Methodological note on provisional 2018-2019 cause of death data

France transmitted to Eurostat provisional causes of death data for the year 2018 and 2019. Around 62% of these provisional CoD data were automated coded by the rule-based system IRIS and IRIS muse, and for the remaining, both the sequence of causes of death and the underlying cause of death were predicted by an AI approach involving deep learning algorithms trained on CoD texts from previous years, some with, and some without recourse to IRIS Muse for coding the underlying cause of death. This is the first time AI is used for coding (part of) CoD data.

This report documents the methodology followed. It also provides a performance analysis by comparing AI approach predictions with true labels of 2016 and 2017 CoD data (not used in training). It finally reports results for 2018 and 2019 and compares them with trends of previous and next years. This highlights the categories of CoD for which the results of the prediction are expected to be correct and those which must be interpreted with caution.

1- Context

For 2018 and 2019 data, due to lack of resources, CépiDc-Inserm (French ONA producer of CoD data) was not able to code 2018 and 2019 death certificates on time and cannot catch up now relying only on manual coding and rule-based system of coding (IRIS) as it does usually. 2018 and 2019 death certificates, which failed to be automated coded by the rule-based system of coding (IRIS) were predicted by deep learning algorithms trained on all available data from 2011 on. Table 1 reports the number of certificates coded by rule-based automated coding system IRIS, by AI, as said before no data were coded manually up to now.

Years\ Type of coding	Manual coding	AI prediction or AI mixed with IRIS	Fully rule-based automated coding with IRIS only	Total
2018 - Numbers	176	218769	375855	594800
2018 - %	0	37	63	100
2019 - Numbers	91	229068	369619	598778
2019 - %	0	38	62	100

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 - Scope : all received certificates for 2018 and 2019.

2- Approach

The AI algorithms used are seq-to-seq translation models of *transformer* type see Vaswani *et al* (2017). *Transformers* have the advantage of taking into account the links between the words of the sentence thanks to their “attention” mechanism. They rely on highly parallelized computation, which allows rapid training that can be fully implemented on conventional infrastructure. The algorithms are implemented with tensorflow and keras, deep learning open libraries which are maintained over time. This work made by the statistical service of the health ministry (DREES) is based on and extends previous works carried out at CépiDc, see Falissard *et al.* (2020) and Falissard (2021).

2.1 Model main specifications

- *Feature engineering*

Input sequences are the concatenation of texts written on each line of the death certificate separated by the label of the line and some additional features. Additional features include 5-year age group, sex and year of death. (not for the moment the type of certificate, neither the new variables on apparent circumstances of death with are only available after 2018 for death certificates which use the new model of certificate)

The input sequence is then composed of age_group sex year of death sep_line1 text written on line 1 sep_line2 text written on line2 …. and ending by sep_underlyingcause

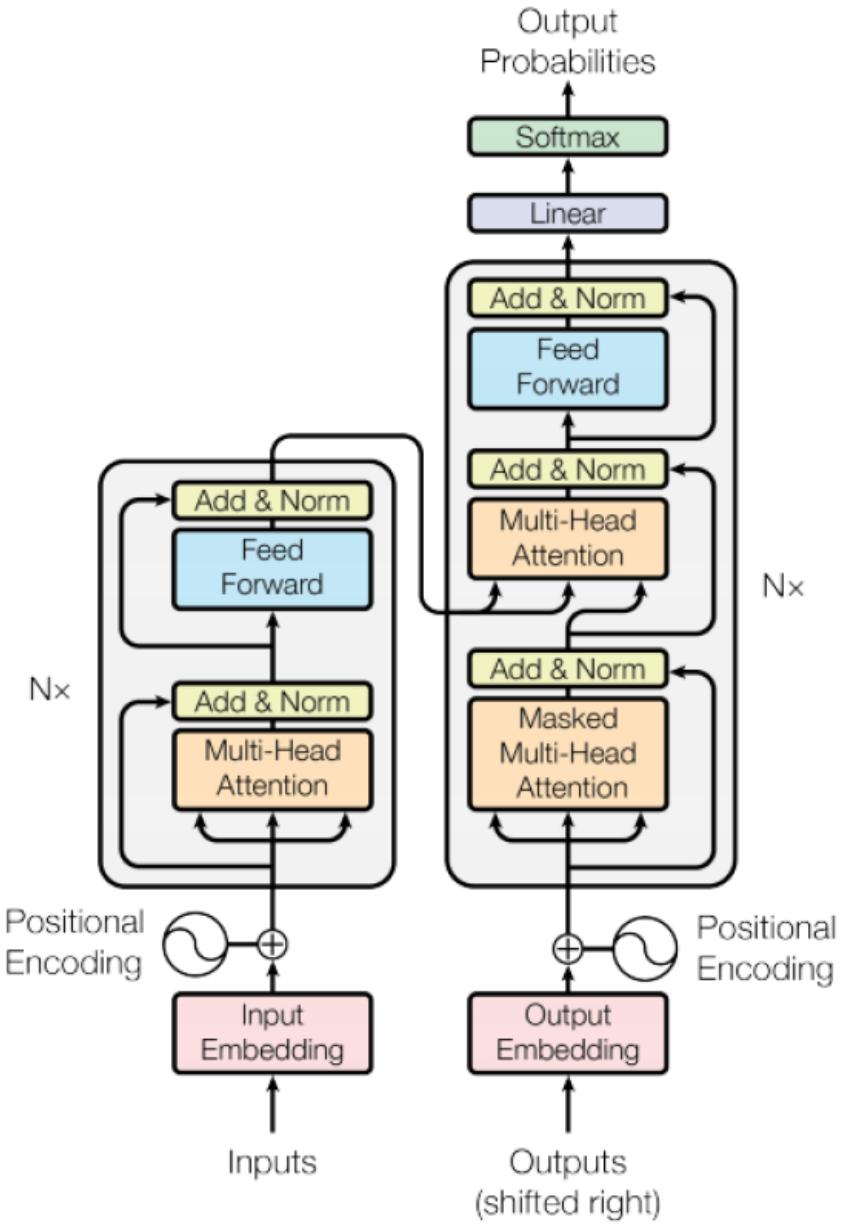
The target variable is the following sequence age_group sex year of death sep_line1 ICDcod1 ICDcod12 sep_line2 ICDcod2 … sep_underlyingcause ICD_code_underlying_cause.

The tokenizer TextVectorization is used to split these sequences into words (no use of subwords nor bigrams of words). The dictionary is made of all the different words obtained from this split - it contains 78982 words for the input sequence and 5935 words for the target sequence.

- *Model architecture*

The architecture of the transformers used is of encoder/decoder type. More precisely, transformers involve one group of layers composed of encoder/decoder + positional embedding + attention mechanism + one layer of dropout (to control overfitting) for both input and target

sequences, and finish by a softmax layer.



Transformer architecture - from Vaswani et al 2017.

- *Training*

For predicting 2018 and 2019 CoD data, the train dataset is composed of all death certificates coded in ICD10 for years of death 2011 to 2017 and 2020 (except COVID ones) plus death certificates of years 2018 and 2019 that are fully automated coded by the rule-based system IRIS.

For testing models performance on 2016 and 2017 CoD data, models are also trained on only 2011 to 2015 data plus death certificates of years 2016 and 2017 that are fully automated coded by the rule-based system IRIS.
Training is done using as loss function the sparse categorical cross-entropy and as optimizer Adam. A call back approach allows one to choose the optimal epoch as the one with the higher accuracy on validation data (20% of the train data set).

Detailed hyperparameters are reported in the appendix.

- *Prediction*. Predictions are made through a greedy search approach: each token is predicted once and enters into the prediction of the next token.

2.2 Combination of three approaches to determine the underlying cause of death

Once the sequence of ICD codes appearing on the death certificate is predicted, we can choose as the underlying cause of death directly the model prediction for this part, or we can apply the rule-based system IRIS/muse to the transformer sequence prediction when it yields a result. When it does not, we can either use the transformer prediction for the underlying cause of death, or the prediction of an additional model. The approach retained mixes the three possibilities.

The underlying cause of death obtained by applying the rule-based system IRIS/MUSE to the sequence of cases predicted by the algorithm has the advantage to ensure the maximum of homogeneity in coding, but this approach does not always provide an unambiguous result. Hence, we re-train an additional transformer model (same structure as previously presented) on a specific train sample composed of death certificates for which IRIS/MUSE fail to yield the underlying cause of death. This is a sort of transfer learning approach allowing one to adapt the train sample to the specific case of IRIS/MUSE rejection. Finally, for these provisional data, we adopted specific rules to choose between the three predictions, based on the precision/recall and accuracy measures for each category obtained on 2016 and 2017 data.

However, even with this, counting predictions on 2016 and 2017 data show some systematic under-estimation for 5 groups of pathologies/conditions (tuberculosis, hepatitis, pregnancies, congenital diseases and homicides). This conducts us to add specific decision rules leading to retain these conditions/pathologies as underlying cause of death either if they are predicted by at least by one approach or if they appear once in the sequence of causes prediction. This reduces but does not completely solve the under-estimations.

At the end, the prediction of the underlying cause of death is obtained after applying a series of rules of decision detailed in the appendix. These rules are set by analyzing results on 2016 and 2017 data, which may induce some overfitting. This issue will be corrected for definitive data.

3- Performance analysis

We assess models performance and overall performance of the automated coding by comparing predictions of the AI approach and real values of underlying cause of death such as coded by coders. We do so on 2016 and 2017 data coded manually, which are NOT used in the training sample, as usual in machine learning.

Results show 88.9% of correct predictions of the underlying cause of death at the European short list grouping level on 2016 and 2017 data that should have been coded manually, see Table 2. If we now add rule-based automated coding, and assume that for these death certificates the coding is correct then the complete coding process ends up to a correct underlying cause of death at the 86 European short list level in 95.4% of the cases, and 93% for the detailed ICD10 4-digit level.

	ICD 4-digit level	86 European short list level
Death certificates on which the AI approach is used (ie those that were manually coded)	83.4%	88.9%
All death certificates	93.0%	95.4%

Table 2 - % of correct predictions of the underlying CoD on 2016-2017 data

Table 3 reports precision, recall and F-measures per category of the European short list. Precision stands for the % of correct labels among predictions of a given category by the AI approach (or AI and rule-based automated coding approach for the 3 first panels), recall stands for the % of correct predictions by the model among all data coded in the given category, F-measure is the harmonic average of the two.

	All 2016-2017 death certificates			2016-2017 death certificates that were manually coded		
	precision	recall	F-measure	precision	recall	F-measure
01.1- Tuberculosis	93%	86%	**89%**	92%	85%	**88%**
01.2- AIDS (HIV diseases)	79%	76%	**78%**	75%	71%	**73%**
01.3- Viral hepatitis	83%	82%	**83%**	70%	69%	**70%**
01.4- Other infectious and parasitic diseases	89%	92%	**91%**	76%	83%	**79%**
02.1.01-Malignant neoplasms of lip, oral cavity, pharynx	98%	96%	**97%**	96%	93%	**95%**
02.1.02-Malignant neoplasms of oesophagus	99%	98%	**99%**	97%	97%	**97%**
02.1.03-Malignant neoplasms of stomach	99%	99%	**99%**	97%	97%	**97%**
02.1.04-Malignant neoplasms of colon, rectum, anus	98%	99%	**99%**	96%	97%	**97%**
02.1.05-Malignant neoplasms of liver and intrahepatic bile ducts	98%	98%	**98%**	94%	96%	**95%**
02.1.06-Malignant neoplasms of pancreas	99%	99%	**99%**	98%	98%	**98%**
02.1.07-Malignant neoplasms of larynx	97%	95%	**96%**	95%	91%	**93%**
02.1.08-Malignant neoplasms of trachea, bronchus, lung	99%	99%	**99%**	97%	97%	**97%**
02.1.09- Malignant neoplasms of skin	97%	98%	**97%**	95%	95%	**95%**
02.1.10-Malignant neoplasms of breast	98%	99%	**98%**	96%	97%	**96%**
02.1.11-Malignant neoplasms of cervix uteri	98%	98%	**98%**	97%	95%	**96%**
02.1.12-Malignant neoplasms of other and unspecified parts of uterus	98%	99%	**98%**	97%	97%	**97%**
02.1.13-Malignant neoplasms of ovary	99%	99%	**99%**	98%	97%	**97%**
02.1.14-Malignant neoplasms of prostate	98%	98%	**98%**	94%	95%	**95%**
02.1.15-Malignant neoplasms of kidney	97%	98%	**98%**	94%	96%	**95%**
02.1.16-Malignant neoplasms of bladder	99%	98%	**98%**	97%	95%	**96%**
02.1.17-Malignant neoplasms of brain and central nervous system	99%	97%	**98%**	98%	94%	**96%**
02.1.18-Malignant neoplasms of thyroid	98%	96%	**97%**	97%	93%	**95%**
02.1.19-Hodgkin disease and lymphomas	98%	96%	**97%**	95%	91%	**93%**
02.1.20- Leukaemia	98%	96%	**97%**	95%	92%	**94%**
02.1.21-Other malignant neoplasms of lymphoid and haematopoietic tissue	97%	96%	**97%**	95%	91%	**93%**
02.1.22-Other malignant neoplasms	96%	97%	**96%**	92%	93%	**93%**
02.2-Non-malignant neoplasms (benign and uncertain)	93%	94%	**93%**	87%	88%	**88%**
03 Diseases of the blood and blood-forming organs and certain disorders inv	79%	82%	**80%**	64%	68%	**66%**
04.1- Diabetes mellitus	93%	93%	**93%**	86%	85%	**85%**
04.2- Other endocrine, nutritional and metabolic diseases	91%	90%	**91%**	79%	78%	**78%**
05.1- Dementia	95%	97%	**96%**	85%	92%	**88%**
05.2- Alcohol abuse (including alcohol psychosis)	94%	93%	**94%**	86%	85%	**86%**
05.3 - drug dependence, toxicomania	87%	84%	**85%**	79%	75%	**77%**
05.4 - Other mental and behavioural disorders	88%	91%	**90%**	76%	82%	**79%**
06.1- Parkinson's disease	98%	98%	**98%**	95%	94%	**94%**
06.2 - Alzheimer's disease	99%	99%	**99%**	95%	96%	**96%**
06.3- Other diseases of the nervous system and the sense organs	94%	93%	**94%**	87%	86%	**86%**
07.1.1-Acute myocardial infarction	95%	97%	**96%**	88%	91%	**90%**
07.1.2-Other ischaemic heart diseases	95%	94%	**94%**	88%	87%	**88%**
07.2-Other heart diseases	96%	96%	**96%**	87%	87%	**87%**
07.3-Cerebrovascular diseases	94%	94%	**94%**	86%	88%	**87%**
07.4- Other diseases of the circulatory system	92%	92%	**92%**	84%	83%	**84%**
08.1 - Influenza	99%	97%	**98%**	97%	94%	**95%**
08.2 - Pneumonia	96%	96%	**96%**	87%	85%	**86%**
08.3.1 - Asthma	95%	96%	**96%**	87%	90%	**88%**
08.3.2-Other chronic lower respiratory diseases	96%	96%	**96%**	90%	90%	**90%**
08.4- Other diseases of the respiratory system	93%	95%	**94%**	78%	82%	**80%**
09.1 - Ulcer of stomach, duodenum, jejunum	94%	92%	**93%**	89%	86%	**88%**
09.2 - Cirrhosis, fibrosis, and chronic hepatitis	97%	97%	**97%**	93%	93%	**93%**
09.3- Other diseases of the digestive system	95%	93%	**94%**	89%	86%	**87%**
10 Diseases of the skin and subcutaneous tissue	89%	87%	**88%**	80%	77%	**78%**
11.1- Rheumatoid arthritis and osteoarthristis	90%	89%	**89%**	84%	81%	**82%**
11.2- Other diseases of the musculoskeletal system/connective tissue	84%	82%	**83%**	76%	74%	**75%**
12.1-Diseases of kidney and ureter	91%	92%	**92%**	79%	81%	**80%**
12.2- Other diseases of the genitourinary system	92%	86%	**89%**	85%	77%	**81%**
13 Complications of pregnancy, childbirth and puerperium	78%	72%	**75%**	75%	67%	**71%**
14 Certain conditions originating in the perinatal period	95%	94%	**95%**	94%	94%	**94%**
15 Congenital malformations and chromosomic abnormalities	92%	84%	**88%**	87%	76%	**81%**
16.1- Sudden infant death syndrome	93%	93%	**93%**	93%	93%	**93%**
16.2- Unknown and unspecified causes	98%	99%	**98%**	92%	95%	**93%**
16.3- Other symptoms, signs, ill-defined causes	99%	99%	**99%**	94%	92%	**93%**
17.1.1 - Transport accidents	98%	96%	**97%**	97%	94%	**95%**
17.1.2 - Accidental falls	95%	94%	**95%**	93%	91%	**92%**
17.1.3 - Drowning and accidental submersion	97%	97%	**97%**	92%	91%	**91%**
17.1.4 - Accidental poisoning	82%	86%	**84%**	74%	79%	**76%**
17.1.5 - Other accidents	90%	89%	**89%**	83%	81%	**82%**
17.2 - Suicide and intentional self-harm	98%	97%	**98%**	96%	94%	**95%**
17.3- Homicide, assault	87%	87%	**87%**	83%	83%	**83%**
17.4-Event of undetermined intent	80%	72%	**76%**	78%	70%	**74%**
17.5- Other external causes of injury and poisoning	43%	41%	**42%**	34%	32%	**33%**

Table 3 Precision, recall and F measures of AI coding approach on 2016-2017 data

This table underlines which code predictions should be looked at with caution. Those with F-measure below 90% are tuberculosis, AIDS, viral hepatitis, diseases of blood, toxicomania, other mental and behavioural disorders, diseases of skin, rheumatoid arthritis, other diseases of the musculoskeletal system, of the genitourinary system, complications of pregnancies, congenital malformations, accidental poisoning, other accidents, homicide, event of undetermined intentions, other external causes. These CoD may indeed concern

- rare cases resulting in small counts.
- “other” types of CoD, for which the imprecision may result from non”Other” categories imprecision
- relatively important counts or categories of special interest for public health, including infectious diseases and diseases of the digestive system, which will have to be closely looked at in future developments.

	2 016				2 017			
	real codes	prediction	significance of the difference	prediction/ real code - 1 (in%)	real codes	prediction	significance of the difference	prediction/ real code - 1 (in%)
01 Infectious and parasitic diseases	10 550	10 785	****	2,2	11 657	12 057	*****	3,4
02 Neoplasms	171 743	171 663		0,0	171 780	171 684		- 0,1
03 Diseases of the blood and blood forming organs and certain	2 305	2 477	*****	7,5	2 585	2 603		0,7
04 Endocrine, nutritional and metabolic diseases	21 321	21 283		-0,2	22 195	21 928	***	- 1,2
05 Mental and behavioural disorders	26 042	26 232	*	0,7	25 953	26 994	*****	4,0
06 Diseases of the nervous system and the sense organs	38 939	39 005		0,2	39 622	39 516		- 0,3
07 Diseases of the circulatory sytem	144 160	144 124		0,0	144 254	144 834	**	0,4
08 Diseases of the respiratory system	41 436	41 638		0,5	44 842	44 957		0,3
09 Diseases of the digestive system	24 264	23 954	****	-1,3	24 259	23 926	****	- 1,4
10 Diseases of the skin and subcutaneous tissue	1 491	1 511		1,3	1 630	1 549	****	- 5,0
11 Diseases of the musculoskeletal system / connective tissue	4 166	4 014	****	-3,6	4 014	4 019		0,1
12 Diseases of the genitourinary system	10 147	10 202		0,5	10 878	10 635	****	- 2,2
13 Complications of pregnancy, childbirth and puerperium	40	39		-2,5	41	35		- 14,6
14 Certains conditions originating in the perinatal period	1 510	1 532		1,5	1 693	1 649	*	- 2,6
15 Congenital malformations and chromosomic abnormalities	1 701	1 547	*****	-9,1	1 645	1 532	*****	- 6,9
16 Symptoms, signs and ill-defined causes	39 968	40 074		0,3	42 724	42 613		- 0,3
17 External causes	38 848	38 551	**	-0,8	39 777	39 018	*****	- 1,9
Total	578 631	578 631			589 549	589 549		

Note : significance degrees of counting differentials come from equality tests assuming real occurrences were Poisson distributed. * pval<.3, ** pval<.2, *** pval<.1, ****pval<.05, ***** pval<.01.

Table 4 - comparison of predictions and real codes distributions on the 17 ICD10 chapters, all 2016 and 2017 death certificates

	2016				2017			
	real codes	prediction	signif. Of diff.	pred./real code - 1 (in%)	real codes	prediction	signif. Of diff.	pred./real code - 1 (in%)
01.1- Tuberculosis	407	375	**	-7,9	404	380	*	- 5,9
01.2- AIDS (HIV diseases)	344	301	****	-12,5	245	266	**	8,6
01.3- Viral hepatitis	593	611		3,0	786	751	*	- 4,5
01.4- Other infectious and parasitic diseases	9 206	9 498	*****	3,2	10 222	10 660	*****	4,3
02.1.01-Malignant neoplasms of lip, oral cavity, pharynx	3 947	3 864	**	-2,1	3 812	3 769		- 1,1
02.1.02-Malignant neoplasms of oesophagus	3 912	3 903		-0,2	3 876	3 871		- 0,1
02.1.03-Malignant neoplasms of stomach	4 621	4 626		0,1	4 628	4 642		0,3
02.1.04-Malignant neoplasms of colon, rectum, anus	18 075	18 138		0,3	18 041	18 072		0,2
02.1.05-Malignant neoplasms of liver and intrahepatic bile duct	8 814	8 846		0,4	8 584	8 667		1,0
02.1.06-Malignant neoplasms of pancreas	11 328	11 318		-0,1	11 507	11 498		- 0,1
02.1.07-Malignant neoplasms of larynx	1 073	1 050		-2,1	1 002	982		- 2,0
02.1.08-Malignant neoplasms of trachea, bronchus, lung	31 959	31 979		0,1	31 487	31 504		0,1
02.1.09- Malignant neoplasms of skin	1 755	1 760		0,3	1 773	1 775		0,1
02.1.10-Malignant neoplasms of breast	12 985	12 995		0,1	13 055	13 081		0,2
02.1.11-Malignant neoplasms of cervix uteri	808	807		-0,1	822	813		- 1,1
02.1.12-Malignant neoplasms of other and unspecified parts of	2 846	2 851		0,2	2 918	2 921		0,1
02.1.13-Malignant neoplasms of ovary	3 505	3 493		-0,3	3 559	3 565		0,2
02.1.14-Malignant neoplasms of prostate	9 043	9 049		0,1	9 229	9 264		0,4
02.1.15-Malignant neoplasms of kidney	3 601	3 623		0,6	3 622	3 668		1,3
02.1.16-Malignant neoplasms of bladder	5 361	5 315		-0,9	5 157	5 104		- 1,0
02.1.17-Malignant neoplasms of brain and central nervous syst	3 976	3 908	*	-1,7	4 106	4 033	*	- 1,8
02.1.18-Malignant neoplasms of thyroid	382	371		-2,9	420	413		- 1,7
02.1.19-Hodgkin disease and lymphomas	4 909	4 783	***	-2,6	4 960	4 867	**	- 1,9
02.1.20- Leukaemia	6 040	5 976		-1,1	6 173	6 085	*	- 1,4
02.1.21-Other malignant neoplasms of lymphoid and haematop	3 449	3 416		-1,0	3 242	3 165	**	- 2,4
02.1.22-Other malignant neoplasms	21 811	21 982	*	0,8	22 191	22 271		0,4
02.2-Non-malignant neoplasms (benign and uncertain)	7 543	7 610		0,9	7 616	7 654		0,5
03 Diseases of the blood and blood-forming organs and certai	2 305	2 477	*****	7,5	2 585	2 603		0,7
04.1- Diabetes mellitus	11 889	11 833		-0,5	11 967	11 898		- 0,6
04.2- Other endocrine, nutritional and metabolic diseases	9 432	9 450		0,2	10 228	10 030	***	- 1,9
05.1- Dementia	19 769	19 870		0,5	19 673	20 585	*****	4,6
05.2- Alcohol abuse (including alcohol psychosis)	2 582	2 572		-0,4	2 466	2 440		- 1,1
05.3 - drug dependence, toxicomania	231	223		-3,5	192	187		- 2,6
05.4 - Other mental and behavioural disorders	3 460	3 567	***	3,1	3 622	3 782	*****	4,4
06.1- Parkinson's disease	6 649	6 625		-0,4	6 833	6 842		0,1
06.2 - Alzheimer's disease	21 122	21 154		0,2	20 976	21 017		0,2
06.3- Other diseases of the nervous system and the sense orga	11 168	11 226		0,5	11 813	11 657	**	- 1,3
07.1.1-Acute myocardial infarction	14 163	14 313	*	1,1	14 115	14 322	***	1,5
07.1.2-Other ischaemic heart diseases	19 077	18 991		-0,5	19 135	19 076		- 0,3
07.2-Other heart diseases	53 344	53 236		-0,2	53 805	53 722		- 0,2
07.3-Cerebrovascular diseases	32 343	32 373		0,1	31 902	32 497	*****	1,9
07.4- Other diseases of the circulatory system	25 233	25 211		-0,1	25 297	25 217		- 0,3
08.1 - Influenza	966	960		-0,6	2 509	2 472		- 1,5
08.2 - Pneumonia	13 332	13 307		-0,2	13 942	13 771	**	- 1,2
08.3.1 - Asthma	936	948		1,3	917	932		1,6
08.3.2-Other chronic lower respiratory diseases	10 445	10 412		-0,3	10 771	10 769		-
08.4- Other diseases of the respiratory system	15 757	16 011	****	1,6	16 703	17 013	****	1,9
09.1 - Ulcer of stomach, duodenum, jejunum	870	846		-2,8	867	858		- 1,0
09.2 - Cirrhosis, fibrosis, and chronic hepatitis	6 941	6 925		-0,2	6 799	6 810		0,2
09.3- Other diseases of the digestive system	16 453	16 183	****	-1,6	16 593	16 258	*****	- 2,0
10 Diseases of the skin and subcutaneous tissue	1 491	1 511		1,3	1 630	1 549	****	- 5,0
11.1- Rheumatoid arthritis and osteoarthristis	567	579		2,1	580	549	*	- 5,3
11.2- Other diseases of the musculoskeletal system/connective	3 599	3 435	*****	-4,6	3 434	3 470		1,0
12.1-Diseases of kidney and ureter	7 592	7 664		0,9	8 124	8 158		0,4
12.2- Other diseases of the genitourinary system	2 555	2 538		-0,7	2 754	2 477	*****	- 10,1
13 Complications of pregnancy, childbirth and puerperium	40	39		-2,5	41	35		- 14,6
14 Certain conditions originating in the perinatal period	1 510	1 532		1,5	1 693	1 649	*	- 2,6
15 Congenital malformations and chromosomic abnormalities	1 701	1 547	*****	-9,1	1 645	1 532	*****	- 6,9
16.1- Sudden infant death syndrome	177	177		0,0	141	141		-
16.2- Unknown and unspecified causes	11 595	11 779	***	1,6	12 763	12 785		0,2
16.3- Other symptoms, signs, ill-defined causes	28 196	28 118		-0,3	29 820	29 687		- 0,4
17.1.1 - Transport accidents	3 285	3 202	**	-2,5	3 146	3 099		- 1,5
17.1.2 - Accidental falls	7 831	7 655	****	-2,2	8 308	8 295		- 0,2
17.1.3 - Drowning and accidental submersion	966	954		-1,2	925	929		0,4
17.1.4 - Accidental poisoning	1 810	1 874	**	3,5	1 733	1 831	****	5,7
17.1.5 - Other accidents	13 816	13 863		0,3	14 328	13 894	*****	- 3,0
17.2 - Suicide and intentional self-harm	8 626	8 507	*	-1,4	8 406	8 351		- 0,7
17.3- Homicide, assault	326	316		-3,1	284	292		2,8
17.4-Event of undetermined intent	795	844	***	6,2	1 112	873	*****	- 21,5
17.5- Other external causes of injury and poisoning	1 393	1 336	**	-4,1	1 535	1 454	****	- 5,3
Total	578 631	578 631			589 549	589 549		

Note : significance degrees of counting differentials come from equality tests assuming real occurrences were Poisson distributed. * pval<.3, ** pval<.2, *** pval<.1, ****pval<.05, ***** pval<.01.

Table 5 - comparison of predictions and real codes distributions on the 86 European short-list codes, all 2016 and 2017 death certificates

Kolmogorov Smirnov tests for equal distributions between real and prediction are also performed, and the underlying khi2 statistic used in choosing the model. At the end, we have

	86 European short list level
Death certificates on which the AI approach is used (ie those that were manually coded)	336
All death certificates	173

4- Results and trend comparisons

The next two tables report countings and standardized mortality rates per CoD as in the European short list as observed for 2016, 2017 and 2020 and provisional for 2018 and 2019, stressing which counting may be either or underestimated in those two years. Figures may differ from the ones published in Eurostat database because they do not cover exactly the same scope (here we focus on French residents died in France including some overseas collectivities which are excluded from the NUTS FR, and we use all certificates received). However, tendencies and main messages can be directly extended to figures published in the Eurostat database.

Countings	2015	2016	2017	Provisional 2018	Provisional 2019	2020	Risk
01.1- Tuberculosis	434	403	402	372	401	295	**underestimation**
01.2- AIDS (HIV diseases)	390	334	237	267	244	201	
01.3- Viral hepatitis	600	587	773	456	445	351	
01.4- Other infectious and parasitic diseases	9 797	9 180	10 193	10 242	10 694	10 208	**overestimation**
02.1.01-Malignant neoplasms of lip, oral cavity, pharynx	3 924	3 936	3 809	3 665	3 474	3 636	**underestimation**
02.1.02-Malignant neoplasms of oesophagus	3 893	3 902	3 865	3 766	3 775	3 630	
02.1.03-Malignant neoplasms of stomach	4 559	4 602	4 612	4 541	4 420	4 258	
02.1.04-Malignant neoplasms of colon, rectum, anus	17 658	18 029	17 996	17 324	17 337	17 197	
02.1.05-Malignant neoplasms of liver and intrahepatic bile d	8 518	8 776	8 551	8 510	8 566	8 727	
02.1.06-Malignant neoplasms of pancreas	10 921	11 300	11 467	11 745	12 182	12 476	
02.1.07-Malignant neoplasms of larynx	1 091	1 069	1 000	942	882	827	
02.1.08-Malignant neoplasms of trachea, bronchus, lung	32 150	31 877	31 402	31 172	31 105	30 935	
02.1.09- Malignant neoplasms of skin	1 850	1 748	1 767	1 755	1 796	1 756	
02.1.10-Malignant neoplasms of breast	12 580	12 936	13 013	13 081	12 967	13 008	
02.1.11-Malignant neoplasms of cervix uteri	763	801	817	857	775	770	
02.1.12-Malignant neoplasms of other and unspecified parts	2 755	2 838	2 903	2 882	2 884	2 845	
02.1.13-Malignant neoplasms of ovary	3 491	3 495	3 545	3 315	3 466	3 341	
02.1.14-Malignant neoplasms of prostate	8 919	9 022	9 212	9 310	9 408	9 178	
02.1.15-Malignant neoplasms of kidney	3 640	3 597	3 612	3 455	3 332	3 483	
02.1.16-Malignant neoplasms of bladder	5 230	5 349	5 146	5 287	5 209	5 345	
02.1.17-Malignant neoplasms of brain and central nervous s	3 885	3 964	4 087	3 721	3 944	4 035	**underestimation**
02.1.18-Malignant neoplasms of thyroid	412	378	420	431	388	362	
02.1.19-Hodgkin disease and lymphomas	4 843	4 869	4 936	4 649	4 745	4 875	**underestimation**
02.1.20- Leukaemia	5 936	6 016	6 134	5 905	5 923	6 165	
02.1.21-Other malignant neoplasms of lymphoid and haema	3 385	3 433	3 230	3 234	3 319	3 283	
02.1.22-Other malignant neoplasms	21 315	21 738	22 106	23 436	24 033	23 018	
02.2-Non-malignant neoplasms (benign and uncertain)	7 441	7 527	7 587	7 781	7 832	7 656	
03 Diseases of the blood and blood-forming organs and cer	2 207	2 291	2 570	3 219	3 188	2 801	**overestimation**
04.1- Diabetes mellitus	12 268	11 848	11 927	11 756	12 063	12 264	
04.2- Other endocrine, nutritional and metabolic diseases	9 357	9 407	10 189	10 496	11 116	11 334	
05.1- Dementia	19 309	19 755	19 661	21 058	20 744	18 596	
05.2- Alcohol abuse (including alcohol psychosis)	2 594	2 577	2 460	2 698	2 739	2 472	
05.3 - drug dependence, toxicomania	160	230	189	211	241	229	
05.4 - Other mental and behavioural disorders	3 344	3 452	3 608	3 948	4 127	4 091	**overestimation**
06.1- Parkinson's disease	6 192	6 642	6 826	6 897	6 865	7 012	
06.2 - Alzheimer's disease	20 872	21 111	20 962	20 396	19 194	18 244	
06.3- Other diseases of the nervous system and the sense or	10 944	11 128	11 782	12 170	12 573	12 360	
07.1.1-Acute myocardial infarction	14 659	14 031	13 976	13 438	13 258	12 922	**underestimation**
07.1.2-Other ischaemic heart diseases	19 310	18 985	19 053	19 032	18 482	18 170	
07.2-Other heart diseases	53 623	53 184	53 652	53 935	50 940	48 061	
07.3-Cerebrovascular diseases	32 176	32 213	31 776	31 834	31 763	31 112	
07.4- Other diseases of the circulatory system	25 019	25 117	25 165	24 612	24 450	24 498	
08.1 - Influenza	1 915	961	2 501	2 297	2 776	871	
08.2 - Pneumonia	13 371	13 305	13 920	14 162	14 264	11 559	
08.3.1 - Asthma	891	929	914	863	855	721	
08.3.2-Other chronic lower respiratory diseases	10 746	10 416	10 747	11 058	10 899	9 372	
08.4- Other diseases of the respiratory system	15 811	15 722	16 675	16 492	16 407	16 188	**overestimation**
09.1 - Ulcer of stomach, duodenum, jejunum	853	867	862	782	724	837	
09.2 - Cirrhosis, fibrosis, and chronic hepatitis	7 056	6 914	6 775	6 671	6 630	6 776	
09.3- Other diseases of the digestive system	16 081	16 396	16 533	16 021	16 368	17 362	**underestimation**
10 Diseases of the skin and subcutaneous tissue	1 379	1 489	1 623	1 628	1 724	1 639	
11.1- Rheumatoid arthritis and osteoarthristis	555	565	578	565	534	583	
11.2- Other diseases of the musculoskeletal system/connect	3 651	3 589	3 424	2 942	3 236	3 440	
12.1-Diseases of kidney and ureter	7 637	7 572	8 105	8 072	8 714	8 580	
12.2- Other diseases of the genitourinary system	2 461	2 550	2 752	2 941	3 098	3 512	
13 Complications of pregnancy, childbirth and puerperium	40	40	41	30	18	41	**overestimation**
14 Certain conditions originating in the perinatal period	1 571	1 501	1 685	1 616	1 607	1 443	
15 Congenital malformations and chromosomic abnormalit	1 694	1 675	1 624	1 477	1 539	1 502	**underestimation**
16.1- Sudden infant death syndrome	165	176	139	165	109	114	
16.2- Unknown and unspecified causes	25 361	27 198	29 680	30 797	34 612	34 657	**overestimation**
16.3- Other symptoms, signs, ill-defined causes	29 163	28 069	29 700	31 899	32 437	33 001	
17.1.1 - Transport accidents	3 199	3 186	3 054	2 664	2 564	2 144	**underestimation**
17.1.2 - Accidental falls	7 684	7 781	8 262	8 996	9 082	9 073	**underestimation**
17.1.3 - Drowning and accidental submersion	904	920	884	879	770	668	
17.1.4 - Accidental poisoning	2 042	1 800	1 725	1 175	1 316	1 505	**overestimation**
17.1.5 - Other accidents	13 991	13 694	14 202	13 019	13 982	14 272	
17.2 - Suicide and intentional self-harm	9 118	8 592	8 367	8 882	8 622	8 986	
17.3- Homicide, assault	336	312	281	435	494	472	
17.4-Event of undetermined intent	873	785	1 102	1 062	1 187	1 552	
17.5- Other external causes of injury and poisoning	844	1 391	1 525	2 583	2 560	1 361	**underestimation**
COVID-19	-	-	-	-	-	69 238	
Total	**591 806**	**592 072**	**604 298**	**607 974**	**612 417**	**667 496**	

Table 7 Trends in countings per CoD of the 86 European short list, with indication of risk of under/over estimation for the provisionary 2018 and 2019 data. French residents died in France, these figures may slightly differ from those published on Eurostat database because of some scope differentials.

Standardized mortality rates	2015	2016	2017	Provisional 2018	Provisional 2019	2020	Risk
01.1- Tuberculosis	0,6	0,6	0,6	0,5	0,6	0,4	underestimation
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	0,7	0,7	0,5	
01.4- Other infectious and parasitic diseases	14,5	13,2	14,1	14,0	14,4	13,5	overestimation
02.1.01-Malignant neoplasms of lip, oral cavity, pharynx	6,4	6,3	6	5,7	5,3	5,5	underestimation
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	6,9	6,6	6,3	
02.1.04-Malignant neoplasms of colon, rectum, anus	27,5	27,4	26,8	25,4	24,9	24,3	
02.1.05-Malignant neoplasms of liver and intrahepatic bile ducts	14,1	14,2	13,6	13,4	13,1	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	48,9	47,9	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	16,6	16,1	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 ute	3,6	3,7	3,7	3,6	3,6	3,4	
02.1.13-Malignant neoplasms of ovary	4,7	4,7	4,7	4,3	4,4	4,2	
02.1.14-Malignant neoplasms of prostate	17,5	17,2	17,1	16,9	16,6	15,9	
02.1.15-Malignant neoplasms of kidney	5,9	5,7	5,7	5,3	5,0	5,1	
02.1.16-Malignant neoplasms of bladder	9	9	8,5	8,4	8,2	8,3	
02.1.17-Malignant neoplasms of brain and central nervous system	6,1	6,2	6,3	5,7	6,0	6	underestimation
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	7,0	6,9	7	underestimation
02.1.20- Leukaemia	9,4	9,3	9,3	8,8	8,7	8,8	
02.1.21-Other malignant neoplasms of lymphoid and haematopoie	5,3	5,3	4,9	4,7	4,8	4,7	
02.1.22-Other malignant neoplasms	33,6	33,5	33,4	34,8	35,0	32,9	
02.2-Non-malignant neoplasms (benign and uncertain)	11,4	11,2	11	11,0	10,9	10,4	
03 Diseases of the blood and blood-forming organs and certain di	3,2	3,3	3,6	4,5	4,4	3,7	overestimation
04.1- Diabetes mellitus	18,4	17,4	17	16,4	16,5	16,5	
04.2- Other endocrine, nutritional and metabolic diseases	13,4	13	13,4	13,7	14,2	14,3	
05.1- Dementia	26,3	25,6	24,4	25,6	24,4	21,2	
05.2- Alcohol abuse (including alcohol psychosis)	4,2	4,1	3,9	4,2	4,3	3,8	
05.3 - drug dependence, toxicomania	0,3	0,4	0,3	0,3	0,4	0,4	
05.4 - Other mental and behavioural disorders	5	4,9	5,1	5,4	5,6	5,4	overestimation
06.1- Parkinson's disease	9,7	10,2	10,2	10,1	9,9	9,9	
06.2 - Alzheimer's disease	26,8	26,2	25,2	24,0	22,0	20,5	
06.3- Other diseases of the nervous system and the sense organs	16,9	16,7	17,4	17,7	18,0	17,4	
07.1.1-Acute myocardial infarction	22,7	21,3	20,8	19,5	19,0	18,3	underestimation
07.1.2-Other ischaemic heart diseases	30,5	29,1	28,3	27,7	26,5	25,6	
07.2-Other heart diseases	77,6	74,1	72,5	70,8	65,0	60,5	
07.3-Cerebrovascular diseases	46	44,9	42,9	42,2	41,3	39,7	
07.4- Other diseases of the circulatory system	36,1	34,9	33,9	32,5	31,3	31	
08.1 - Influenza	2,7	1,4	3,4	3,1	3,7	1,2	
08.2 - Pneumonia	20	19,2	19,3	19,3	18,7	15,3	
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	16,4	15,8	13,5	
08.4- Other diseases of the respiratory system	24	23	23,6	23,0	22,2	21,8	overestimation
09.1 - Ulcer of stomach, duodenum, jejunum	1,3	1,3	1,2	1,1	1,0	1,1	
09.2 - Cirrhosis, fibrosis, and chronic hepatitis	11,4	11,1	10,7	10,4	10,2	10,3	
09.3- Other diseases of the digestive system	23,7	23,5	23,1	21,9	22,0	22,9	underestimation
10 Diseases of the skin and subcutaneous tissue	1,9	2	2,1	2,1	2,1	2	
11.1- Rheumatoid arthritis and osteoarthristis	0,7	0,7	0,7	0,7	0,6	0,7	
11.2- Other diseases of the musculoskeletal system/connective tiss	5,3	5,1	4,8	4,0	4,3	4,5	
12.1-Diseases of kidney and ureter	11,5	11	11,4	11,2	11,8	11,3	
12.2- Other diseases of the genitourinary system	4	3,9	4,1	4,2	4,3	4,8	
13 Complications of pregnancy, childbirth and puerperium	0,1	0,1	0,1	0,0	0,0	0,1	overestimation
14 Certain conditions originating in the perinatal period	1	1	1,1	1,1	1,1	1	
15 Congenital malformations and chromosomic abnormalities	2,1	2,1	2	1,8	1,9	1,8	underestimation
16.1- Sudden infant death syndrome	0,1	0,1	0,1	0,1	0,1	0,1	
16.2- Unknown and unspecified causes	38,2	39,8	42,2	43,1	47,1	46,3	overestimation
16.3- Other symptoms, signs, ill-defined causes	41,3	38,3	38,9	40,7	40,3	39,8	
17.1.1 - Transport accidents	5	5	4,8	4,1	4,0	3,3	underestimation
17.1.2 - Accidental falls	11,4	11,2	11,5	12,2	12,0	11,7	underestimation
17.1.3 - Drowning and accidental submersion	1,4	1,4	1,4	1,3	1,2	1	
17.1.4 - Accidental poisoning	3,1	2,7	2,6	1,7	1,9	2,2	overestimation
17.1.5 - Other accidents	20,9	19,9	20	17,8	18,8	18,9	
17.2 - Suicide and intentional self-harm	14,8	13,9	13,4	14,1	13,6	14,1	
17.3- Homicide, assault	0,5	0,5	0,4	0,7	0,8	0,7	
17.4-Event of undetermined intent	1,4	1,3	1,8	1,7	1,9	2,4	
17.5- Other external causes of injury and poisoning	1,2	2,1	2,2	3,8	3,6	1,9	underestimation
COVID-19	0	0	0	0,0	0,0	92,9	

Table 8 Trends in standardized mortality rates per CoD of the 86 European short list, with indication of risk of under/over estimation for the provisionary 2018 and 2019 data.

5- Conclusion and next steps

These provisional data for 2018 and 2019 are to be considered with caution for the categories at risk of under or overestimation as presented in the last two tables. Further work is planned in 2023 to have a particular look at these categories. No analysis have been done yet on regional, age-based and sex-based comparisons and relying statistics at fine level of CoD should also be for the moment considered with caution. However, for wide groups of CoD and national tendencies, they can be used and will give Eurostat the possibility to aggregate 2018 and 2019 CoD at the EU level .

Next steps involve having manual coding/ checking of some well chosen subsamples of 2018 and 2019 data, retrain if necessary algorithms with these new informations for providing final estimates for 2018 and 2019 by mid-2023. France also work on including AI coding as part of its usual production process.

6- References

-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

- 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/

Annex A - Transformer specifications and hyperparameters

For transformer algorithm with and without transfer learning

batch_size = 200

buffer_size = 5 000

d_model = 514

latent_dim = 2 048

num_heads = 8

num_layers = 1

dropout = 0.1

epoch = 100

Optimizer : Adam

Loss :Sparse categorical crossentropy

Metric for validation: Accuracy

Metrics for test: Precision, Recall, F_measure, full accuracy

Train/test samples

For performance analysis on 2016/2017 data :

- the train sample includes all death certificates of years 2011-2015 + those fully automated coded by the rule-based system IRIS for 2016 and 2017 (ie 3,447,459 certificates)
- Test on certificates of 2016 and 2017 manually coded (ie 487,047 certificates)

For transfer learning

- Train: all certificates predicted to be rejected by IRIS MUSE by a simple binary classification model (ie 422 456 certificates)
- Test : all certificates of 2016-2017 rejected by IRIS MUSE (ie 78 753 certificates)

For 2018/2019 predictions:

- the train sample includes all death certificates of years 2011-2017 + 85% of those of 2020, which do not concern COVID (ie 5,173,106 certificates)
- Test on 15% of 2020 certificates, which do not concern COVID (ie 447,837 certificates)
- Predictions on 2018 and 2019 not fully coded by IRIS 447,837 certificates

For transfer learning

- Train: all certificates predicted to be rejected by IRIS MUSE by a simple binary classification model (ie 1,877,493 certificates)

- Test : all certificates of 2016-2017 rejected by IRIS MUSE (ie 78 753 certificates)
- Predictions on 76,230 certificates

Annex B - Decision rules for determining the underlying cause of death

```
# step 1 : applying best model per category

 Synt <- ifelse (  TL_86 %in% c("17.3") ,  iris_tl2 , NA)

 Synt <- ifelse (  IRIS_86 %in% c("05.1","05.2") ,  iris_keras4 ,  Synt)

 Synt <- ifelse ( KERAS4_86 %in%
c("01.4","02.2","04.1","04.2","07.1","07.2","07.3","07.4","08.4","17.1","17.3","17.4","17.5") ,  keras4 ,
Synt)

 Synt <- ifelse (  IRIS_86 %in% c("06.1","06.2","06.3","09.1","09.2","09.3","11.1","11.2","16.2","16.3") ,
iris_keras4 ,  Synt)

 Synt <- ifelse (  TL_86 %in% c("01.2","08.1","08.2","08.3","10  ","12.1","12.2","13  ","15  ","16.1") ,
iris_tl2 ,  Synt)

 Synt <- ifelse (  IRIS_86 %in% c("05.3","14  ") ,  iris_keras4 ,  Synt)

 Synt <- ifelse ( KERAS4_86 %in% c("01.3","02.1","05.4","17.2") ,  keras4 ,  Synt)

 Synt <- ifelse (  IRIS_86 %in% c("03  ") ,  iris_keras4 ,  Synt)


# step 2 : specific treatment for 5 underestimates

        # which ones

 Tub <- ifelse( KERAS4_86 =="01.1",1,0) +ifelse(  IRIS_86 =="01.1",1,0)

 Hep <- str_count( keras_code, "b18")+ str_count( keras_code,"b19")

 Gros <- str_count( keras_code, "o8")+ str_count( keras_code,"o9")

 Cong <- str_count( keras_code, "q8")+ str_count( keras_code,"q9")

 Hom <- ifelse( KERAS4_86 =="17.3",1,0) +ifelse(  IRIS_86 =="17.3",1,0)  + ifelse(  TL_86
=="17.3",1,0)

        # what to do

 Synt <- ifelse ( Tub > 0 , "B909" ,  Synt)

 Synt <- ifelse ( Gros> 0 , "O95" ,   Synt)
```

```
 Synt <- ifelse ( Cong> 1 , "Q600" ,   Synt)

 Synt <- ifelse ( Hom > 0 , "X99" ,   Synt)

 Synt <- ifelse (is.na( Synt) &  Hep > 0, "B182" ,  Synt)


# step 3: decision by default

 Synt <- ifelse (is.na( Synt) ,  iris_keras4 ,  Synt)
```