Analyse de la base SIDEP
Note technique

27/08/2020

Cette note vise à décrire les choix méthodologiques concernant l'analyse de la base SIDEP pour la construction des indicateurs produits par Santé publique France.

1. Nettoyage des variables

Variable TYPE_ANALYSE
Le référentiel contient des modalités relatives aux tests PCR « 94500-6 », « 94534-5 » et « 94534-7 », et d'autres modalités relatives aux tests sérologiques. Les données sont filtrées sur les tests PCR.

Les dates
- Mise à NA si date_heure_prelevement > date/heure de l'extraction (date_integration_sidep)
- Mise à NA si date_heure_prelevement < 2020-03-09
- Pas de contrôle sur date_heure_valide_analyse (champ mal défini pour l'instant)

Résultats des tests
- Mis à NA si la variable resultat non dans la liste des références (voir liste des variables en point 5)

FINESS
- Seule la colonne FINESS juridique est remplie, mais elle contient à la fois des codes finess géographiques et des finess juridiques (par comparaison avec les données Etalab : https://www.data.gouv.fr/fr/datasets/finess-extraction-du-fichier-des-etablissements/)
- Les Finess juridiques correspondent à des SEL (« laboratoire » avec x sites de prélèvements) et les Finess géographiques aux sites de prélèvements
- Jointure avec la base Etalab pour identifier :
  - Les finess selon leur origine (géographique ou juridique)
  - Les finess invalides (i.e. qui ne sont pas présents dans la base d'Etalab)
- FINESS à NA sinon
- Ajout d'une colonne finess_all pour le décompte des établissements connectés

IRIS
La validité des numéros IRIS est évaluée en les comparants à une liste de référence
https://www.insee.fr/fr/information/2017499
https://www.insee.fr/fr/information/2028040
https://geoservices.ign.fr/documentation/diffusion/telechargement-donnees-libres.html#contoursiris

Code postal
La validité des codes postaux est évaluée en les comparants à une liste de référence
https://www.data.gouv.fr/fr/datasets/base-officielle-des-codes-postaux/
maj : 9 avril 2019.
Dans le cas où un code postal est à cheval sur deux départements, le département attribué au code postal est celui dans lequel réside le plus grand nombre d'habitants de ce code postal.

Département

Le département correspond à celui de l'IRIS du patient, ou du code postal du patient si l'IRIS est invalide. Dans le cas où le code postal est également invalide (non dans liste), mais de codé sur 5 digits et que ses deux/trois premiers digits correspondent à un département, le département est tout également défini à partir du code postal.

Dans le cas où ni l'IRIS, ni le FINESS ne permettent de déterminer le département, une jointure est réalisée sur le finess :

- Si le finess est géographique, prise en compte du département du finess géographique.
- Si le finess est juridique, prise en compte du département finess si l'ensemble des établissements rattachés au finess sont dans le même département.

A noter que cette dernière correction du département par le département du laboratoire n'est plus réalisée depuis le 12 aout 2020 si le prélèvement est réalisé dans un aéroport (code prescripteur « 291991222 » – variable rpps_adeli_prescripteur).

Age (age révolu)

- age=age_annee si age_annee>0
- age=age_mois/12 si age_annee==0
- age=NA si age_annee >120 ou age_annee <0

Sexe

- Mis à NA si ni F ni M

Apparition des symptômes

- Mis à NA si non dans la liste des références (voir liste des variables en point 5)
- Mis à NA si "U"

Profession de santé

- Mis à NA si non dans O ou N

Lieu de résidence

- Mis à NA si non dans la liste des références (voir liste des variables en point 5)
- Mis à NA si "U"

2. Suppression de lignes

- Suppression des lignes avec des résultats de test non définis (non dans P, N, X ou I).
- Une seule ligne conservée si pseudonyme + date/heure de prélèvement identique. La ligne la plus récemment créée dans SI-DEP (variable date d'intégration dans SI-DEP) est conservée (car pouvant correspondre à une correction du laboratoire ayant entraîné la réémission du résultat).
- Suppression des lignes avec date à NA

3. Définition des indicateurs

Couverture du dispositif

- Décompte du nombre de finess distincts qui se sont connectés entre le début du dispositif et le jour j, en distinguant si le finess est présent ou non dans Etalab (finess connu ou inconnu). Dans le cas où le finess retrouvé dans Etalab est de type juridique, l'ensemble des établissements qui lui sont rattachés (tels que dénombrés dans Etalab), s'ils ne sont pas par ailleurs retrouvés dans SIDEP, sont comptabilisés.
- ATTENTION : le total national intègre les prélèvements dont on ne connait pas le code géographique (i.e. finess inconnu dans ETALAB)

Pression épidémique
- Nombre d'individus (identifiant pseudonyme_hash) avec un résultat positif.
- Filtre sur les pcr positives (resultat=="P")
- Sélection de la première date avec pcr positive si plusieurs prélèvements positifs pour un même patient
- Une ligne par patient

Taux de positivité
- Nombre d'individus (identifiant pseudonyme_hash) avec un résultat positif ou négatif.
- Filtre sur les pcr valides (résultat P ou N)
- Si plusieurs prélèvements sont rapportés pour un même patient:
  - Sélection de la première date pour les pcr ayant le même résultat (par exemple première date si plusieurs pcr négatives)
  - Si pcr discordantes chez un même patient (N et P), la première pcr positive est conservée.

Taux de positivité bruts et standardisé
La même sélection de données que pour les taux de positivité est effectuée.

La taux bruts sont calculés en rapportant la nombre de cas positifs à la taille de la population, INSEE, de l'année 2020.

Les taux sont par ailleurs standardisés (standardisation directe), sur l'âge (tranches d'âge de 10 ans) et le sexe, la population de France métropolitaine de 2020 servant de population de référence[1].

Capacité analytique
- Nombre de prélèvements avec un résultat valide.
- Filtre sur les pcr valides (résultat P ou N)
- ATTENTION : on décompte ici les nombres de tests. Les décomptes sont donc plus élevés que ceux de PRESSION EPIDEMIQUE et TAUX DE POSITIVITE

Performance de la campagne de diagnostic
- Nombre de prélèvements dans chaque catégorie de la variable apparition_symptomes
- Filtre sur la variable apparition_symptomes (voir liste des variables en point 5)

4. Note sur les ventilations

Dates
- Les dates utilisées pour les ventilations sont les dates de prélèvement (jour-mois-année).

Tableaux départementaux et régionaux

[1] Pour une zone géographique *j* en le taux standardisé T_j est calculé en utilisant la formule :

$$T_j = \sum_{a,s} \frac{P_{asj}}{Pop_{asj}} Pop_{as}^{Fr}$$

où P_{asj} est le nombre de cas positifs pour la classe d'âge *a*, le sexe *s* et la zone géographique *j*, Pop_{asj} la population correspondante, et Pop_{as}^{Fr} la population française correspondante.

- Les indicateurs départementaux et régionaux sont calculés à partir de l'ensemble des données pour lesquelles il existe un département valide
- Les indicateurs nationaux sont calculés à partir de l'ensemble des données, même celles n'ayant pas un code département valide
- ATTENTION : les totaux des indicateurs départementaux et régionaux diffèrent donc des totaux nationaux

Croisements âge et sexe
- Pour les indicateurs par âge et sexe, une modalité "Totale" est ajoutée dans les tables pour les variables âges et sexe.
- La valeur des indicateurs dans les modalités "Totale" comprennent les données avec des valeurs manquantes (i.e. sur âge ou sexe)
- Attention : la somme des valeurs par âge pour un sexe donné ne correspondent donc pas nécessairement à la ligne age="Total" pour ce même sexe

Croisements typologie_lieu_residence et profession_sante_patient
- Pour les résultats ventilés par typologie_lieu_residence ou profession_sante_patient, les données sont au préalable filtrées sur ces variables.
- Les totaux dans les tables correspondent donc aux totaux des enregistrements pour lesquels la variable est renseignée.

5. Liste des variables

Nom	Libellé	Détails
pseudonyme_hash	Identifiant patient pseudonymisé	
sexe	Sexe	M ou F
age_annee	Age en année	
age_mois	Age en mois	
code_postal	Code postal de résidence du patient	
numero_iris	Code IRIS de résidence du patient	
typologie_lieu_residence	Lieu de résidence actuelle du patient	6 modalités valides : I hébergement individuel H hospitalisé E résident en EHPAD C en milieu carcéral A autre structure d'hébergement collectif U ne sait pas
profession_sante_patient	Le patient est-il un professionnel intervenant dans le système de santé ?	3 modalités valides : O oui N non U ne sait pas
apparition_symptomes	Date d'apparition des symptômes	7 modalités valides : ASY asymptomatique S01 symptômes apparus le jour ou la veille du prélèvement S24 symptômes apparus 2, 3 ou 4 jours avant le prélèvement

		S57 symptômes apparus 5, 6 ou 7 jours avant le prélèvement SS2 symptômes apparus entre 8 et 15 jours avant le prélèvement SS3 symptômes apparus plus de deux semaines avant le prélèvement U ne sait pas
num_dossier	Numéro de dossier/prélèvement	
date_heure_prelevement	Date, heure et minute du prélèvement	
finess_geographique_lbm	FINESS géographique	Champs vide au 15/07/2020
finess_juridique_lbm	FINESS juridique	FINESS géographique, ou, à défaut, juridique, du laboratoire responsable du prélèvement et de l'analyse (et non du laboratoire éventuellement sous-traitant)
type_analyse	Type d'analyse réalisée	5 modalités valides au 15/07/2020 94500-6 Résultat 94534-5 Reprise calcul 94533-7 Reprise calcul 94563-4 IgG anti-SARS-CoV-2 94564-2 COVID IgM 94762-2 Anticorps totaux Covid-19
resultat	Résultat du test	4 modalités valides P positif N négatif I indéterminé X prélèvement non conforme
date_heure_valide_analyse	Date, heure et minute de validation de l'analyse	Au 15/07/2020, cette variable est non exploitable (correspond à ~ 90% à la variable date_heure_prelevement)
date_integration_sidep	Date et heure de validation de l'analyse	Variable intégrée à la base SIDEP le 19/05/2020
rpps_adeli_prescripteur	Code prescripteur	Variable intégrée à la base SIDEP le 11/08/2020