Comparaison entre approches statistiques et réseaux de neurones pour identifier les patients en errance diagnostique à partir du Système National des Données de Santé

Corentin Faujour, MSc[1,2], Stéphane Bouée, MD[1], Corinne Emery, MSc[1], Anne-Sophie Jannot[2,3], PU-PH.

1- CEMKA, Bourg-la-Reine, France, 2- Université Paris Cité, Inria, Inserm, HeKA, F-75015 Paris, France, 3- Banque Nationale de Données Maladies Rares (BNDMR), AP-HP, Paris, France

CONTEXTE

Les bases médico-administratives comme le SNDS permettent de reconstruire des **parcours de soins longitudinaux** à partir d'événements médicaux codés et datés.

L'objectif de ce travail est d'exploiter ces parcours de soins pour repérer **plus précocement** des patients susceptibles d'être en **errance diagnostique**.

Les approches statistiques reposent souvent sur des variables agrégées, comme le comptage de codes, qui **ne conservent pas l'ordre ni la temporalité des événements**.

Les réseaux séquentiels, tels que les **LSTM** ou les **Transformers**, permettent de modéliser les trajectoires comme des **séquences d'événements**.

OBJECTIFS

Les objectifs de ce travail sont :

- D'évaluer l'apport de l'ordre et des délais entre événements pour la détection de signaux précoces avant diagnostic.
- D'étudier la capacité prédictive et l'interprétabilité de trois modèles :
  - statistiques basés sur la **fréquence des codes.**
  - séquentiels intégrant **l'ordre des événements.**
  - séquentiels intégrant **l'ordre et les délais entre événements**.

MÉTHODES

Des trajectoires ont été **simulées à partir de paramètres estimés dans le SNDS** pour les populations suivantes :

- **22 000 patients SLA** identifiés par délivrance de riluzole (>=2)
- **22 000 témoins** appariés sur l'âge et le sexe.

Les événements simulés comprennent :

- Des événements de fond
- Des événements liés à la pathologie, **plus fréquents à l'approche du diagnostic chez les cas**.

Les performances ont été évaluées par l'AUC à **différents horizons avant diagnostic** (classification binaire).

CONCLUSION

La prise en compte de **l'ordre des événements** améliore la classification des trajectoires de soins simulées, avec un gain moyen d'environ **+0,06 AUC** par rapport aux approches statistiques basées sur la fréquence des codes.

L'ajout des **délais entre événements** n'apporte **pas de gain notable** dans ce cadre, suggérant que la modélisation de la temporalité reste un enjeu méthodologique ouvert.

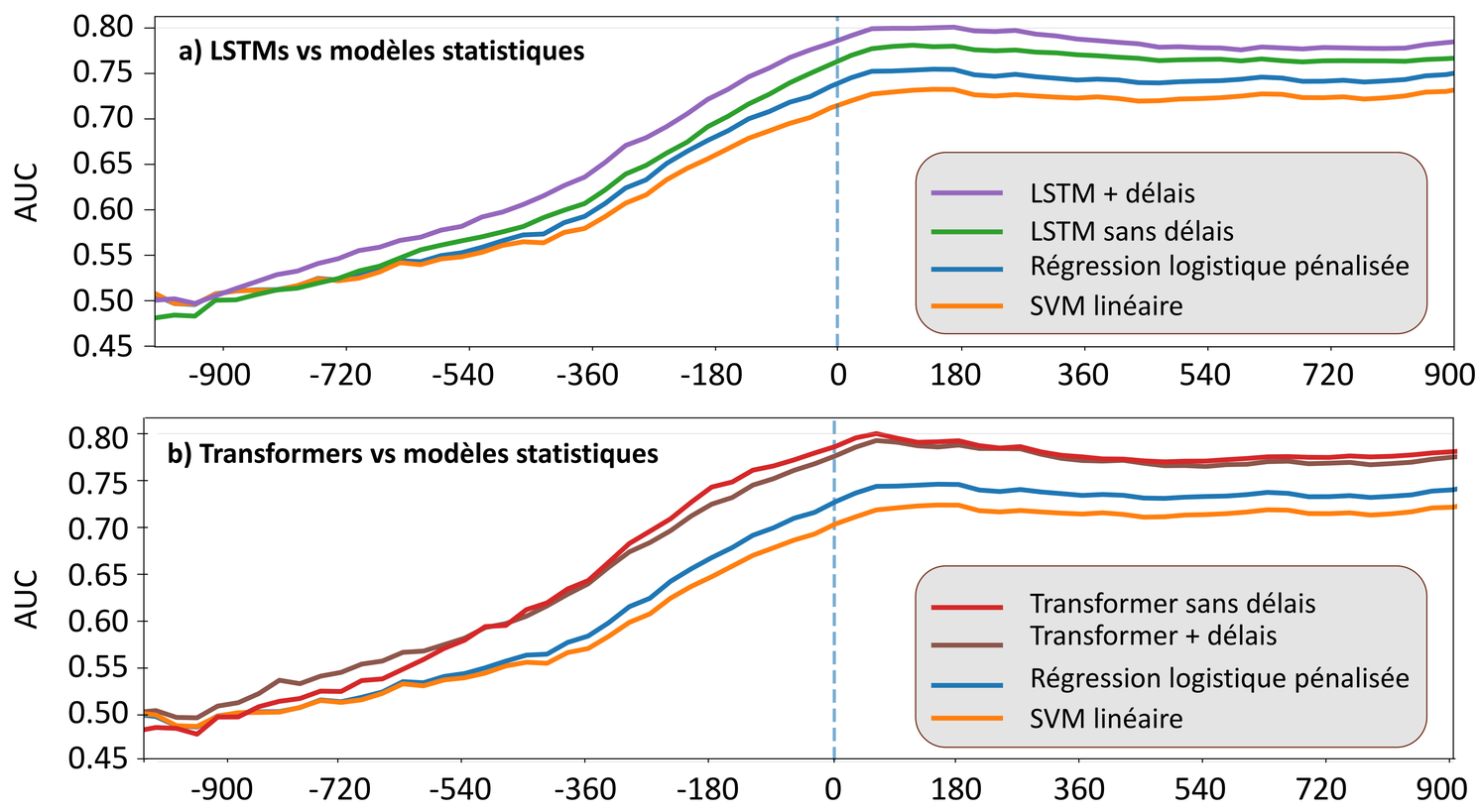
Ces approches ouvrent des perspectives pour les **études en vie réelle** visant à mieux caractériser les **parcours de soins avant diagnostic** et à identifier des **signaux d'errance** à partir des données du SNDS.

RÉSULTATS

Modéliser l'ordre des événements améliore la capacité prédictive (Figure 1)

- Les modèles séquentiels surpassent les modèles basés sur la fréquence des codes avec un gain moyen d'environ **+0,06 AUC**
- L'ajout **des délais entre événements n'apporte pas de gain notable** aux réseaux séquentiels.
- À **180 jours avant diagnostic**, les performances restent modérées : **AUC = 0,65 – 0,75** selon les modèles.

FIGURE 1. Capacité prédictive des réseaux séquentiels (LSTM et Transformers) vs modèles statistiques en fonction de l'horizon de prédiction.



Interprétabilité des réseaux séquentiels

- Les **événements liés à la pathologie** sont **correctement identifiés** comme les plus contributifs, par la régression logistique comme par le Transformer (Figure 2).

FIGURE 2. Identification des événements les plus contributifs à la prédiction.

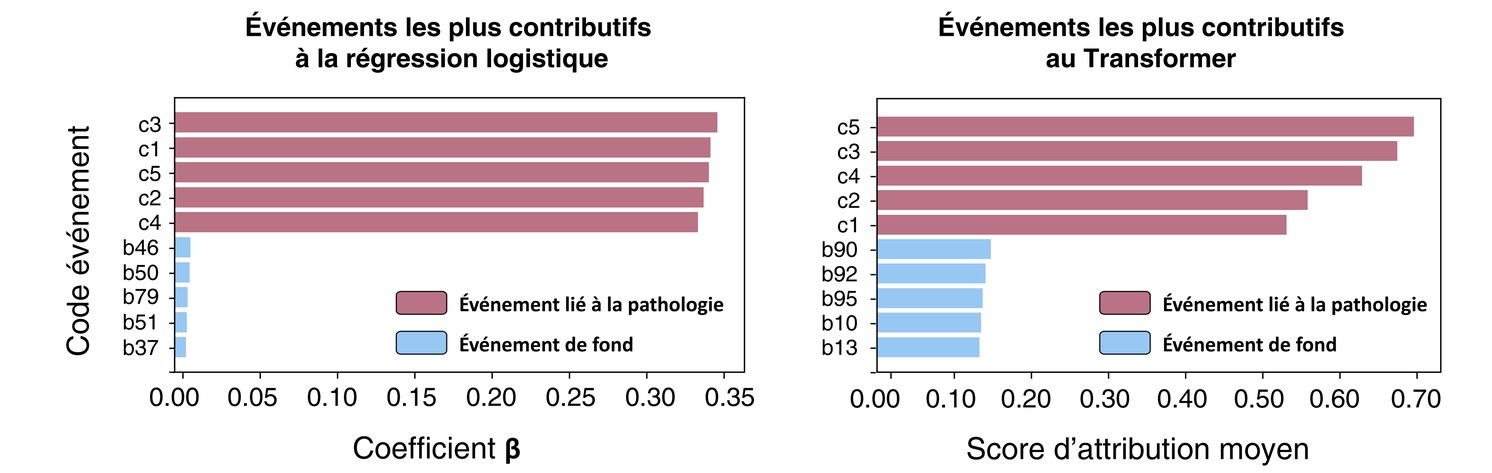
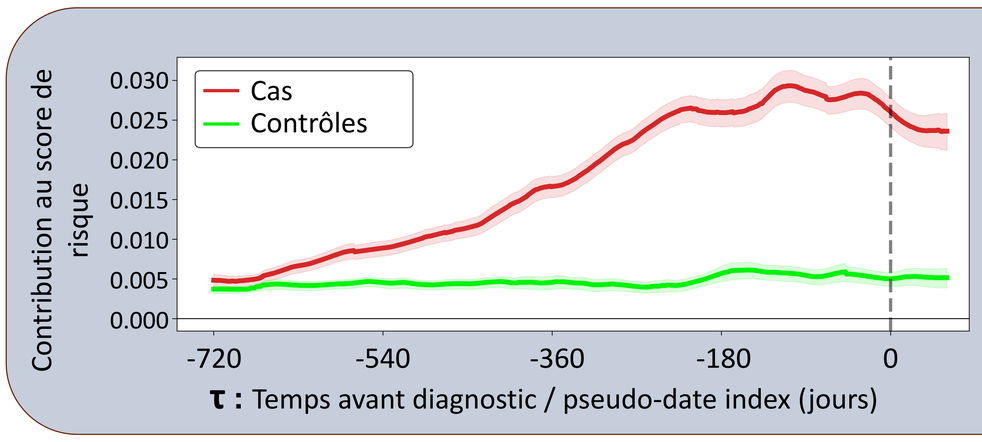


FIGURE 3. Interprétation temporelle : suivi à l'échelle de la cohorte



- Les contributions sont agrégées le long de l'axe temporel afin de **suivre l'évolution du signal prédictif**.
- Chez les cas, le signal augmente **plusieurs mois avant le diagnostic**, tandis qu'il reste faible chez les témoins.

Lien vers l'article ***Lien vers le poster***

Contact: *corentin.faujour@cemka.fr*

Financement : *CIFRE (ANRT)*
ANR France 2030, reference 22-PESN-0013.