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

Methodological note on final 2018-2019 cause of death data

Combining a deep-learning-based approach, rule-based automated expert system and targeted manual coding for ICD-10 cause of death coding of French death certificates in 2018 - 2019

Elisa Zambetta, Nirintsoa Razakamanana, Aude Robert, François Clanché(*), Cecilia Rivera, Diane Martin, Zina Hebbache, Rémi Flicoteaux(**) and Elise Coudin
CépiDc-INSERM, (*) DREES, (**) APHP

September 2023

Document de travail du CépiDc N°2 -english version

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

Abstract

The ICD-10 coding of French cause of death (CoD) data for 2018 and 2019 combines fully-automatic batch coding by the rule-based system expert IRIS/MUSE, predictions by deep learning algorithms, and manual coding targeted at certificates of special interest for public health and research. This paper presents the supervised learning approach retained, including its use in targeting certificates sent to manual coding, and evaluates its performance. Compared to a traditional coding campaign relying only on IRIS/MUSE automatic batch coding and manual coding, the present campaign reaches 93.4% of accuracy for coding the underlying cause at the finest ICD-10 level and 95.5% at the European Short List level, with only 3% of manual coding.
The paper details also CoD categories for which differentials with a traditional coding campaign remain.

key -words: causes of death, mortality, ICD-10

Table des matières

1 Introduction

Causes of death (CoD) are usually coded from death records either by automated rule-based expert systems or manually by assisted coding using the same expert systems. The entire process requires significant human resources if expert systems are unable to automatically code a sufficient number of certificates, especially since determining the underlying cause according to ICD rules can be complex. In France, in 2018 and 2019, 38% of death certificates could not be automatically coded by IRIS/MUSE, the expert coding system, and a complementary traditional coding campaign based on assisted coding could not be carried out due to a lack of human resources. A new approach introducing neural network predictions (seq-to-seq algorithms) trained on previously coded data was therefore developed and applied. Thus, the 2018 and 2019 coding campaign combines three coding methods:[1] the use of predictions from seq-to-seq algorithms allows 34% of certificates to be coded, manual coding targeted at certificates of particular interest for public health (AIDS, maternal and infant deaths, research database) and those for which AI predictions have a low confidence index for 3% of certificates and automatic coding by batch from the rules system (Iris/Muse), 62%.

Table 1 shows the countings for each coding method and compares them with the provisional data released in December 2022.

Years\ Type of coding	Manual coding	AI-based coding	Fully rule-based automated coding with IRIS/MUSE	Total
Final 2018 Counts	18142	200217	376305	594664
Final 2018 %	3%	34%	63%	100%
Provisional 2018 - %	*0%*	*37%*	*63%*	*100%*
Final 2019 Counts	18805	196291	383611	598707
Final 2019 %	3%	33%	64%	100%
Provisional 2019 - %	*0%*	*38%*	*62%*	*100%*

Note: Missing certificates are excluded (around 15000 per year) added to the final data with R99 CoD.

Table 1 - Number of certificates per type of coding - Comparison between final and provisional data - Scope : all received certificates for 2018 and 2019.

Certificates coded manually (in assisted coding) are presented in Appendix A1. The reader is referred to the Report of statistics on causes of death 2018 and 2019 (CépiDc-Inserm working document, n°3, see the CépiDc website) for more details on the 2018, 2019 campaign (collection, coding, variables).

[1] Provisional data disseminated in December 2022 relied only on expert-system batch automated coding and AI automated coding. The manual coding phases were conducted between February and June 2023.

2 3-method coding campaign

The campaign combining the three coding methods is based on a loop between AI, expert system and manual coding. First, Transformer-type seq-to-seq algorithms are trained to predict the sequence of CoD and the underlying cause based on already coded data (including batch coding): yellow phase in Figure 1. An indicator of confidence in the prediction of the algorithms is also calculated for each certificate. This allows certificates for which the prediction is less certain to be targeted and sent for manual coding, thus complementing those manually coded due to public health reasons (step 1, pink phase in Figure 1). In the second step, the training databases are updated with the new manual codings, and some of the algorithms are retrained on these data (step 2, blue phase). In the final step, a specific algorithm (BiLSTM) performs a classification task and chooses between the different UC code proposals from different versions of the algorithms. In the end, ICD-coded data for 2018 and 2019 correspond to the AI-coded certificates, plus those batch-coded by the expert system and those for which manual coding was performed (step 3, green phase). All the elements of this process are described in detail below.

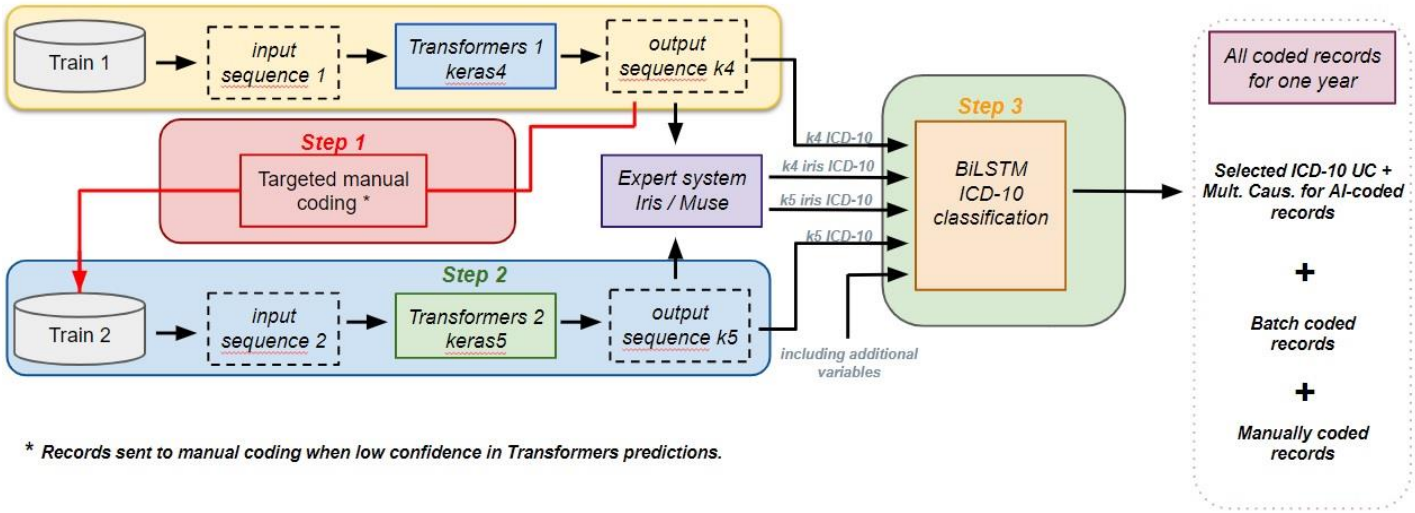


Figure 1 - 3-method coding campaign and loop between the 3 coding methods

3 CoD predictions with deep learning

The approach adopted is based on supervised learning. The algorithms used are neural networks and Transformer-type seq-to-seq translation models (see Vaswani et al 2017, Falissard et al. 2022). The same type of algorithms were used to code the provisional data (see Clanché et al. 2023). They are implemented using TensorFlow and Keras. They are used here both to predict multiple causes, to give a proposal for the underlying cause (which may or may not be retained), and to target the certificates that need to be manually coded as a priority (AI targeted manual coding). In practice, two models are used that differ in some of their features: the one used to predict provisional data (k4, see the report French metadata on provisional 2018 and 2019 CoD data, and Clanché et al., 2023) and an improved model (k5).

3.1 Model main specifications

- *Feature engineering/data pipeline*

The model input sequences are concatenations of the texts on each line of the certificate, separated by the line label, plus some additional variables. The additional variables systematically include gender, age group and year of death. They then differ depending on the model.

The first model (k4), used to predict provisional data, does not include any other additional variables.

The second model (k5) includes, in addition to the previous variables, the type of certificate (electronic or paper), the form of the certificate (1997 or 2017 version)[2], and the manner of death, a new variable introduced in the 2017 certificate versions to better identify external causes.

For k5, the input sequence is then composed as

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

The output sequence has the same structure as the input one, except that ICD codes replace texts/words and the manner of death is not repeated. The output sequence ends with the ICD-code of the underlying cause.

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

Example:

input sequence : *Paperback CertificateVersion2017 Women 55yo year2017 sepLine1 cardiorespiratory arrest sepLine2 pleural effusion sepLine3 lung metastases sepLine4 breast cancer sepLine7 natural death sepUC*

output sequence : *[start] Paperback CertificateVersion2017 Women 55yo year2017 sepLine1 r092 sepLine2 j90 sepLine3 c780 sepLine4 c509 sepLine7 sepUC c509 [end]*

"Tokenizer" is used to cut texts into tokens (words). The input dictionary contains 117,443 tokens and the output one, 6155 tokens.

- *Model architecture*

Transformer algorithms are of encoder/decoder type. Inputs are represented by their embedding in a vector space of finite size (512) and by the position of words in the sentence (positional encoding). The Transformers model encoder applies the same layers several times to the input sequence, combining a multi-headed attention mechanism model (to account for links between words) and a fully connected feed-forward network that captures position, followed by normalization. The decoder also repeats the

[2] The death certificate forms have changed as of 2018, but the use of the new form was gradual.

same layers on the output sequence, interposing a model of the attention mechanism at the encoder output. Each set of layers also ends with a fully connected feed-forward network and a normalization step. The decoder output then passes through a linear transformation and a softmax function to convert the decoder output into predicted probabilities of the next word. The algorithm contains 96,000,000 parameters (weights). See Appendix A2 for an illustration of the network architecture and the k5 model codes.

For k4 model specification, the reader is referred to previous documents such as the note French metadata on provisional CoD data, and Clanché et al. 2023. In the following, we focus on k5.

- *Training set*

The models are trained on already labelled death certificates, i.e. for which the sequence of multiple CoDs and the underlying cause are known

The training base for the k5 model contains 5,317,843 certificates, and consists of

-all labeled data from 2011 to 2015 (automatic and manual coding),

-all automatically (batch) coded certificates for 2016 and 2017, plus 300,000 observations randomly selected from those manually coded for 2016 and 2017

-all automatically (batch) coded certificates for 2018 and 2019, plus half of all manually coded certificates as of June 8, 2023 (50% share regardless of the coded sample)

-78% of 2020 batch coding and 56% of manual coding, always randomly drawn

-96% of 2021 automatic coding and 40% of manual coding as of June 8, 2023 (excluding EDP, left as test).

	Train		Test		To be predicted
	manual coding	automatic coding	manual coding	automatic coding	
2011-2015	2764209		0	0	0
2016-2017	299984	681122	187056	11	0
2018-2019	17534	745466	17740	0	412561
2020	156331	291795	121461	84026	0
2021	25836	389566	38830	18291	181023*
Total	5371843		467415		
*some of them will be manually coded till the end of 2023					

The validation set consists of 20% of the training set, randomly selected once for all before training.

- *Test*

The test, constructed with already coded certificates not included in the training set, contains 467,415 observations, of which 365,087 are manually coded.[3]

[3] In practice, there is an overlap between the test of k5 and the training set of k4, so we will check the performance on the sole intersection of tests when necessary, but the results reported in the document generally concern the test of k5 and iris5, the models which will ultimately be the most widely selected.

- *Training strategy*

The k5 model was first trained on an initial train/validation base comprising almost 5.3 million observations at the beginning of 2023. Then, the weights were re-estimated in a fine-tuning step consisting of 10 optimization epochs on the entire train/validation base including 42,328 observations from 2018, 2019, 2021 manually coded during the first half of 2023 corresponding to part of the manual recovery targeted for 2018 and 2019 and for 2021. This strategy is the result of a trade-off between the model's full training times (several days) and the completeness of the learning base.

3.2 Underlying cause determination

The output predicted by the model provides two suggestions for the underlying cause. It is indeed possible to use the underlying cause directly predicted by the algorithm, which is at the last position in the sentence. It is also possible to apply the IRIS/MUSE expert coding system to the sequence of multiple causes predicted by the algorithm, and to use the underlying cause to which it leads, when there is one. In addition, the two models k4 and k5 can propose different underlying causes, and different sequences of causes, which can lead to different underlying cause proposals when IRIS/MUSE is applied. Therefore, there are potentially 4 underlying cause suggestions- those coming directly from the k4 and k5 algorithms, and those after IRIS/MUSE is run on the sequences of causes predicted by k4 and k5, i.e. iris4 and iris5; as well as two predicted sequences of multiple causes. Note that if IRIS/MUSE does not conclude, the underlying cause directly predicted by the algorithm is used. In this case, there is only one suggestion per algorithm.

A “surmodel” is used, also based on supervised learning. This “surmodel” responds to a 5-class classification problem, indicating which of the preceding models will be retained to provide the underlying cause and, by extension, the multiple causes, or if none of the models leads to a good prediction (6% of the cases in the train). In the latter case, we use the iris5 prediction.

The input sequences of the surmodel include the codes for ICD-10 and European short-list (86 categories) for the underlying cause and for the multiple causes predicted by k4 and k5, the probabilities associated with the outputs of k4 and k5, the probability differences between the two most probable underlying causes (discriminating power), the type of certificate (electronic or paper), the manner of death, the number of multiple causes on the certificate (indicator of certificate complexity), and the number of times the models predict the same code for the underlying cause (indicator of reliability of this proposal). The input sequence is

“k4_UC k5_UC k4iris_UC k5iris_UC k4_86 k5_86 k5iris_86 k4iris_86 k4_multiple_causes k5_multiple_causes certificat_type age MannerOfDeath proba_max_k4 proba_diff_k4 proba_max_k5 proba_diff_k5 nb_causes_k4 nb_causes_k5 nb_equal”

The algorithm chosen is a bidirectional long-term short-term memory (BiLSTM, see Graves et al 2005, Baldi et al. 1999). Training is performed on the train intersection common to k4 and k5. The pre-processing, model architecture and codes are reported in Appendix A3. The number of times the same code is proposed and the codes predicted from the underlying cause at European shortlist level bring the most explanatory power to the model (Shapley values, see Appendix A3).

4 Using AI to target certificates to send to manual coding

In addition to certificates of particular public health interest or research interest, manual coding focuses on certificates for which AI predictions have a low confidence level. The targeting approach aims to achieve a given level of precision (90% or 92%) in each European shortlist category, ensuring that 3-method campaign codes match those of a traditional campaign 90% or 92% of the time.

For this, a confidence score is computed for each certificate. This confidence score allows us to prioritize certificates to be sent to manual coding. It depends on the underlying cause code predicted by k4, by k5 and on the variables the most discriminant to capture the certificate complexity. See Appendix A4 for a detailed presentation of the underlying linear probability model.

We focus on certificates in European shortlist categories for which we estimate, based on deaths in 2016 and 2017, that the precision, i.e. the number of correctly predicted underlying causes over the number of predicted causes in the category, does not reach 90% (P1), and then 92.5% (P2). We then simulate the additional manual coding rate that would be required to achieve these precisions, if the certificates with the lowest confidence indicators were sent for manual coding by order. These rates are then applied to the 2018/2019 counts. In practice, it was possible to manually code all certificates classified as P1 for 2018 and 2019, 64% of those classified as P2 for 2018, and 82% of those classified as P2 for 2019. Table 2 shows counts and proportions manually coded for each of the 12 problematic categories according to year.

	P1 (90%)	+P2 (92.5%)	% of targeted manual coding in 2018	% of targeted manual coding in 2019	average %
01.3- Viral hepatitis	101	76	0,30	0,32	0,31
01.4- Other infectious and parasitic diseases	408	408	0,08	0,09	0,09
03- Diseases of the blood and blood-forming organs	966	580	0,35	0,37	0,36
04.2- Autres maladies endocrinienness, nutritionnelles et metaboliques		418	0,03	0,04	0,04
05.3 - drug dependence, toxicomania	27	40	0,20	0,22	0,21
05.4 - Other mental and behavioural disorders	199	598	0,15	0,17	0,16
10 Diseases of the skin and subcutaneous tissue	201	201	0,16	0,18	0,17
11.1- Rheumatoid arthritis and osteoarthristis		30	0,03	0,04	0,04
11.2- Other diseases of the musculoskeletal system/connective tissue	759	570	0,30	0,32	0,31
12.1-Diseases of kidney and ureter		335	0,03	0,04	0,04
12.2- Other diseases of the genitourinary system		158	0,03	0,04	0,04
17.1.4 - Accidental poisoning	709	304	0,45	0,47	0,46
17.1.5 - Other accidents		1 517	0,06	0,08	0,07
17.3- Homicide, assault	37	186	0,21	0,25	0,23
17.4-Event of undetermined intent	237	158	0,21	0,23	0,22
17.5- Other external causes of injury and poisoning	3 114	389	0,86	0,88	0,87
Total	6 758	5 967			

Note: Columns 1 and 2: if we manually recode the 101 2018/2019 certificates for which the underlying cause predicted by k4 is viral hepatitis (01.3) and for which the confidence indicators are the lowest, we would reach an overall precision (including batch or other manual coding) of 90% for this category if we refer to the simulations built on the years 2016 and 2017. By coding the following 76, we would achieve 92.5%. The overall precision of a category is obtained by assuming that the certificates automatically coded by Iris/Muse and those manually coded are correct.

The last three columns report the % of data actually manually coded. 30% of the certificates that the k4 model classified as 01.3 were taken over manually in 2018, and 32% in 2019. In each case, these were the certificates with the lowest confidence indicators among those that k4 classified in this category.

Table 2 : number of certificates in 2018 /2019 to be coded manually to achieve a precision of 90% / 92.5% in total (i.e. taking in to account batch coding and all manual coding) and % of targeted manual coding achieved in practice for 2018 and 2019 data.

5 Performance analysis

5.1 Building a Reference Test Population

The test set, which consists of annotated observations that have been excluded from training, allows us to evaluate performance, i.e. the accuracy/consistency between the coding that would have been obtained in a conventional coding campaign combining batch coding and assisted manual coding and that of the 3-method approach.

This set includes 365,087 manually coded certificates for which multiple and underlying causes are also predicted by AI. This set is not representative of the distribution by cause of manual coding in a given year because it over-represents sensitive deaths and AI-targeted low confidence samples in certain years. It also over-represents Echantillon Demographique Permanent deaths, and is therefore unsuitable for evaluating manual coding targeted at these deaths.

To assess the accuracy between the final data of 2018 and 2019 and what would have been obtained after a conventional coding campaign, we limit this test set to respect the proportions of sensitive deaths and EDP deaths as observed in the total population of deaths, to also respect the proportion of targeted manual coding as performed in 2018 and 2019, and we complete the set in the right proportions of automatically batch-coded deaths. Thus, in the first stage, we focus only on the randomly drawn samples only (2016, 2017, 2020 manually coded test sets, and manually coded test random samples for 2021), i.e. 332,183 observations. The second stage consists of completing this base with proportional draws in automatic batch coding for each sub-sample.

We then obtain a reference test population of 797,651 observations that is representative of the distribution of causes of death over the years 2016, 2017, 2020 and 2021. The proportion of automatic batch coding in this population is 58%, which is slightly lower than the actual proportion of automatic batch coding in 2018 /2019. The consequence of this slight underestimation of automatic coding will therefore be a slight underestimation of coding accuracy.

We then simulate the contributions of targeted manual coding, assuming that the coded underlying cause is correct for certificates related to EDP, sensitive deaths, and AI-targeted manual coding. Appendix A5 details how to sample the batch to simulate a representative population, and how to identify these groups in the reference test population.

5.2 Overall accuracy

On the part of the reference test population that would have been manually coded in a conventional coding campaign, the underlying cause obtained by combining the "surmodel" prediction and targeted

manual coding matches the underlying cause coded by the coding team at the finest ICD level in 84.1% of cases. It falls into the same category in the European shortlist in 89.3% of the time. Table 4 reports the accuracy of the different models, combined or not with IRIS/MUSE and with targeted manual coding as performed in 2018 and 2019. At the finest ICD level, the k5 model prediction is correct in 78.5% of cases. Applying IRIS/MUSE to the sequence of causes predicted by k5 when it gives an unambiguous answer gains one point of accuracy. The performance of the k4 model, the one used for provisional data, is less good. However, the two models are complementary, since by combining them through the surmodel, the accuracy reaches 81.9%. Taking into account the targeted manual coding, the accuracy increases by another 2 points to 84.1%. The evaluation of each step in the targeted manual coding process will be described in detail below.

Manual	K5	K5IrisMuse	K4	K4IrisMuse	Surmodel	Surmodel+ Manual coding	Nobs
	ICD-10 4 digit level accuracy						
All	**0,785**	**0,796**	**0,768**	**0,769**	**0,819**	**0,841**	**332183**
2016	0,777	0,783	0,803	0,795	0,811	0,835	93144
2017	0,774	0,779	0,802	0,793	0,809	0,832	93912
2020	0,798	0,815	0,738	0,748	0,834	0,855	121461
2021	0,792	0,813	0,649	0,674	0,812	0,831	23666
	European short-list level accuracy						
All	**0,856**	**0,861**	**0,830**	**0,829**	**0,878**	**0,894**	**332183**
2016	0,851	0,853	0,867	0,857	0,874	0,890	93144
2017	0,848	0,849	0,866	0,857	0,870	0,886	93912
2020	0,865	0,874	0,794	0,801	0,889	0,903	121461
2021	0,860	0,873	0,736	0,754	0,873	0,888	23666

Reading: In 78.5% of cases, the underlying cause directly predicted by k5 exactly matches the manually coded one at the finest ICD level. In 85.6% of cases, the underlying cause predicted by k5 falls into the same Eurostat shortlist category as the manually coded underlying CoD.

Table 4: Accuracy of underlying cause predicted by deep learning (k4 or k5), combination of deep learning and IRIS/MUSE, surmodel combined or not with manual coding.

For the European shortlist, the surmodel gained 1.7 points of accuracy compared to iris5 (k5 combined with IRIS/MUSE), while the targeted manual coding gained 1.6 points. In total, the accuracy reaches 89.4%. Finally, the performance is stable over the years.

If we now take into account the fact that in 2018 and 2019 about 62-63% of the deaths are coded by batch, and that for these certificates the coding does not change compared to a conventional campaign, we obtain a perfect match in 93.4% of the cases at the finest ICD level and in 95.6% of the cases at the European shortlist level (Table 5).

Manual+batch	K5	K5IrisMuse	K4	K4IrisMuse	Surmodel	ınualCoding	Nobs
	ICD-10 4 digit level accuracy						
All	**0,910**	**0,915**	**0,903**	**0,904**	**0,925**	**0,934**	**797 651**
2016	0,906	0,909	0,917	0,914	0,921	0,931	221 807
2017	0,907	0,908	0,918	0,914	0,921	0,930	226 856
2020	0,914	0,921	0,889	0,893	0,929	0,938	285 784
2021	0,922	0,930	0,869	0,878	0,929	0,937	63 204
	European short-list level accuracy						
All	**0,940**	**0,942**	**0,929**	**0,929**	**0,949**	**0,956**	**797 651**
2016	0,937	0,938	0,944	0,940	0,947	0,954	221 807
2017	0,937	0,937	0,944	0,941	0,946	0,953	226 856
2020	0,942	0,947	0,912	0,916	0,953	0,959	285 784
2021	0,947	0,952	0,901	0,908	0,952	0,958	63 204

Reading: In 91.5% of cases, the 4-position UC obtained by batch coding where possible or by k5 prediction combined with IRIS/MUSE (iris5) is the same as that which would have been obtained by a conventional coding campaign combining batch and assisted manual coding only. This results in an accuracy of 94.2% for the European shortlist level.

Table 5: Accuracy of UC predicted by deep learning (k4 or k5), a combination of deep learning and IRIS/MUSE, surmodel combined or not with manual coding, and the UC coded in the general population (including batch).

5.3 Precision, recall and count differentials

Tables 6 and 7 show the precisions, recalls, F-measures and predicted counts per category at the European shortlist level, for the surmodel and when targeted manual coding is also taken into account. Precision is the proportion of correct predictions relative to all predictions in the category; recall is the proportion of observations correctly predicted by the model relative to all observations actually in the category; F-measure is the harmonic mean of the two.

Across the entire Reference Test Population, the combination of batch, "surmodel" and targeted manual coding campaign achieves very high levels of consistency (in terms of precision and recall) with a conventional coding campaign for most categories, with an average F-measure per category of 0.94. F-measures remain below 0.9 for 10 out of the 71 shortlist categories: viral hepatitis, blood and hematopoietic diseases, pharmacology, skin diseases, rheumatoid arthritis, other musculoskeletal diseases, genitourinary diseases, accidental intoxications, undetermined intentions and other external causes. This means that trends and counts in these categories should be interpreted with caution. In particular, we stress both statistically significant discrepancies and significant volume discrepancies (Poisson test) for :

03, blood diseases, underestimation of 7% of the expected number of deaths

11.2, other diseases of the musculoskeletal system, underestimation of 4% of the expected number of deaths

17.1.4, accidental poisoning, underestimation of 8% of the expected number of deaths

17.5, other external causes, underestimation of 37%.

Test that would have been manually coded	Real codes	Surmodel						Surmodel+ targeted manual coding					
		Precision	Recall	F-measure	Predictions	Pred./Real codes -1	Sign. of diff	Precision	Recall	F-measure	Predictions	Pred./Real codes -1	Sign. of diff
01.1- Tuberculosis	424	0,906	0,816	0,859	382	-9,9%	***	0,939	0,877	0,907	396	-6,6%	*
01.2- AIDS (HIV diseases)	260	0,784	0,685	0,731	227	-12,7%	***	0,974	1,000	0,987	267	2,7%	
01.3- Viral hepatitis	334	0,670	0,725	0,696	361	8,1%	*	0,769	0,796	0,782	346	3,6%	
01.4- Other infectious and parasitic diseases	5737	0,802	0,777	0,789	5560	-3,1%	***	0,835	0,815	0,825	5603	-2,3%	**
02.1.01-Malignant neoplasms of lip, oral cavity, pharynx	2761	0,938	0,895	0,916	2636	-4,5%	***	0,943	0,900	0,921	2634	-4,6%	***
02.1.02-Malignant neoplasms of oesophagus	2447	0,957	0,954	0,956	2438	-0,4%		0,959	0,957	0,958	2441	-0,2%	
02.1.03-Malignant neoplasms of stomach	2359	0,945	0,933	0,939	2330	-1,2%		0,951	0,937	0,944	2325	-1,4%	
02.1.04-Malignant neoplasms of colon, rectum, anus	9820	0,953	0,952	0,953	9811	-0,1%		0,956	0,955	0,956	9815	-0,1%	
02.1.05-Malignant neoplasms of liver and intrahepatic bile ducts	4782	0,937	0,926	0,932	4725	-1,2%		0,943	0,930	0,936	4717	-1,4%	
02.1.06-Malignant neoplasms of pancreas	5411	0,971	0,969	0,970	5395	-0,3%		0,974	0,971	0,972	5393	-0,3%	
02.1.07-Malignant neoplasms of larynx	654	0,895	0,875	0,885	639	-2,3%		0,905	0,884	0,894	639	-2,3%	
02.1.08-Malignant neoplasms of trachea, bronchus, lung	15882	0,952	0,951	0,951	15864	-0,1%		0,954	0,954	0,954	15880	0,0%	
02.1.09- Malignant neoplasms of skin	1168	0,916	0,930	0,923	1185	1,5%		0,921	0,933	0,927	1184	1,4%	
02.1.10-Malignant neoplasms of breast	6828	0,950	0,950	0,950	6824	-0,1%		0,954	0,954	0,954	6827	0,0%	
02.1.11-Malignant neoplasms of cervix uteri	524	0,931	0,929	0,930	523	-0,2%		0,946	0,937	0,942	519	-1,0%	
02.1.12-Malignant neoplasms of other and unspecified parts of uterus	1664	0,940	0,912	0,926	1613	-3,1%		0,948	0,916	0,932	1608	-3,4%	*
02.1.13-Malignant neoplasms of ovary	1785	0,953	0,947	0,950	1774	-0,6%		0,956	0,950	0,953	1773	-0,7%	
02.1.14-Malignant neoplasms of prostate	4825	0,944	0,938	0,941	4795	-0,6%		0,948	0,942	0,945	4796	-0,6%	
02.1.15-Malignant neoplasms of kidney	2147	0,943	0,908	0,925	2068	-3,7%	**	0,949	0,913	0,931	2067	-3,7%	**
02.1.16-Malignant neoplasms of bladder	2952	0,937	0,943	0,940	2972	0,7%		0,943	0,945	0,944	2960	0,3%	
02.1.17-Malignant neoplasms of brain and central nervous system	2205	0,932	0,926	0,929	2190	-0,7%		0,938	0,929	0,933	2185	-0,9%	
02.1.18-Malignant neoplasms of thyroid	267	0,916	0,861	0,888	251	-6,0%		0,916	0,861	0,888	251	-6,0%	
02.1.19-Hodgkin disease and lymphomas	3253	0,932	0,942	0,937	3289	1,1%		0,943	0,949	0,946	3276	0,7%	
02.1.20- Leukaemia	3643	0,937	0,948	0,942	3685	1,2%		0,944	0,953	0,948	3677	0,9%	
02.1.21-Other malignant neoplasms of lymphoid and haematopoietic tissue	1959	0,930	0,917	0,923	1933	-1,3%		0,940	0,928	0,933	1934	-1,3%	
02.1.22-Other malignant neoplasms	15015	0,855	0,884	0,869	15514	3,3%	****	0,864	0,892	0,878	15503	3,3%	****
02.2-Non-malignant neoplasms (benign and uncertain)	5261	0,842	0,851	0,846	5312	1,0%		0,856	0,864	0,860	5310	0,9%	
03 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	2033	0,727	0,628	0,674	1756	-13,6%	****	0,821	0,720	0,767	1781	-12,4%	****
04.1- Diabetes mellitus	7313	0,892	0,867	0,879	7108	-2,8%	***	0,902	0,877	0,889	7114	-2,7%	***
04.2- Other endocrine, nutritional and metabolic diseases	5987	0,801	0,777	0,789	5809	-3,0%	***	0,826	0,806	0,816	5846	-2,4%	**
05.1- Dementia	8407	0,848	0,911	0,879	9029	7,4%	****	0,858	0,919	0,888	9010	7,2%	****
05.2- Alcohol abuse (including alcohol psychosis)	1510	0,790	0,825	0,807	1576	4,4%	**	0,810	0,848	0,829	1582	4,8%	**
05.3 - drug dependence, toxicomania	199	0,722	0,613	0,663	169	-15,1%	***	0,840	0,739	0,786	175	-12,1%	**
05.4 - Other mental and behavioural disorders	2493	0,793	0,785	0,789	2466	-1,1%		0,836	0,826	0,831	2464	-1,2%	
06.1- Parkinson's disease	2915	0,913	0,927	0,920	2959	1,5%		0,919	0,932	0,926	2954	1,3%	
06.2 - Alzheimer's disease	7994	0,928	0,942	0,935	8117	1,5%	*	0,934	0,946	0,940	8098	1,3%	
06.3- Other diseases of the nervous system and the sense organs	7316	0,838	0,832	0,835	7258	-0,8%		0,855	0,853	0,854	7295	-0,3%	
07.1.1-Acute myocardial infarction	6433	0,883	0,905	0,894	6595	2,5%	***	0,889	0,913	0,901	6607	2,7%	***
07.1.2-Other ischaemic heart diseases	11020	0,871	0,870	0,870	11013	-0,1%		0,881	0,881	0,881	11027	0,1%	
07.2-Other heart diseases	23508	0,857	0,861	0,859	23625	0,5%		0,869	0,875	0,872	23660	0,6%	
07.3-Cerebrovascular diseases	18752	0,884	0,896	0,890	19014	1,4%	**	0,894	0,905	0,900	18986	1,2%	**
07.4- Other diseases of the circulatory system	14925	0,849	0,834	0,842	14669	-1,7%	***	0,865	0,854	0,859	14749	-1,2%	*
08.1 - Influenza	760	0,908	0,933	0,920	781	2,8%		0,920	0,941	0,930	777	2,2%	
08.2 - Pneumonia	4640	0,824	0,839	0,832	4726	1,9%		0,840	0,854	0,847	4718	1,7%	
08.3.1 - Asthma	425	0,847	0,821	0,834	412	-3,1%		0,859	0,842	0,850	417	-1,9%	
08.3.2-Other chronic lower respiratory diseases	5630	0,872	0,897	0,885	5788	2,8%	***	0,882	0,904	0,893	5769	2,5%	**
08.4- Other diseases of the respiratory system	6656	0,786	0,766	0,776	6482	-2,6%	***	0,805	0,785	0,795	6496	-2,4%	***
09.1 - Ulcer of stomach, duodenum, jejunum	598	0,847	0,855	0,851	603	0,8%		0,867	0,880	0,873	607	1,5%	
09.2 - Cirrhosis, fibrosis, and chronic hepatitis	4084	0,896	0,909	0,902	4144	1,5%		0,907	0,917	0,912	4131	1,2%	
09.3- Other diseases of the digestive system	11248	0,853	0,852	0,852	11244	0,0%		0,869	0,873	0,871	11298	0,4%	
10 Diseases of the skin and subcutaneous tissue	1185	0,754	0,776	0,765	1220	3,0%		0,823	0,822	0,823	1183	-0,2%	
11.1- Rheumatoid arthritis and osteoarthristis	435	0,806	0,736	0,769	397	-8,7%	**	0,843	0,777	0,809	401	-7,8%	*
11.2- Other diseases of the musculoskeletal system/connective tissue	3136	0,742	0,726	0,734	3069	-2,1%		0,835	0,787	0,811	2956	-5,7%	****
12.1-Diseases of kidney and ureter	4459	0,807	0,778	0,792	4300	-3,6%	***	0,830	0,802	0,816	4310	-3,3%	***
12.2- Other diseases of the genitourinary system	2352	0,808	0,789	0,799	2296	-2,4%		0,838	0,816	0,827	2292	-2,6%	
13 Complications of pregnancy, childbirth and puerperium	51	0,909	0,392	0,548	22	-56,9%	****	1,000	1,000	1,000	51	0,0%	
14 Certain conditions originating in the perinatal period	1762	0,930	0,945	0,938	1790	1,6%		0,991	1,000	0,995	1778	0,9%	
15 Congenital malformations and chromosomic abnormalities	1384	0,866	0,734	0,795	1173	-15,2%	****	0,916	0,842	0,877	1272	-8,1%	****
16.1- Sudden infant death syndrome	174	0,910	0,931	0,920	178	2,3%		0,972	0,983	0,977	176	1,1%	
16.2- Unknown and unspecified causes	4747	0,812	0,866	0,838	5062	6,6%	****	0,824	0,876	0,849	5044	6,3%	****
16.3- Other symptoms, signs, ill-defined causes	6428	0,834	0,878	0,855	6764	5,2%	****	0,845	0,888	0,866	6751	5,0%	****
17.1.1 - Transport accidents	2283	0,943	0,912	0,927	2209	-3,2%	*	0,949	0,924	0,936	2224	-2,6%	
17.1.2 - Accidental falls	8520	0,911	0,933	0,922	8720	2,3%	***	0,920	0,939	0,929	8700	2,1%	**
17.1.3 - Drowning and accidental submersion	395	0,825	0,896	0,859	429	8,6%	**	0,845	0,914	0,878	427	8,1%	*
17.1.4 - Accidental poisoning	1610	0,786	0,726	0,755	1488	-7,6%	****	0,889	0,796	0,840	1442	-10,4%	****
17.1.5 - Other accidents	12758	0,855	0,844	0,850	12598	-1,3%	*	0,881	0,870	0,876	12596	-1,3%	*
17.2 - Suicide and intentional self-harm	4999	0,925	0,920	0,923	4972	-0,5%		0,940	0,933	0,937	4963	-0,7%	
17.3- Homicide, assault	382	0,827	0,586	0,686	271	-29,1%	****	0,919	0,916	0,917	381	-0,3%	
17.4-Event of undetermined intent	1404	0,689	0,644	0,666	1312	-6,6%	***	0,813	0,716	0,761	1236	-12,0%	****
17.5- Other external causes of injury and poisoning	1570	0,485	0,320	0,386	1034	-34,1%	****	0,831	0,471	0,601	891	-43,2%	****
18- COVID	12936	0,945	0,967	0,956	13240	2,4%	****	0,949	0,970	0,959	13222	2,2%	***
Total	332183				332183						332183		

Note: significance levels of counting differentials come from equality tests assuming real occurrences were Poisson distributed., * pval<.2, ** pval<.1, *** pval<.05, **** pval<.01

Table 6 : Performance et predicted counts by surmodel and surmodel combined with targeted manual coding evaluated on certificates of the test reference population that would have been coded manually in a conventional coding campaign.

	Real codes	Surmodel						Surmodel+ targeted manual coding					
All test reference population		Precision	Recall	F-measure	Predictions	Pred./Real codes -1	Sign. of diff	Precision	Recall	F-measure	Predictions	Pred./Real codes -1	Sign. of diff
01.1- Tuberculosis	476	0,917	0,836	0,875	434	-8,8%	**	0,946	0,891	0,918	448	-5,9%	
01.2- AIDS (HIV diseases)	332	0,836	0,753	0,792	299	-9,9%	**	0,979	1,000	0,990	339	2,1%	
01.3- Viral hepatitis	560	0,797	0,836	0,816	587	4,8%		0,860	0,879	0,869	572	2,1%	
01.4- Other infectious and parasitic diseases	12936	0,914	0,901	0,907	12759	-1,4%	*	0,928	0,918	0,923	12802	-1,0%	
02.1.01-Malignant neoplasms of lip, oral cavity, pharynx	4996	0,966	0,942	0,954	4871	-2,5%	**	0,969	0,945	0,957	4869	-2,5%	**
02.1.02-Malignant neoplasms of oesophagus	4797	0,978	0,976	0,977	4788	-0,2%		0,979	0,978	0,979	4791	-0,1%	
02.1.03-Malignant neoplasms of stomach	5790	0,978	0,973	0,975	5761	-0,5%		0,980	0,974	0,977	5756	-0,6%	
02.1.04-Malignant neoplasms of colon, rectum, anus	23061	0,980	0,980	0,980	23052	0,0%		0,981	0,981	0,981	23056	0,0%	
02.1.05-Malignant neoplasms of liver and intrahepatic bile ducts	11426	0,974	0,969	0,971	11369	-0,5%		0,976	0,971	0,973	11361	-0,6%	
02.1.06-Malignant neoplasms of pancreas	15433	0,990	0,989	0,989	15417	-0,1%		0,991	0,990	0,990	15415	-0,1%	
02.1.07-Malignant neoplasms of larynx	1271	0,947	0,935	0,941	1256	-1,2%		0,951	0,940	0,946	1256	-1,2%	
02.1.08-Malignant neoplasms of trachea, bronchus, lung	40493	0,981	0,981	0,981	40475	0,0%		0,982	0,982	0,982	40491	0,0%	
02.1.09- Malignant neoplasms of skin	2241	0,956	0,963	0,960	2258	0,8%		0,958	0,965	0,962	2257	0,7%	
02.1.10-Malignant neoplasms of breast	16601	0,980	0,979	0,980	16597	0,0%		0,981	0,981	0,981	16600	0,0%	
02.1.11-Malignant neoplasms of cervix uteri	1048	0,966	0,965	0,965	1047	-0,1%		0,973	0,969	0,971	1043	-0,5%	
02.1.12-Malignant neoplasms of other and unspecified parts of uterus	3630	0,973	0,960	0,966	3579	-1,4%		0,976	0,961	0,969	3574	-1,5%	
02.1.13-Malignant neoplasms of ovary	4424	0,981	0,979	0,980	4413	-0,2%		0,982	0,980	0,981	4412	-0,3%	
02.1.14-Malignant neoplasms of prostate	11882	0,977	0,975	0,976	11852	-0,3%		0,979	0,976	0,978	11853	-0,2%	
02.1.15-Malignant neoplasms of kidney	4626	0,974	0,957	0,966	4547	-1,7%		0,977	0,960	0,968	4546	-1,7%	
02.1.16-Malignant neoplasms of bladder	6874	0,973	0,976	0,974	6894	0,3%		0,975	0,977	0,976	6882	0,1%	
02.1.17-Malignant neoplasms of brain and central nervous system	5232	0,971	0,969	0,970	5217	-0,3%		0,974	0,970	0,972	5212	-0,4%	
02.1.18-Malignant neoplasms of thyroid	490	0,956	0,924	0,940	474	-3,3%		0,956	0,924	0,940	474	-3,3%	
02.1.19-Hodgkin disease and lymphomas	6393	0,965	0,970	0,968	6429	0,6%		0,971	0,974	0,972	6416	0,4%	
02.1.20- Leukaemia	7856	0,971	0,976	0,973	7898	0,5%		0,974	0,978	0,976	7890	0,4%	
02.1.21-Other malignant neoplasms of lymphoid and haematopoietic tissue	4290	0,968	0,962	0,965	4264	-0,6%		0,973	0,967	0,970	4265	-0,6%	
02.1.22-Other malignant neoplasms	29282	0,925	0,940	0,932	29781	1,7%	****	0,929	0,945	0,937	29770	1,7%	****
02.2-Non-malignant neoplasms (benign and uncertain)	10175	0,918	0,923	0,920	10226	0,5%		0,925	0,930	0,927	10224	0,5%	
03 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	3491	0,851	0,783	0,816	3214	-7,9%	****	0,902	0,837	0,868	3239	-7,2%	****
04.1- Diabetes mellitus	16008	0,951	0,939	0,945	15803	-1,3%	*	0,956	0,944	0,950	15809	-1,2%	*
04.2- Other endocrine, nutritional and metabolic diseases	13704	0,915	0,903	0,909	13526	-1,3%	*	0,925	0,915	0,920	13563	-1,0%	
05.1- Dementia	25311	0,947	0,971	0,959	25933	2,5%	****	0,951	0,973	0,962	25914	2,4%	****
05.2- Alcohol abuse (including alcohol psychosis)	3230	0,900	0,918	0,909	3296	2,0%		0,909	0,929	0,919	3302	2,2%	
05.3 - drug dependence, toxicomania	308	0,831	0,750	0,788	278	-9,7%	**	0,901	0,831	0,865	284	-7,8%	*
05.4 - Other mental and behavioural disorders	4907	0,895	0,891	0,893	4880	-0,6%		0,917	0,912	0,914	4878	-0,6%	
06.1- Parkinson's disease	8866	0,971	0,976	0,974	8910	0,5%		0,973	0,978	0,975	8905	0,4%	
06.2 - Alzheimer's disease	25747	0,977	0,982	0,980	25870	0,5%		0,979	0,983	0,981	25851	0,4%	
06.3- Other diseases of the nervous system and the sense organs	15541	0,924	0,921	0,923	15483	-0,4%		0,932	0,931	0,931	15520	-0,1%	
07.1.1-Acute myocardial infarction	18023	0,957	0,966	0,962	18185	0,9%		0,960	0,969	0,964	18197	1,0%	*
07.1.2-Other ischaemic heart diseases	24438	0,942	0,941	0,942	24431	0,0%		0,946	0,946	0,946	24445	0,0%	
07.2-Other heart diseases	67415	0,950	0,952	0,951	67532	0,2%		0,954	0,956	0,955	67567	0,2%	
07.3-Cerebrovascular diseases	41319	0,947	0,953	0,950	41581	0,6%	*	0,952	0,957	0,954	41553	0,6%	
07.4- Other diseases of the circulatory system	33025	0,932	0,925	0,929	32769	-0,8%	*	0,939	0,934	0,937	32849	-0,5%	
08.1 - Influenza	1668	0,957	0,969	0,963	1689	1,3%		0,963	0,973	0,968	1685	1,0%	
08.2 - Pneumonia	16322	0,949	0,954	0,952	16408	0,5%		0,954	0,958	0,956	16400	0,5%	
08.3.1 - Asthma	1077	0,941	0,929	0,935	1064	-1,2%		0,945	0,938	0,941	1069	-0,7%	
08.3.2-Other chronic lower respiratory diseases	13006	0,944	0,955	0,950	13164	1,2%	*	0,948	0,958	0,953	13145	1,1%	
08.4- Other diseases of the respiratory system	21100	0,934	0,926	0,930	20926	-0,8%		0,939	0,932	0,936	20940	-0,8%	
09.1 - Ulcer of stomach, duodenum, jejunum	1081	0,915	0,920	0,917	1086	0,5%		0,926	0,933	0,930	1090	0,8%	
09.2 - Cirrhosis, fibrosis, and chronic hepatitis	8986	0,952	0,959	0,955	9046	0,7%		0,957	0,962	0,960	9033	0,5%	
09.3- Other diseases of the digestive system	22147	0,925	0,925	0,925	22143	0,0%		0,933	0,935	0,934	22197	0,2%	
10 Diseases of the skin and subcutaneous tissue	2067	0,857	0,872	0,864	2102	1,7%		0,899	0,898	0,898	2065	-0,1%	
11.1- Rheumatoid arthritis and osteoarthristis	726	0,888	0,842	0,864	688	-5,2%	*	0,909	0,866	0,887	692	-4,7%	
11.2- Other diseases of the musculoskeletal system/connective tissue	4537	0,823	0,811	0,817	4470	-1,5%		0,888	0,853	0,870	4357	-4,0%	****
12.1-Diseases of kidney and ureter	10646	0,921	0,907	0,914	10487	-1,5%	*	0,930	0,917	0,924	10497	-1,4%	*
12.2- Other diseases of the genitourinary system	4029	0,889	0,877	0,883	3973	-1,4%		0,906	0,893	0,899	3969	-1,5%	
13 Complications of pregnancy, childbirth and puerperium	54	0,920	0,426	0,582	25	-53,7%	****	1,000	1,000	1,000	54	0,0%	
14 Certain conditions originating in the perinatal period	2048	0,940	0,953	0,946	2076	1,4%		0,992	1,000	0,996	2064	0,8%	
15 Congenital malformations and chromosomic abnormalities	2105	0,917	0,825	0,869	1894	-10,0%	****	0,946	0,896	0,920	1993	-5,3%	***
16.1- Sudden infant death syndrome	179	0,913	0,933	0,923	183	2,2%		0,972	0,983	0,978	181	1,1%	
16.2- Unknown and unspecified causes	20174	0,953	0,968	0,961	20489	1,6%	***	0,957	0,971	0,964	20471	1,5%	***
16.3- Other symptoms, signs, ill-defined causes	40404	0,972	0,981	0,976	40740	0,8%	**	0,974	0,982	0,978	40727	0,8%	*
17.1.1 - Transport accidents	3678	0,965	0,945	0,955	3604	-2,0%		0,968	0,953	0,961	3619	-1,6%	
17.1.2 - Accidental falls	11146	0,932	0,948	0,940	11346	1,8%	**	0,938	0,953	0,946	11326	1,6%	**
17.1.3 - Drowning and accidental submersion	1090	0,933	0,962	0,948	1124	3,1%		0,941	0,969	0,955	1122	2,9%	
17.1.4 - Accidental poisoning	2163	0,844	0,796	0,819	2041	-5,6%	****	0,920	0,848	0,883	1995	-7,8%	****
17.1.5 - Other accidents	18254	0,899	0,891	0,895	18094	-0,9%		0,917	0,909	0,913	18092	-0,9%	
17.2 - Suicide and intentional self-harm	11281	0,967	0,965	0,966	11254	-0,2%		0,973	0,970	0,972	11245	-0,3%	
17.3- Homicide, assault	499	0,879	0,683	0,769	388	-22,2%	****	0,938	0,936	0,937	498	-0,2%	
17.4-Event of undetermined intent	1709	0,748	0,707	0,727	1617	-5,4%	***	0,850	0,767	0,806	1541	-9,8%	****
17.5- Other external causes of injury and poisoning	1847	0,594	0,422	0,493	1311	-29,0%	****	0,871	0,551	0,675	1168	-36,8%	****
18- COVID	35680	0,980	0,988	0,984	35984	0,9%	*	0,981	0,989	0,985	35966	0,8%	*
Total	797651				797651						797651		

Note: significance levels of counting differentials come from equality tests assuming real occurrences were Poisson distributed., * pval<.2, ** pval<.1, *** pval<.05, **** pval<.01

Table 7 : Performance et predicted counts by surmodel and surmodel combined with targeted manual coding evaluated on all Test Reference Population (including batch coded certificates).

There is also a (smaller) risk of overestimation for 02.1.22 other malignant tumours and 05.1, dementia.

As expected, the targeted manual coding improves consistency with a conventional campaign. In particular, especially for targeted categories: i.e. sensible deaths - 01.2, HIV/Aids, 13 pregnancy

complications, 14 15 et 16.1 perinatality, congenital malformations child sudden death, which are representative of young child deaths ; i.e. also AI-targeted categories - 01.3 viral hepatitis, 03 blood diseases, 05.3 pharmacology, 11.2 other diseasees of the musculoskeletal system, 17.1.4 accidental poisoning , 17.3 homicides and assaults, 17.4 undetermined intentions and 17.5 other external causes; and categories for which a special final manual coding was done at the end of the campaign : risk of tuberculosis (01.1), homicides and assaults (17.3) and pharmacology (05.3).

The retained approach of targeted manual coding based on the simulated short-list category precisions also seems to improve the recall measures. Finally, we reach 90% of precision for each European shortlist category except for 01.3 viral hepatitis;[4] for 17.4 undetermined intention et 17.5 other external causes.[5]

5.4 Details on performance gains of each step of the targeted manual coding

This part details the gains in accuracy /performance of each step on the targeted manual coding.

	Test population that should have been coded manually			All Test Reference Population	
	ICD-10 4 digit level	European short-list	% Manual coding	ICD-10 4 digit level	European short-list
Surmodel	0,819	0,878	-	0,925	0,949
+ special interest deaths	0,823	0,880	0,016	0,926	0,950
+ random sample	0,831	0,885	0,044	0,930	0,952
+ low AI confidence sample	0,840	0,893	0,022	0,934	0,955
+ last verifications	0,841	0,894	0,002	0,934	0,956
Nb obs.	332 183	332 183		797 651	797 651

Reading: 81.9% of the UCs predicted by the surmodel match the ICD-10 coded UC, 82.3% when including deaths of special public health interest, which represent 1.6% of the test population that would have been manually coded in a traditional campaign.

Table 8 - Accuracy of the underlying cause predicted by the surmodel combined with each targeted manual coding step.

The targeted manual coding improves the accuracy by 2.2 points on the test population that would have been manually coded in a traditional coding campaign, increasing this accuracy from 81.9% to 84.1%. However, the performance contributions of each step differ. If we relate the increase in accuracy to the percentage these certificates represent in total, we see that coding a sensitive death is 1.6 times more effective than coding a randomly selected certificate, and coding a certificate targeted by AI 2.4 times more effective. This can provide information on the proportions of manual coding to be allocated to these different stages, without neglecting the contribution to the quality of the training dataset and taking into account the importance of a human view on death certificates of particular interest for public health use.

[4] Counts of viral hepatitis were overestimated in 2017 (corrected 2017 numbers are expected by the end of 2023). This can explain the discrepancies shown.

[5] For those two categories, discrepancies could be related the introduction of the 2017 death certificate form. The latter introduced by the end of 2017/ beginning of 2018, asks the certifier to report the manner of death. They could also be related to an improved data collection since 2018 with medical legal institutes providing data directly from the internal IT systems.

5.5 *Comparison with provisional data*

To compare the final data with the provisional data disseminated in the winter 2022-2023, we simulate the coding that we would have obtained by applying the approach used for the provisional data on the reference test population (without COVID since k4 did not predict COVID). This approach uses the predictions of the k4 model (trained on a smaller sample than that used for k5), runs IRIS/MUSE on these predictions, and performs an *ad hoc* synthesis to choose between the two underlying cause proposals.[6]

	Accuracy - all test population	
	ICD-10 level	Eur. Short List Level
Final data - surmodel	0,922	0,947
Prov. data - retained model	*0,914*	*0,943*
Final data - surmodel + targeted manual coding	0,932	0,954
Prov. data - retained model + targeted manual coding	*0,926*	*0,950*

Reading: 91.4% of the certificates of the reference test population (excluding COVID) would have been correctly predicted if we had used the same strategy (k4 model, iris4 and summary model) as for the provisional data.

Tableau 9 : Comparison of UC accuracies resulting from the strategy adopted for the provisional data, and that for the final data.

At the most detailed level of the ICD and across the entire test population, the provisional data would result in an accuracy of 91.4%. For the final data, which combines the surmodel and targeted manual coding, this reaches to 93.2% (excluding COVID), i.e. +1.8 points, which breaks down into +1.2 points of gain coming from the targeted manual coding and +0 .6 points coming from the sophistication of the AI models.

6 Final results for 2018 and 2019- counts and standardized mortality rates

In the following two tables, counts and standardized mortality rates for 2018 and 2019 (final data) are compared with those for 2015, 2016, 2017 and 2020. The tables also indicate the categories of the European shortlist for which the F-measures are below 0.9, and whether this may entail a risk of under- or overestimation of counts and rates, as well as the categories for which an over-or underestimation was found based on significance level of the Poisson tests.

[6] Strictly speaking, the synthesis model used for the provisional data also mobilized the predictions of an additional model re-estimated only on the observations for which Iris/Muse did not propose a unique underlying cause code. The (weak) contribution of this model was not reproduced in this simulation.

Countings	2015	2016	2017	2018 def	2019 def	2020	Risk
01.1- Tuberculosis	434	403	402	351	347	295	
01.2- AIDS (HIV diseases)	390	334	237	241	237	202	
01.3- Viral hepatitis	600	587	773*	423	390	351	overest. (F)
01.4- Other infectious and parasitic diseases	9797	9180	10193	10289	10879	10208	
02.1.01-Malignant neoplasms of lip, oral cavity, pharynx	3924	3936	3809	3713	3527	3636	underest. (P)
02.1.02-Malignant neoplasms of oesophagus	3893	3902	3865	3772	3784	3630	
02.1.03-Malignant neoplasms of stomach	4559	4602	4612	4614	4474	4258	
02.1.04-Malignant neoplasms of colon, rectum, anus	17658	18029	17996	17327	17360	17197	
02.1.05-Malignant neoplasms of liver and intrahepatic bile ducts	8518	8776	8551	8536	8579	8727	
02.1.06-Malignant neoplasms of pancreas	10921	11300	11467	11774	12199	12476	
02.1.07-Malignant neoplasms of larynx	1091	1069	1000	946	868	827	
02.1.08-Malignant neoplasms of trachea, bronchus, lung	32150	31877	31402	31054	30957	30935	
02.1.09- Malignant neoplasms of skin	1850	1748	1767	1762	1825	1756	
02.1.10-Malignant neoplasms of breast	12580	12936	13013	12958	12834	13008	
02.1.11-Malignant neoplasms of cervix uteri	763	801	817	858	779	769	
02.1.12-Malignant neoplasms of other and unspecified parts of uterus	2755	2838	2903	2910	2886	2845	
02.1.13-Malignant neoplasms of ovary	3491	3495	3545	3373	3495	3341	
02.1.14-Malignant neoplasms of prostate	8919	9022	9212	9271	9302	9178	
02.1.15-Malignant neoplasms of kidney	3640	3597	3612	3443	3325	3483	
02.1.16-Malignant neoplasms of bladder	5230	5349	5146	5331	5218	5345	
02.1.17-Malignant neoplasms of brain and central nervous system	3885	3964	4087	3812	4064	4035	
02.1.18-Malignant neoplasms of thyroid	412	378	420	433	388	362	
02.1.19-Hodgkin disease and lymphomas	4843	4869	4936	4670	4766	4875	
02.1.20- Leukaemia	5936	6016	6134	6008	6012	6165	
02.1.21-Other malignant neoplasms of lymphoid and haematopoietic tissue	3385	3433	3230	3296	3352	3283	
02.1.22-Other malignant neoplasms	21315	21738	22106	22739	23338	23018	overest. (P)
02.2-Non-malignant neoplasms (benign and uncertain)	7441	7527	7587	7691	7741	7656	
03 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	2207	2291	2570	2876	2784	2802	underest. (F)
04.1- Diabetes mellitus	12268	11848	11927	11419	11424	12264	
04.2- Other endocrine, nutritional and metabolic diseases	9357	9407	10189	10517	10981	11333	
05.1- Dementia	19309	19755	19661	21306	21003	18595	overest. (P)
05.2- Alcohol abuse (including alcohol psychosis)	2594	2577	2460	2680	2672	2472	
05.3 - drug dependence, toxicomania	160	230	189	219	241	229	underest. (F)
05.4 - Other mental and behavioural disorders	3344	3452	3608	3809	3926	4090	
06.1- Parkinson's disease	6192	6642	6826	6912	6828	7013	
06.2 - Alzheimer's disease	20872	21111	20962	20457	19251	18243	
06.3- Other diseases of the nervous system and the sense organs	10944	11128	11782	12275	12730	12359	
07.1.1-Acute myocardial infarction	14659	14031	13976	13450	13270	12922	
07.1.2-Other ischaemic heart diseases	19310	18985	19053	19028	18461	18170	
07.2-Other heart diseases	53623	53184	53652	54918	50894	48060	
07.3-Cerebrovascular diseases	32176	32213	31776	31780	31969	31112	
07.4- Other diseases of the circulatory system	25019	25117	25165	24477	24034	24497	
08.1 - Influenza	1915	961	2501	2297	2795	871	
08.2 - Pneumonia	13371	13305	13920	14313	14518	11559	
08.3.1 - Asthma	891	929	914	847	840	719	
08.3.2-Other chronic lower respiratory diseases	10746	10416	10747	10910	10787	9373	
08.4- Other diseases of the respiratory system	15811	15722	16675	16741	16571	16186	
09.1 - Ulcer of stomach, duodenum, jejunum	853	867	862	819	815	837	
09.2 - Cirrhosis, fibrosis, and chronic hepatitis	7056	6914	6775	6749	6715	6777	
09.3- Other diseases of the digestive system	16081	16396	16533	16830	17355	17363	
10 Diseases of the skin and subcutaneous tissue	1379	1489	1623	1519	1656	1639	amb. sign
11.1- Rheumatoid arthritis and osteoarthristis	555	565	578	585	528	583	underest. (F)
11.2- Other diseases of the musculoskeletal system/connective tissue	3651	3589	3424	3194	3459	3440	underest. (F)
12.1-Diseases of kidney and ureter	7637	7572	8105	7695	8333	8579	
12.2- Other diseases of the genitourinary system	2461	2550	2752	2950	3122	3511	underest. (F)
13 Complications of pregnancy, childbirth and puerperium	40	40	41	39	32	41	
14 Certain conditions originating in the perinatal period	1571	1501	1685	1622	1558	1443	
15 Congenital malformations and chromosomic abnormalities	1694	1675	1624	1489	1600	1502	underest (P)
16.1- Sudden infant death syndrome	165	176	139	184	132	114	
16.2- Unknown and unspecified causes	25361	27198	29680	30442	34736	34657	
16.3- Other symptoms, signs, ill-defined causes	29163	28069	29700	31385	32448	32999	overest (P)
17.1.1 - Transport accidents	3199	3186	3054	2692	2568	2144	
17.1.2 - Accidental falls	7684	7781	8262	8902	9008	9073	overest (P)
17.1.3 - Drowning and accidental submersion	904	920	884	857	719	668	
17.1.4 - Accidental poisoning	2042	1800	1725	1366	1236	1505	underest. (F)
17.1.5 - Other accidents	13991	13694	14202	13240	14085	14271	
17.2 - Suicide and intentional self-harm	9118	8591	8367	8868	8822	8986	
17.3- Homicide, assault	336	312	281	437	474	472	
17.4-Event of undetermined intent	873	785	1102	1275	1394	1552	underest. (F)
17.5- Other external causes of injury and poisoning	844	1391	1525	1855	1713	1361	underest. (F)
18- COVID	0	0	0	0	0	69249	

Note : over/underest (F) denotes risk of over/underestimation of countings and F <.90 ; over/underest (P) denotes risk of over/underestimation of countings indicated by Poisson tests of differentials are significant at 5%.

Table 10 : Counts per CoD of the European shortlist from 2015 to 2020, with indication of risk of over/underestimation in 2018 and 2019 (final data).

Standardized mortality rates	2015	2016	2017	2018 def	2019 def	2020	Risk
01.1- Tuberculosis	0,6	0,6	0,6	0,5	0,5	0,4	
01.2- AIDS (HIV diseases)	0,6	0,5	0,4	0,4	0,4	0,3	
01.3- Viral hepatitis	0,9	0,9	1,2	0,6	0,6	0,5	overest. (F)
01.4- Other infectious and parasitic diseases	14,6	13,2	14,2	14,1	14,7	13,6	
02.1.01-Malignant neoplasms of lip, oral cavity, pharynx	6,4	6,3	6	5,8	5,4	5,5	underest. (P)
02.1.02-Malignant neoplasms of oesophagus	6,5	6,3	6,2	5,9	5,9	5,5	
02.1.03-Malignant neoplasms of stomach	7,3	7,3	7,2	7	6,7	6,3	
02.1.04-Malignant neoplasms of colon, rectum, anus	27,5	27,4	26,8	25,4	25	24,4	
02.1.05-Malignant neoplasms of liver and intrahepatic bile ducts	14,1	14,2	13,6	13,4	13,2	13,1	
02.1.06-Malignant neoplasms of pancreas	17	17,3	17,2	17,4	17,6	17,8	
02.1.07-Malignant neoplasms of larynx	1,9	1,8	1,7	1,5	1,4	1,3	
02.1.08-Malignant neoplasms of trachea, bronchus, lung	53,1	51,7	50,1	48,7	47,6	46,9	
02.1.09- Malignant neoplasms of skin	2,9	2,7	2,7	2,7	2,7	2,6	
02.1.10-Malignant neoplasms of breast	16,8	16,9	16,8	16,5	16	16	
02.1.11-Malignant neoplasms of cervix uteri	1,1	1,1	1,1	1,2	1,1	1,1	
02.1.12-Malignant neoplasms of other and unspecified parts of uterus	3,6	3,7	3,7	3,6	3,6	3,5	
02.1.13-Malignant neoplasms of ovary	4,7	4,7	4,7	4,4	4,5	4,2	
02.1.14-Malignant neoplasms of prostate	17,5	17,2	17,1	16,8	16,4	15,9	
02.1.15-Malignant neoplasms of kidney	5,9	5,7	5,7	5,3	4,9	5,1	
02.1.16-Malignant neoplasms of bladder	9	9	8,5	8,5	8,2	8,3	
02.1.17-Malignant neoplasms of brain and central nervous system	6,1	6,2	6,3	5,8	6,1	6	
02.1.18-Malignant neoplasms of thyroid	0,6	0,6	0,6	0,6	0,5	0,5	
02.1.19-Hodgkin disease and lymphomas	7,6	7,5	7,4	7	7	7	
02.1.20- Leukaemia	9,4	9,3	9,3	8,9	8,8	8,9	
02.1.21-Other malignant neoplasms of lymphoid and haematopoietic tissue	5,3	5,3	4,9	4,8	4,8	4,7	
02.1.22-Other malignant neoplasms	33,6	33,5	33,5	33,8	34,1	33	overest. (P)
02.2-Non-malignant neoplasms (benign and uncertain)	11,4	11,2	11	10,9	10,7	10,4	
03 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	3,2	3,3	3,6	4	3,8	3,7	underest. (F)
04.1- Diabetes mellitus	18,4	17,4	17	16	15,7	16,5	
04.2- Other endocrine, nutritional and metabolic diseases	13,4	13	13,5	13,7	14,1	14,4	
05.1- Dementia	26,3	25,6	24,4	25,9	24,8	21,4	overest. (P)
05.2- Alcohol abuse (including alcohol psychosis)	4,2	4,1	3,9	4,2	4,2	3,8	
05.3 - drug dependence, toxicomania	0,3	0,4	0,3	0,4	0,4	0,4	underest. (F)
05.4 - Other mental and behavioural disorders	5	4,9	5,1	5,2	5,3	5,4	
06.1- Parkinson's disease	9,7	10,2	10,2	10,2	9,9	10	
06.2 - Alzheimer's disease	26,8	26,2	25,2	24	22,1	20,7	
06.3- Other diseases of the nervous system and the sense organs	16,9	16,7	17,4	17,9	18,2	17,5	
07.1.1-Acute myocardial infarction	22,7	21,3	20,8	19,6	19	18,4	
07.1.2-Other ischaemic heart diseases	30,5	29,1	28,3	27,7	26,5	25,8	
07.2-Other heart diseases	77,6	74,1	72,5	72,3	65,1	60,9	
07.3-Cerebrovascular diseases	46	44,9	42,9	42,1	41,7	39,9	
07.4- Other diseases of the circulatory system	36,1	34,9	33,9	32,4	30,9	31,1	
08.1 - Influenza	2,7	1,4	3,4	3,2	3,7	1,2	
08.2 - Pneumonia	20	19,2	19,4	19,6	19,1	15,4	
08.3.1 - Asthma	1,2	1,2	1,2	1,1	1,1	0,9	
08.3.2-Other chronic lower respiratory diseases	17,4	16,4	16,4	16,2	15,6	13,6	
08.4- Other diseases of the respiratory system	24	23	23,6	23,4	22,5	21,9	
09.1 - Ulcer of stomach, duodenum, jejunum	1,3	1,3	1,2	1,1	1,1	1,1	
09.2 - Cirrhosis, fibrosis, and chronic hepatitis	11,4	11,1	10,7	10,5	10,3	10,3	
09.3- Other diseases of the digestive system	23,7	23,6	23,2	23,1	23,4	23	
10 Diseases of the skin and subcutaneous tissue	1,9	2	2,1	1,9	2,1	2	amb. sign
11.1- Rheumatoid arthritis and osteoarthristis	0,7	0,7	0,7	0,7	0,6	0,7	underest. (F)
11.2- Other diseases of the musculoskeletal system/connective tissue	5,3	5,1	4,8	4,4	4,6	4,5	underest. (F)
12.1-Diseases of kidney and ureter	11,6	11	11,4	10,6	11,3	11,4	
12.2- Other diseases of the genitourinary system	4	3,9	4,1	4,2	4,4	4,8	underest. (F)
13 Complications of pregnancy, childbirth and puerperium	0,1	0,1	0,1	0,1	0,1	0,1	
14 Certain conditions originating in the perinatal period	2	1,9	2,2	2,1	2,1	2	
15 Congenital malformations and chromosomic abnormalities	2,4	2,4	2,3	2,1	2,3	2,2	underest (P)
16.1- Sudden infant death syndrome	0,2	0,2	0,2	0,2	0,2	0,2	
16.2- Unknown and unspecified causes	38,3	39,9	42,3	42,7	47,5	46,6	
16.3- Other symptoms, signs, ill-defined causes	41,4	38,3	38,9	39,9	40,4	40,1	overest (P)
17.1.1 - Transport accidents	5	5	4,8	4,2	4	3,3	
17.1.2 - Accidental falls	11,4	11,2	11,5	12	11,9	11,8	overest (P)
17.1.3 - Drowning and accidental submersion	1,4	1,4	1,4	1,3	1,1	1	
17.1.4 - Accidental poisoning	3,1	2,7	2,6	2	1,8	2,2	underest. (F)
17.1.5 - Other accidents	20,9	19,9	20	18,2	19	19	
17.2 - Suicide and intentional self-harm	14,8	13,9	13,4	14,1	13,9	14,1	
17.3- Homicide, assault	0,5	0,5	0,4	0,7	0,7	0,7	
17.4-Event of undetermined intent	1,4	1,3	1,8	2	2,2	2,4	underest. (F)
17.5- Other external causes of injury and poisoning	1,2	2,1	2,2	2,7	2,4	1,9	underest. (F)
18- COVID	0	0	0	0	0	93,4	

Note : over/underest (F) denotes risk of over/underestimation of countings and F <.90 ; over/underest (P) denotes risk of over/underestimation of countings indicated by Poisson tests of differentials are significant at 5%.

Table 11 : standardized mortality rate per CoD of the European shortlist, with indication of risk of over/underestimation for 2018 - 2019 final data.

Viral hepatitis (01.3) is slightly overestimated in 2018 and 2019 (following an error detected in 2017, 2% on the test numbers), but this does not affect the decreasing trend since 2015 (with 2017 corrected) in numbers and (very slowly) in SMRs.

Among the tumors, oral cancers (02.1.01) are probably slightly underestimated (-3% on the test numbers), the increase in the number of deaths between 2019 and 2020 must therefore be interpreted with caution, an increase that will not be not found on SMRs.

Other malignant tumors (02.1.22) are likely to be overestimated (2% of the test numbers), which could contribute to the apparent increase in the SMRs in 2018 and 2019 . This apparent increase is therefore not interpretable.

Blood diseases (03) are slightly underestimated (-7% of the test numbers). The increase in numbers and SMRs may be greater between 2017 and 2018.

Dementia (05.1) may be overestimated (2% of the test numbers) and the decrease in numbers and SMRs may be smaller between 2019 and 2020.

Drug dependence (05.3) may be underestimated (-8% of the test numbers). As the numbers in this category are very low, the rates (reported per 100,000 persons) are not affected.

Skin diseases (10) may not always be well identified but this does not lead to errors in test numbers and rates (the errors compensate for each other).

Rheumatoid arthritis (11.1) and other diseases of the osteoarticular system (11.2) may be underestimated (-5% and -4% of test numbers), so, changes in numbers and SMRs should be interpreted with caution.

Other diseases of the genitourinary system (12.2) may be underestimated (-2%). The upward trend in numbers is confirmed.

Congenital malformations (15) may be underestimated (-5% on the test numbers). The downward trend in numbers and rates since 2019 may be stronger.

Unknown and unspecified causes (16.2) may be slightly overestimated (2% of the test sample).

Accidental falls (17.1.2) may be slightly overestimated (2% of the test population). The increase in numbers between 2019 and 2020 is therefore perhaps more accentuated in reality than when reading the series.

Accidental poisoning (17.1.4) may be underestimated (-8% on test numbers). The drop in the rate between 2019 and 2020 is probably a little steeper.

Undetermined intentions (17.4) may be underestimated (-6%). The potential underestimation is greatest for other external causes (17.5) (-37% on the test population). Trends in this category should not be interpreted. It must be aggregated to categories with much higher numbers.

Finally, it should be noted that codes I460 and I469 (unspecified cardiac arrest), which were included in Other cardiovascular diseases until 2018, have been all coded as R99 (unknown causes) from 2019 on.

7 Conclusion

The final data for 2018 and 2019 were produced using the approach presented. The combination of the three coding methods, and in particular the targeting by AI of samples sent to human coders, appears to be effective. This illustrates how AI, automated and human coding methods are mutually enriching. However, limitations (risks of under- or over-estimation) appear for certain categories of ICD codes, with the advantage of being quantifiable. These limitations encourage us to increase the amount of targeted manual rework for 2021 data. They also encourage us to integrate the quality of multiple cause coding in targeting samples to send to manual coding. France continues to work on including AI coding as part of its usual CoD data production process. The transition to ICD 11 remains an open question.

8 References

- Clanché, Razakamanana, Coudin, Robert, "Les statistiques provisoires sur les causes de décès en 2018 et 2019, une nouvelle méthode de codage faisant appel à l'intelligence artificielle", Drees Méthode n°8

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

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

- "Report on provisional 2018 and 2019 CoD data partly predicted by deep learning", French metadata on Eurostat website

-Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz and Polosukhin, Illia. "Attention is all you need." Paper presented at the meeting of the Advances in Neural Information Processing Systems, 2017. https://arxiv.org/abs/1706.03762?context=cs

-Baldi, Pierre & Brunak, Søren & Frasconi, Paolo & Soda, Giovanni & Pollastri, Gianluca. (1999). Exploiting the Past and the Future in Protein Secondary Structure Prediction. Bioinformatics (Oxford, England). 15. 937-46. 10.1093/bioinformatics/15.11.937.

-Graves, A. and Schmidhuber, J., "Framewise phoneme classification with bidirectional LSTM networks," Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., Montreal, QC, Canada, 2005, pp. 2047-2052 vol. 4, doi: 10.1109/IJCNN.2005.1556215.

- Martin, D., et al. « Rapport de production – Années de décès 2018 et 2019 – Données définitives », document de travail du CépiDc n°3, septembre 2023. https://www.cepidc.inserm.fr/documentation/rapport-de-production-annees-de-deces-2018-et-2019-donnees-definitives-document-de-travail-du-cepidc-n32023

Appendix A1- details on samples manually coded in 2018 and 2019

Assisted manual coding (also called manual coding or targeted manual coding) in 2018 and 2019 includes:

1. The permanent demographic sample not coded in batch:[7] 9892 certificates in 2018; 9705 in 2019

2. Deaths of special public health interest deaths not coded by Iris/Muse: 3272 (year 2018); 3171 (year 2019).

Deaths that require a high level of verification to ensure public health surveillance. When automatically coded by IRIS/MUSE, these certificates are usually checked by the coding team. Deaths of special public health interest in 2018 correspond to all mentions of AIDS/HIV on the certificate, maternal deaths and all deaths of persons younger than 15 years. Deaths of special public health interest in 2019 correspond to all mentions of AIDS/HIV on the certificate, maternal deaths, all deaths under 28 days of age, mentions of violence in deaths of persons younger than 15 years, deaths of persons younger than 15 years with mention of an interest code (see below), certificates with mention of a P code and all deaths of persons younger than 15 years that are not automatically coded by IRIS/MUSE.

Underlying cause or on the certificate with IRIS/MUSE coding	**Description**	**excepted ..**
G%	Diseases of the nervous system and the sense organs	G12%, G40%, G41%, G70%, G71%, G72%, G80%, G93%
F%	Mental and behavioural disorders	
J%	Diseases of the respiratory system	J09%, J10%, J11%, J12%, J21%, J35%, J45%, J46% , J840
K%	Diseases of the digestive system	K35%,K65%
L%	Diseases of the skin and subcutaneous tissue	
C01-C98	Tumours	C222, C40%, C41%, C49%, C62%, C64%, C71%, C72%,C91%,C92%,C93%,C94%,C95%

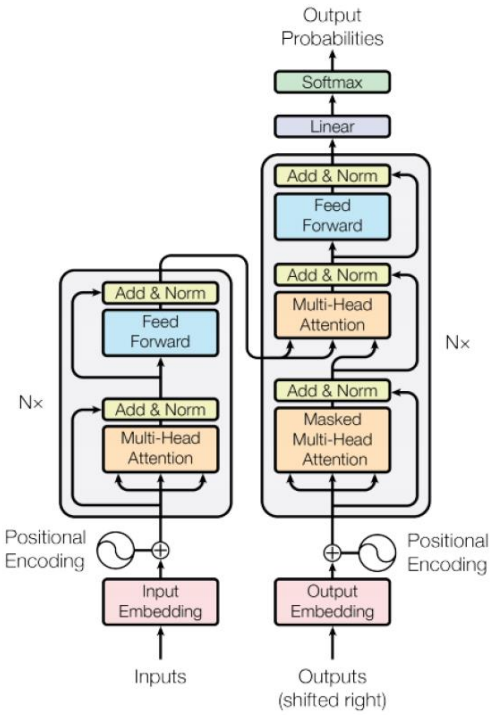
3. Samples from the categories with AI predictions with low confidence: 3116 (2018); 3202 (2019) priority 1 (needed to achieve at least 90% precision for each category of the Eurostat shortlist); some of the samples in priority 2 - the target numbers are 2646 (2018) and 2877 (2019) (chosen so as to achieve 92.5% of precision), and 1717 for 2018 and 2357 for 2019 will eventually be manually coded .

4. Simulation based on testing the performance of the entire coding strategy still showed insufficient performance for tuberculosis, homicides and drug dependence and toxicomania. A final phase of manual coding concerned for homicide: when the underlying cause selected by the surmodel is not homicide, but at least one of the models predicts homicide as underlying cause or multiple causes, or the manner of death mentions it, 256 certificates were manually coded in 2018; 244 in 2019. for drug dependence and toxicomania and tuberculosis: when the underlying cause selected by the surmodel does not fall into the category but at least one model predicts it as the underlying cause, i.e. for drug dependence 46 in 2018 and 60 in 2019 and for tuberculosis 55 in 2018 and 89 in 2019. For more details see the Production Report on 2018 and 2019 data.

[7] The permanent demographic sample ("échantillon démographique permanent", EDP) is a demographic panel of 4% of the population, selected by day of birth (https://www.insee.fr/fr/metadonnees/source/serie/s1166).

Appendix A2- Architecture and codes of Transformer k5

1. Architecture



Transformer architecture - from Vaswani et al 2017.

2. Keras5 Code

```
sequence_length = 100

batch_size = 256

buffer_size = 5000

embed_dim = 514

latent_dim = 2048

num_heads = 8

dropout = 0.2"""

## Create vocabulary : Train + Test subset

tab_vocab = pd.concat([tab_finale_train, tab_finale_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
# Output text
text_vectorization_tar.fit_on_texts(tar_texts)
voc_output = text_vectorization_tar.word_index
inp_vocab_size = len(voc_input)
tar_vocab_size = len(voc_output)
Split data in Training and validation split
val_samples = tab_finale_train.sample(frac=0.2, replace=False, random_state=7, ignore_index=True)
print("Shape val :", val_samples.shape)
val_certifs = val_samples['NumCertificat'].to_list()
train_samples = tab_finale_train[~tab_finale_train['NumCertificat'].isin(val_certifs)]
print("Shape train :", train_samples.shape)
## Tokenize 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=5000):
        super(CustomSchedule, self).__init__()
        self.d_model = d_model
        self.d_model = tf.cast(self.d_model, tf.float32)
        self.warmup_steps = warmup_steps
    def __call__(self, step):
        arg1 = tf.math.rsqrt(step)
        arg2 = step * (self.warmup_steps ** -1.5)
        return tf.math.rsqrt(self.d_model) * tf.math.minimum(arg1, arg2)
    def get_config(self):
        config = {
            'd_model': self.d_model,
            'warmup_steps': self.warmup_steps,
        }
        return config
learning_rate = CustomSchedule(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
"""
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)
history = model.fit(train_dataset,
                    epochs=60,
                    validation_data=val_dataset,
                    callbacks=model_checkpoint_callback)
```

Appendix A3- Details of the classification surmodel for selecting the underlying cause among the different proposals

To select the underlying cause among the 4 possible different model outputs – direct predictions of the underlying cause by k4 and k5, and application of the IRIS/MUSE expert system to the sequences of multiple causes predicted by k4 and k5 - iris4 and iris5, we use also a supervised learning surmodel. This surmodel responds to a classification problem in 5 classes, determining among the previous models the one we will select to predict the underlying cause, according to the characteristics of the certificates. The algorithm chosen for this model is a BiLSTM (Bidirectional Long Term Short Term memory, see Graves et al 2005, Baldi et al. 1999), a model classically used in sequence analysis and which proves to be the most efficient among the algorithms tested. Other models (LSTM, FastText, XGboost as well as a dedicated Transformer) were also tested but proved to be less efficient.

1- Train sets

The surmodel is trained on the intersection of the train sets of k4 and k5, keeping only the manually coded certificates. Since the distinction by type of coding (batch or manual) has only been recorded since 2016, the data is limited ot the years 2016 and following years, this corresponds to 482,149 certificates. The test sample is the same as the one presented above which is used to evaluate the k5 model. Table 1 shows the distribution of certificates by year for the training and test bases.

Manual	Train	Test
All	482 149	332 183
2016	149 841	93 144
2017	150 143	93912
2020	156 330	121 461
2021	25835	23 666

Table 1: Distribution of Certificates in the Database

2 - Surmodel

The surmodel aims to choose the correct underlying cause among the four model proposals. The surmodel predicts five classes: "k4", "k4_iris", "k5", "k5_iris" and "pas_orig", indicating the origin of the proposition to be selected. The fifth class "pas_orig" indicates that none of the models predicted the correct underlying cause. In this case, we will select the proposal from iris5, which is our main/reference model. Table 2 shows the proportion of the five classes in the data. Iris5 most often provides the correct underlying cause. This comes from the fact that when several models correctly predict the same underlying cause, the iris5 class (reference model) is affected first.

Classes	Train	Test
Keras5_iris	86%	83,5%
keras4_iris	4%	5,9%
keras5	3%	2,6%
Keras4	0,9%	1,2%
Pas_orig	6,08%	6,9%

Table 2: Proportion of classes in the database

2.1 - Data processing

The surmodel input sequences are concatenations of the ICD-10 code of the underlying cause predicted by k5, k4, k4_iris, k5_iris, as well as the aggregation into 86 positions of the European shortlist, the list of multiple causes predicted by k4 and k5, electronic/paper, age group, manner of death, probabilities related to the underlying cause prediction (k4 and k5), differences in probabilities between the two most probable underlying causes codes estimated by k4 and by k5 (discriminatory power of the models), the number of multiples causes, an indicator of the number of times the 4 models (k4, k5, iris4 and iris5) produce similar results (indicator of the reliability of the propositions).

The model input sequence is as follows:

“keras4_UC keras5_UC keras4iris_UC keras5iris_UC keras4_UC86 keras5_UC86 keras5iris_UC86 keras4iris_UC86 keras4_list_multiple_causes keras5_list_multiple_causes electronic/paper-back age MannerOfDeath proba_max4 proba_diff4 proba_max5 proba_d iff5 nb_causes_k4 nb_causes_k5 nb_equal”

Figure 3 reports the Shapley values of the explanatory variables. The Shapley value measures for each explanatory variable its importance in the prediction (SHAP package: “Shapley Additive Explanations”). The variables that play the most in the prediction here are the consistency indicator between the models number of times the 4 models (k4, k5, iris4 and iris5) produce similar results, the UC codes aggregated at the European shortlist level, the manner of death, the sequence of multiple causes and probabilities of models k4 and k5.

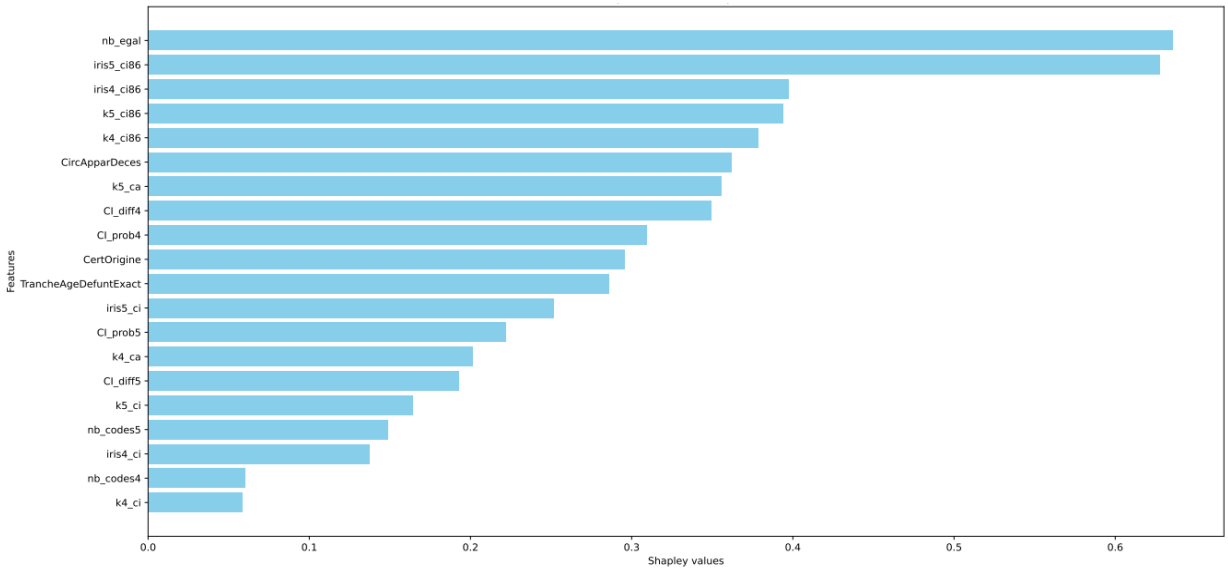


Figure A3.1 : Bar plot for feature importance

The input sequence is upstream converted into digital vectors to be used as input to the model, the steps are as follows:

Tokenization: the sentence is divided into words called tokens

Indexing: each token is associated with a unique index in a dictionary of words.

Subscript sequence transformations: the sentence is then represented as a sequence of subscripts corresponding to the tokens.

Padding: to ensure that all sequences have the same length values are added to fill the shorter sequences and truncate the longer sequences.

The input of the model first passes through an embedding layer, where each word is represented by a vector of fixed dimension. When training the model, the word representation vectors are adjusted by the model to capture semantic relationships between words, meaning that words with similar meanings will have close vectors in projection space . Then, the BiLSTM layer makes it possible to extract important sequential information in the sequence and represents it in the form of characteristic vectors: features. The Fully Connected layer classifies.

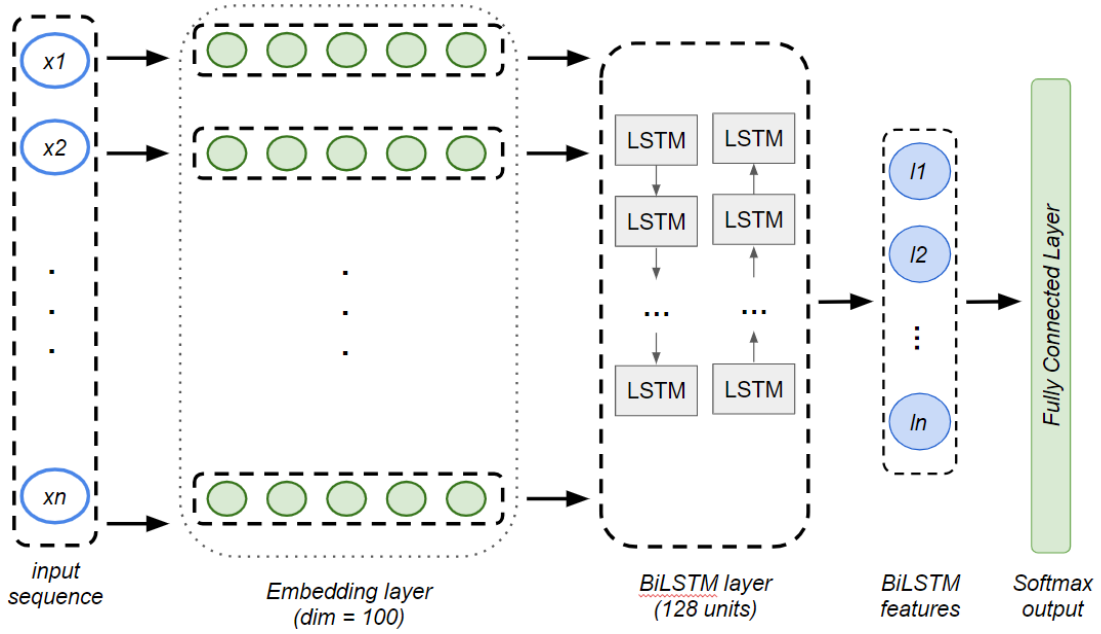


Figure A3.2 : Text classification BiLSTM network

2.3 - Hyper-parameters and loss

The loss function used is cross-entropy. The surmodel was trained using the hyper-parameters summarized in Table 3. The Adam algorithm was used to minimize the loss function. A dynamic adaptation approach of the learning rate during epochs was used to improve convergence and optimize learning performance.

Hyperparameters	Value
Sequence length	390
Optimizer	Adam
Batch size	128
Vocabulary size	4 028
Embedding dimension	100

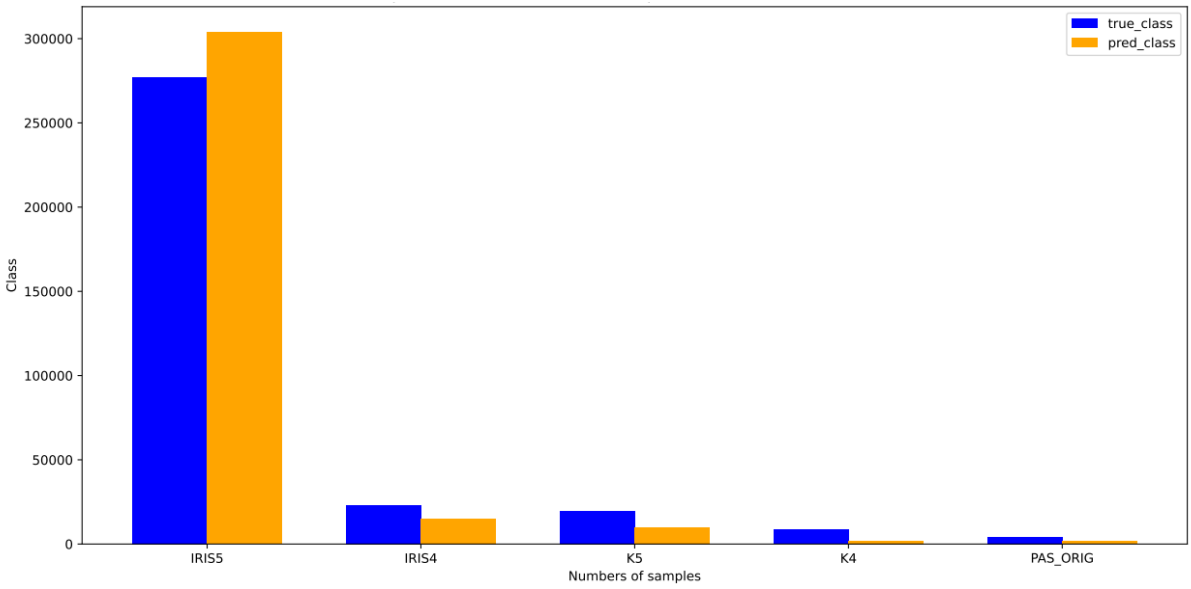
Table 3: Selected hyperparameters

3 - Results and performance analysis of the surmodel

The final performance of the surmodel is 85.6% (see: table 4). In 85.6% of the cases, the suremodel predicts the correct “class”/”origin”. Figure 2 illustrates the distribution of predictions per class on the test set. The model mainly predicts iris5. The codes of keras4 and pas_orig are very rarely retained.

Train	Validation	Test
88,29%	88,02%	85,6%

Table 4: BiLSTM performance results



At the ICD code level (the ICD code of the underlying cause predicted by the model/the origin and retained by the surmodel), the surmodel predicts the correct ICD-10 code for the underlying cause in 81.9% of the cases on the test set. This increases the performance by 2.4 points, given that iris 5 has an accuracy of 79.5%.

4. Codes

```
## Create vocabulary : Train + Val subset
"""
# Création du vocabulaire
texts = tab['text'].to_list()
tokenizer = Tokenizer()
# Input text
tokenizer.fit_on_texts(texts)
sequences = tokenizer.texts_to_sequences(texts)
vocab_size = len(tokenizer.word_index) + 1
max_sequence_length = max([len(seq) for seq in sequences])
print("max_sequence_length : ", max_sequence_length)
# Sequence to numerical
x_train_pad = tokenizer.texts_to_sequences(x_train)
x_train_pad = pad_sequences(x_train_pad, maxlen=max_sequence_length, padding="post", truncating="post")
x_val_pad = tokenizer.texts_to_sequences(x_val)
x_val_pad = pad_sequences(x_val_pad, maxlen=max_sequence_length, padding="post", truncating="post")
```

```
"""
## Encode labels
"""
# Encode the categorical labels
label_encoder = LabelEncoder()
labels = tab['origine']
label_encoder.fit(labels)
num_classes = len(label_encoder.classes_)
print("Nb classes :", num_classes)
y_train_cat = label_encoder.transform(y_train)
y_train_cat = to_categorical(y_train_cat, num_classes=num_classes)
y_val_cat = label_encoder.transform(y_val)
y_val_cat = to_categorical(y_val_cat, num_classes=num_classes)
"""
## Create deep learning model
"""
max_sequence_length = 390
cum_sch = 256
batch_size = 128
def bilstm1(vocab_size, num_classes, sequence_length):
  # Création du modèle
  model = Sequential()
  model.add(Embedding(input_dim=vocab_size, output_dim=100, input_length=sequence_length))
  model.add(Bidirectional(LSTM(128)))
  model.add(Dense(num_classes, activation='softmax'))
  return model
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)
model = bilstm1(vocab_size, num_classes, max_sequence_length)
class CustomSchedule(tf.keras.optimizers.schedules.LearningRateSchedule):
  def __init__(self, d_model, warmup_steps=5000):
    super(CustomSchedule, self).__init__()
    self.d_model = d_model
    self.d_model = tf.cast(self.d_model, tf.float32)
    self.warmup_steps = warmup_steps
```

```
    def __call__(self, step):
        arg1 = tf.math.rsqrt(step)
        arg2 = step * (self.warmup_steps ** -1.5)
        return tf.math.rsqrt(self.d_model) * tf.math.minimum(arg1, arg2)
    def get_config(self):
        config = {
            'd_model': self.d_model,
            'warmup_steps': self.warmup_steps,
        }
        return config
cum_sch = 256
learning_rate = CustomSchedule(cum_sch)
optimizer = tf.keras.optimizers.Adam(learning_rate,
                                     beta_1=0.9,
                                     beta_2=0.98,
                                     epsilon=1e-9)
model.compile(optimizer=optimizer, loss='categorical_crossentropy', metrics=['accuracy'])
# Load weights : charger les poids précédemment optimisés pour fine-tuner le modèle
model.load_weights(filepath=checkpoint_filepath).expect_partial()
model.fit(x_train_pad, y_train_cat, batch_size=batch_size, epochs=100,
          validation_data=(x_val_pad, y_val_cat),
          callbacks=model_checkpoint_callback, shuffle=True)
```

Appendix A4- AI-targeting manual coding samples

Estimation of a “confidence indicator” in the algorithm prediction

This score, calculated for each certificate, reflects the probability estimate of perfect match between the underlying cause predicted by deep learning and the underlying cause that the coding team would have coded: the higher it is (closer to 1) the more the AI predicted cause is likely to match the underlying cause coded by the coding team. To do this, we estimate on part of data for 2016 and 2017 a linear probability model that explains ICD-10 codes equality between the underlying cause coded by the coding team and the one predicted by the model k4 by some explanatory variables. The variables entering this model are:

- Underlying cause code grouped in European shortlist categories predicted after running IRIS/MUSE (by far the most explanatory),

- proxies of the length and complexity of the certificate text (number of words in the certificate, with a polynomial up to order 3, number of codes in the sequence),

- whether or not IRIS/MUSE reach an unambiguous underlying cause,

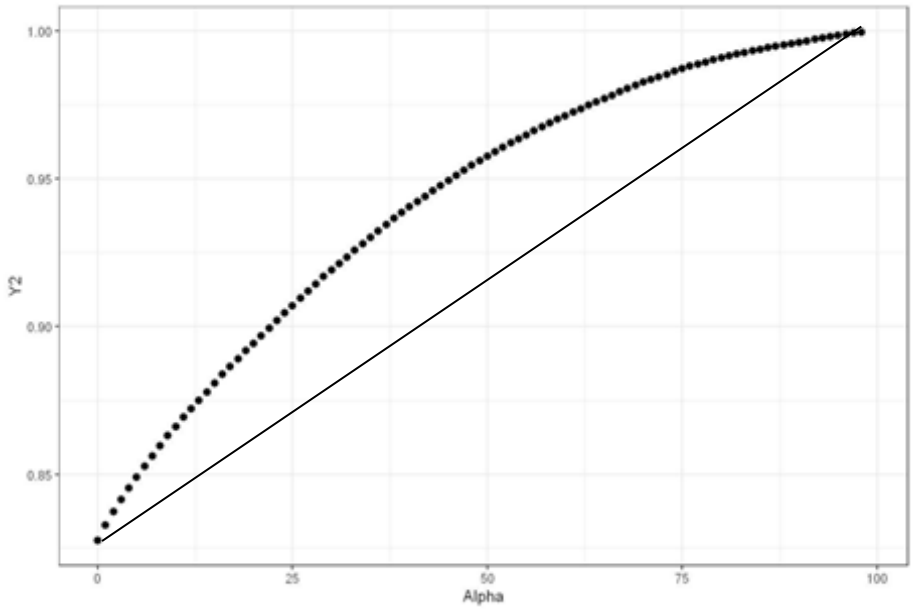
- whether the codes proposed by the deep learning model, IRIS/MUSE and a model with over-sampling of cases rejected by MUSE that was also used for provisional data, are equal

- as well as two scores given by the deep learning algorithm (the probability associated with the code of the underlying cause predicted by the model and the difference between this probability and the probability of the second most probable underlying cause according to the algorithm). This last variable captures the discriminatory power of the algorithm.

- Sex and age group are also included in the model.

The adjusted R2 of the model is around 20%.

Probability estimate provides a “confidence indicator” in the consistency between the AI prediction and manual coding. This indicator is then calculated on the rest of the 2016 and 2017 data, which were coded manually but not used to estimate the functional form of the indicator (classic approach in machine learning to separate train and test to avoid overfitting). We then simulate the impact of a targeted manual coding on the $\alpha \%$ of the data presenting the lowest confidence scores (see graph below).



On all 2017/2016 deaths in the test, without targeted manual coding (alpha=0), the ICD-10 code predicted by AI after running IRIS/MUSE equals the one coded by the coding team in 82% of cases. If 25% (alpha=25) of the certificates with the lowest confidence indicators are coded manually, then bringing the accuracy for these certificates to 1, the overall consistency/accuracy reaches 91%. It would have been necessary to code manually more than 40% of the certificates to reach this level if they had been randomly selected.

Estimation of the proportion of certificates to send to manual coding per predicted category

This reasoning enables one to compute the proportion of certificates to send to manual coding per (predicted) category of the European shortlist. We focus only on the predicted categories grouped at the level of the European shortlist for which the precision between prediction by deep learning and manual coding are the lowest. Hence, we calculate for each of them on the basis of 2016 and 2017 test data, the precision level needed in order to reach a total precision of 90% and 92.5% when automatically coded certificates and certificates already coded (EDP, special interest for public health deaths) are assumed to be coded correctly.

We define eff_codes (for 2018 and 2019): the number of certificates already coded with a underlying cause in the category (this coding being obtained automatically or coded manually) eff_noncodes (for 2018 and 2019): the number of certificates that are not encoded and for which the deep learning algorithm predicts a underlying cause in the category. Some of them will be ultimately coded manually, the question is how many and eff_tot is the sum of the two

The coding rate in the category is denoted $a = \frac{eff\ code}{eff\ tot}$

The overall precision is $P_t = (1-a)P_{ia} + a$ with P_{ia} the precision for the non-coded in the category

To ensure that P_t* the expected threshold on the overall precision is attained, the precision $P_{ia}*$ on the non-coded should be

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

The precision for the non-coded can be seen as a function of the targeted manual coding rate in the category, going from the simulated precision in the category if there is no additional manual coding (estimated for 2016/2017) up to 1 if we consider that the entire category is sent to manual coding (see previous graph). Inverting this function for $P_{ia}*$ yields the targeted manual coding rate to be perform on the category, focusing on the certificates for which the confidence indicator is the lowest:

$Manual\ coding\ rate\ per\ cat = {P_{ia}}^{-1}\ \ P_{ia}*$ to be applied then to the 2018 2019 counts.

	P1 (90%)	+P2 (92.5%)
01.3- Viral hepatitis	101	76
01.4- Other infectious and parasitic diseases	408	408
03- Diseases of the blood and blood-forming organs	966	580
04.2- Autres maladies endocrinienness, nutritionnelles et metaboliques		418
05.3 - drug dependence, toxicomania	27	40
05.4 - Other mental and behavioural disorders	199	598
10 Diseases of the skin and subcutaneous tissue	201	201
11.1- Rheumatoid arthritis and osteoarthristis		30
11.2- Other diseases of the musculoskeletal system/connective tissue	759	570
12.1-Diseases of kidney and ureter		335
12.2- Other diseases of the genitourinary system		158
17.1.4 - Accidental poisoning	709	304
17.1.5 - Other accidents		1 517
17.3- Homicide, assault	37	186
17.4-Event of undetermined intent	237	158
17.5- Other external causes of injury and poisoning	3 114	389
Total	6 758	5 967

Reading: by coding manually the 101 certificates from 2018/2019 whose AI predicted underlying cause is viral hepatitis (01.3) and with the lowest confidence indicators, according to the simulations for the years 2016 and 2017, we would achieve an overall precision of 90% for this category, 92.5% if we manually code the following 76 ones. Overall precision is computed by assuming that IRIS/MUSE automatically coded and manually coded certificates are correctly coded.

Table D.1: number of certificates 2018/2019 to be manually coded to reach a coding precision of 90% / 92.5% per European shortlist category

Appendix A5- Details on the strategy followed for performance evaluation of the targeted manual coding campaign

Reference test population

We focus on samples from the test randomly selected in order to respect the distribution of causes of death in the population (testunifauto, 332183 observations, see below), i.e.:

-Test 2016/2017 x manual coding

-Test 2020 x manual coding

-Test 2021 x (ECH1 (selection of 100 collection batches), ECH2 (deaths at the beginning of the quarter), ECH4 (randomly selected sample of certificates rejected by IRIS/MUSE)

The second step consists of completing these data with proportional selections in automatic batch coding. The completed database is called simulrep1819. In detail,

-Test 2016/2017 being a random selection from all the manually coded certificates → we will select in the same proportions (38.41%) randomly in batch 16/17. We thus obtain 128663 observations to add to the reference population from the batch for 2016 and 132944 for 2017. -Test 2020 being a random sample of all the manually coded certificates, we draw from the 2020 batch coded certificates with the same sampling rate (43.72%). We obtain 164323 observations.

-Test 2021 x ECH1 - ECH1 covers some batch and manual coded certificates – we select in the batch part of ECH1 the same proportion as the train/test ratio set on the manual coding. We obtain 21341 observations selected from the ECH1 batch.

- Test 2021 x ECH2 - Deaths that occurred at the beginning of the quarter. We complete in the same proportions as the train/test ratio of deaths having occurred on the same days coded automatically by batch. We obtain 8006 additional observations.

-Test 2021 x ECH4 - Rejects from the automatic IRIS/MUSE batch. We select from the 2021 batch coded certificates (excluding ECH1 and excluding ECH2) with the same sampling rate corrected also to respect the train/test proportion. We obtain 10191 additional observations.

In total we have 797,651 observations, with an automatic coding proportion of 58%.

Assessing the performance of the targeted manual coding strategy of the final 2018 and 2019 data

To make it possible to measure the magnitudes of the contributions consistent with the fact that the EDP and the sensitive death samples were coded manually for 2018 and 2019, we construct indicators in the reference test population identifying these cases. For the EDP, we identify the deceased born on January 2,3,4,5 or April 1,2,3,4, July or October and whose certificates have not already been coded by batch (14715), which corresponds to the definition of the EDP. For deaths of public health special interest, we apply the program used to identify them on the 2018 and 2019 data. To simulate the

impact of targeted manual coding, we rely on the predictions of the confidence scores from k5 and the iris5 cause predictions. We apply the same share of manual coding in the 12 categories as what was done in practice in 2018 and 2019. As we have not coded all the priorities 2 we apply the proportion actually coded (small overestimation because we are targeting the lowest confidence rates). We calculate three targeted manual coding indicators, one on average (P1+75% of P2), one specific to the proportions coded for 2018 (P1+65% of P2) and a last specific to the proportions coded for 2019 (P1+82% of P2), always targeting observations with the lowest confidence scores. There are 7414 in the reference test population according to the average targeted manual coding indicator (7751 according to the manual coding indicator as carried out for 2019 and 7042 according to the manual coding indicator as carried out for 2018). To simulate the contribution of each of the manual coding steps, it is then sufficient to consider that the observations of the reference test population selected according to these indicators are correctly coded.

In order to measure the targeted manual coding performance on the entire population, in accordance with the fact that the deaths of special public health interest were manually coded, we construct indicators in the reference test population that identity these cases. For the permanent demographic sample, we identify the deceased born on January 2,3,4,5 or April 1,2,3,4, July or October and whose certificates have not yet been coded by batch (14715) this which corresponds to the definition of the research database. For deaths of special public health interest, we apply the identification rules used to identify them on the 2018 and 2019 data. To simulate the impact of AI-targeted manual coding, we rely on the confidence scores of the k5 and iris5 cause predictions. We apply the same proportion of manual coding to the 12 targeted categories as we did for 2018 and 2019. Since we did not code all of Priority 2, we apply the proportion that was actually coded (which results in a slight overestimate because we target the lowest confidence scores). We calculate three indicators of manual coding, one on average (P1+75% of P2), one specific to the proportions coded in 2018 (P1+65% of P2) and in 2019 (P1+82% of P2) respectively, always targeting the observations with the lowest confidence scores. There are 7414 in the reference test population according to the average indicator, 7751 according to the 2019 indicator, and 7042 according to the 2018 indicator. To simulate the contribution of each of the manual coding step, we consider that the observations of the reference test population selected according to these indicators are correctly coded.

Performance comparison between the final data and provisional data strategies

		Prov. Data				Prov. Data+ manual coding				Final data incl. manual coding			
All test reference population	Real codes	F-measure	Predictions	Pred. / real codes - 1	sign. Of diff	F-measure	Predictions	Pred. / real codes - 1	sign. Of diff	F-measure	Predictions	Pred. / real codes - 1	sign. Of diff
01.1	476	86,5%	484	1,7%		91,9%	453	-4,8%		91,8%	448	-5,9%	
01.2	332	75,0%	271	-18,4%	****	99,0%	339	2,1%		99,0%	339	2,1%	
01.3	560	80,7%	629	12,3%	****	85,9%	606	8,2%	**	86,9%	572	2,1%	
01.4	12936	87,5%	14239	10,1%	****	89,5%	14001	8,2%	****	92,4%	12780	-1,2%	*
02.1.01	4996	95,6%	4831	-3,3%	***	95,9%	4834	-3,2%	***	95,7%	4868	-2,6%	**
02.1.02	4797	97,8%	4770	-0,6%		98,0%	4773	-0,5%		97,9%	4791	-0,1%	
02.1.03	5790	97,8%	5677	-2,0%	*	98,0%	5682	-1,9%	*	97,7%	5755	-0,6%	
02.1.04	23061	97,8%	23031	-0,1%		97,9%	23027	-0,1%		98,1%	23051	0,0%	
02.1.05	11426	97,2%	11295	-1,1%		97,4%	11295	-1,1%		97,3%	11358	-0,6%	
02.1.06	15433	98,8%	15349	-0,5%		98,9%	15363	-0,5%		99,0%	15411	-0,1%	
02.1.07	1271	94,2%	1265	-0,5%		94,4%	1268	-0,2%		94,6%	1256	-1,2%	
02.1.08	40493	97,8%	40235	-0,6%		98,0%	40276	-0,5%		98,2%	40488	0,0%	
02.1.09	2241	96,0%	2227	-0,6%		96,2%	2229	-0,5%		96,2%	2257	0,7%	
02.1.10	16601	97,7%	16710	0,7%		97,9%	16708	0,6%		98,1%	16595	0,0%	
02.1.11	1048	96,5%	1033	-1,4%		97,1%	1033	-1,4%		97,1%	1043	-0,5%	
02.1.12	3630	96,7%	3547	-2,3%	*	97,0%	3548	-2,3%	*	96,9%	3573	-1,6%	
02.1.13	4424	98,0%	4351	-1,7%		98,1%	4355	-1,6%		98,1%	4411	-0,3%	
02.1.14	11882	97,4%	11938	0,5%		97,5%	11944	0,5%		97,8%	11853	-0,2%	
02.1.15	4626	96,5%	4557	-1,5%		96,7%	4561	-1,4%		96,8%	4546	-1,7%	
02.1.16	6874	97,1%	6822	-0,8%		97,3%	6817	-0,8%		97,6%	6882	0,1%	
02.1.17	5232	97,4%	5101	-2,5%	**	97,6%	5108	-2,4%	**	97,2%	5211	-0,4%	
02.1.18	490	94,6%	466	-4,9%		94,7%	467	-4,7%		94,0%	474	-3,3%	
02.1.19	6393	96,2%	6359	-0,5%		97,1%	6400	0,1%		97,3%	6411	0,3%	
02.1.20	7856	96,9%	7726	-1,7%	*	97,5%	7774	-1,0%		97,6%	7886	0,4%	
02.1.21	4290	96,0%	4161	-3,0%	***	96,6%	4188	-2,4%	*	97,0%	4264	-0,6%	
02.1.22	29282	92,8%	30535	4,3%	****	93,3%	30481	4,1%	****	93,7%	29760	1,6%	****
02.2	10175	92,1%	10285	1,1%		92,7%	10290	1,1%		92,8%	10220	0,4%	
3	3491	78,1%	3765	7,8%	****	85,5%	3400	-2,6%	*	86,8%	3237	-7,3%	****
04.1	16008	94,1%	15905	-0,6%		94,6%	15905	-0,6%		95,0%	15795	-1,3%	**
04.2	13704	90,1%	13519	-1,3%	*	91,2%	13515	-1,4%	*	92,1%	13548	-1,1%	*
05.1	25311	95,2%	25870	2,2%	****	95,5%	25847	2,1%	****	96,2%	25898	2,3%	****
05.2	3230	91,3%	3287	1,8%		92,3%	3306	2,4%	*	91,9%	3301	2,2%	
05.3	308	80,8%	274	-11,0%	**	87,7%	269	-12,7%	***	86,5%	284	-7,8%	*
05.4	4907	88,7%	5021	2,3%	*	90,9%	4992	1,7%		91,5%	4875	-0,7%	
06.1	8866	97,3%	8884	0,2%		97,5%	8891	0,3%		97,6%	8895	0,3%	
06.2	25747	98,1%	25644	-0,4%		98,3%	25660	-0,3%		98,2%	25828	0,3%	
06.3	15541	91,8%	15552	0,1%		92,8%	15581	0,3%		93,1%	15513	-0,2%	
07.1.1	18023	95,9%	18330	1,7%	***	96,2%	18316	1,6%	***	96,4%	18192	0,9%	
07.1.2	24438	93,8%	24432	0,0%		94,2%	24444	0,0%		94,7%	24431	0,0%	
07.2	67415	94,6%	66860	-0,8%	***	95,1%	66914	-0,7%	**	95,5%	67533	0,2%	
07.3	41319	94,6%	41710	0,9%	**	95,1%	41661	0,8%	**	95,5%	41534	0,5%	
07.4	33025	92,2%	32950	-0,2%		93,0%	32989	-0,1%		93,7%	32834	-0,6%	
08.1	1668	95,9%	1631	-2,2%		96,4%	1635	-2,0%		96,9%	1682	0,8%	
08.2	16322	94,5%	15902	-2,6%	****	95,0%	15946	-2,3%	****	95,7%	16374	0,3%	
08.3.1	1077	94,1%	1054	-2,1%		94,6%	1056	-1,9%		94,2%	1067	-0,9%	
08.3.2	13006	94,9%	12973	-0,3%		95,3%	12983	-0,2%		95,4%	13131	1,0%	
08.4	21100	92,6%	20773	-1,5%	***	93,3%	20762	-1,6%	***	93,7%	20908	-0,9%	*
09.1	1081	92,0%	1062	-1,8%		93,4%	1071	-0,9%		93,0%	1090	0,8%	
09.2	8986	95,5%	8976	-0,1%		95,9%	8992	0,1%		96,0%	9026	0,4%	
09.3	22147	92,4%	21497	-2,9%	****	93,4%	21667	-2,2%	****	93,5%	22191	0,2%	
10	2067	84,9%	2005	-3,0%	*	88,1%	1996	-3,4%	*	89,8%	2065	-0,1%	
11.1	726	83,2%	769	5,9%	*	86,0%	772	6,3%	**	88,7%	692	-4,7%	
11.2	4537	80,6%	4115	-9,3%	****	87,1%	4223	-6,9%	****	87,0%	4356	-4,0%	****
12.1	10646	90,6%	10560	-0,8%		91,6%	10575	-0,7%		92,4%	10496	-1,4%	*
12.2	4029	84,6%	3518	-12,7%	****	86,7%	3593	-10,8%	****	90,0%	3966	-1,6%	
13	54	55,2%	33	-38,9%	****	100,0%	54	0,0%		100,0%	54	0,0%	
14	2048	92,6%	1975	-3,6%	*	99,7%	2060	0,6%		99,6%	2064	0,8%	
15	2105	80,0%	2252	7,0%	****	86,7%	2313	9,9%	****	92,0%	1993	-5,3%	***
16.1	179	92,7%	175	-2,2%		99,4%	179	0,0%		97,8%	181	1,1%	
16.2	20174	95,5%	21086	4,5%	****	95,9%	20985	4,0%	****	96,4%	20462	1,4%	***
16.3	40404	97,6%	40706	0,7%	*	97,8%	40683	0,7%	*	97,8%	40711	0,8%	*
17.1.1	3678	95,1%	3550	-3,5%	***	95,7%	3573	-2,9%	**	96,1%	3618	-1,6%	
17.1.2	11146	93,5%	11162	0,1%		94,2%	11166	0,2%		94,6%	11315	1,5%	*
17.1.3	1090	93,3%	1134	4,0%	*	94,4%	1130	3,7%		95,5%	1122	2,9%	
17.1.4	2163	79,0%	2124	-1,8%		86,6%	2170	0,3%		88,3%	1994	-7,8%	****
17.1.5	18254	88,8%	18217	-0,2%		90,6%	18217	-0,2%		91,4%	18083	-0,9%	
17.2	11281	95,0%	11232	-0,4%		96,1%	11240	-0,4%		97,2%	11245	-0,3%	
17.3	499	76,5%	565	13,2%	****	94,2%	486	-2,6%		93,7%	498	-0,2%	
17.4	1709	61,9%	1159	-32,2%	****	73,4%	1312	-23,2%	****	80,6%	1541	-9,8%	****
17.5	1847	49,7%	1869	1,2%		65,7%	1622	-12,2%	****	67,5%	1166	-36,9%	****
			761971				761971				761292		
		90,7%				93,7%				94,2%			
											679 predicted as COVID		

Table A5.1: Precision, recall comparisons on Reference Test Population (COVID excluded) between strategies followed for provisional and final data