RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Drees
DIRECTION DE LA RECHERCHE, DES ÉTUDES, DE L'ÉVALUATION ET DES STATISTIQUES

Le 29 octobre 2021

Statistiques issues des données appariées entre SI-VIC, SI-DEP, VAC-SI : amélioration de la méthode et révision des statistiques produites

Présentation méthodologique associée à la publication du 29 octobre 2021

Plan

Pourquoi devient-il nécessaire de revoir la méthode de production des statistiques exploitant les données appariées ?

Les choix méthodologiques concernant les observations non appariées entre les trois bases de données mobilisées dépendent de l'information disponible, qui est maintenant plus large que lors des premières publications

Les statistiques issues des données appariées visent à informer sur la nature du statut vaccinal (information disponible dans VAC-SI) :

- des tests (PCR et TAg) renseignés dans SI-DEP,
- des événements hospitaliers (entrées et décès) renseignés dans SI-VIC.

Pour ce faire, il faut rechercher, pour chaque test ou événement, l'information sur la situation vaccinale du patient concerné dans la base VAC-SI. Une variable identifiante, le pseudonyme, est la clé utilisée dans cette recherche entre les trois bases (encadré ci-dessous). Pour les évènements pour lesquels on identifie que la personne associée a reçu une injection, on attribue les statuts vaccinaux complets ou primo-doses suivant les définitions mentionnées dans l'annexe méthodologique de chaque publication. Pour les autres événements, pour lesquels il n'était pas possible d'identifier au moins une injection du patient concerné, le patient était conventionnellement considéré comme non vacciné, comme indiqué dans l'annexe méthodologique des publications de la DREES depuis le 20 août. Ce choix permettait de pouvoir établir des statistiques sur l'ensemble du champ des tests et des événements hospitaliers renseignés dans SI-VIC et SI-DEP[1], mais reposait sur l'hypothèse que le taux d'appariement fondé sur le pseudonyme entre SI-VIC ou SI-DEP et VAC-SI était proche de 100 %. C'est pourquoi, consciente des risques de non-appariement liés à des différences dans les traits identifiants renseignés dans les différentes bases (encadré), la DREES a choisi de mettre davantage en avant des statistiques portant sur les tests PCR et sur les hospitalisations concernant des personnes avec tests PCR positif identifié, la qualité des données saisies – y compris pour les données identifiantes – étant supérieure à ce que l'on observe pour les tests antigéniques.

[1] À l'exception des observations pour lesquelles l'âge du patient n'est pas connu dans les bases pseudonymisées.

Encadré : La construction d'un pseudonyme commun dans les bases SI-VIC, SI-DEP et VAC-SI

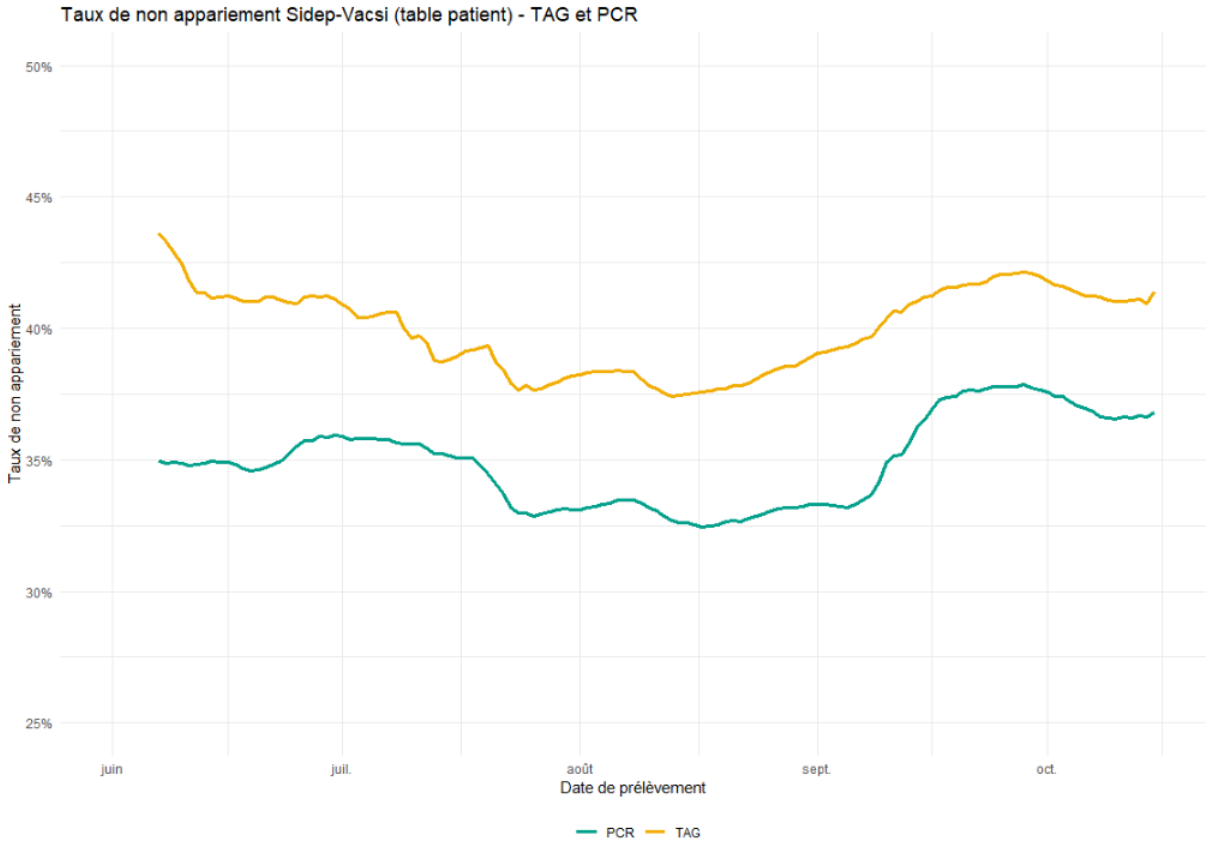
Dans chacune des trois bases, lors de leur constitution et avant transmission à la DREES, les seules informations sur le nom patronymique, premier prénom, sexe et date de naissance des personnes hospitalisées, testées ou vaccinées servent à la constitution d'un pseudonyme, chaîne de caractères non signifiante identifiant de façon unique chaque personne. Ces informations nominatives sont ensuite supprimées des bases avant transmission à la DREES pour exploitation statistique. Seul le pseudonyme est transmis. Pour que l'appariement sur ce pseudonyme entre tests, hospitalisations et vaccinations soit possible, il faut que les données identifiantes renseignées dans chacune des bases et utilisées en entrée de l'algorithme soient identiques. Cela peut ne pas être le cas pour plusieurs raisons : erreur de saisie des identités, renseignement du nom marital et pas du nom patronymique, écarts dans la saisie pour les caractères accentués ou spéciaux (traits d'union par exemple). Certaines de ces différences sont « gommées » par une étape de normalisation des traits d'identité réalisée en amont de la constitution du pseudonyme, pour les trois bases séparément. Des analyses méthodologiques menées en septembre ont cependant montré que l'étape de normalisation pouvait être améliorée afin d'augmenter le taux d'appariement. Des travaux sont actuellement en cours de réalisation sur ce sujet.

L'expertise du taux d'appariement entre la base VAC-SI et les deux autres n'a pu être menée qu'en septembre dernier, en mobilisant cette fois une autre information disponible dans VAC-SI, concernant l'éligibilité des patients à la vaccination[2]. À partir de cette information, il est possible de rechercher pour tous les tests et événements dont le patient ne présente aucun enregistrement d'injection (à partir de son pseudonyme) si ce patient est cependant effectivement présent comme éligible dans la table de patients de VAC-SI.

Il s'avère que la proportion de non appariés avec cette table de patients est importante (graphique 1), et trop élevée pour que soit valable la convention de considérer comme non vaccinées les personnes dont aucune injection n'a été retrouvée dans VAC-SI. En outre, cette convention a créé un biais croissant dans le temps, au fur et à mesure que la couverture vaccinale augmentait, la part de réellement non vaccinés parmi les non-appariés diminuant mécaniquement.

[2] Plus précisément, la DREES reçoit VAC-SI sous la forme de plusieurs tables, et notamment une table recensant l'ensemble des injections réalisées depuis le début de la vaccination (premières, deuxièmes et troisièmes doses, rappels depuis le mois de septembre), et une table concernant l'ensemble des patients susceptibles d'être vaccinés (ciblés par l'Assurance maladie), qu'ils le soient effectivement ou non. Les premières exploitations menées par la DREES ne mobilisaient que la table vaccination.

Graphique 1 : Taux de non-appariement des tests PCR et antigéniques entre SI-DEP et VAC-SI (table patient)



Sources : SI-DEP, VAC-SI, calculs DREES. Données en moyenne mobile glissante sur 7 jours.

Les analyses de la répartition des observations non appariées sont importantes afin de les traiter de façon la plus adéquate possible. En particulier, il est crucial de savoir si les observations non appariées se retrouvent dans des proportions similaires selon toutes les dimensions d'information disponibles dans les bases ou bien si, au contraire, elles sont plus ou moins fréquentes que dans l'ensemble des observations pour certaines catégories de population. Autrement dit, il convient de savoir si le défaut d'appariement n'est pas lié statistiquement avec une information renseignée comme l'âge, la région de résidence, le sexe notamment.

Les analyses menées sur les taux d'appariement sur les différentes catégories de population ont révélé que les défauts d'appariements sont légèrement plus importants chez les femmes, les personnes âgées, et les populations d'outre-mer.

Des investigations complémentaires menées par l'ensemble des parties prenantes, notamment ceux chargés de la pseudonymisation des tables, ont permis d'identifier certaines des causes de ce défaut d'appariement. Ainsi :

- à partir de tests sur des identités fictives, il est apparu que les algorithmes ne traitaient pas identiquement les personnes ayant un espace ou un tiret dans leur nom ou prénom. Sur ce point, des modifications de l'algorithme de traitement du pseudonyme dans VAC-SI sont en cours d'implémentation. Elles permettront de réduire le défaut d'appariement constaté entre VAC-SI et les deux autres bases appariées ;

- les données SI-DEP sont transmises à la DREES dès réception dans le système sans nécessairement que la donnée ait été « vérifiée » par appel au téléservice INS (identité nationale de santé), qui vise à « nettoyer » les données d'entrée, par exemple en réattribuant un nom patronymique lorsque c'est le nom marital qui a été donné. Ce « nettoyage » des bases n'est par ailleurs réalisé qu'en cas d'utilisation de la carte vitale du patient pour la saisie dans SI-DEP, ce qui n'est pas forcément très fréquent notamment s'agissant des tests antigéniques en pharmacie. Le défaut d'appariement qui en résulte n'est donc pas le fait d'un problème dans les systèmes d'information, mais résulte de la saisie des données en entrée. Des consignes ont de nouveau été transmises aux professionnels de santé pour tenter d'améliorer la qualité des variables renseignées dans les systèmes d'information.

Dès lors, on fait l'hypothèse qu'à âge, sexe, région donnée, le phénomène de non appariement est aléatoire et indépendant du statut vaccinal.

Enfin, et plus marginalement, l'apparition des vaccinations de rappel à partir d'août 2021 a conduit à une autre modification des données publiées, afin que la date de vaccination considérée soit bien celle de la dose ayant donné lieu au statut vaccinal complet et non celle de la dose de rappel.

Les limites sur la qualité de l'appariement entre les bases sur la vaccination et celles les tests ou les hospitalisations nécessitent de revoir les choix pour les observations « non appariées »

Ce constat d'une forte proportion d'observations dont le statut vaccinal est inconnu, ainsi que l'identification des raisons du non-appariement, conduit la DREES à revoir comme suit le processus d'élaboration des statistiques issues des appariements.

Les observations pour lesquelles le patient est retrouvé dans la table des injections conservent le même statut vaccinal (complet ou primo-dose). Parmi les observations restantes, celles dont le patient est retrouvé comme éligible dans VAC-SI (mais sans injection), sont considérées comme relevant d'un non-vacciné. Si ce n'est pas le cas, les observations en question sont dites « non appariées ».

Il est possible de répartir ces observations restant non appariées entre les différents statuts vaccinaux : cela permet d'obtenir, pour l'ensemble du champ statistique des événements SI-VIC et SI-DEP - c'est-à-dire les tests, les hospitalisations et les décès hospitaliers – et non la seule sous-partie appariée, une vision de leur décomposition par statut vaccinal. Plus précisément, les observations appariées sont surpondérées par strate afin d'utiliser leurs caractéristiques vaccinales pour représenter l'ensemble de la strate. Ces groupes élémentaires d'observations appelés strates sont définies par certaines caractéristiques identiques et on considère que le fait d'être apparié ou non est aléatoire au sein de ces groupes. Les strates sont constituées à partir de la semaine de prélèvement ou de survenue, de l'âge décennal, du sexe, du département de résidence et, s'agissant des tests, du caractère symptomatique ou non du patient au moment de son test[3]. On repondère ainsi, strate par strate, les

[3] S'agissant des données sur les hospitalisations, le reclassement est fait aux niveaux de stratification successifs suivants :
- tranche d'âge de 10 ans, sexe, département et semaine calendaire ;
- si ce niveau ne permet pas de catégoriser les non-appariés (aucun effectif apparié dans la classe), tranche d'âge de 10 ans, département et semaine calendaire ;
- si ce niveau ne permet pas de catégoriser les non-appariés (aucun effectif apparié dans la classe), tranche d'âge de 10 ans, région et semaine calendaire ;

observations appariées pour que chaque strate retrouve son « poids d'origine » (nombre de tests, nombre d'hospitalisations)[4]. Ce choix conduit à ce que chaque observation soit pondérée d'un poids supérieur ou égal à 1 et, dans la plupart des cas, non entier, conduisant à des valeurs décimales (arrondies au centième) dans les données mises en ligne.

Toutes les observations pondérées par strate sont ensuite agrégées sur l'ensemble des strates. L'intérêt de travailler par strate fine plutôt que d'appliquer des pondérations pour l'ensemble de la population est de limiter le biais lié aux hétérogénéités de taux d'appariement existant selon les âges, sexes, régions de résidence (et caractère symptomatique le cas échéant, comme présenté en annexe).

Ces statistiques fournissent des répartitions de tests et d'événements hospitaliers, mais il est également nécessaire de disposer d'une décomposition de l'ensemble de la population selon les différents statuts vaccinaux afin de mesurer des entrées hospitalières ou des tests à taille de population comparable (pour 1 million ou 100 000 habitants). Comme indiqué dans l'ensemble des précédents encadrés méthodologiques des publications de la DREES, pour les personnes vaccinées, les dénombrements issus de la table des injections de VAC-SI fournissent une estimation de taille de population. En revanche, le nombre de personnes non vaccinées est obtenu en retranchant les personnes vaccinées de l'estimation de population résidente produite par l'Insee. En effet, le nombre de personnes éligibles dans la table VAC-SI est très vraisemblablement sur-estimé[5].

En définitive, les graphiques suivants montrent le résultat du « reclassement » des individus non appariés entre les différents statuts vaccinaux.

- si ce niveau ne permet pas de catégoriser les non-appariés (aucun effectif apparié dans la classe), région et semaine calendaire ;
- si ce niveau ne permet pas de catégoriser les non-appariés (aucun effectif apparié dans la classe), semaine calendaire ;
- si ce niveau ne permet pas de catégoriser les non-appariés (aucun effectif apparié dans la classe), ensemble de la période.

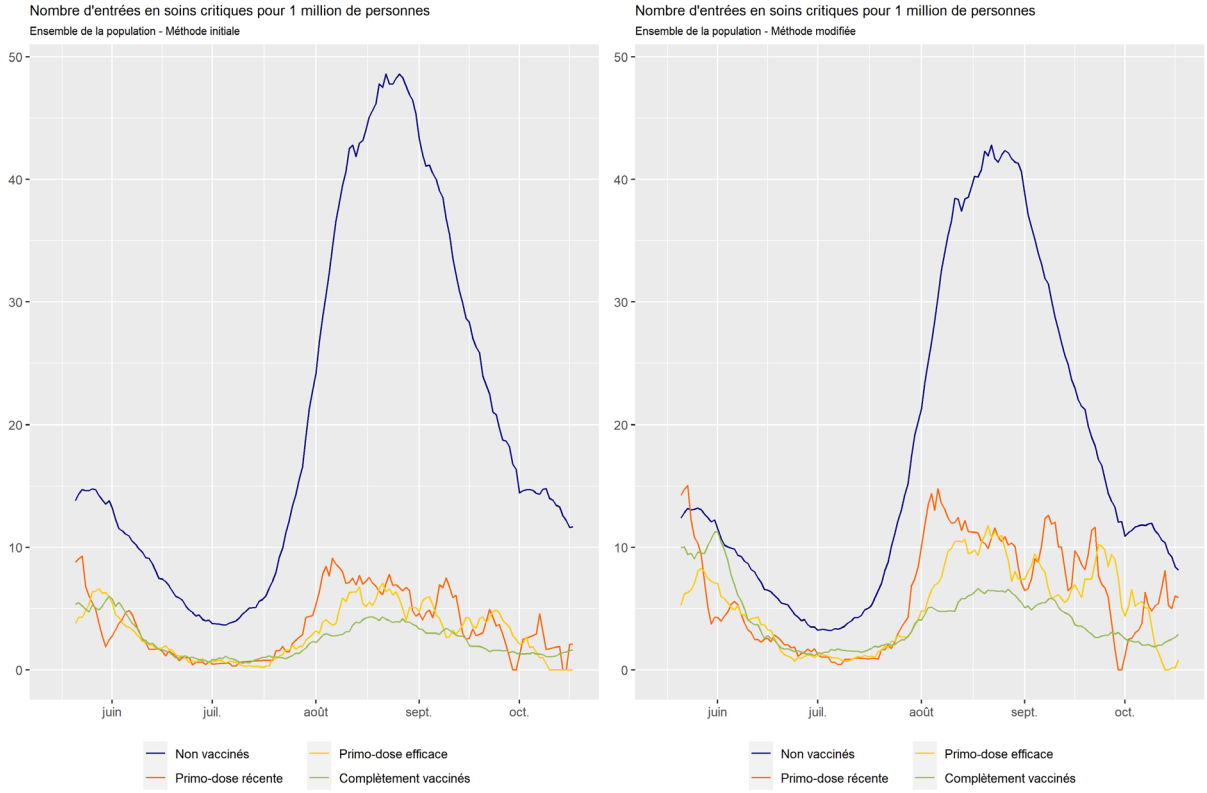
S'agissant des données sur les tests, le reclassement est fait aux niveaux de stratification successifs suivants :
- tranche d'âge de 10 ans, sexe, département, existence de symptômes et semaine calendaire ;
- si ce niveau ne permet pas de catégoriser les non-appariés (aucun effectif apparié dans la classe), tranche d'âge de 10 ans, département, existence de symptômes et semaine calendaire ;
- si ce niveau ne permet pas de catégoriser les non-appariés (aucun effectif apparié dans la classe), tranche d'âge de 10 ans, région, existence de symptômes et semaine calendaire ;
- si ce niveau ne permet pas de catégoriser les non-appariés (aucun effectif apparié dans la classe), région, existence de symptômes et semaine calendaire ;
- si ce niveau ne permet pas de catégoriser les non-appariés (aucun effectif apparié dans la classe), existence de symptômes et semaine calendaire.

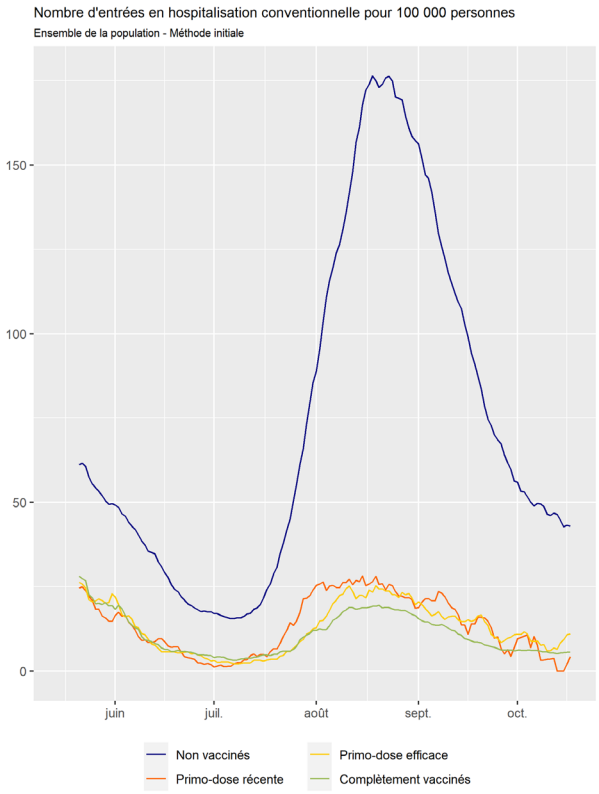
Les variables reclassées sont les suivantes :
- pour les tests : nombre de tests positifs et nombre de tests négatifs ;
- pour les hospitalisations (distinguées selon leur nature : hospitalisation conventionnelle, soins critiques, *etc.*) : nombre d'admissions avec PCR positive, nombre d'admissions avec PCR négative, nombre d'admissions sans test PCR.

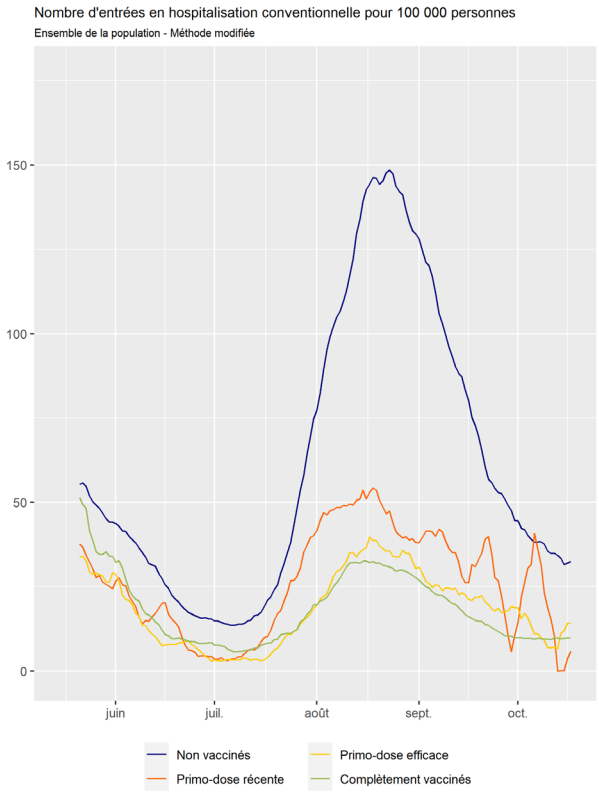
[4] Par exemple si 240 tests sont observés pour une strate donnée dont 200 sont identifiés par appariement, 150 relevant de vaccinés et 50 relevant de non-vaccinés, le ratio 240 / 200 est appliqué à ces appariés. Si bien qu'on aboutit à 180 tests provenant des vaccinés et 60 de non-vaccinés.

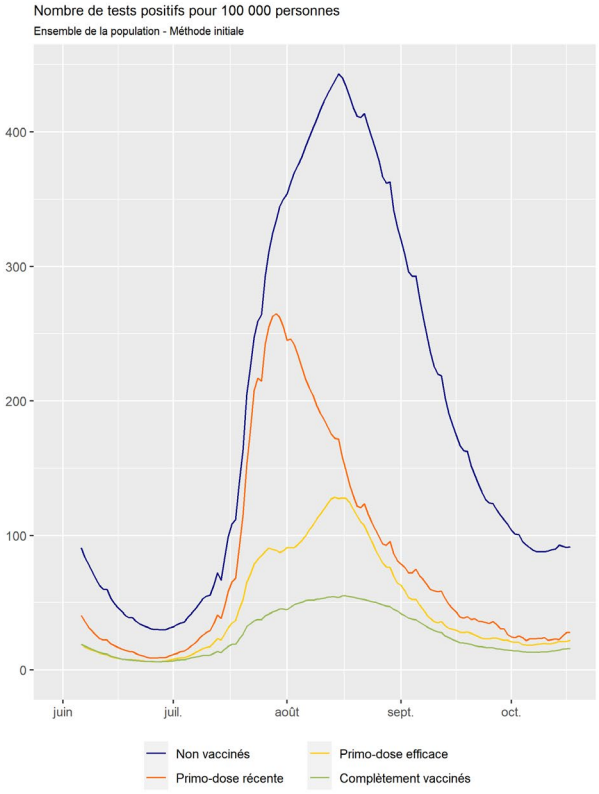
[5] en raison de doublons, de personnes décédées continuant d'appartenir à la table patients ou autres problèmes de cohérence de champ.

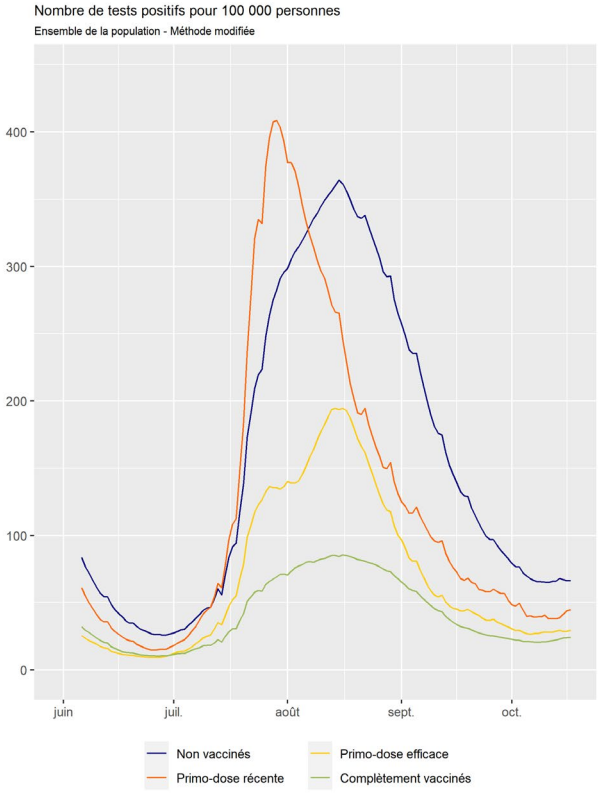
Graphiques 2 : nombres d'entrées hospitalières et de tests positifs à taille de population comparable pour chaque statut vaccinal – comparaison entre ancienne et nouvelle méthode de traitement des données non-appariées











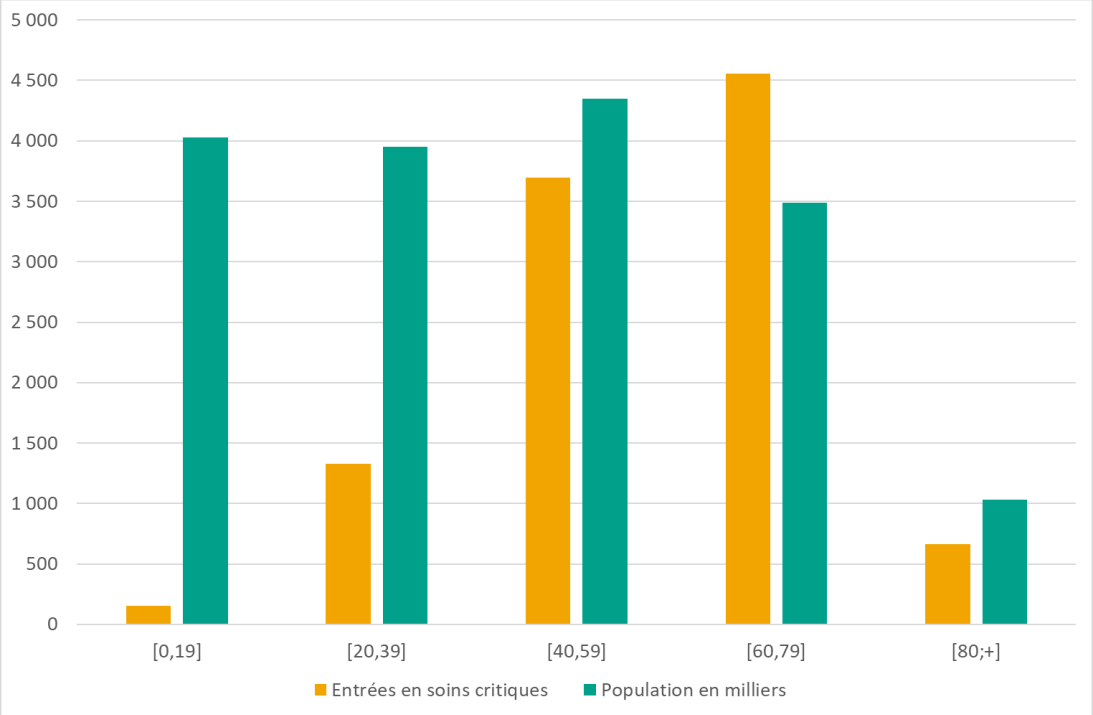
Sources : SI-VIC, SI-DEP, VAC-SI, calculs DREES. Données en moyenne mobile glissante sur 7 jours.

Pourquoi convient-il de centrer les messages issus des appariements sur les adultes ?

Le faible nombre d'événement hospitaliers des moins de 20 ans conduit à préférer se restreindre aux adultes de plus de 20 ans pour les principaux indicateurs analysés

Dans un souci d'exhaustivité des analyses, l'ensemble de la population, quel que soit l'âge des personnes, avait été retenu comme champ d'élaboration des principaux indicateurs présentés dans les publications hebdomadaires. Cependant, en raison de très faibles nombres d'hospitalisés de moins de 20 ans, l'inclusion de cette catégorie des plus jeunes perturbe l'analyse des statistiques produites en les rendant difficilement comparables aux statistiques de risques relatifs disponibles par ailleurs. Le graphique ci-dessous illustre bien cette singularité de la classe d'âge des moins de 20 ans, qui représentent environ un quart de la population française mais à peine 5 % des admis en soins critiques depuis le 31 mai.

Graphique 3 : Nombre d'entrées en soins critiques et population par âge



Sources : SI-VIC, VAC-SI, Insee, calculs DREES.

Note : entrées hospitalières entre le 31 mai et le 10 octobre 2021, estimations de population au 1er janvier 2021.

En effet, le risque relatif (*RR)* de l'ensemble de la population (*pop*) entre les personnes non-vaccinées et complètement vaccinées peut par exemple s'écrire comme suit pour les entrées en soins critiques (*SC)* en distinguant les jeunes de moins de 20 ans (*J*) et les adultes (*A*) à partir de cet âge, qu'ils soient vaccinés (complètement vaccinés *CV*) ou non (*NV*).

$$RR\ en\ SC\ = \frac{\left(\frac{SC_{NV}}{pop_{NV}}\right)}{\left(\frac{SC_{CV}}{pop_{CV}}\right)} = \frac{\left(\frac{SC^{J}_{NV} + SC^{A}_{NV}}{pop^{J}_{NV} + pop^{A}_{NV}}\right)}{\left(\frac{SC^{J}_{CV} + SC^{A}_{CV}}{pop^{J}_{CV} + pop^{A}_{CV}}\right)}$$

Parmi les termes concernant les jeunes (en vert), alors que les effectifs de population sont d'ampleur comparable à ceux des adultes, les effectifs d'hospitalisés sont bien moindres et négligeables au regard des effectifs adultes, si bien que le risque relatif est proche de :

$$RR\ en\ SC\ = \frac{\left(\frac{SC^{A}_{NV}}{pop^{J}_{NV} + pop^{A}_{NV}}\right)}{\left(\frac{SC^{A}_{CV}}{pop^{J}_{CV} + pop^{A}_{CV}}\right)} = \frac{SC^{A}_{NV}}{SC^{A}_{CV}} \times \frac{pop^{J}_{CV} + pop^{A}_{CV}}{pop^{J}_{NV} + pop^{A}_{NV}}$$

La population des jeunes complètement vaccinés pop^{J}_{CV} étant initialement nulle, le risque relatif *RR* est tiré à la baisse avant la hausse de la couverture vaccinale[6] et cet effet s'atténue ensuite, ce qui conduit à une hausse du risque relatif qui reflète bien moins des changements épidémiologiques que des variations de couverture vaccinale.

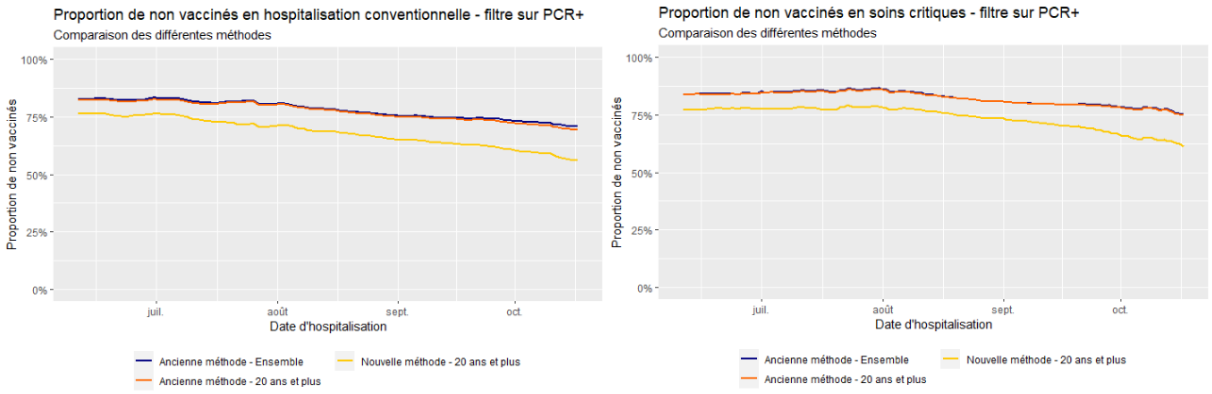
Dans tous les cas, même si la DREES fait désormais le choix éditorial de mettre l'accent sur des indicateurs et analyses issues des appariements portant sur le périmètre des adultes de 20 ans et plus, l'ensemble des données pour toutes les tranches vingtennales de la population (y compris les moins de 20 ans) reste disponible comme précédemment dans les mises à disposition chaque semaine sur le site de la DREES.

Quels sont les effets de ces révisions sur les statistiques produites jusqu'à présent ?

Les révisions des données précédemment évoquées (répartition de la population non appariée au prorata de la structure vaccinale observée sur la population appariée) conduisent, toutes choses égales par ailleurs, à revoir à la hausse le nombre d'évènements (tests positifs ou hospitalisations) chez les personnes vaccinées et à le revoir à la baisse chez les personnes non vaccinées, par rapport à l'hypothèse sous-jacente précédente. Dès lors, le taux de personnes non vaccinées parmi les événements considérés est révisé à la baisse, de même que le risque relatif des événements.

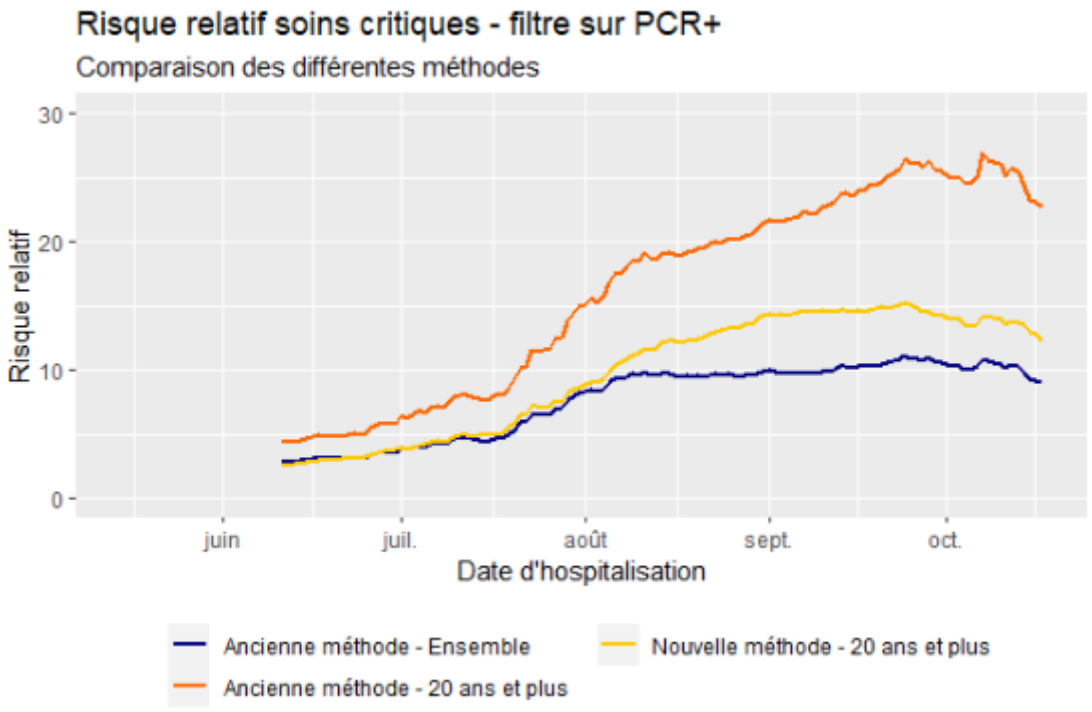
Les graphiques ci-dessous montrent l'impact de la révision sur la part des personnes non vaccinées parmi les personnes hospitalisées. Le chiffre de « 8 personnes hospitalisées sur 10 ne sont pas vaccinées » reste valable jusqu'à la mi-août s'agissant des hospitalisations en soins critiques.

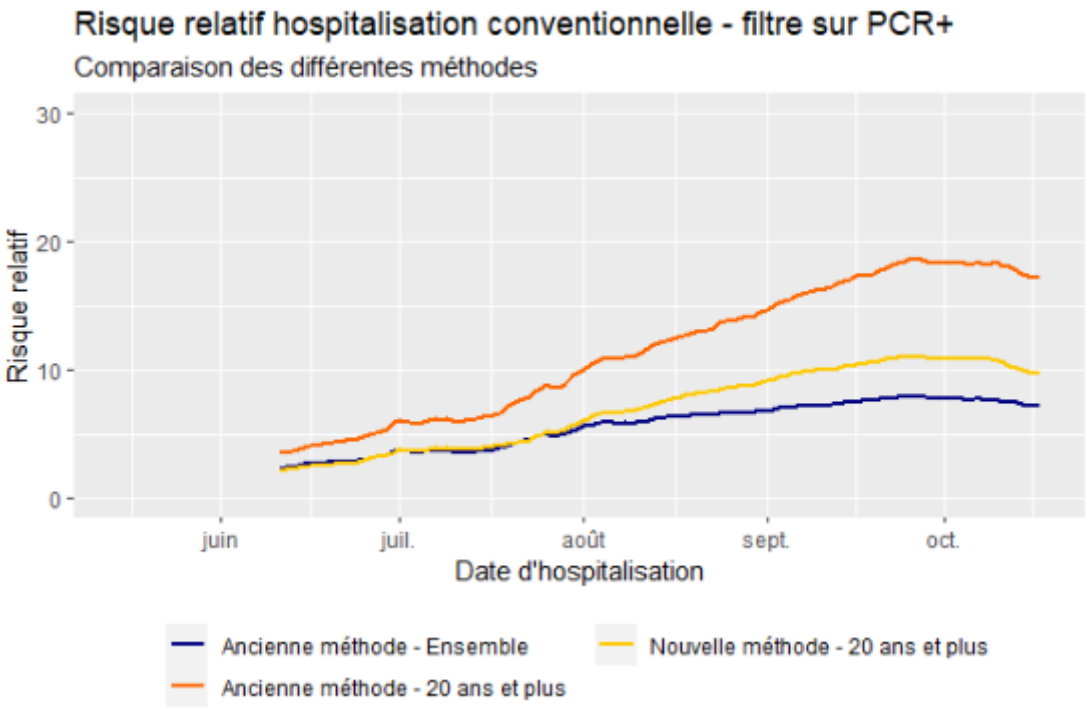
[6] Les mineurs de 12 ans et plus peuvent accéder à la vaccination depuis le 15 juin 2021.

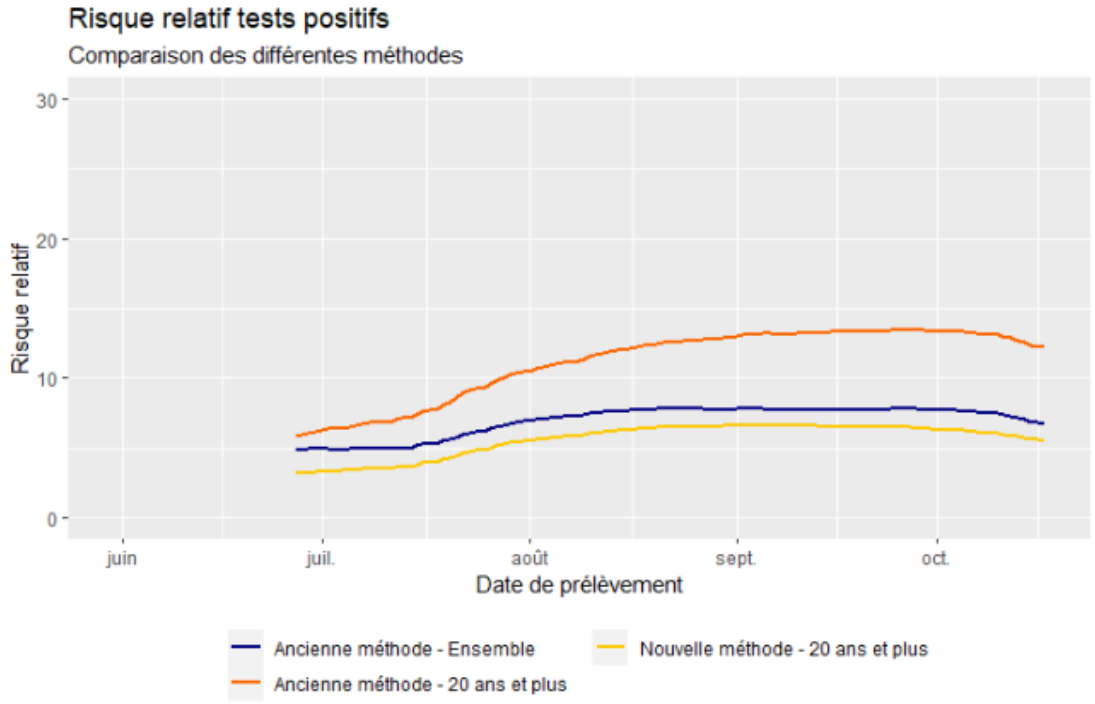


Afin d'analyser des indicateurs plus homogène durant la période considérée, les statistiques principales portent désormais sur la population de 20 ans et plus. Le risque relatif à ne pas être vacciné est plus élevé sur ce champ que sur celui calculé pour l'ensemble de la population.

Graphiques 4 : risque relatif d'entrées hospitalières et de test positif selon la méthode et le champ de population retenu





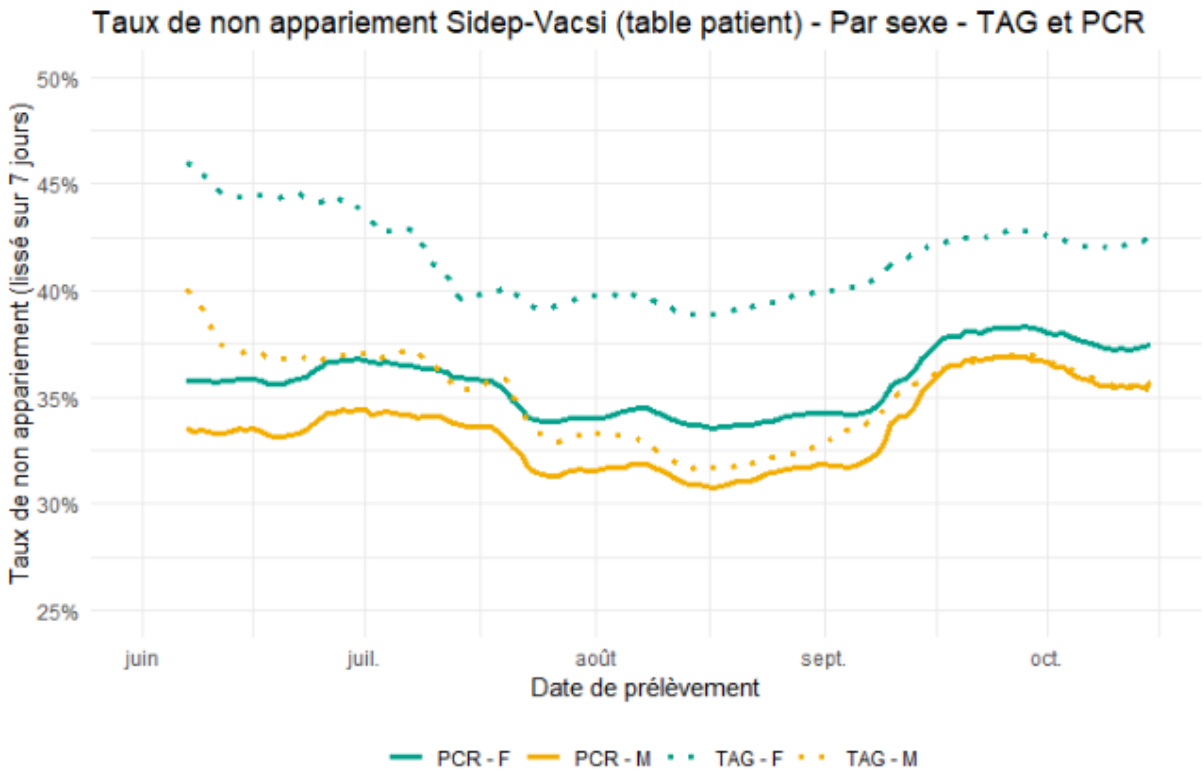


Sources : SI-VIC, SI-DEP, VAC-SI, calculs Drees.

Pour la population adulte (20 ans et plus), il y a près de 10 fois plus de chances d'être hospitalisé (y compris en hospitalisation conventionnelle) en n'étant pas vacciné qu'en l'étant. Le risque relatif est moins favorable sur les tests PCR positifs, légèrement supérieur à 5.

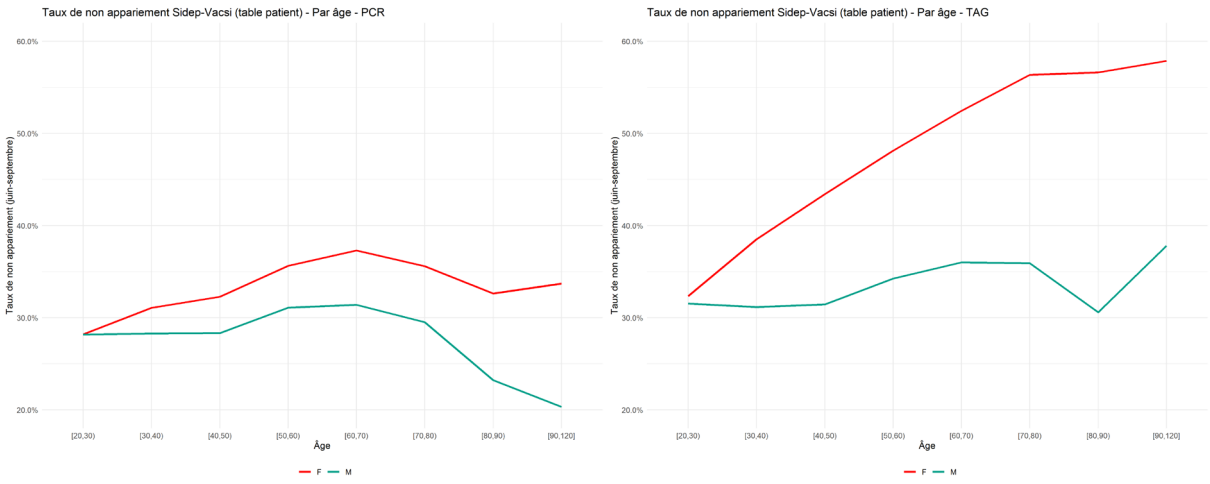
Annexe – Statistiques descriptives sur le taux de non-appariement SI-VIC-VAC-SI

Selon le sexe



Sources : SI-DEP, VAC-SI, calculs Drees.

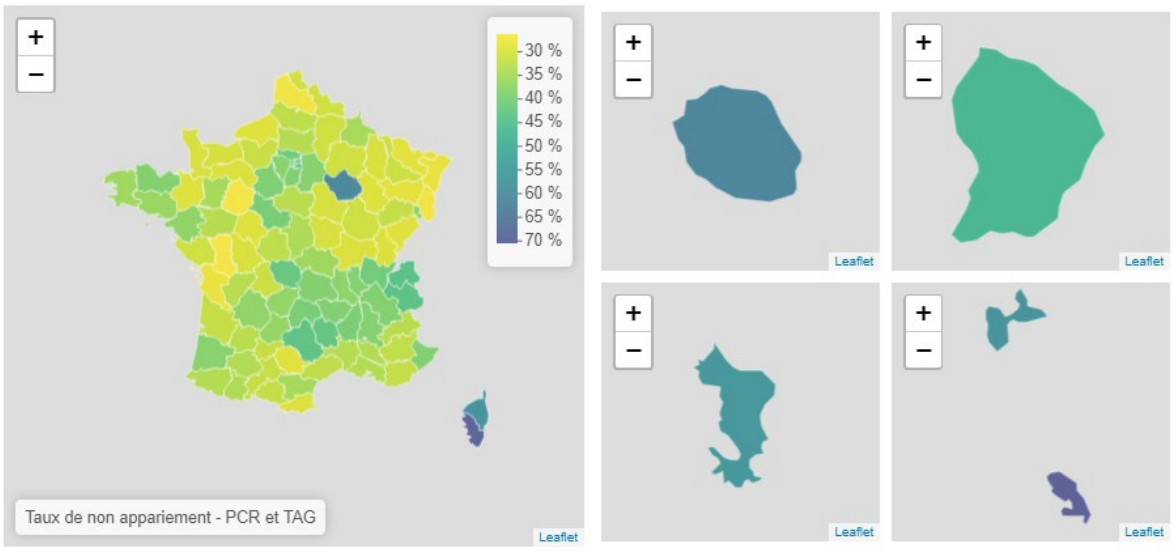
Selon l'âge et le sexe, par type de test



Sources : SI-DEP, VAC-SI, calculs Drees.

Note : dans la perspective d'étendre le champ des statistiques produites sur les tests, les investigations méthodologiques ont aussi porté sur les tests antigéniques (TAg).

Selon le département



Sources : SI-DEP, VAC-SI, calculs Drees. Ensemble des tests (PCR et TAG), données du mois de juillet 2021.

Selon le caractère symptomatique ou non