Annonce de séminaire

Séminaire CépiDc-Drees - Causes de décès 2018-2019 - données provisoires codées en partie par *deep learning*, apports et limites.

Intervenants (entre autres) : Drees (François, Clanché et Daniel Razakamanana) - CépiDc-Inserm (Elise Coudin, Aude Robert)

Date : lundi 30 janvier, 14h
Lieu : Salle Trocadéro, 2ème étage Biopark - sonner à l'interphone Aviesan 2ème étage.
Institut de Santé Publique BIOPARK- Bâtiment A, 8, rue de la Croix de Jarry, 75013 Paris

Résumé :
Le CépiDc Inserm a transmis à Eurostat fin 2022 les données définitives sur les causes de décès de l'année 2020, ainsi que des données provisoires sur les causes de décès pour les années 2018 et 2019. Alors que la production des données 2020 a été tout-à-fait classique combinant codage par système expert sur règles déterministes (Iris, Muse) et codage manuel par l'équipe du CépiDc, 38% des données provisoires pour 2018 et 2019 ont été codées selon la classification internationale des maladies CIM 10 en partie par l'application d'algorithmes d'intelligence artificielle. La séquence des causes de décès et la cause initiale du décès ont dans ce cas été prédites par une approche de *deep learning* où des algorithmes de type *transformers* ont étaient entraînés sur les textes des certificats de décès des années précédentes (et suivante). Ces travaux menés par la Drees en collaboration étroite avec le CépiDc mettent en application et étendent les travaux précédents de Falissard et al 2020 et Falissard 2021.

Ces données provisoires 2018 et 2019 seront révisées pour obtenir des données définitives mi-2023.

Ce séminaire vise à présenter rapidement la méthodologie suivie, présenter la performance de la méthode et ses limites sur certaines catégories ciblées de façon à accompagner l'usage qui pourrait être fait de ces données provisoires au niveau individuel (via le SNDS) ou macro. On présentera donc les résultats de l'analyse de performance obtenue en comparant les prédictions de l'approche IA avec les véritables étiquettes des données 2016 et 2017 lorsque celles-ci ne sont pas utilisées dans l'entraînement. On présentera aussi les résultats pour 2018 et 2019 obtenus en les comparant aux tendances des années précédentes et suivantes. On mettra en évidence les catégories de codes pour lesquelles l'approche est performante et celles qui souffrent de sous ou sur-estimations systématiques.

L'enjeu de ce séminaire est double - échanger sur la méthodologie pour préparer les améliorations sur les données définitives mais surtout échanger sur l'utilisation potentielle de ces données au niveau individuel par les différents acteurs regroupés (recherche, acteurs de la santé et de la statistique publiques).

Pour assister à ce séminaire, merci de s'inscrire sur

https://docs.google.com/forms/d/e/1FAIpQLSdZO0wlsoq_G1oYVjveiptrovXI6WN-oXlpLvAaR76HYqVNvg/viewform?usp=sf_link