilisée en début de campagne 2021 contient 5,3 millions de certificats. Elle est complétée en fin de campagne de la moitié* des certificats de 2021 codés manuellement pendant la campagne 2021 pour obtenir le codage final. Sur une machine avec une unité de traitement graphique de 48 Gigaoctets de RAM (GPU Nvidia RTX A6000), l'entraînement dure environ quatre jours pour trois millions de certificats.

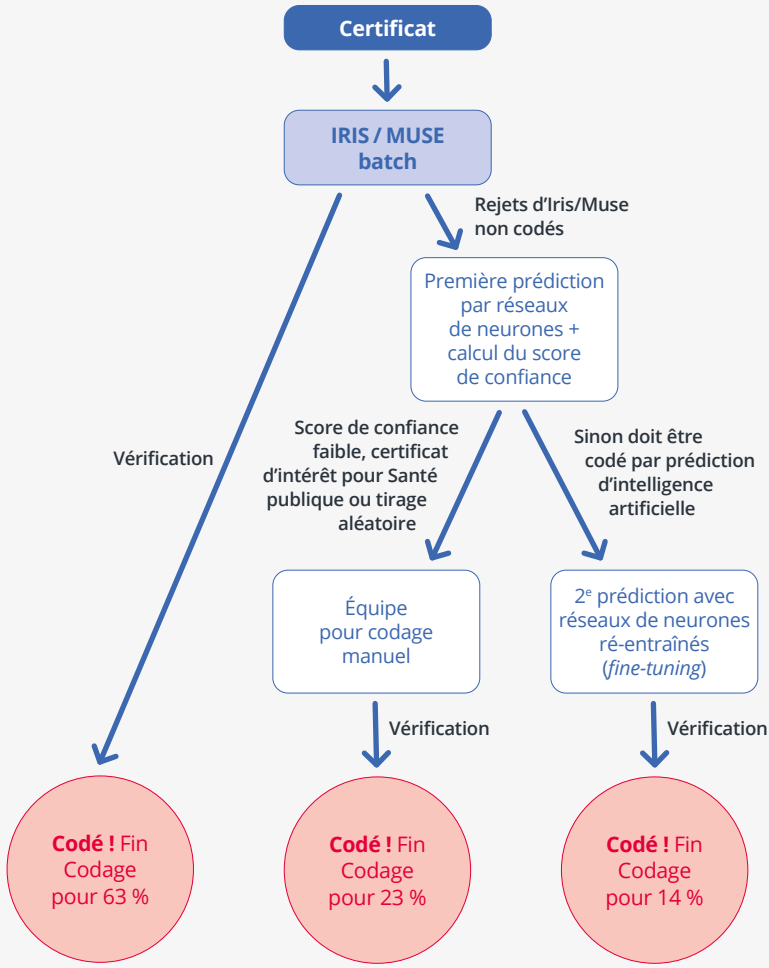
* *L'autre moitié va servir à tester les modèles d'IA.*

Articuler les modes de codage et cibler l'expertise humaine

La campagne de production articule entre eux les trois modes de codage de façon à maximiser la qualité de la production dans ses différentes dimensions (***figure 5***). Les certificats qui ne sont pas codés entièrement automatiquement par batch vont être affectés soit à un codage manuel assisté, soit à un codage par prédiction de réseaux de neurones profonds en respectant les principes suivants :

- par campagne, le volume de certificats pouvant être codés manuellement est fixé à l'avance, en fonction des ressources humaines disponibles dans l'équipe de codage, de façon à respecter les délais de diffusion de la base statistique (93 000 certificats en 2021) ;

Figure 5 - Circuit de codage d'un certificat pendant la campagne de production



> ***La campagne de production articule entre eux les trois modes de codage de façon à maximiser la qualité de la production dans ses différentes dimensions.***

- les situations pour lesquelles les prédictions des RNP sont moins vraisemblables, repérées sur la base d'un score de confiance modélisé et estimé pour chaque certificat, seront codées manuellement de façon ordonnée en commençant par les plus mauvaises. Il en sera de même pour certains cas d'intérêt pour la santé publique pour lesquels l'OMS requiert un suivi attentif : décès d'enfants en dessous d'un certain âge, morts maternelles, virus de l'immunodéficience humaine (VIH) ou syndrome d'immunodéficience acquise (SIDA) ;
- en complément, pour assurer l'entraînement et le contrôle réguliers des RNP, des échantillons tirés aléatoirement sont envoyés en reprise manuelle ;
- en fin de campagne, une partie de ces certificats codés manuellement, ainsi qu'une partie de ceux codés par batch automatique, sont utilisées pour ré-entrainer les RNP et obtenir la prédiction finale des certificats codés par RNP.

Ainsi, l'expertise humaine n'est plus mobilisée en bout de chaîne pour traiter les rejets du système de règles, mais dès le début pour alimenter les modèles et se concentrer sur les cas les plus complexes.

Au total, pour les décès en 2021, le batch aura codé 63 % des certificats, 23 % auront été codés en utilisant des prédictions de RNP (combinées ou non avec le système de règles), et 14 % auront demandé une intervention manuelle.

▶ Une évaluation statistique de la campagne annuelle de production

Introduire des RNP dans une production statistique nécessite de les « superviser », c'est-à-dire d'évaluer régulièrement leur performance en comparant leurs résultats à ceux des autres méthodes et de les réestimer régulièrement pour prendre en compte les nouveautés de vocabulaire, de description des pathologies ou les modifications dans les règles de codage.

La performance des RNP s'évalue en comparant la séquence de sortie prédite par le modèle avec la séquence véritablement observée sur un jeu de données déjà codées qui n'a pas été utilisé dans l'entraînement. Ce principe, essentiel en apprentissage statistique, permet de pallier le risque de sur-apprentisssage de ces modèles qui par nature comprennent plus de paramètres à estimer qu'il n'y a d'observations dans la base d'entraînement.

La stratégie d'évaluation de la qualité d'une campagne de production repose aussi sur ce principe. Sur un jeu de test de référence n'entrant pas dans les entraînements des RNP et représentatif de la distribution des causes de décès en population générale, on peut estimer la cohérence de la campagne de codage d'une année donnée en comparaison

à une campagne traditionnelle en simulant l'articulation des modes de codage telle qu'elle a été réalisée lors de la production. Ainsi, en 2021, en prenant en compte le fait que 63 % des certificats ont été codés par batch (codage par rapport à une campagne classique identique), dans 95,7 % des cas, le code de cause initiale issu de la campagne à trois modes de codage est le même que celui que l'on aurait obtenu par une campagne classique de codage. Dans 97,3 % des cas, ce code apparaît dans la même rubrique de la shortlist (sélection) européenne[23] que celui obtenu par une campagne classique de codage. En outre, la campagne à trois modes de codage respecte les délais impartis en ne reposant que sur 14 % de codage manuel contre 37% dans une campagne traditionnelle.

▶ Un exemple d'usage de l'IA pour la production de statistique publique

L'usage des RNP entraînés sur l'historique des données déjà annotées dans la production de la statistique sur les causes de décès a permis de rattraper le retard de production, accumulé au cours du temps, en réduisant drastiquement le nombre de certificats devant être codés manuellement (Clanché et alii, 2023 ; Zambetta et alii, 2023). Ainsi, en septembre 2023, la France a été en mesure de fournir les données définitives des années 2018 et 2019 avec seulement 3 % de codage manuel[24], ciblé évidemment, tout en garantissant une cohérence de cause initiale par rapport à une campagne traditionnelle à 93,4 % au niveau le plus fin de la CIM, et de 95,6 % au niveau de la *shortlist* européenne. Depuis, le calendrier régulier de production est raccourci, visant une fourniture des données 2023 dix-huit mois après la fin de l'année. Cet exemple témoigne de l'effet positif de l'intégration des procédures d'IA dans une production statistique régulière. Cette intégration est accessible et ouvre de nouvelles possibilités. Par exemple, un premier traitement purement automatique (RNP et batch) peut fournir des données provisoires[25] codées dès réception *(fast estimates* (premières estimations) utiles pour la veille sanitaire). Les précautions mises en œuvre pour guider le choix de la procédure s'appuient sur le code des bonnes pratiques de la statistique européenne : actualité, ponctualité, rapport coût-efficacité, etc. En outre, les architectures des réseaux ont été choisies pour leur simplicité. Leur entraînement et leur inférence peuvent être réalisés sur des infrastructures conventionnelles. Il a été décidé de ne pas s'appuyer sur des modèles pré-entraînés, ni très complexes, pour garder entièrement le contrôle de la procédure statistique, depuis les données d'entraînement jusqu'aux modèles. On assure ainsi réplicabilité et transparence et on limite les risques de biais dans un souci d'impartialité et d'objectivité. Enfin, l'évaluation statistique qui permet de contrôler les modèles est inhérente à ces procédures. Leurs erreurs peuvent être mesurées, analysées, documentées.

> ***Les précautions mises en œuvre pour guider le choix de la procédure s'appuient sur le code des bonnes pratiques de la statistique européenne : actualité, ponctualité, rapport coût-efficacité, etc.***

23 Principales marges de diffusion de la statistique européenne (Eurostat, 2012).
24 Ce taux manuel pour 2018-2019 est faible comparé à celui pour 2021 (14 %) du fait d'un rattrapage pour les années 2018 et 2019, c'est-à-dire qu'elles ont été traitées en parallèle de la campagne courante, courant 2022 et 2023.
25 Ainsi la diffusion des causes de décès en 2021 et en 2022 a été accompagnée de premières estimations sur les grandes causes de décès l'année suivante (Cadillac et alii, 2023, 2024).

Plusieurs axes de travaux se dessinent pour la suite : tout d'abord, sur l'explicabilité de ces modèles de façon à conforter la confiance des utilisateurs dans les statistiques produites à l'aide de l'intelligence artificielle, puis sur le maintien des modèles et des bases d'apprentissage (entraînement et test). Pour la statistique sur les causes de décès, le prochain défi sera aussi le passage à la version 11 de la CIM. Les modèles actuels sont totalement adhérents à la version de la nomenclature sur laquelle ils ont été entraînés, en l'occurrence la CIM 10. Passer à la CIM 11 implique d'adapter ou de changer les modèles, tout en maintenant la stratégie générale de la campagne de codage présentée ici.

▶ Fondements juridiques

- Règlement (CE) n° 1338/2008 du Parlement européen et du Conseil du 16 décembre 2008 relatif aux statistiques communautaires de la santé publique et de la santé et de la sécurité au travail. In : *site de l'Union européenne*. [en ligne]. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/ALL/?uri=CELEX:32008R1338*.

- Règlement (UE) n° 328/2011 de la Commission du 5 avril 2011 portant application du règlement (CE) n° 1338/2008 du Parlement européen et du Conseil relatif aux statistiques communautaires de la santé publique et de la santé et de la sécurité au travail, en ce qui concerne les statistiques sur les causes de décès. In : *site de l'Union européenne*. [en ligne]. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/ALL/?uri=CELEX%3A32011R0328*.

- Article L2223-42 du Code général des collectivités locales. In : *site de Légifrance*. [en ligne]. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000023711931/2024-04-25/*.

- Article R2213-1 du Code général des collectivités locales. In : *site de Légifrance*. [en ligne]. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000006395864*.

- Loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques. In : *site de Légifrance*. [en ligne]. Mis à jour le 25 mars 2019. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000888573*.

- Arrêté du 17 juillet 2017 relatif aux deux modèles du certificat de décès. In : *site de Légifrance*. [en ligne]. Mis à jour le 11 mai 2020. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000035388290*.

▶ Bibliographie

- AUBENQUE, Maurice, DAMIANI, Paul et DERUFFE, Louise, 1978. La mortalité par cause en France de 1925 à 1974. In : *Journal de la Société statistique de Paris*. Tome 119, n° 3, pp. 276-295. [en ligne]. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *http://www.numdam.org/item/JSFS_1978__119_3_276_0.pdf*.

- BABET, Damien, DELTOUR, Quentin, FARIA, Thomas et HIMPENS, Stéphanie, 2023. Les réseaux de neurones appliqués à la statistique publique : méthodes et cas d'usages. In : *Document de travail*. [en ligne]. Février 2023. Insee. N° M2023/01. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/6801092*.

- BOUVIER-COLLE, Marie-Hélène, VALLIN, Jacques et HATTON, Françoise, 1990. Mortalité et causes de décès en France. In : *Les éditions de l'INSERM. Collection Grandes enquêtes en santé publique et épidémiologie*. [en ligne]. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://gallica.bnf.fr/ark:/12148/bpt6k3325350b*.

- CADILLAC, Manon, FOUILLET, Anne, RIVERA, Cecilia, CLANCHÉ, François et COUDIN, Élise, 2023. Grandes causes de décès en France en 2021 : une année encore fortement marquée par le Covid-19. In : *Études et Résultats*. [en ligne]. Décembre 2023. N° 1288. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://drees.solidarites-sante.gouv.fr/publications-communique-de-presse/etudes-et-resultats/grandes-causes-de-deces-en-france-en-2021-une*.

- CADILLAC, Manon, FOUILLET, Anne, RIVERA, Cecilia et COUDIN, Élise, 2022. Les causes de décès en France en 2022 : recul du Covid-19 et hausse des maladies respiratoires. In : *Études et Résultats*. [en ligne] Octobre 2024. N° 1312. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://drees.solidarites-sante.gouv.fr/publications-communique-de-presse/etudes-et-resultats/241008_ER_les-causes-de-deces-2022*.

- CLANCHÉ, François, RAZAKAMANANA, Nirintsoa, COUDIN, Élise et ROBERT, Aude, 2023. Les statistiques provisoires sur les causes de décès en 2018 et 2019 - Une nouvelle méthode de codage faisant appel à l'intelligence artificielle. In : *Drees Méthodes*. [en ligne]. Mars 2023. N° 8. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://drees.solidarites-sante.gouv.fr/publications-jeux-de-donnees-communique-de-presse/drees-methodes/les-statistiques-provisoires-sur*.

- DESROSIÈRES, Alain, 1993. *La politique des grands nombres. Histoire de la raison statistique*. La Découverte. ISBN 978-2707165046.

- EUROSTAT, 2012. Causes of death, Eurostat Shortlist. In : *site de Statistics Finland*. [en ligne]. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://stat.fi/en/luokitukset/kuolinsyyt/kuolinsyyt_11_20140101*.

- EUROSTAT, 2024. Causes of death statistics. In : *site d'Eurostat*. [en ligne]. Mars 2024. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Causes_of_death_statistics*.

- FALISSARD, Louis, MORGAND, Claire, ROUSSEL, Sylvie, IMBAUD, Claire, GHOSN, Walid, BOUNEBACHE, Karim et REY, Grégoire, 2020. A Deep Artificial Neural Network-Based Model for Prediction of Underlying Cause of Death From Death Certificates: Algorithm Development and Validation. In : *JMIR Medical Informatics*. [en ligne]. Avril 2020. Volume 8, n° 4. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://doi.org/10.2196/1712*.

- PITA FERREIRA, Patricia, GODINHO SIMÕES, Diogo, PINTO DE CARVALHO, Constança, DUARTE, Francisco, FERNANDES, Eugénia, CASACA CARVALHO, Pedro, LOFF, José Francisco, SOARES, Ana Paula, ALBUQUERQUE, Maria João, PINTO-LEITE, Pedro, PERALTA-SANTOS, André, 2022. Real-Time Classification of Causes of Death Using Artificial Intelligence - Sensitivity Analysis. In : *European Journal of Public Health*. [en ligne]. Octobre 2022. Volume 32, Supplément 3. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://doi.org/10.1093/eurpub/ckac129.574*.

- FOUILLET, Anne, GHOSN, Walid, RIVERA, Cecilia, CLANCHÉ, François et COUDIN, Élise, 2023. Grandes causes de mortalité en France en 2021 et tendances récentes. In : *Bulletin épidémiologique hebdomadaire*. [en ligne]. 19 décembre 2023. N° 26. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://beh.santepubliquefrance.fr/beh/2023/26/2023_26_1.html*.

- FOUILLET, Anne, CADILLAC, Manon, RIVERA, Cecilia, et COUDIN, Élise, 2024. Grandes causes de mortalité en France en 2022 et tendances récentes. In : *Bulletin épidémiologique hebdomadaire*. [en ligne]. 8 octobre 2024. N° 18. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *http://beh.santepubliquefrance.fr/beh/2024/18/2024_18_1.html*.

- HEBBACHE, Zina, BOULET, Pierre, ROBERT, Aude, ZAMBETTA, Elisa, RAZAKAMANA, Daniel, COUDIN, Élise et MARTIN, Diane, 2024. Rapport de production : année de décès 2021. In : *Document de travail du CépiDc*. [en ligne]. Mars 2024. N° 8. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://www.cepidc.inserm.fr/documentation/rapport-de-production-annee-de-deces-2021*.

- IRIS INSTITUTE, 2024. Iris software. In : *site de Federal Institute for Drugs and Medical Devices*. [en ligne]. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://www.bfarm.de/EN/Code-systems/Collaboration-and-projects/Iris-Institute/Iris-software/_node.html*.

- LU, Tsung-Hsueh, TSAU, Shih-Ming et WU, Tzu-Chin, 2005. The Automated Classification of Medical Entities (ACME) system objectively assessed the appropriateness of underlying cause-of-death certification and assignment. In: *Journal of Clinical Epidemiology*. [en ligne]. Décembre 2005. Volume 58, n° 12, pp. 1277-1281. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *http://doi.org/10.1016/j.jclinepi.2005.03.017*.

- NATIONAL CENTER FOR HEALTH STATISTICS, 2023. MedCoder. In: *site du National Center for Health Statistics*. [en ligne]. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://www.cdc.gov/nchs/nvss/medcoder.htm*.

- NATIONAL CENTER FOR HEALTH STATISTICS, 2015. Mortality Medical Data System. In : *site du National Center for Health Statistics*. [en ligne]. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://www.cdc.gov/nchs/nvss/mmds.htm*.

- NAVARRA, Simone, CAPPELLA, Marisa, JOHANSSON, Lars Age, PELIKAN, László, FROVA, Luisa et GRIPPO, Francesco, 2016. Decision Table Editor: a web application for the management of the international tables for mortality coding. In: *Istat working papers*. [en ligne]. N° 6. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://www.istat.it/it/files/2016/04/IWP_06_2016.pdf*.

- ORGANISATION MONDIALE DE LA SANTÉ, 2008. Classification statistique internationale des maladies et des problèmes de santé connexes, dixième révision, Volume 2. In : *site de l'Organisation mondiale de la santé*. [en ligne]. Édition 2008. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://icd.who.int/browse10/Content/statichtml/ICD10Volume2_fr_2008.pdf*.

- PAVILLON, Gérard et LAURENT, Françoise, 2003. Certification et codification des causes médicales de décès. In : *Bulletin Epidémiologique Hebdomadaire*. [en ligne]. 8 juillet 2003, mis à jour le 30 août 2019, n° 30-31, p. 134-8. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://www.santepubliquefrance.fr/docs/certification-et-codification-des-causes-medicales-de-deces*.

- REY, Grégoire, 2016. Les données des certificats de décès en France : processus de production et principaux types d'analyse. In : *La Revue de Médecine Interne*. [en ligne]. Octobre 2016. Volume 37, n° 10, pp. 685-693. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *http://DOI.org/10.1016/j.revmed.2016.01.011*.

- VALLIN, Jacques et MESLÉ, France, 1988. Les causes de décès en France de 1925 à 1978. In : *Collection : Cahiers*. N° 115. ISBN 978-2-7332-0115-2.

- VASWANI, Ashish, SHAZEER, Noam, PARMAR, Niki, USZKOREIT, Jakob, JONES, Llion, GOMEZ, Aidan N., KAISER, Łukasz et POLOSUKHIN, Illia, 2017. Attention Is All You Need. In : *Advances in Neural Information Processing Systems*. [en ligne]. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf*.

- ZAMBETTA, Élisa, RAZAKAMANANA, Nirintsoa, ROBERT, Aude, CLANCHÉ, François, RIVERA, Cecilia, MARTIN, Diane, HEBBACHE, Zina, FLICOTEAUX, Rémi et COUDIN Élise, 2023. Codage des causes de décès de 2018 et 2019 en CIM10 - Approche combinant deep learning, système expert et codage manuel ciblé. In : *Document de travail du CépiDc N° 2*. [en ligne]. Septembre 2023. [Consulté le 6 novembre 2024]. Disponible à l'adresse : *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*.

- ZAMBETTA, Élisa, RAZAKAMANANA, Nirintsoa, ROBERT, Aude, CLANCHÉ, François, RIVERA, Cecilia, MARTIN, Diane, HEBBACHE, Zina, FLICOTEAUX, Rémi et COUDIN Élise, 2024. 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. In : *International Journal of Medical Informatics*. Août 2024. Volume188.