EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Combining deep neural networks, a rule-based expert system and targeted manual coding for ICD-10 coding causes of death of French death certificates

Diane Martin (PharmD) - Inserm - CépiDc

Production department manager

Elisa Zambetta, Nirintsoa Razakamanana, Aude Robert, François Clanché, Cecilia Rivera, Zina Hebbache, Rémi Flicoteaux, Elise Coudin

INSERM
Centre d'Epidémiologie sur les causes médicales de décès (CépiDc)
Paris - France

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

Inserm CépiDc

The conference is partly financed by the European Union

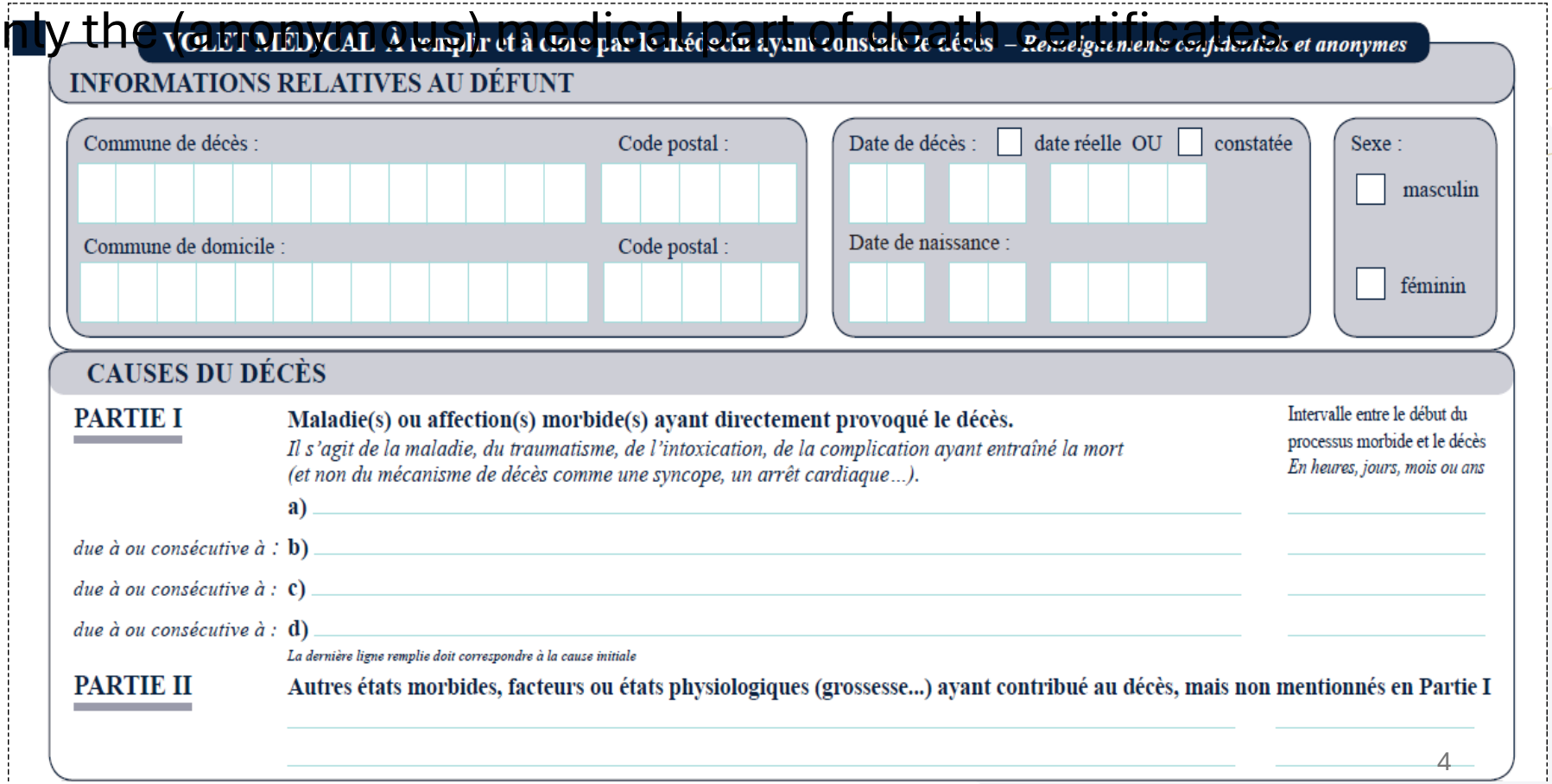
Introduction

- INSERM in charge of producing **statistics on causes of death (CoD) for France**
  - Around **650 000 deaths** per year
- Health surveillance, Official statistics, Research
- CoD statistics governed by European regulation n°328/2011
  - provide data for year N before the end of year N+2
  - In compliance with International Classification of Diseases mortality rules (ICD-WHO)
- Delays in dissemination > Inter-ministerial mission in 2021
  - Catch-up and renovation of the production process > specific project with dedicated resources (with the Health Statistical Service - DREES)
  - Integrating new tools into regular production : deep neural networks

The conference is partly financed by the European Union

Death certificate - medical part

- **International standard (ICD-WHO)** completed by a physician certifying death
- INSERM receives only the (anonymous) medical part of death certificates



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

INFORMATIONS RELATIVES AU DÉFUNT

Commune de décès : Code postal :

Commune de domicile : Code postal :

Date de décès : ☐ date réelle OU ☐ constatée

Date de naissance :

Sexe : ☐ masculin ☐ féminin

CAUSES DU DÉCÈS

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

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

a) _____

due à ou consécutive à : b) _____

due à ou consécutive à : c) _____

due à ou consécutive à : d) _____

La dernière ligne remplie doit correspondre à la cause initiale

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

The conference is partly financed by the European Union

Death certificate - medical part

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

LIEU DU DÉCÈS

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

MORT SUBITE S'agit-il d'un décès brutal et inattendu, évocateur de mort subite* ?

☐ oui ☐ non ☐ ne sait pas

** décès non traumatique (adulte, enfant, nourrisson) avec mode de survenue brutal (en moins d'une heure ou probablement) et inattendu (exclusion des maladies chroniques au stade terminal)*

CIRCONSTANCES APPARENTES DU DÉCÈS

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

EN CAS DE MORT VIOLENTE (accidentelle, délictuelle, suicidaire, criminelle)
Précisez le lieu de survenue de l'événement déclencheur :

- ☐ Domicile
- ☐ Lieu de sport
- ☐ Voie publique
- ☐ Commerce
- ☐ Local industriel, chantier
- ☐ Exploitation agricole
- ☐ Établissement accueillant du public
- ☐ Autre lieu ou indéterminé

GROSSESSE La femme décédée était-elle enceinte ?

- ☐ non, pas au cours de l'année précédant le décès
- ☐ pas au moment du décès, mais grossesse terminée depuis 42 jours ou moins
- ☐ pas au moment du décès, mais grossesse terminée depuis plus de 42 jours et moins d'1 an
- ☐ oui, au moment du décès
- ☐ ne sait pas

La grossesse a-t-elle contribué au décès ? ☐ oui ☐ non ☐ ne sait pas

ACTIVITÉ PROFESSIONNELLE Le décès est-il survenu lors d'une activité professionnelle* ?

☐ oui ☐ non ☐ ne sait pas

** toute activité source de revenu (y compris au domicile), les trajets domicile-travail, les déplacements professionnels, etc.*

RECHERCHE DE LA CAUSE DU DÉCÈS
Une recherche de la cause du décès a-t-elle été demandée ?

☐ oui, recherche médicale ☐ oui, recherche médico-légale ☐ non

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

SIGNATURE *Nom lisible et cachet obligatoire du médecin*

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

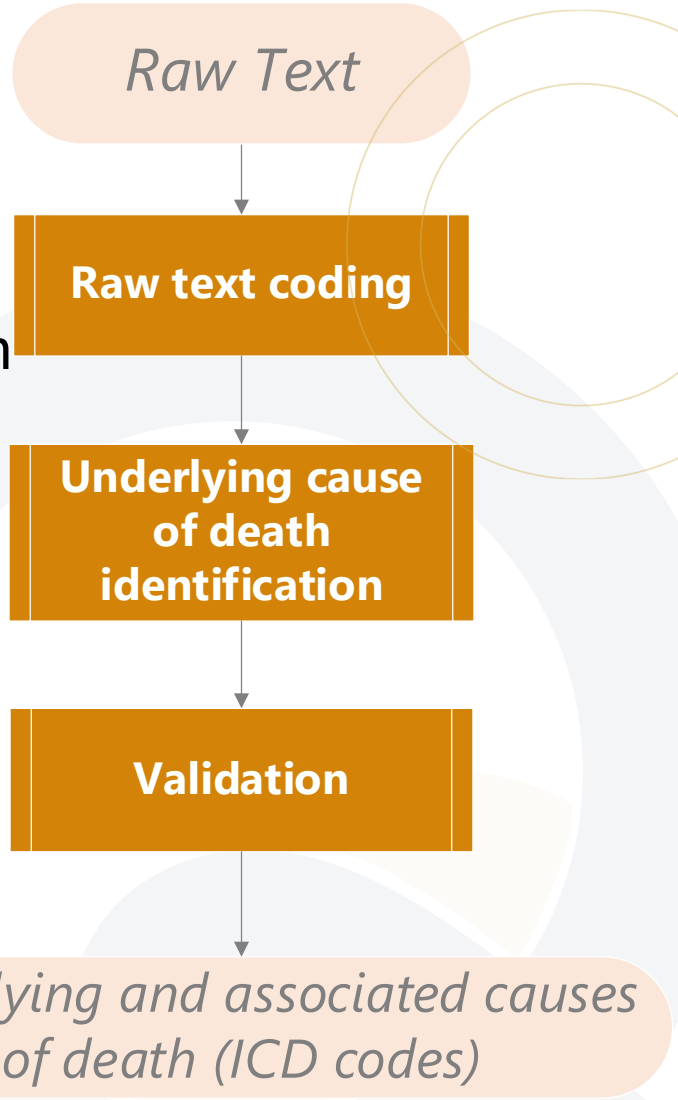
The conference is partly financed by the European Union

Coding causes of death

- Code all causes mentioned on the certificate using ICD-10

  (raw text to ICD-10 codes)

- Determine, among theses codes, the underlying cause of death
  - By applying the mortality decision rules provided by the ICD
  - take into account the sequence of causes



Underlying cause of death (UC):
the disease or injury which initiated the train of morbid events leading directly to death, or the circumstances of the accident or violence which produced the fatal injury

Associated causes:
causes mentioned on the death certificate, excluding the underlying cause

The conference is partly financed by the European Union

3- coding mode campaign

- Expert system for automated batch coding (63%) – IRIS/Muse
  1. Iris codes all causes mentionned using ICD dictionary
  2. Muse determines the UC code using the codes identified in step 1 according to causal sequence

- Assisted coding by the coding team (3% 2018-2019 ; 14% 2021)
  - Deaths of specific interest for Public Health (maternal deaths, HIV, neonatal or infant deaths, etc.)
  - Randomly selected certificates to test and maintain DNNs
  - Targeted certificates to ensure sufficient precision for each ICD European shortlist category

The number of certificates to be coded manually is fixed in advance accounting for human resources available in order to respect the delay of production : around 100 000 certificates per year

- Deep neural network (DNN) predictions trained on past labelled data (34% 2018-2019;23% 2021)
  - Determination of all ICD-10 codes of all causes mentioned and UC code

Approach: deep neural networks

*Coding all causes of death and determining which is the underlying cause of death is seen as a **translation task***

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]*

- Seq-to-seq algorithms, Transformers architecture, state of the art, achievable on a conventional infrastructure
- 96 millions of parameters to estimate
- Supervised learning, training done on 5.3 millions of already coded death certificates (2011–2021) (no data, nor algo from outside)
- Training take 4 days on one machine with one GPU of 48GO of RAM, can predict 120,000 certificates in 24hours

UC code can be determined by

1. applying IRIS/Muse on the sequence of codes predicted by the DNN or
2. by directly using the code predicted by DNN

The conference is partly financed by the European Union

Targeting certificates to be coded by the coding team

Objective : optimize the allocation of limited human resources to achieve a given level of precision in each dissemination margin [Eurostat shortlist of ICD codes]

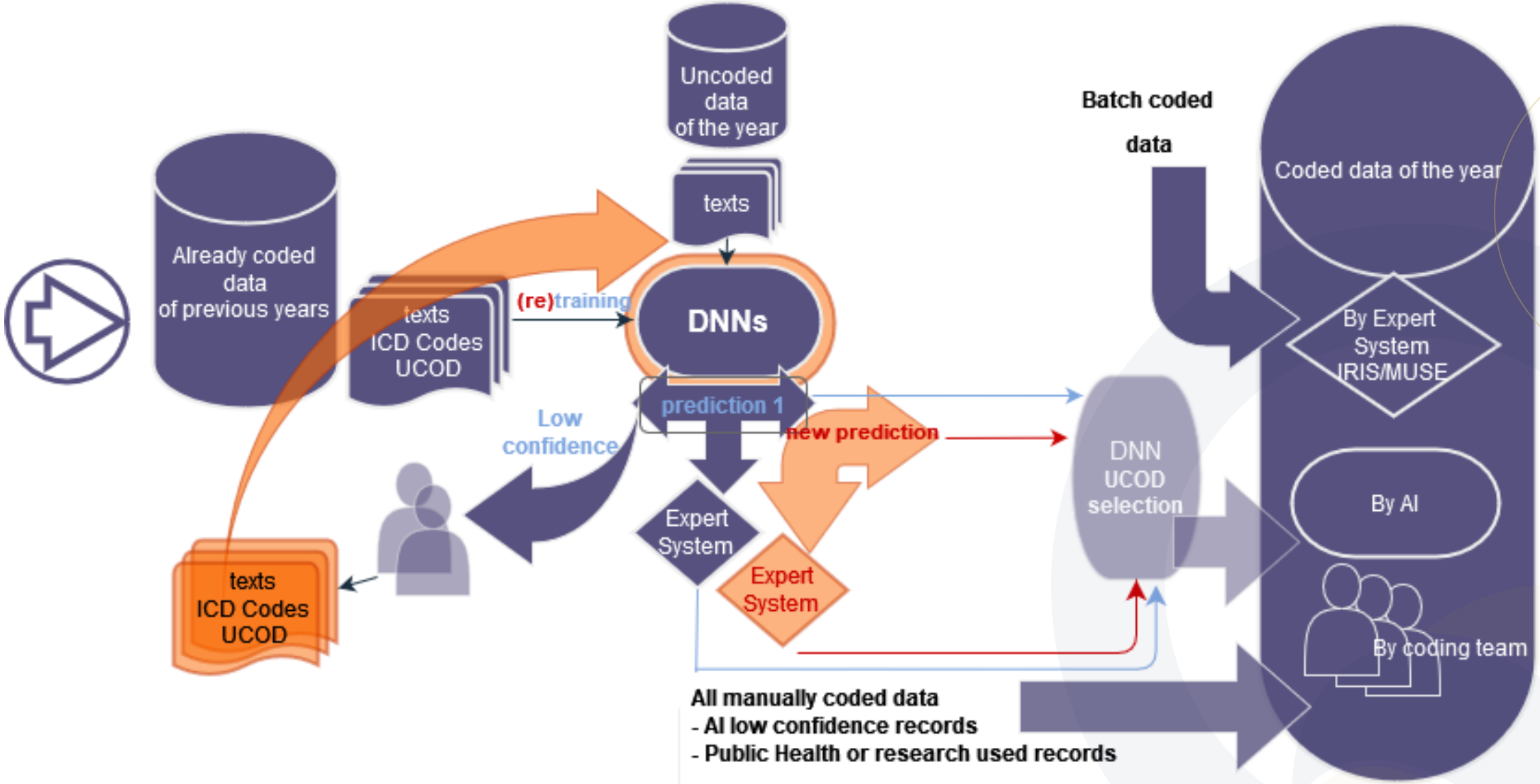
The coding team code around 100 000 certificates per year among the 240 000 certificates not automatically coded by the expert system for a given year

- Computation of a confidence indicator for each certificate :
  - Based on the probability estimate that the UC code predicted by DNN is the same as the one that the coding team would have coded, conditional on individual characteristics of the certificate

- Send by order to the coding team certificates with the lowest confidence indicators within a given European shortlist category where the overall precision is below a given threshold in order to achieve XX% of precision in each category

- in practice - several batches of certificates are sent in order to the coding team - to achieve 90 %, 92% of precision (for 2018/2019); then 94, 95,...97% for 2021 (and 2022).

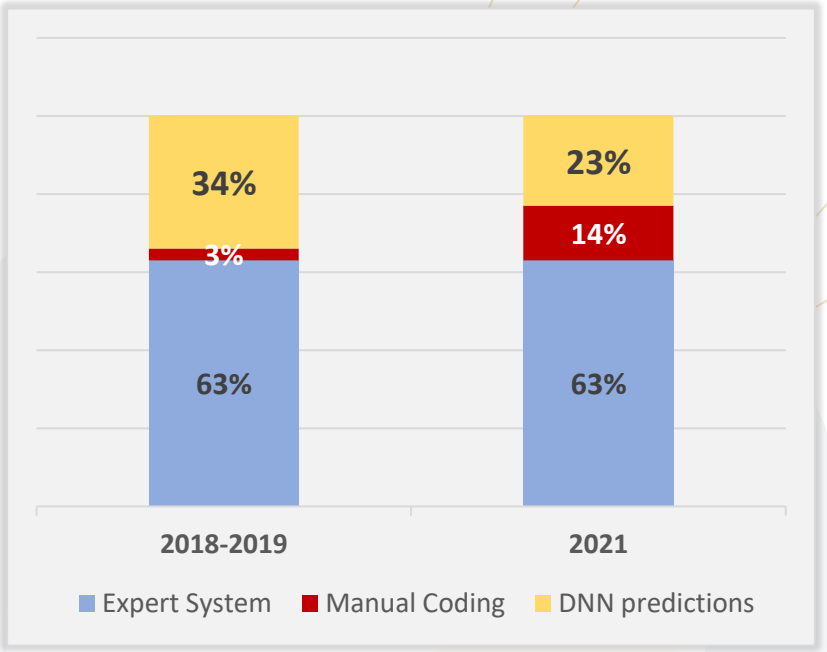
3- coding mode campaign



The conference is partly financed by the European Union

3- coding mode campaign

- Catch up : 2018-2019 death certificates
  - 63% fully automatically coded by IRIS/Muse : 380,000/year
  - 34% coded by DNN prediction (some combined to IRIS/Muse 200,000/year
  - 3% manually coded (assisted by IRIS/Muse and partly targeted) 18000/year



- Regular production : 2021 death certificates
  - 63% fully automatically coded by expert system IRIS/Muse : 406,000
  - 23% coded by DNN prediction (some combined to IRIS/Muse):149,000
  - 14% manually coded (assisted by IRIS/Muse and partly targeted): 93,000

The conference is partly financed by the European Union

Performance analysis and how to ensure statistical accuracy

- A reference test population, with certificates coded in the "traditional“ way (manual coding + expert System), but not used to train the AI models.
- Global indicators:
  - “Accuracy” : % certificates for which UC code identified by the 3-mode campaign is the same as that which would have been identified with the traditional campaign
  - 2021 : 95.7% of accuracy at the ICD code level; 97.3% at ICD European shortlist level
  - 2018-2019 : 93.4% of accuracy at the ICD code level ; 95.6% at ICD European Shortlist level
- Category-based indicators – for exemple « Alzheimer's disease »:
  - precision index : 98.9%
  - recall index : 98.3%
  - difference in countings : 25747 (true) VS 25851 (predicted) -> +0.4% (NS)
- At the end of the process, some category based indicators are estimated to be below 90% in 2018-2019 ; only 2 between 90% and 92% in 2021
- These indicators are used at every stage of the production process

The conference is partly financed by the European Union

Discussion and on-going

Lessons learnt

- Integrating AI in production is achievable in a fully controlled way (no data / algo from outside)
- Replicability/explainability
- New skills in the team (data scientist and data engineer) that have to be maintained (different for rule-based automatisation)
- Change in work
    - cultural change - more statistical monitoring
    - need for more expertise from coding team
    - new work organisation

→ conducting change

Opportunities

- Timeliness and having first estimates

-Monitor statistically significant issues [past errors, …]

-Focus coding team expertise on important matters

-ICD-10 code a large amount of data (useful for transition between classifications /bridge coding ICD9/10 ou ICD10/11)

Questions still open

-% of manual coding needed for maintaining/monitor the algo →open question as DNNs can loose rapidly performance (<> rule-based automatisation)

-Maintaining the quality of the training set [data]/ having a Gold standard base

-Interaction with the coding team: showing or not AI predictions to coding team ?

-ICD11 ?

The conference is partly financed by the European Union

References

- Methods
  - Zambetta E., Razakamanana N., Robert A., Clanché F., Rivera C., Martin D., Hebbache Z., Flicoteaux R., Coudin E.,Combining deep neural networks, a rule-based expert system and targeted manual coding for ICD-10 coding causes of death of French death certificates from 2018 to 2019, International Journal of Medical Informatics,2024, 105462, ISSN 1386-5056,https://doi.org/10.1016/j.ijmedinf.2024.105462
  - Zambetta,E., Razakamanana, N., Robert A., Clanché F., Martin D., Hebbache Z., Flicoteaux R. and Coudin E. Combining a Transformer-based approach, rule-based expert system, and targeted manual coding for ICD-10 cause of death coding of French death certificates in 2018 and 2019 – CEPIDC Working Paper n°2, 2023. DT_CEPIDC_N2_english_version.pdf (inserm.fr)
  - Hebbache Z., Robert A., Clanché F., Coudin E., Martin D., « Rapport de Production, Année de décès 2018-2019 ». DT_CEPIDC_N3_Rapport de production 2018-2019_0.pdf (inserm.fr)
  - Hebbache Z., Boulet P., Robert A., Zambetta E., Razakamanana N., Coudin E., Martin D., « Rapport de Production, Année de décès 2021 ». DT_CEPIDC_N4_Rapport de production 2021.pdf (inserm.fr)
  - Clanché F., Razakamanana N., Coudin E., Robert A., "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. 2023 (PDF) DREES MÉTHODES N° 8 • mars 2023 (researchgate.net)

- Analyses
  - Fouillet A, Ghosn W, Rivera C, Clanché F, Coudin E. Grandes causes de décès en 2021 et tendances récentes. Bulletin épidémiologique hebdomadaire. 2023 Dec;(26):554–69. Grandes causes de décès en France en 2021| Santé publique France (santepubliquefrance.fr)
  - Cadillac, Manon, Fouillet, Anne, Rivera C, Clanché F, Coudin E. Grandes causes de décès en France en 2021 : une année encore fortement marquée par le Covid-19. Etudes et Résultats. 2023 Dec;1288. Grandes causes de décès en France en 2021 : une année encore fortement marquée par le Covid-19 | Direction de la recherche, des études, de l'évaluation et des statistiques (solidarites-sante.gouv.fr)

The conference is partly financed by the European Union

Thank you for attention

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

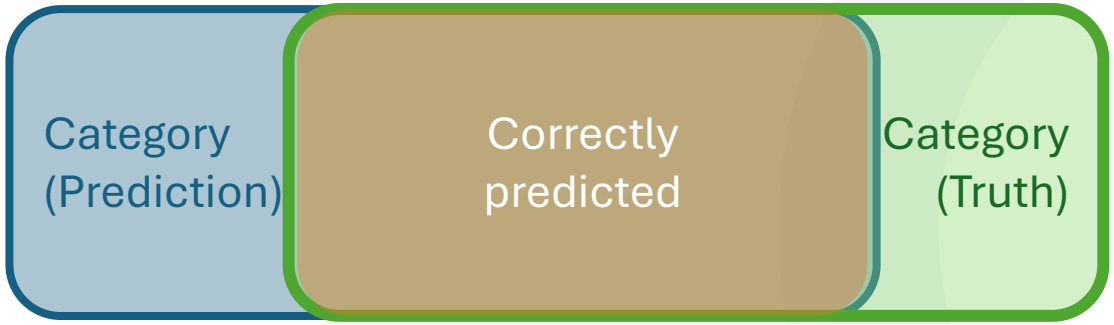
INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

Definitions

- **Precision:** proportion of observations correctly predicted by the model in the category relative to all predictions in the category
- **Recall:** proportion of observations correctly predicted by the model in the category relative to all observations actually in the category



- **F-measure :** harmonic mean of precision and recall.

Accuracy

2021 campaign	K5	K5IrisMuse	K4	K4IrisMuse	UC code choice	UC code choice + targeted manual coding	N. Obs
Test population that should have been coded manually							
ICD-10 4 digit level	0,791	0,801	0,786	0,799	0,815	0,897	332183
European Shorlist level	0,86	0,865	0,856	0,863	0,876	0,934	332183
All test reference population (Manual coding + expert system)							
ICD-10 4 digit level	0,913	0,917	0,911	0,916	0,923	0,957	797651
European Shorlist level	0,942	0,944	0,94	0,943	0,948	0,973	797651

2018-2019 campaign	K5	K5IrisMuse	K4	K4IrisMuse	UC code choice	UC code choice + targeted manual coding	N. Obs
Test population that should have been coded manually							
ICD-10 4 digit level	0,785	0,796	0,768	0,769	0,816	0,841	332183
European Shorlist level	0,856	0,861	0,83	0,829	0,878	0,894	332183
All test reference population (Manual coding + expert system)							
ICD-10 4 digit level	0,91	0,915	0,903	0,904	0,925	0,934	797651
European Shorlist level	0,94	0,942	0,929	0,929	0,949	0,956	797651

Accuracy by campaign step in 2021

	Test population that should have been coded manually			All test reference population	
	Counts	ICD-10 4 digit level	European shortlist	ICD-10 4 digit level	European shortlist
Surmodel	223904	0,815	0,815	0,923	0,923
+ special interest deaths	4197	0,820	0,878	0,925	0,949
+ random sample	72805	0,860	0,905	0,942	0,961
+ low AI confidence sample	31060	0,896	0,934	0,957	0,972
+ last verifications	217	0,897	0,934	0,957	0,973
Nb obs	332 183	332 183	332 183	797 651	797 651

Precision/recall / Predicted counts (shortlist level) 2021

		Expert system + AI pred + targeted manual coding						
UCOD European shortlist	Real UCOD	Precision	Recall	F-measure	Pred. UCOD	Pred/Real UCOD - 1	Sign of diff	F-measure 2018-2019
01.1- Tuberculosis	476	0,974	0,929	0,951	454	- 0,046		0,918
01.2- AIDS (HIV diseases)	332	0,988	1,000	0,994	336	0,012		0,990
01.3- Viral hepatitis	560	0,952	0,913	0,932	537	- 0,041		0,869
01.4- Other infectious and parasitic diseases	12936	0,968	0,953	0,961	12744	- 0,015	**	0,923
02.1.01-Malignant neoplasms of lip, oral cavity, pharynx	4996	0,977	0,961	0,969	4915	- 0,016		0,957
02.1.02-Malignant neoplasms of oesophagus	4797	0,982	0,986	0,984	4816	0,004		0,979
02.1.03-Malignant neoplasms of stomach	5790	0,982	0,987	0,984	5819	0,005		0,977
02.1.04-Malignant neoplasms of colon, rectum, anus	23061	0,987	0,987	0,987	23053	- 0,000		0,981
02.1.05-Malignant neoplasms of liver and intrahepatic bile ducts	11426	0,982	0,983	0,983	11442	0,001		0,973
02.1.06-Malignant neoplasms of pancreas	15433	0,992	0,993	0,993	15438	0,000		0,990
02.1.07-Malignant neoplasms of larynx	1271	0,965	0,965	0,965	1270	- 0,001		0,946
02.1.08-Malignant neoplasms of trachea, bronchus, lung	40493	0,986	0,992	0,989	40714	0,005		0,982
02.1.09- Malignant neoplasms of skin	2241	0,966	0,975	0,971	2262	0,009		0,962
02.1.10-Malignant neoplasms of breast	16601	0,985	0,990	0,987	16693	0,006		0,981
02.1.11-Malignant neoplasms of cervix uteri	1048	0,980	0,980	0,980	1048	-		0,971
02.1.12-Malignant neoplasms of other and unspecified parts of uterus	3630	0,982	0,976	0,979	3608	- 0,006		0,969
02.1.13-Malignant neoplasms of ovary	4424	0,986	0,986	0,986	4424	-		0,981
02.1.14-Malignant neoplasms of prostate	11882	0,984	0,985	0,984	11902	0,002		0,978
02.1.15-Malignant neoplasms of kidney	4626	0,980	0,976	0,978	4606	- 0,004		0,968
02.1.16-Malignant neoplasms of bladder	6874	0,982	0,984	0,983	6890	0,002		0,976
02.1.17-Malignant neoplasms of brain and central nervous system	5232	0,985	0,976	0,980	5187	- 0,009		0,972
02.1.18-Malignant neoplasms of thyroid	490	0,946	0,963	0,954	499	0,018		0,940
02.1.19-Hodgkin disease and lymphomas	6393	0,977	0,980	0,979	6413	0,003		0,972
02.1.20- Leukaemia	7856	0,979	0,978	0,979	7848	- 0,001		0,976

Inserm CépiDc

Precision/recall/ Predicted counts (shortlist level) 2021

02.1.21-Other malignant neoplasms of lymphoid and haematopoietic tissue	4290	0,974	0,977	0,976	4303	0,003		0,970
02.1.22-Other malignant neoplasms	29282	0,967	0,962	0,964	29144	- 0,005		0,937
02.2-Non-malignant neoplasms (benign and uncertain)	10175	0,960	0,958	0,959	10159	- 0,002		0,927
03 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	3491	0,958	0,899	0,927	3276	- 0,062	****	0,868
04.1- Diabetes mellitus	16008	0,975	0,958	0,966	15728	- 0,017	***	0,950
04.2- Other endocrine, nutritional and metabolic diseases	13704	0,963	0,946	0,954	13472	- 0,017	***	0,920
05.1- Dementia	25311	0,963	0,984	0,973	25880	0,022	****	0,962
05.2- Alcohol abuse (including alcohol psychosis)	3230	0,958	0,963	0,960	3248	0,006		0,919
05.3 - drug dependence, toxicomania	308	0,959	0,916	0,937	294	- 0,045		0,865
05.4 - Other mental and behavioural disorders	4907	0,953	0,933	0,943	4804	- 0,021	*	0,914
06.1- Parkinson's disease	8866	0,980	0,987	0,983	8929	0,007		0,975
06.2 - Alzheimer's disease	25747	0,987	0,989	0,988	25800	0,002		0,981
06.3- Other diseases of the nervous system and the sense organs	15541	0,965	0,959	0,962	15446	- 0,006		0,931
07.1.1-Acute myocardial infarction	18023	0,970	0,979	0,974	18199	0,010	*	0,964
07.1.2-Other ischaemic heart diseases	24438	0,962	0,966	0,964	24552	0,005		0,946
07.2-Other heart diseases	67415	0,970	0,974	0,972	67671	0,004		0,955
07.3-Cerebrovascular diseases	41319	0,970	0,972	0,971	41436	0,003		0,954
07.4- Other diseases of the circulatory system	33025	0,965	0,959	0,962	32819	- 0,006		0,937

Precision/recall/ Predicted counts (shortlist level) 2021

08.1 - Influenza	1668	0,981	0,980	0,981	1667	- 0,001		0,968
08.2 - Pneumonia	16322	0,973	0,974	0,974	16335	0,001		0,956
08.3.1 - Asthma	1077	0,966	0,971	0,969	1083	0,006		0,941
08.3.2-Other chronic lower respiratory diseases	13006	0,966	0,974	0,970	13109	0,008		0,953
08.4- Other diseases of the respiratory system	21100	0,964	0,959	0,962	20989	- 0,005		0,936
09.1 - Ulcer of stomach, duodenum, jejunum	1081	0,959	0,962	0,960	1085	0,004		0,930
09.2 - Cirrhosis, fibrosis, and chronic hepatitis	8986	0,971	0,980	0,976	9070	0,009		0,960
09.3- Other diseases of the digestive system	22147	0,962	0,959	0,961	22068	- 0,004		0,934
10 Diseases of the skin and subcutaneous tissue	2067	0,949	0,935	0,942	2036	- 0,015		0,898
11.1- Rheumatoid arthritis and osteoarthristis	726	0,949	0,927	0,938	709	- 0,023		0,887
11.2- Other diseases of the musculoskeletal system/connective tissue	4537	0,929	0,915	0,922	4469	- 0,015		0,870

Precision/recall/ Predicted counts (shortlist level)

12.1-Diseases of kidney and ureter	10646	0,963	0,948	0,955	10483	- 0,015	*	0,924
12.2- Other diseases of the genitourinary system	4029	0,941	0,929	0,935	3974	- 0,014		0,899
13 Complications of pregnancy, childbirth and puerperium	54	1,000	1,000	1,000	54	-		1,000
14 Certain conditions originating in the perinatal period	2048	0,998	1,000	0,999	2053	0,002		0,996
15 Congenital malformations and chromosomic abnormalities	2105	0,966	0,932	0,948	2031	- 0,035	*	0,920
16.1- Sudden infant death syndrome	179	0,983	0,994	0,989	181	0,011		0,978
16.2- Unknown and unspecified causes	20174	0,964	0,983	0,974	20572	0,020	****	0,964
16.3- Other symptoms, signs, ill-defined causes	40404	0,981	0,987	0,984	40630	0,006		0,978
17.1.1 - Transport accidents	3678	0,978	0,961	0,970	3616	- 0,017		0,961
17.1.2 - Accidental falls	11146	0,971	0,975	0,973	11196	0,004		0,946
17.1.3 - Drowning and accidental submersion	1090	0,976	0,972	0,974	1086	- 0,004		0,955
17.1.4 - Accidental poisoning	2163	0,948	0,916	0,932	2090	- 0,034	*	0,883
17.1.5 - Other accidents	18254	0,952	0,949	0,950	18184	- 0,004		0,913
17.2 - Suicide and intentional self-harm	11281	0,983	0,984	0,983	11289	0,001		0,972
17.3- Homicide, assault	499	0,986	0,962	0,974	487	- 0,024		0,937
17.4-Event of undetermined intent	1709	0,936	0,882	0,908	1611	- 0,057	***	0,806
17.5- Other external causes of injury and poisoning	1847	0,903	0,757	0,823	1549	- 0,161	****	0,675
18- COVID	35680	0,987	0,993	0,990	35867	0,005		0,985

Inserm CépiDc