



Traitement automatique du langage pour la surveillance de l'incidence des tentatives de suicide en région parisienne

Présentation : Ariel Cohen





L'article publié

npj | mental health
research

www.nature.com/npjmentalhealth

ARTICLE OPEN



Natural language processing of multi-hospital electronic health records for public health surveillance of suicidality

Romain Bey¹, Ariel Cohen¹, Vincent Trebassen², Basile Dura¹, Pierre-Alexis Geoffroy^{3,4,5,6}, Charline Jean^{1,7,8}, Benjamin Landman², Thomas Petit-Jean¹, Gilles Chatellier^{1,9}, Kankoe Sallah¹⁰, Xavier Tannier¹¹, Aurelie Bourmaud^{9,12,13} and Richard Delorme^{2,14}

There is an urgent need to monitor the mental health of large populations, especially during crises such as the COVID-19 pandemic, to timely identify the most at-risk subgroups and to design targeted prevention campaigns. We therefore developed and validated surveillance indicators related to suicidality: the monthly number of hospitalisations caused by suicide attempts and the prevalence among them of five known risks factors. They were automatically computed analysing the electronic health records of fifteen university hospitals of the Paris area, France, using natural language processing algorithms based on artificial intelligence. We evaluated the relevance of these indicators conducting a retrospective cohort study. Considering 2,911,920 records contained in a common data warehouse, we tested for changes after the pandemic outbreak in the slope of the monthly number of suicide attempts by conducting an interrupted time-series analysis. We segmented the assessment time in two sub-periods: before (August 1, 2017, to February 29, 2020) and during (March 1, 2020, to June 31, 2022) the COVID-19 pandemic. We detected 14,023 hospitalisations caused by suicide attempts. Their monthly number accelerated after the COVID-19 outbreak with an estimated trend variation reaching 3.7 (95%CI 2.1–5.3), mainly driven by an increase among girls aged 8–17 (trend variation 1.8, 95%CI 1.2–2.5). After the pandemic outbreak, acts of domestic, physical and sexual violence were more often reported (prevalence ratios: 1.3, 95%CI 1.16–1.48; 1.3, 95%CI 1.10–1.64 and 1.7, 95%CI 1.48–1.98), fewer patients died ($p = 0.007$) and stays were shorter ($p < 0.001$). Our study demonstrates that textual clinical data collected in multiple hospitals can be jointly analysed to compute timely indicators describing mental health conditions of populations. Our findings also highlight the need to better take into account the violence imposed on women, especially at early ages and in the aftermath of the COVID-19 pandemic.

npj Mental Health Research (2024)3:6; <https://doi.org/10.1038/s44184-023-00046-7>

Publié le 14/02/2024

<https://www.nature.com/articles/s44184-023-00046-7>



1

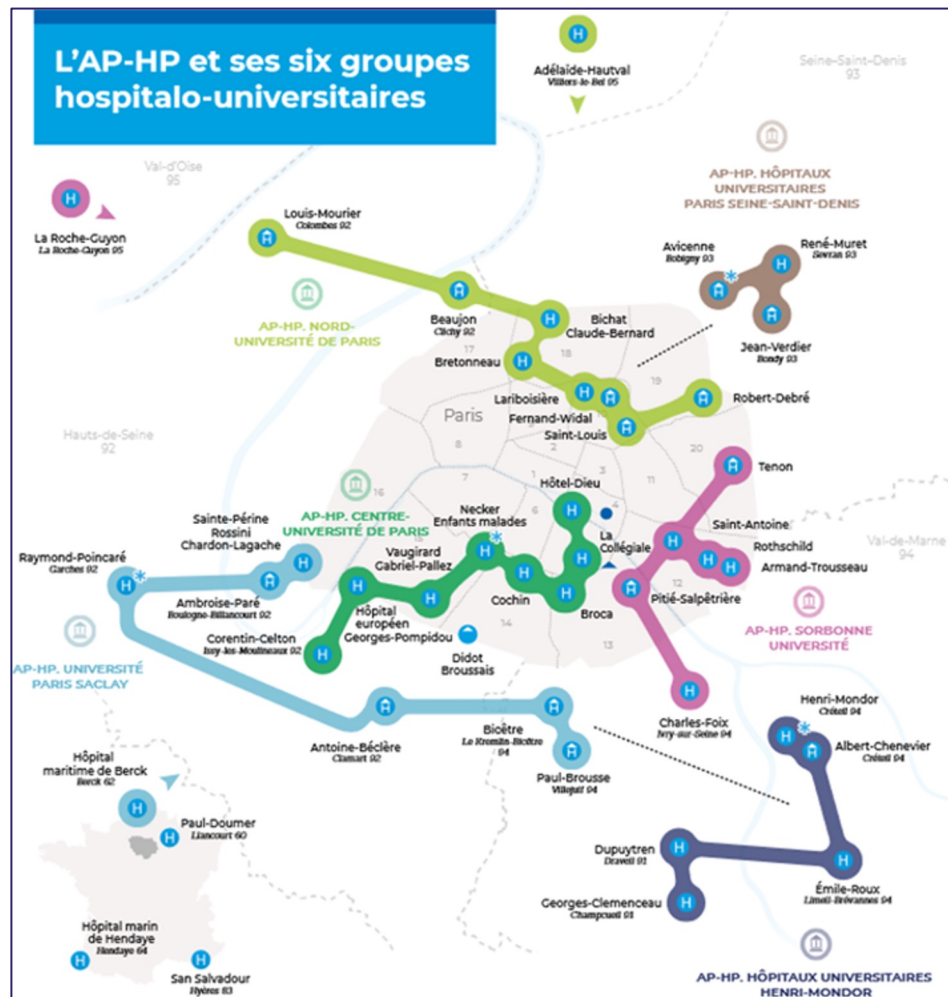
Présentation de l'entrepôt de données de santé de l'AP-HP

3





L'Entrepôt de Données de Santé (EDS) c'est quoi ?



L'Entrepôt de Données de Santé (EDS) de l'AP-HP héberge une base de données centralisées et standardisées contenant les données de vie réelles collectées par les différentes applications des 39 hôpitaux de l'AP-HP.

Ses finalités sont :

- La recherche
- Le pilotage (gestion et médical) - BI
- L'innovation

L'EDS
de l'AP-HP



Les données présentes dans l'EDS

L'Entrepôt de Données de Santé est alimenté par plusieurs types de données :

- Des **données structurées** provenant des bases de données relationnelles des applications sources de l'AP-HP (dossiers administratifs, actes médicaux, données démographiques, etc.)
- Des **données semi-structurées** provenant de métadonnées de fichiers, d'images ou de signaux (DICOM, signaux physiologiques, etc.)
- Des **données non structurées** provenant de fichiers et d'images (documents médicaux au format PDF, images médicales, etc.)

11 millions de
patients

40 millions de
dossiers
administratifs
(séjours)

40 millions
d'actes médicaux

40 millions
de diagnostics

140 millions
de documents
médicaux

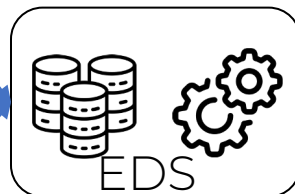
30 millions
d'examens
d'imagerie

2 milliards
d'analyses de
biologie

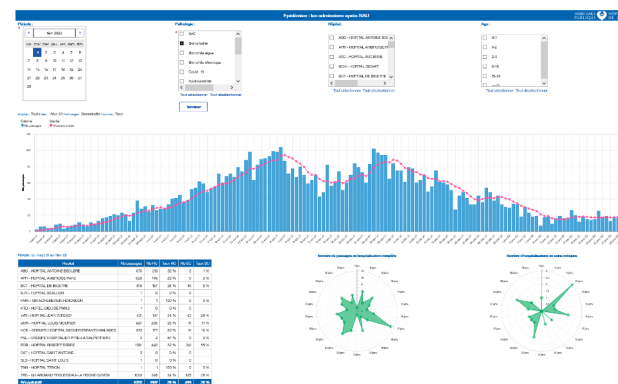


Réutilisation secondaire des données

Réutilisation
secondaire

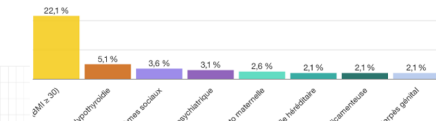


Pilotage



Antécédents médicaux

Autre, Pathologie psychiatrique, Age maternel ≥ 40 ans, Allergie immunologique à la lactose, Obésité (BMI ≥ 30), Thrombose veineuse, Hypertension, Allergie médicamenteuse, Problèmes sociaux, Végétisme général



Top 10 des antécédents en %

npj Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

International electronic health clinical course profiles: the

Gabriel A. Brat^{1,4,5}, Griffin M. Weber^{1,4,5}, Nils Gehlen¹, Lemuel R. Waltman¹, Gilbert S. Omenn¹, Alberto Ma Shaw N. Murphy¹, Sehi L. Yi¹, Mark S. Keller¹, Ricca Alba Gutierrez-Sacristan¹, John J. Holmes^{1,11}, Douglas Douglas A. Murad^{1,2}, Lucia Scudeller^{1,3}, Mauro Bucal Romain Griffier^{1,8}, Sebastian Cossin^{1,9}, Bertrand M. Piotr Sli^{1,3}, Amelia L. M. Tan^{1,3}, Kee Yuan Ngiam^{1,2}, Robert L. Bradford^{1,2}, Mohamad Danial^{1,3}, Christel Di Nina Orlova^{1,2}, Julien Dubiel^{1,2}, Martin Hilka^{1,2}, Anne S. Anita Burgun^{1,2}, Melodie Berraux^{1,2}, Amaud Sandrin^{1,2}, Tobias Grading^{1,2}, Julien Champ^{1,2}, Martin Boeke Damien Leprovost^{1,2}, Thomas Moreau^{1,2}, Gael Varoqui Charlotte Caucheteux^{1,2}, Christian Haverkamp^{1,2}, Gui Tianxi Cai^{1,2} and Isaac S. Kohane^{1,2}

Journal of the American Medical Informatics Association, 2024, 1–11
https://doi.org/10.1093/amia/ocae069

Research and Applications

Research and Applications

Collaborative and private clinical data warehouse language processing pipeline

Thomas Petit-Jean^{1,2}, MSc^{1,2}, Christel Gilles Chatellier^{1,2}, MD^{1,4}, Marie Frank Emmanuelle Kempf^{1,2}, MD, PhD^{6,7}, Ro

¹Innovation and Data Unit, IT Department, Assistance P d'Epidémiologie et de Santé Publique, INSERM, Sorbonne Assistance Publique-Hôpitaux de Paris, Le Kremlin-Bicêtre, Paris, Centre-Université de Paris (APHP-CL) Hôpitaux Universitaires Paris-Saclay, Assistance Publique d'Informatique Médicale et d'Ingénierie des Connaissances, Université, Paris, 75005, France, ²Department of Medical Public-Hôpitaux de Paris, Créteil, 94000, France

³Corresponding author: Thomas Petit-Jean, MSc, Innovation & France (thomas.petitjean@aphp.fr)

Received: 28 April 2023 | Revised: 17 June 2023 | Accepted: 21 June 2023
DOI: 10.1002/ijc.34675

AMIA OXFORD

RESEARCH ARTICLE
Cancer Epidemiology

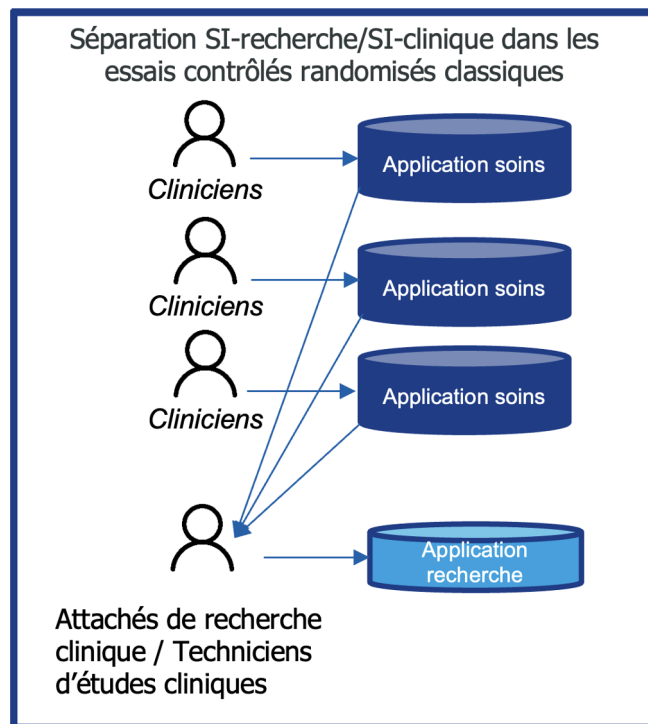
No changes in clinical presentation, treatment strategies and survival of pancreatic cancer cases during the SARS-COV-2 outbreak: A retrospective multicenter cohort study on real-world data

Emmanuelle Kempf^{1,2} | Sonia Priou^{3,4} | Guillaume Lamé⁴ | Alexis Laurent⁵ | Etienne Guével³ | Stylianos Tzedakis⁶ | Romain Bey³ | David Fuks⁶ | Gilles Chatellier⁷ | Xavier Tannier¹ | Gilles Galula⁸ | Rémi Flicoteaux⁹ | Christel Daniel³ | Christophe Tournigand² | on behalf of the Assistance Publique – Hôpitaux de Paris (AP-HP) Cancer Group, a Cancer Research Application on Big Data (CRAB) Initiative

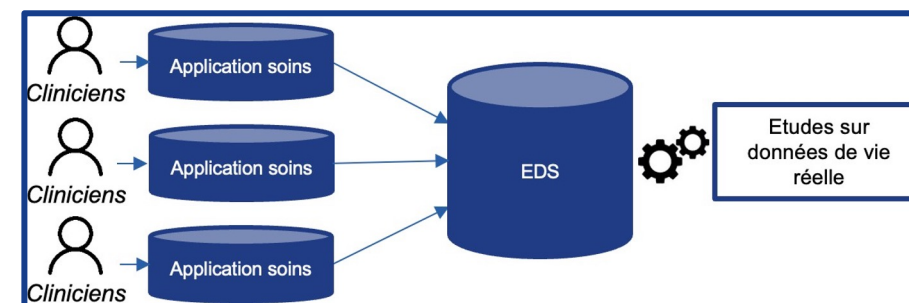


Données de vie réelle vs Essais classiques

RWD



Essais contrôlés randomisés classiques



Etudes sur données de vie réelle



»2

Motivation

8





Motivation

Novembre 2020



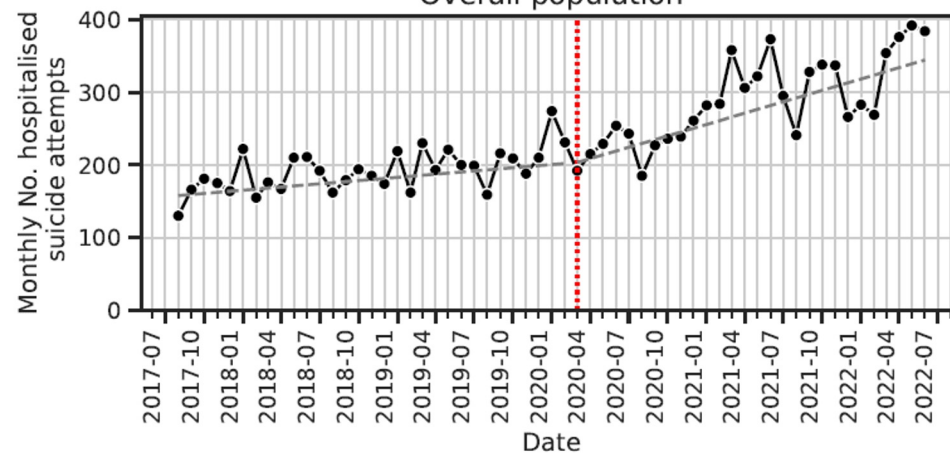
- La santé mentale des franciliens a-t-elle été impactée par le contexte épidémique?
- Y a-t-il une variation du nombre d'hospitalisations pour tentative de suicide depuis le COVID-19 en région parisienne ?
- Dans quelle proportion?
- Y a-t-il des populations plus à risque?
- Quelle est l'influence des facteurs de risque ?
- Peut-on utiliser des grandes bases de données de vie réelle pour fournir de nouveaux outils de surveillance sanitaire en santé mentale?

Motivation



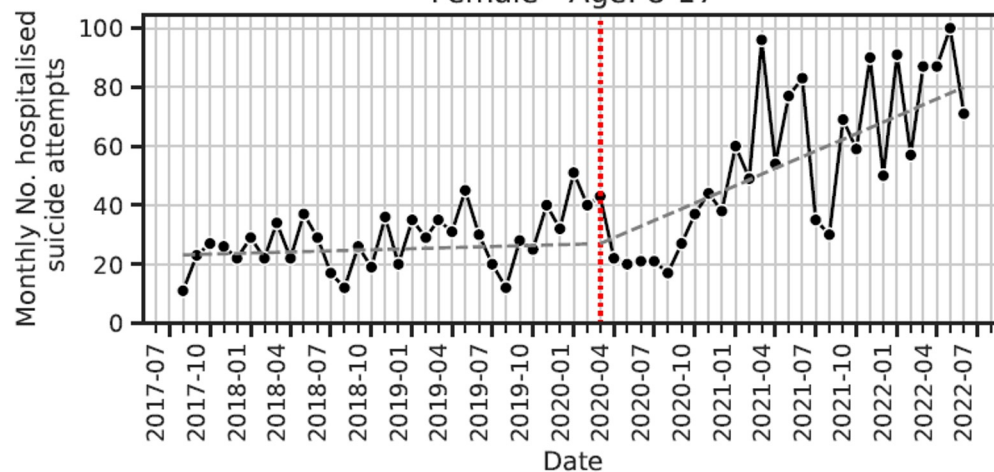
Résultat principal

Overall population

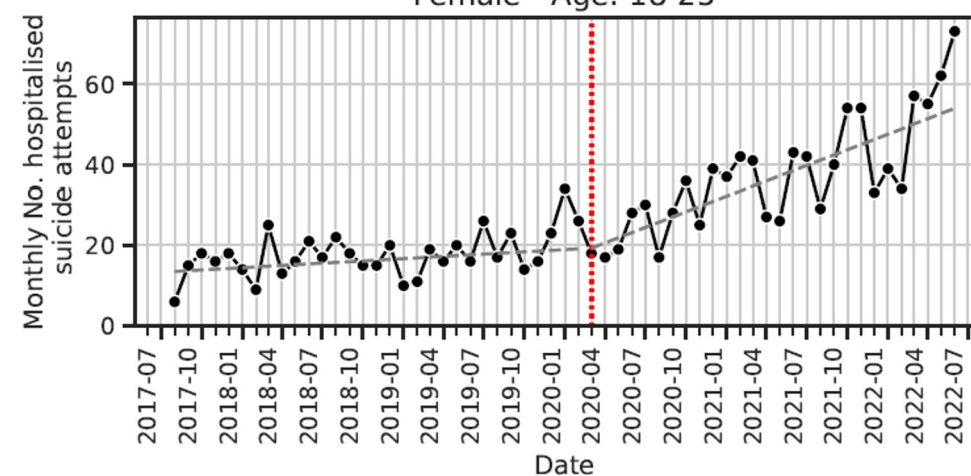


- COVID-19 outbreak
- Model linear component
- Hospitalised suicide attempts

Female - Age: 8-17



Female - Age: 18-25





»3

Comment y arriver ?





Equipe projet



Pôle Innovation & Données – DSN – APHP



Service de Pédiopsychiatrie – Robert Debré



LIMICS – Sorbonne Université



Unité de recherche clinique (URC) - GHU Nord

Equipe
projet



Données du codage PMSI insuffisantes (information structurée)

- Pas de code pour les TS > codes approximatifs X60 – X85
- Latence de 2 mois
- Biais économique

Exemple

X80 : Lésion auto-infligée par saut dans le vide

X65 : Auto-intoxication par l'alcool et exposition à l'alcool

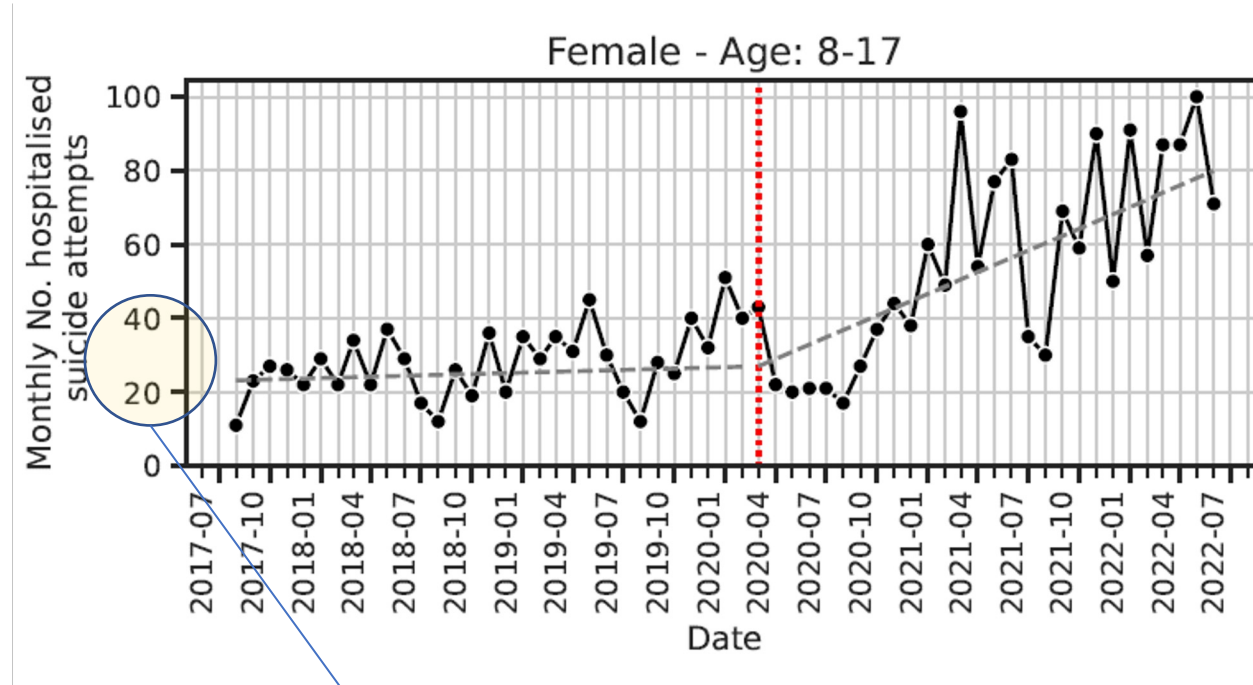
X62 : Auto-intoxication par des narcotiques et psychodysléptiques [hallucinogènes]
et exposition à ces produits, non classés ailleurs

Quelles
données?

Besoin d'aller vers des données textuelles



Des événements rares et diffus



~30 séjours par mois !

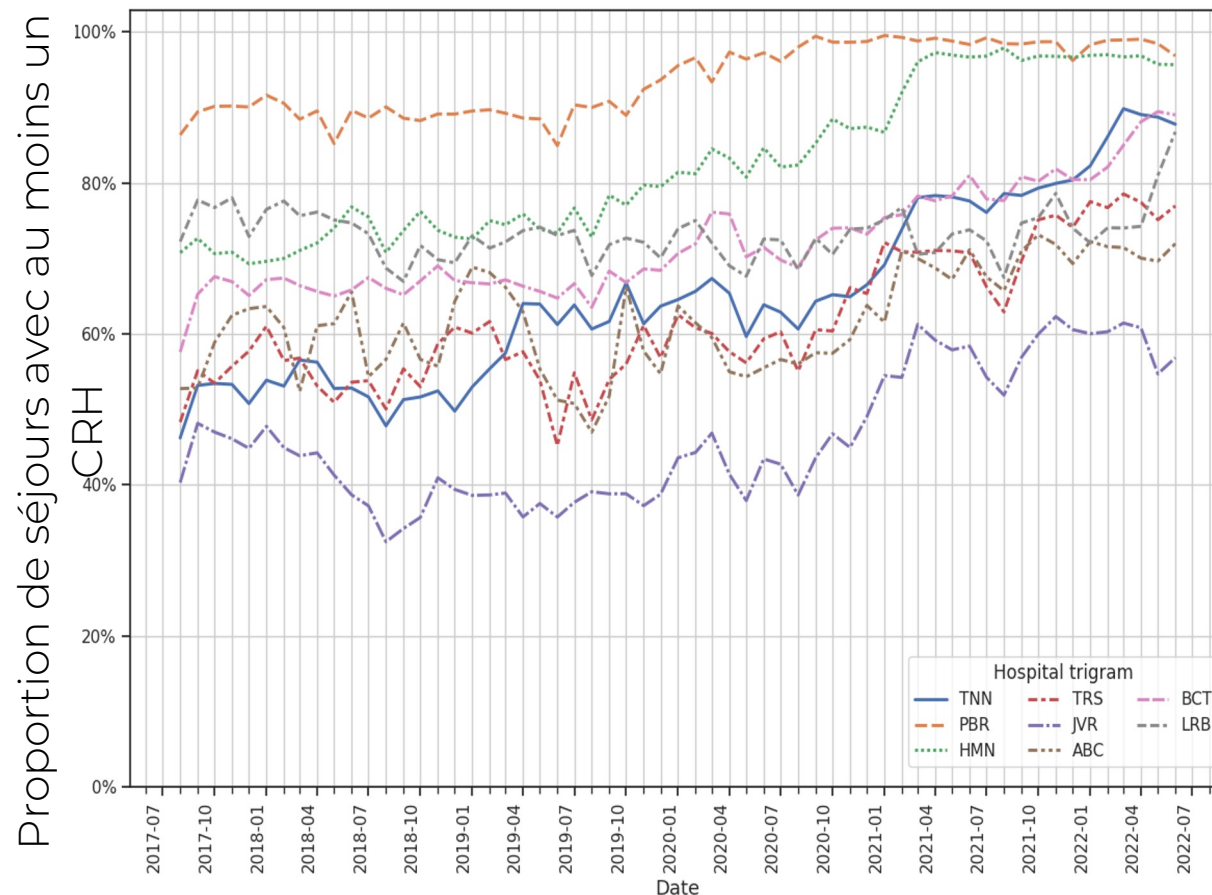
Besoin d'automatisation de la fouille des données textuelles



Données lacunaires

- Déploiement du dossier patient informatisé (DPI)

>> fait par service et échelonné dans le temps, instable



Solution:

- Sélection des 15 hôpitaux
- Etudes de sensibilité (par hôpital, imputation)

Données
manquantes



Déterminer le motif du séjour à partir des CR médicaux (Texte libre !)

- Plusieurs façon de faire référence au même concept
- Fautes de frappe
- Mots polysémiques
- Référence aux antécédents
- Evènements vécus par de proches du patient
- Evènements exprimés sous forme d'hypothèse

Complexité
linguistique

CHANGEMENT DE PTC ARTHROPLASTIE TOTALE DU GENOU GAUCHE – PROTHÈSE SEMI CONTRAINTE TS (STRYKER) - TIBIA TAILLE 2 – FÉMUR TAILLE 2 – TIGES D'EXTENSION TIBIALE (LONGUEUR 100 MM DIAMETRE 13)

MOTIF D'HOSPITALISATION

Patient de 18 ans, a été hospitalisée pour Détresse respiratoire aiguë sur probable pneumopathie à COVID.

Antécédents familiaux :

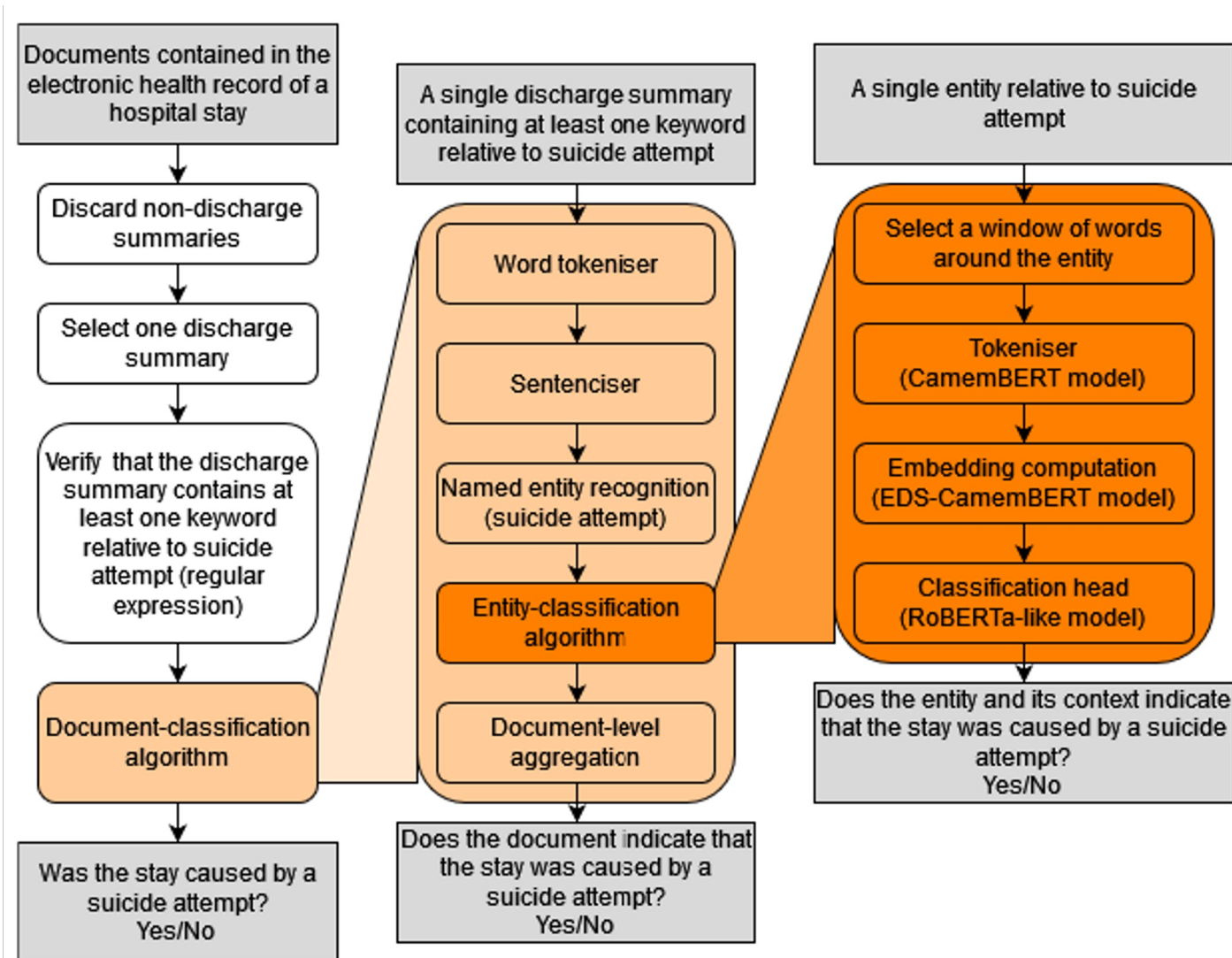
Tentative de suicide de sa mère à l'adolescence suite à un conflit

Commentaires : Admission par la BSPP pour IMV (alcool et médicaments anti-rejets etc) avec

COMPTE RENDU D'HOSPITALISATION du 30/03/2020 au 02/04/2020 Dernière TS en 2010.



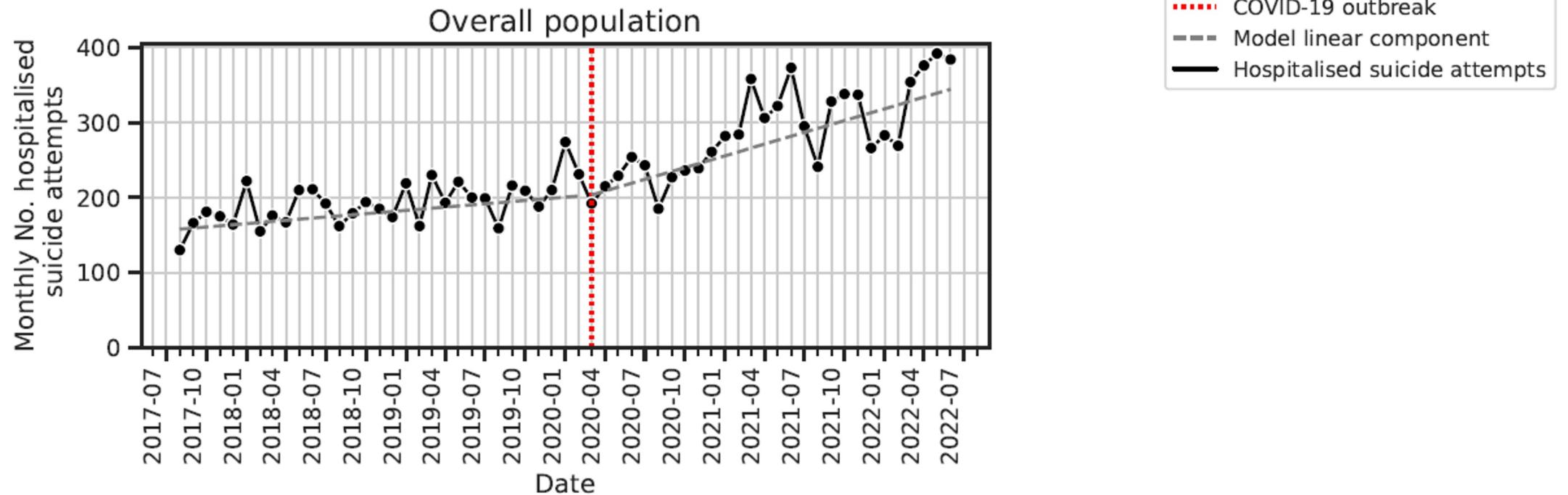
Algorithme de détection des TS



Algorithme



- Etude de cohorte observationnelle rétrospective
- Population: Patients hospitalisés (HC) dans 15 hôpitaux de l'AP-HP
- Série temporelle interrompue



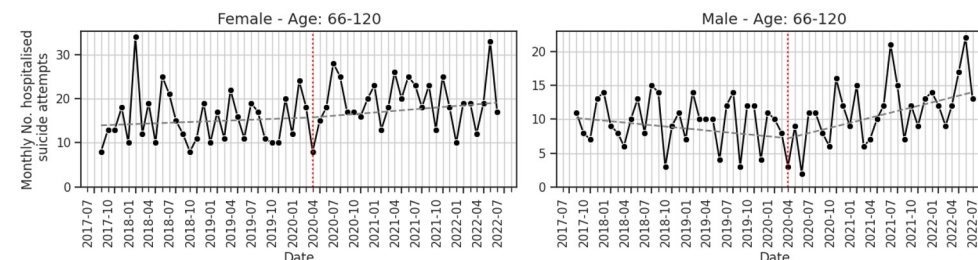
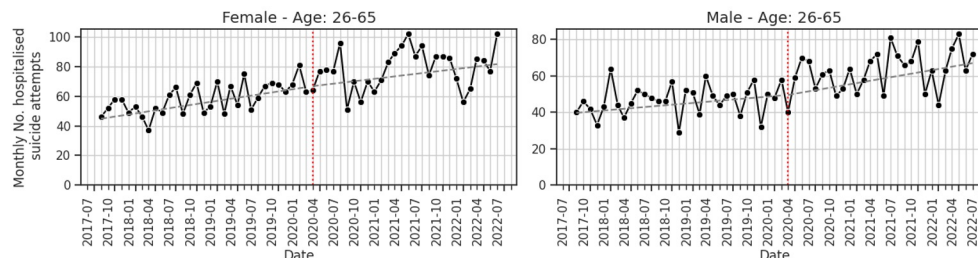
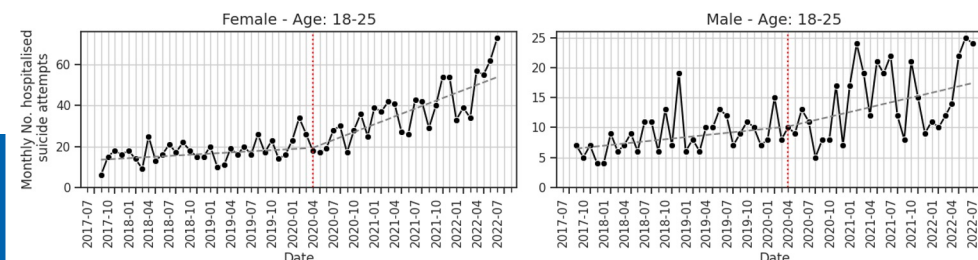
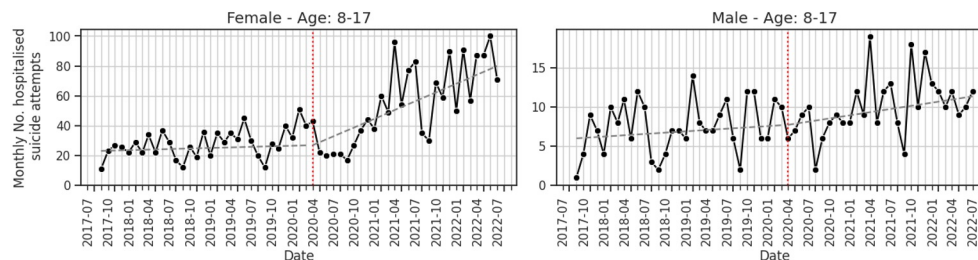
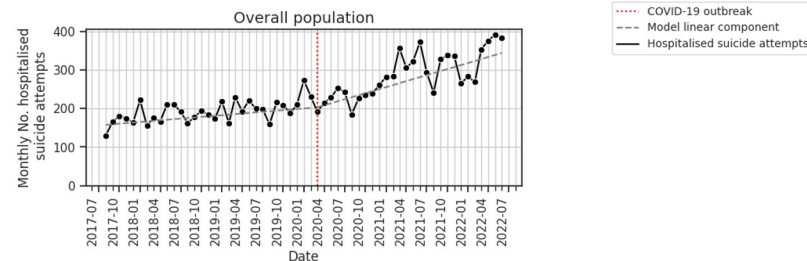
$$\text{Eq. 1 } N_T = \alpha_0 + \alpha_1 T + \alpha_2 (T - \tilde{T}) \times 1\{T \geq \tilde{T}\} + \sum_{m=\text{Jan}}^{\text{Dec}} \beta_m \times 1\{T = m\} + e_T$$



>>4

Résultats

19



Nombre d'hospitalisations mensuelles motivés par une tentative de suicide



Population la plus affectée :

- Femmes 8-17 ans
- Femmes 18-25 ans

**Table 1.** Number of hospitalisations caused by suicide attempts, seasonally-adjusted long-term means and trends, and post-pandemic trend variations.

		No. hospitalisations caused by suicide attempts (%) - pre-pandemic period	No. hospitalisations caused by suicide attempts (%) - post-pandemic period	Long term mean α_0 (95%CI) ^a	Long term trend α_1 (95% CI) ^a	Trend variation after COVID-19 outbreak α_2 (95%CI) ^a
Male	8–17	232 (3.9%)	281 (3.5%)	6.0 (4.4–7.7)	0.1 (–0.0–0.1)	0.1 (–0.1–0.2)
	18–25	271 (4.6%)	405 (5.0%)	6.5 (3.9–9.0)	0.1 (–0.0–0.3)	0.1 (–0.1–0.4)
	26–65	1451 (24.4%)	1755 (21.7%)	39.7 (34.4–45.0)	0.3 (0.0–0.6)	0.3 (–0.2–0.9)
	66–	300 (5.0%)	313 (3.9%)	10.2 (8.0–12.4)	–0.1 (–0.2–0.0)	0.3 (0.1–0.6)
	All Ages	2254 (37.9%)	2754 (34.1%)	62.4 (55.6–69.2)	0.4 (0.0–0.8)	0.9 (0.2–1.6)
Female	8–17	865 (14.5%)	1535 (19.0%)	23.2 (16.8–29.5)	0.1 (–0.2–0.5)	1.8 (1.2–2.5)
	18–25	549 (9.2%)	1045 (13.0%)	13.5 (9.2–17.8)	0.2 (–0.0–0.4)	1.1 (0.7–1.5)
	26–65	1810 (30.4%)	2197 (27.2%)	44.8 (38.4–51.3)	0.7 (0.4–1.0)	–0.2 (–0.8–0.5)
	66–	476 (8.0%)	538 (6.7%)	14.0 (10.6–17.3)	0.1 (–0.1–0.2)	0.1 (–0.3–0.4)
	All Ages	3700 (62.1%)	5315 (65.9%)	95.5 (83.2–107.7)	1.1 (0.4–1.7)	2.8 (1.6–4.1)
Overall		5954 (100.0%)	8069 (100.0%)	157.9 (142.0–173.7)	1.5 (0.6–2.3)	3.7 (2.1–5.3)

^aParameters were computed by ordinary least squares regressions following Eq. (1). All models were controlled for seasonal effects. 95%CI were based on standard errors. Number of hospitalisation stays caused by suicide attempts were used as dependent variables.

$$\text{Eq. 1 } N_T = \alpha_0 + \alpha_1 T + \alpha_2 (T - \tilde{T}) \times 1\{T \geq \tilde{T}\} + \sum_{m=\text{Jan}}^{\text{Dec}} \beta_m \times 1\{T = m\} + e_T$$



Table 3. Prevalence of risk factors in pre- and post-pandemic periods.

Risk factor	Group	Pre-covid prevalence (number of stays)	Post-covid prevalence (number of stays)	Prevalence ratio (95% CI) ^a	<i>p</i> -value ^a
Domestic violence	Overall	0.06 (360)	0.08 (639)	1.3 (1.16–1.48)	<0.0001
	Male	0.05 (106)	0.06 (163)	1.3 (0.99–1.60)	0.059
	Female	0.07 (254)	0.09 (476)	1.3 (1.13–1.51)	0.00035
Physical violence	Overall	0.02 (142)	0.03 (258)	1.3 (1.10–1.64)	0.0047
	Male	0.01 (29)	0.02 (50)	1.4 (0.90–2.22)	0.14
	Female	0.03 (113)	0.04 (208)	1.3 (1.02–1.60)	0.032
Sexual violence	Overall	0.04 (244)	0.07 (567)	1.7 (1.48–1.98)	<0.0001
	Male	0.01 (31)	0.02 (66)	1.7 (1.14–2.66)	0.0098
	Female	0.06 (213)	0.09 (501)	1.6 (1.40–1.91)	<0.0001
Social isolation	Overall	0.06 (383)	0.08 (640)	1.2 (1.09–1.39)	0.00071
	Male	0.07 (159)	0.08 (216)	1.1 (0.91–1.35)	0.31
	Female	0.06 (224)	0.08 (424)	1.3 (1.13–1.54)	0.00049
Suicide attempt history	Overall	0.36 (2160)	0.40 (3199)	1.1 (1.05–1.14)	<0.0001
	Male	0.33 (749)	0.36 (993)	1.1 (1.00–1.17)	0.037
	Female	0.38 (1411)	0.42 (2206)	1.1 (1.03–1.15)	0.0013

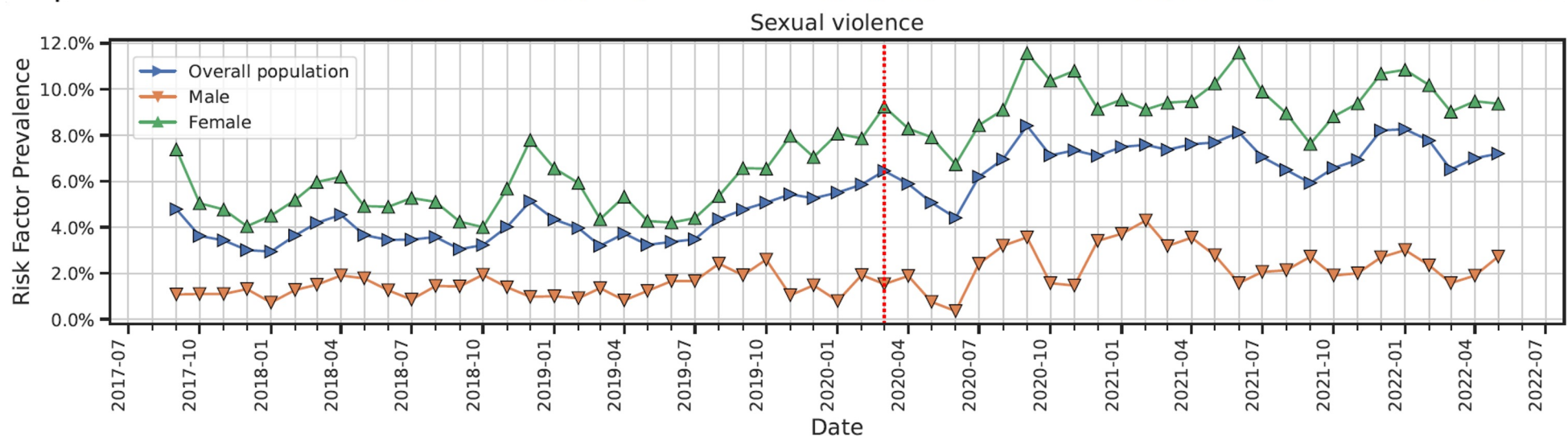




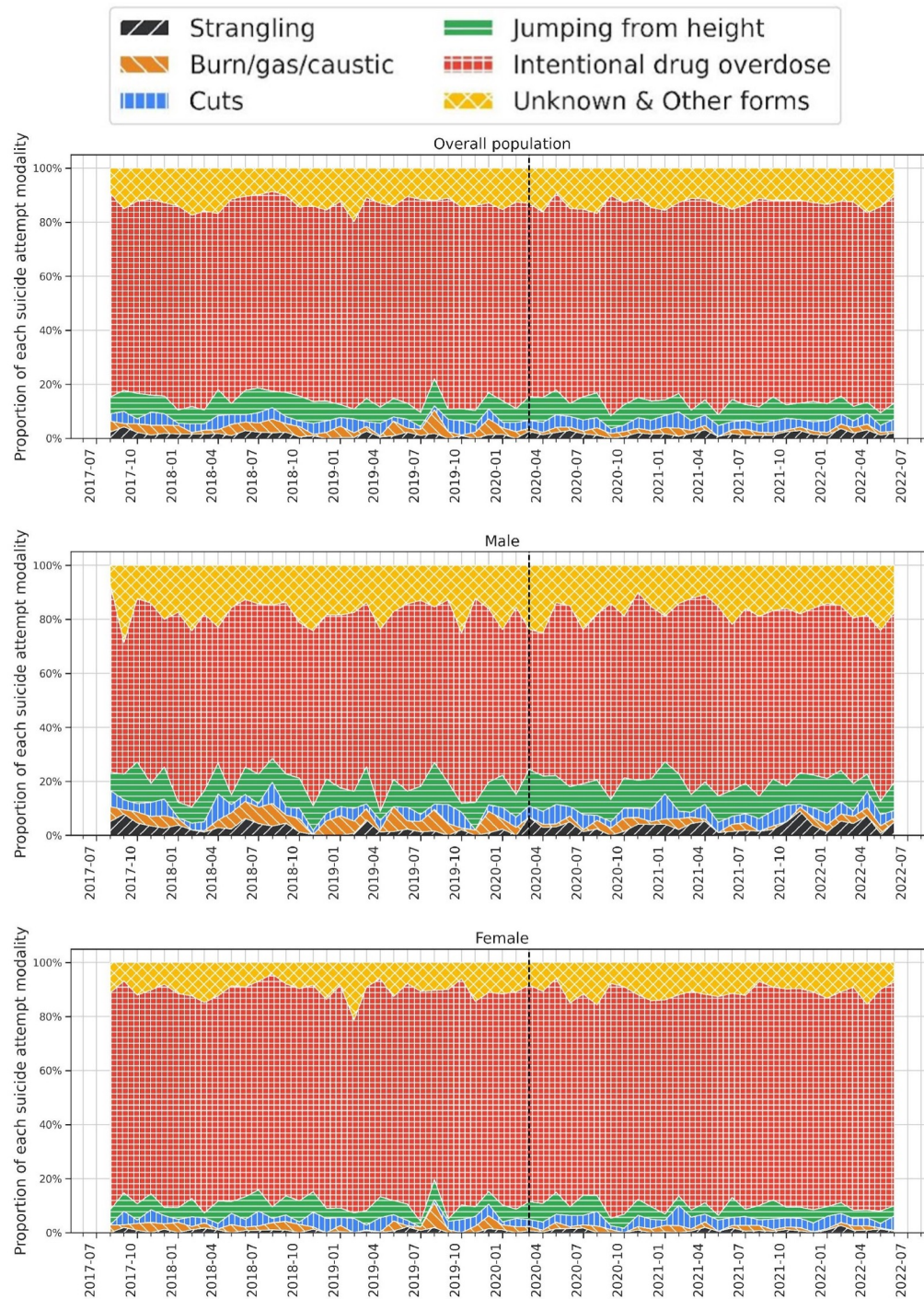
Table 2. Positive predictive values of the natural language processing algorithms.

	Pre-pandemic positive predictive value (95%CI, No. annotated records) ^a	Post-pandemic positive predictive value (95%CI, No. annotated records) ^a
Main algorithm used to detect hospitalisations caused by suicide attempts (hybrid machine learning and rule-based approach)	0.85 (0.76–0.91, 85)	0.86 (0.76–0.92, 77)
Alternative algorithm used to detect hospitalisations caused by suicide attempts (rule-based approach)	0.51 ^b	0.52 ^b
Algorithm used to detect social isolation risk factor	1.00 ^c (0.89–1.00, 30)	0.96 ^c (0.79–0.99, 23)
Algorithm used to detect domestic violence risk factor	0.96 ^c (0.79–0.99, 23)	0.96 ^c (0.80–0.99, 24)
Algorithm used to detect sexual violence risk factor	0.84 ^c (0.70–0.93, 38)	0.96 ^c (0.82–0.99, 28)
Algorithm used to detect physical violence risk factor	0.86 ^c (0.69–0.95, 29)	0.96 ^c (0.82–0.99, 28)
Algorithm used to detect suicide attempt history risk factor	0.93 ^c (0.78–0.98, 29)	0.83 ^c (0.66–0.93, 30)

^aWilson score intervals were used to compute the 95%CI.

^bFor the estimation of the performances of the alternative rule-based SA-classification algorithm, additional stays were drawn randomly among the stays that were classified as SA-caused by the rule-based algorithm but not by the hybrid algorithm—see Supplement. The Wilson score intervals could not be computed using this method.

^cFor the estimation of the performances of the risk factors algorithms, stays were drawn randomly among those that were labelled positive both by the main SA-classification algorithm and by the risk factors-classification algorithms.



- Extraction de la modalité grâce au TAL
- Confirmation des différences connues dans la littérature entre Hommes / Femmes



- Étude observationnelle qui ne prouve pas l'existence de relations de causalité.
- Étude rétrospective qui devrait être complétée par une étude prospective afin de démontrer pleinement l'utilité des indicateurs de santé mentale développés pour faire face aux crises à venir.

Limites

- Les facteurs de risque sont rapportés par les cliniciens.
- Les visites aux urgences qui n'ont pas été suivies d'une hospitalisation ont été écartées en raison de la faible disponibilité des données.



- Constat de l'**augmentation du nombre de TS** mensuelles après le COVID, particulièrement parmi les femmes entre 8 et 25 ans.
- La **violence**, facteur de risque connu, a été plus rapporté après le COVID.
- Cohérence des résultats avec la littérature en utilisant une **méthodologie innovante**.
- Indicateurs de **surveillance sanitaire** grâce au TAL dans un EDS.
- Possibilité de calcul en **temps réel** pour la surveillance épidémiologique.
- Possibilité de détection des **événements rares** parmi les millions d'hospitalisations.

Conclusions



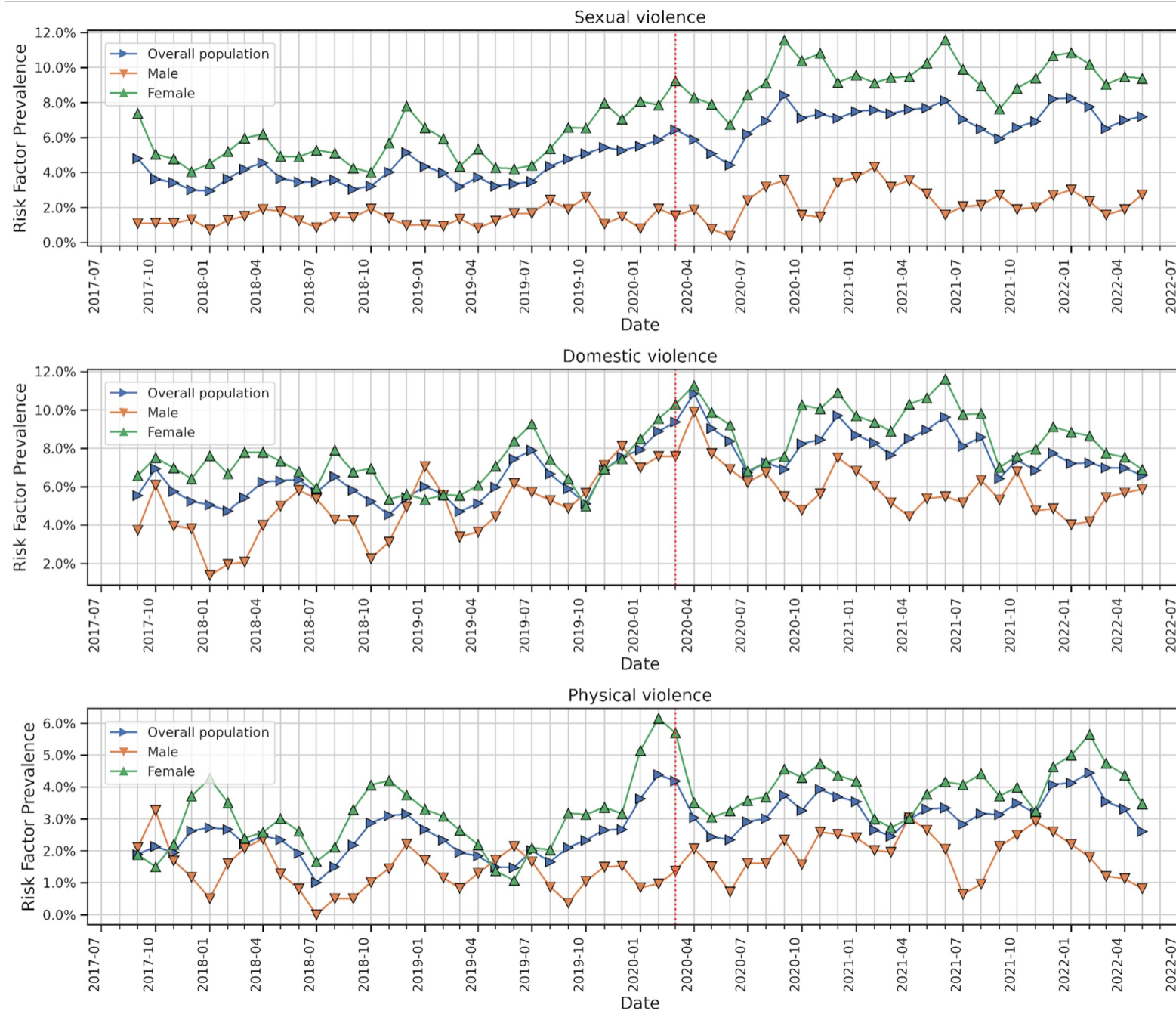
Questions ?

Perspectives ?

Suggestions ?

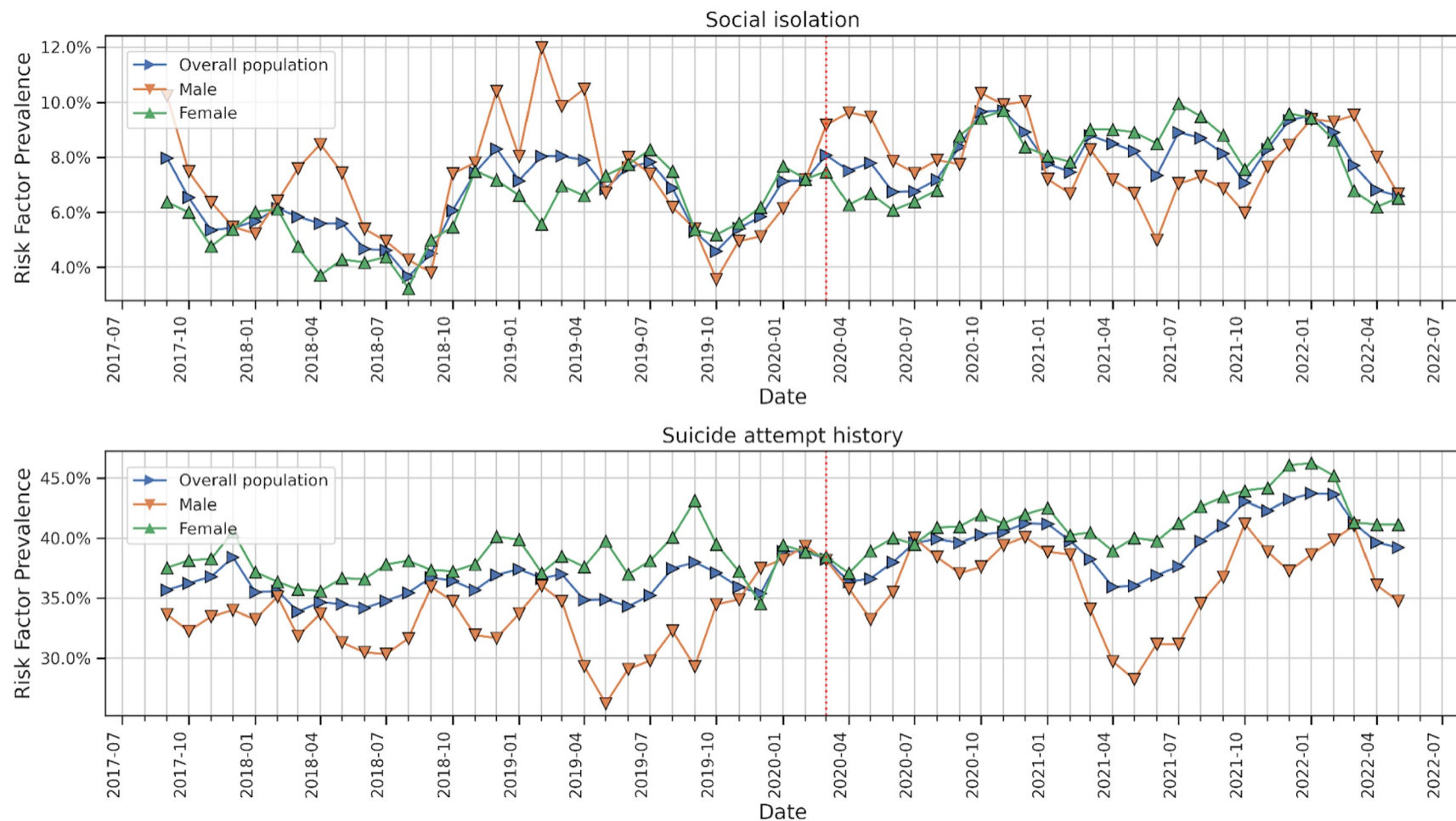


Annexe



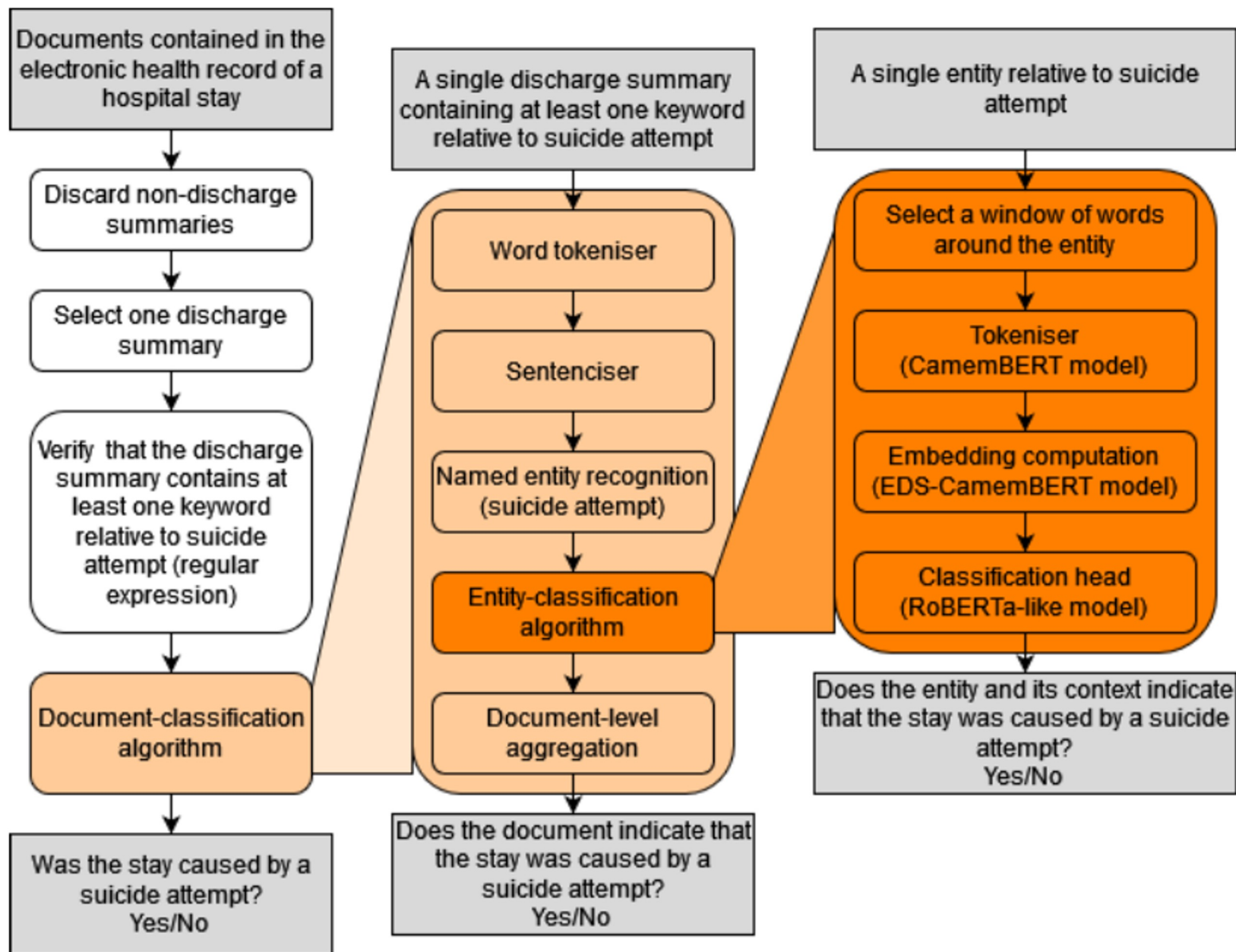


Facteurs de risque





L'algorithme



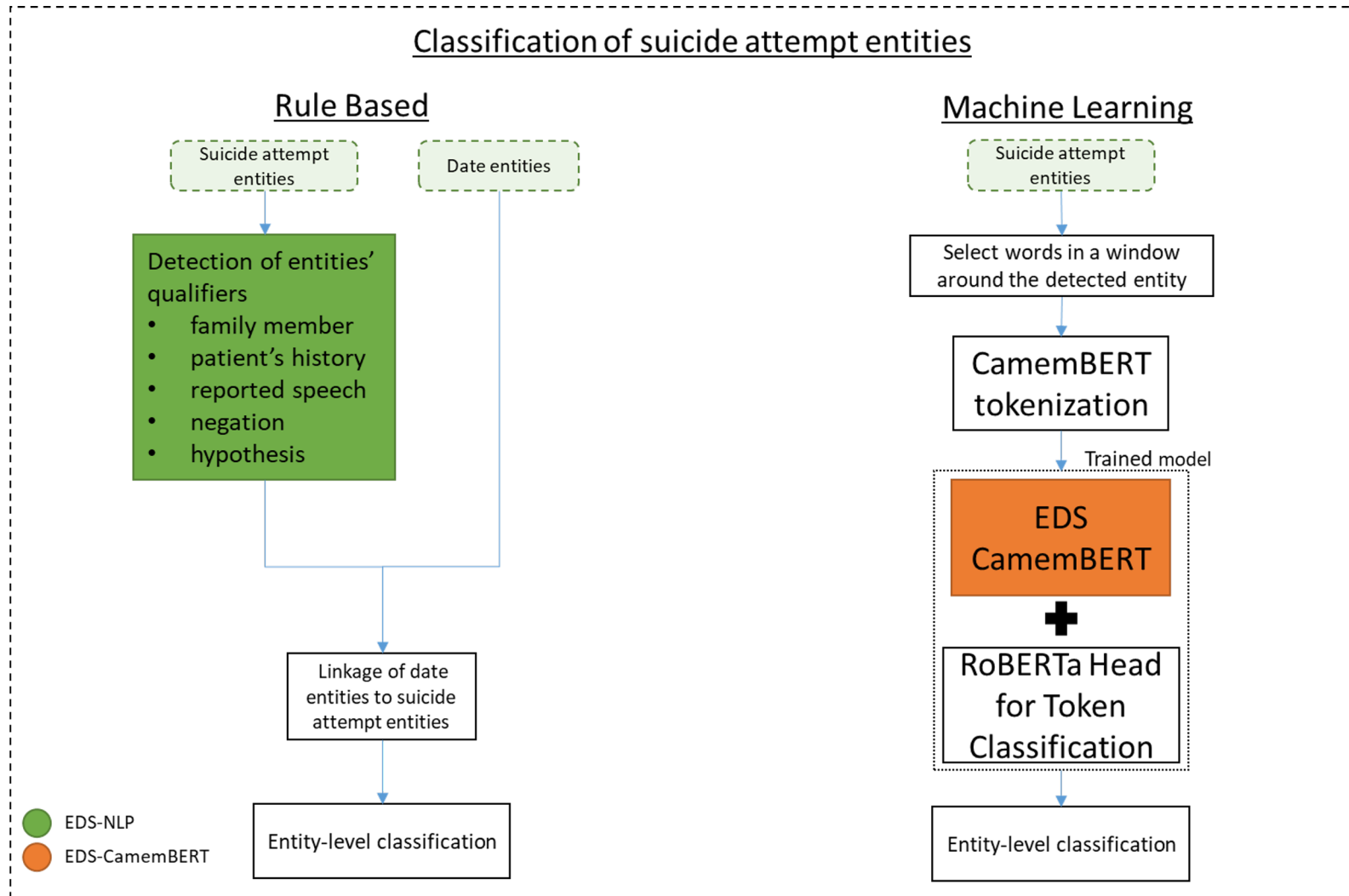
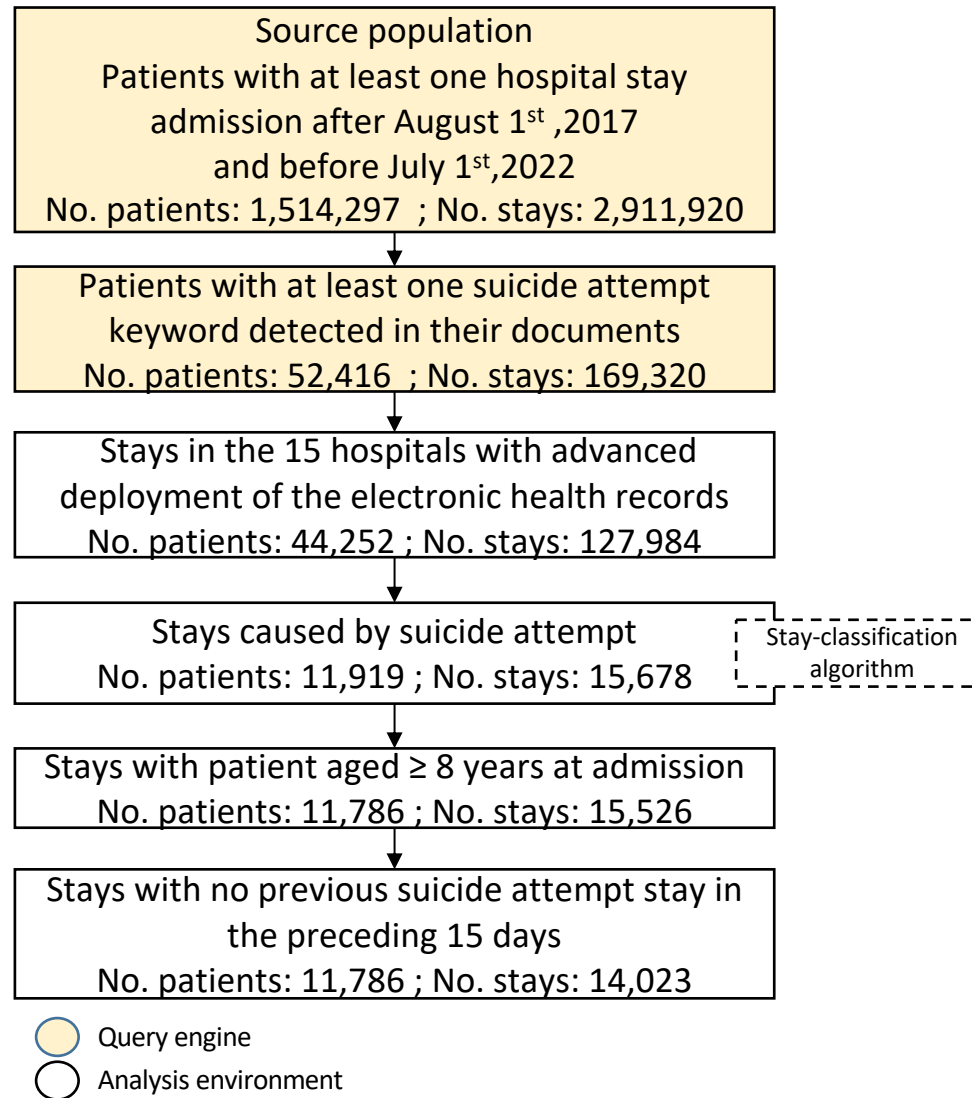


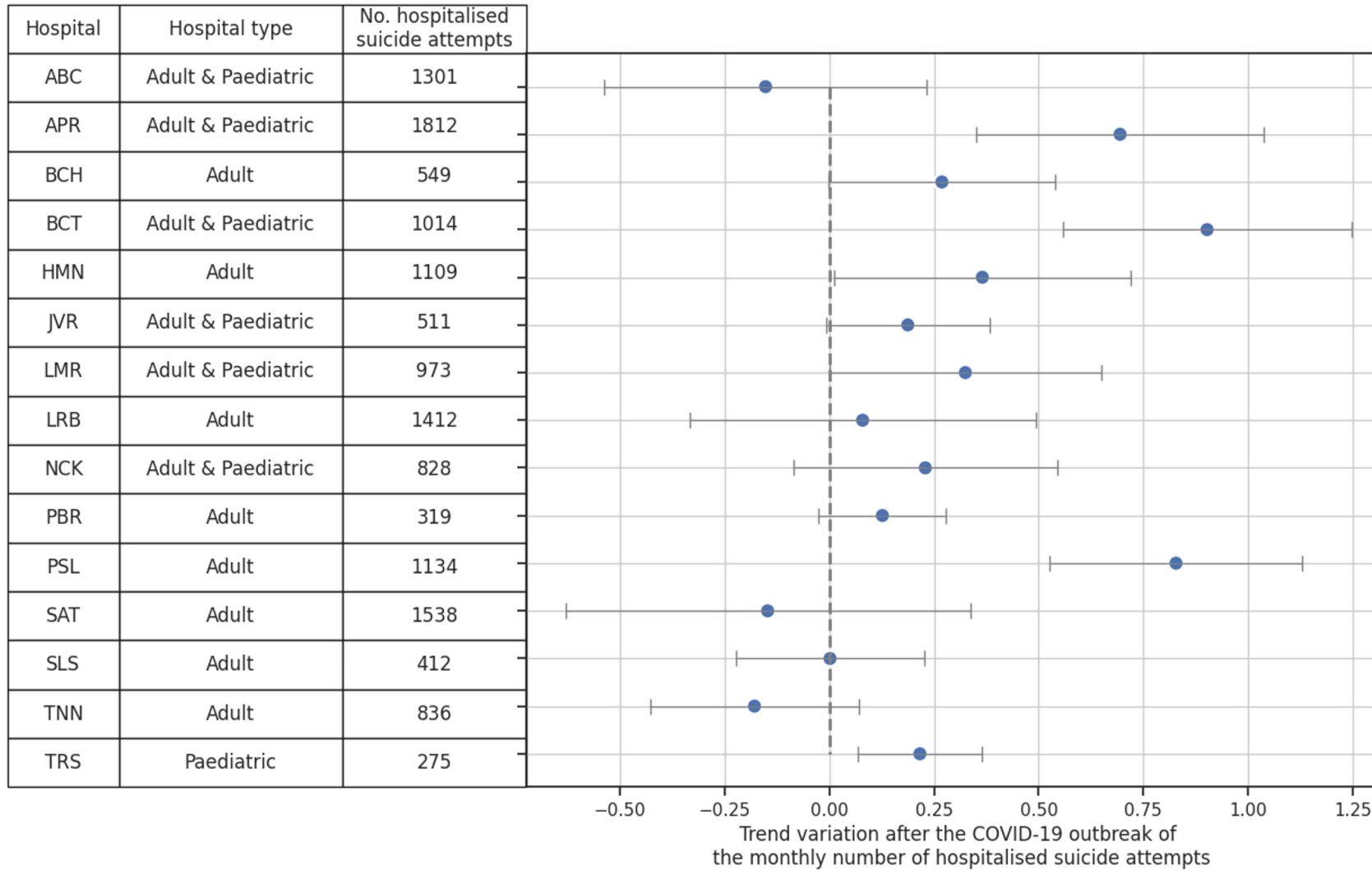


Diagramme
d'exclusion et
d'inclusion





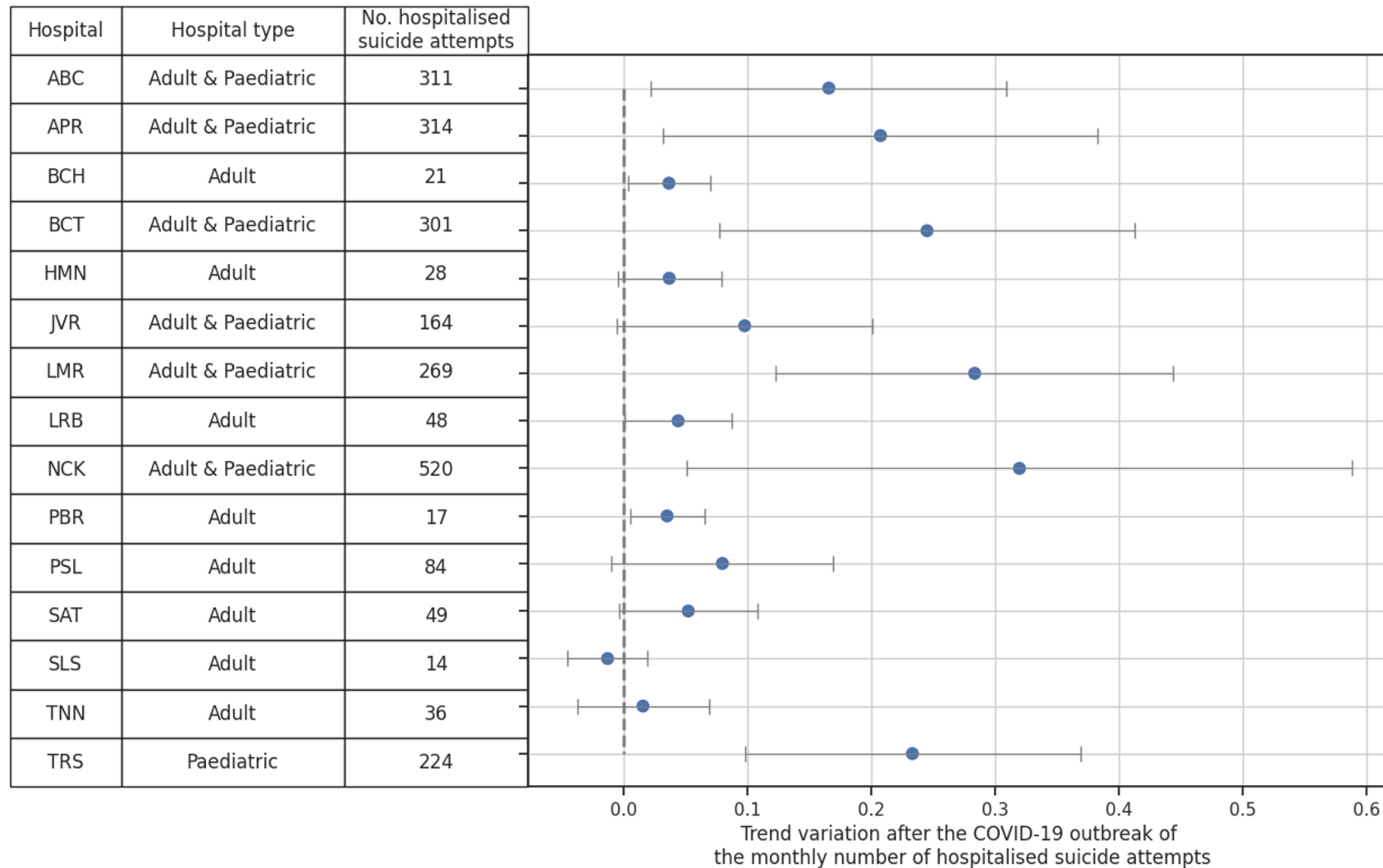
Per-hospital forest plot of trend variations - overall population



Trend variations



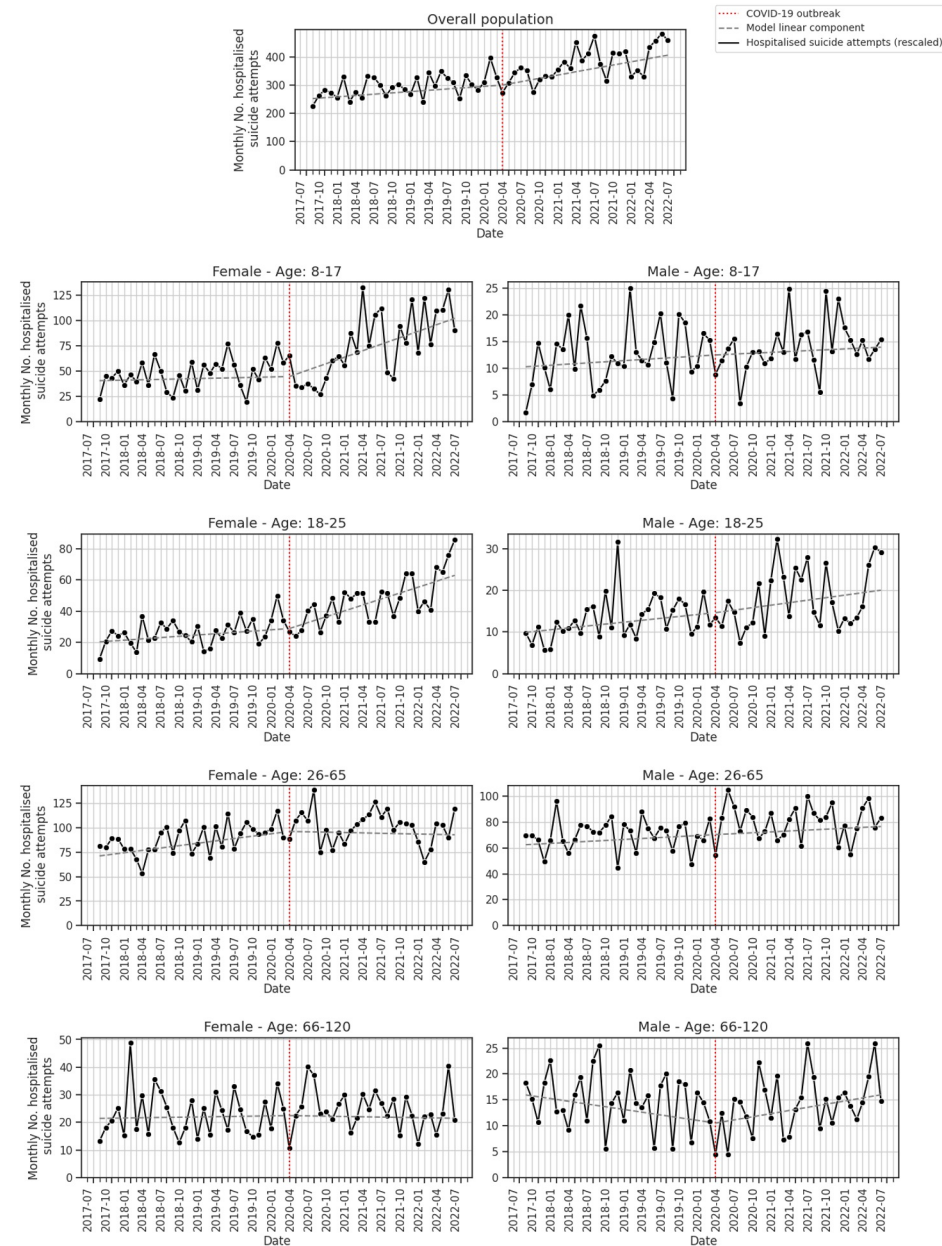
Per-hospital forest plot of trend variations – girls 8-17 years



Trend variations



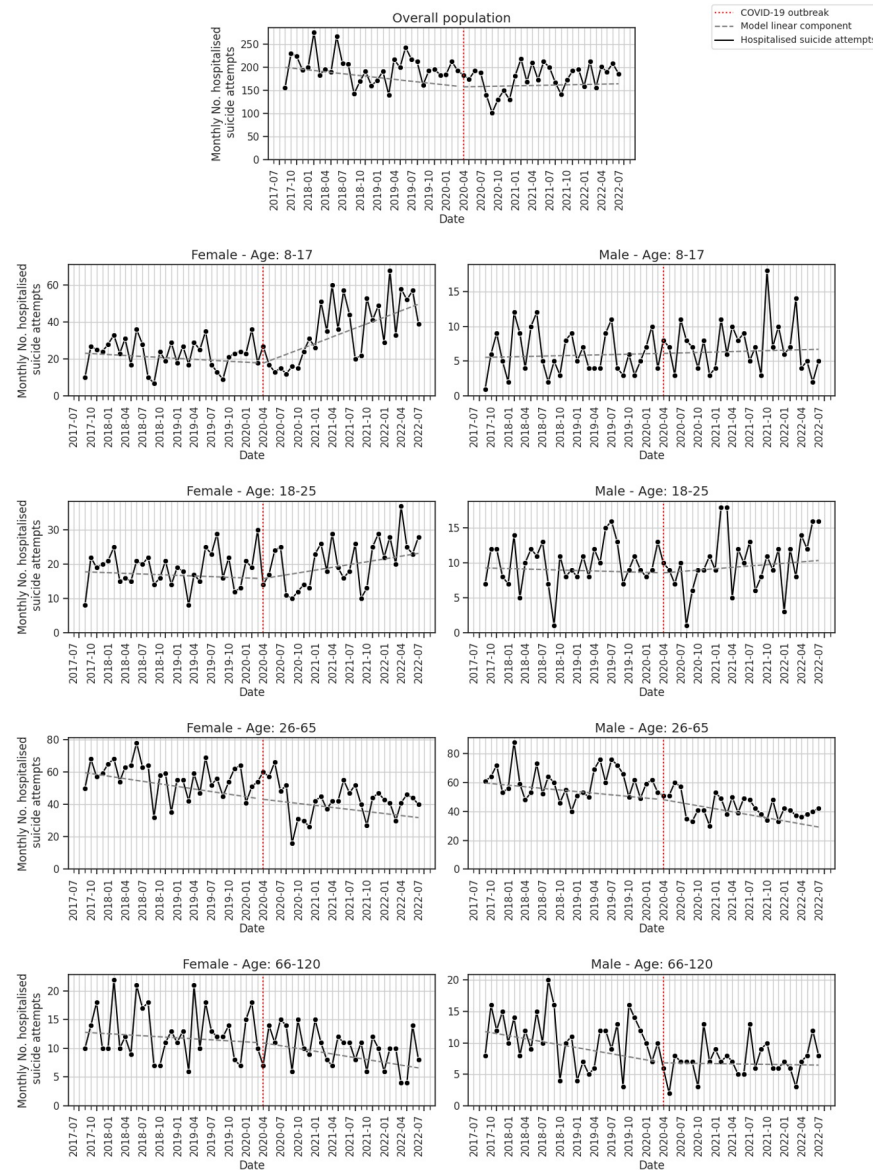
Monthly numbers of hospitalizations caused by suicide attempts - completeness adjusted



Completeness
adjusted



Monthly numbers of hospitalisations caused by suicide attempts - claim data





Open-source

Plusieurs projets matures: <https://aphp.github.io/>



Open
source

Intérêt:

- Mutualisations entre projets de recherche
- Construction progressive de l'outil de recherche sur le temps long
- Mode de collaboration entre cliniciens, chercheurs, méthodologistes, data scientists, data engineers, etc.
- Reproductibilité, auditabilité
- Qualité du code, optimisation computationnelle