Note méthodologique – Affinement de la méthode d'imputation des personnes non appariées avec la base VACSI

08/04/2022

La Direction de la recherche, des études, de l'évaluation et des statistiques (DREES) publie les résultats relatifs à l'épidémie de Covid-19 à partir des données d'appariements de tests, d'entrées hospitalières et de vaccinations. Les statistiques issues de ces données appariées visent à informer sur la nature du statut vaccinal (information disponible dans VACSI) :

- des tests de dépistage 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 VACSI. Une variable non-signifiante, 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. Par ailleurs, la base VACSI comporte des informations sur l'ensemble *a priori* des personnes éligibles à la vaccination[1]. Dès lors, si une personne est trouvée dans ce référentiel mais qu'aucune injection ne lui est associée, elle est considérée comme non vaccinée. Toutefois, le pseudonyme utilisé dans les différentes bases présente certaines limites (en particulier il nécessite une saisie non erronée des traits d'identité, la saisie du nom de naissance plutôt que du nom marital, *etc.*), conduisant à ce qu'une partie des observations de tests ou d'hospitalisations ne sont pas appariées avec la base VACSI : pour celles-ci, il est nécessaire de réaliser une imputation, à partir des caractéristiques vaccinales observées pour des personnes similaires (en fonction de la zone géographique, de l'âge, du sexe, de la période de temps, *etc.*). Le détail de cette imputation a été présenté dans la publication du 29 octobre 2021 (https://drees.solidarites-sante.gouv.fr/sites/default/files/2021-10/211029%20Am%C3%A9liorations%20m%C3%A9thodologiques%20des%20appariements_vf.pdf).

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

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.

L'imputation repose sur l'hypothèse centrale que le phénomène de non-appariement est aléatoire et indépendant du statut vaccinal, pour des personnes similaires selon un jeu de caractéristiques données. Ainsi, les observations non appariées sont réparties entre les différents statuts vaccinaux, en reproduisant la structure vaccinale de la population similaire dont on connait le statut vaccinal. Cette imputation est réalisée par strates, définies – jusqu'à présent - à partir des critères suivants :

- S'agissant des données sur les hospitalisations, le reclassement était 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 ;

[1] Plus précisément, la DREES reçoit VACSI 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.

    - 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 ;
    - 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 était 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.

Des analyses sur les données récentes ont fait apparaître certaines fragilités dans cette méthode d'imputation, en particulier concernant l'ordre de relâchement des variables pour définir les strates de personnes. En effet, cette imputation a un impact d'autant plus fort sur les statistiques produites que les évènements régionaux pour lesquels on dispose de l'information sur le statut vaccinal sont très peu nombreux, que ce soit lié à un moins bon taux d'appariement à la base VACSI, et/ou à des nombres d'évènements (en particulier, hospitalisation) faibles. Partant de ce constat, l'ordre de relâchement des variables utilisées pour la définition des strates a été revu et affiné. Ainsi les variables sont relâchées de façon plus « souple » (par exemple auparavant on passait de la semaine à l'ensemble de la période tandis que le relâchement est plus progressif désormais), et plus précisément dans l'ordre successif suivant :

- sexe ;
- âge décennal (mais l'âge vingtennal est conservé) ;
- semaine (mais le mois est conservé) ;
- mois (mais le trimestre est conservé) ;
- âge vingtennal ;
- trimestre.

Ainsi, si toutes les contraintes sont amenées à être relâchées, ce qui arrive très rarement, les imputations se font à partir du statut vaccinal observé sur les personnes de la même région. Par ailleurs, un critère de taille (au moins 4 personnes) est imposé sur le groupe de personnes servant à imputer le statut vaccinal pour éviter que l'extrapolation soit faite sur un échantillon beaucoup trop faible. Le tableau ci-dessous précise la fréquence des observations selon le fait qu'elles soient appariées ou imputées à partir de différentes variables, en précisant l'écart entre la version antérieure et la version actuelle. En outre, une annexe présente la même information dans le temps.

Tableau : Fréquences des différentes méthodes d'imputations des statuts vaccinaux des tests RT-PCR positifs et des hospitalisations conventionnelles avec tests RT-PCR positifs

		Imputation antérieure au 7 avril 2022					Imputation affinée						
Parts des groupes de variables utilisés pour l'imputation du 28 février 2022 au 27 mars 2022		apparié	région semaine âge décennal sexe	région semaine âge décennal	région semaine	semaine	apparié	région semaine âge décennal sexe	région semaine âge décennal	région semaine âge vingtennal	région mois âge vingtennal	région trimestre âge vingtennal	région trimestre
Tests RT-PCR positifs	Auvergne-Rhône-Alpes	83,7%	16,3%	0,0%	0,0%	0,0%	83,7%	16,2%	0,0%	0,0%	0,0%	0,0%	0,0%
	Bourgogne-Franche-Comté	90,6%	9,4%	0,0%	0,0%	0,0%	90,6%	9,3%	0,0%	0,0%	0,0%	0,0%	0,0%
	Bretagne	90,3%	9,7%	0,0%	0,0%	0,0%	90,3%	9,7%	0,0%	0,0%	0,0%	0,0%	0,0%
	Corse	83,4%	16,5%	0,0%	0,0%	0,0%	83,4%	15,8%	0,5%	0,1%	0,1%	0,0%	0,0%
	Centre-Val de Loire	90,2%	9,8%	0,0%	0,0%	0,0%	90,2%	9,8%	0,0%	0,0%	0,0%	0,0%	0,0%
	Grand Est	90,4%	9,6%	0,0%	0,0%	0,0%	90,4%	9,6%	0,0%	0,0%	0,0%	0,0%	0,0%
	Guadeloupe	55,4%	38,5%	3,2%	2,9%	0,1%	55,4%	20,6%	12,0%	7,1%	3,2%	1,6%	0,1%
	Guyane	72,7%	24,4%	1,6%	1,4%	0,0%	72,7%	13,6%	6,4%	5,0%	2,2%	0,2%	0,0%
	Hauts-de-France	88,7%	11,3%	0,0%	0,0%	0,0%	88,7%	11,3%	0,0%	0,0%	0,0%	0,0%	0,0%
	Île-de-France	83,9%	16,1%	0,0%	0,0%	0,0%	83,9%	16,1%	0,0%	0,0%	0,0%	0,0%	0,0%
	La Réunion	57,9%	42,0%	0,1%	0,1%	0,0%	57,9%	41,6%	0,2%	0,2%	0,1%	0,0%	0,0%
	Martinique	37,7%	61,0%	0,8%	0,4%	0,1%	37,7%	53,9%	5,1%	2,0%	1,3%	0,1%	0,0%
	Mayotte	62,3%	11,5%	8,2%	11,5%	6,6%	62,3%	0,0%	0,0%	1,6%	18,0%	14,8%	3,3%
	Nouvelle-Aquitaine	88,0%	12,0%	0,0%	0,0%	0,0%	88,0%	12,0%	0,0%	0,0%	0,0%	0,0%	0,0%
	Normandie	91,7%	8,3%	0,0%	0,0%	0,0%	91,7%	8,3%	0,0%	0,0%	0,0%	0,0%	0,0%
	Occitanie	88,4%	11,6%	0,0%	0,0%	0,0%	88,4%	11,6%	0,0%	0,0%	0,0%	0,0%	0,0%
	Provence-Alpes-Côte d'Azur	87,4%	12,6%	0,0%	0,0%	0,0%	87,4%	12,6%	0,0%	0,0%	0,0%	0,0%	0,0%
	Pays de la Loire	87,7%	12,3%	0,0%	0,0%	0,0%	87,7%	12,3%	0,0%	0,0%	0,0%	0,0%	0,0%
Hospitalisations conventionnelles avec tests RT-PCR positifs	Auvergne-Rhône-Alpes	83,0%	16,3%	0,5%	0,1%	0,1%	83,0%	12,1%	2,0%	2,3%	0,6%	0,0%	0,0%
	Bourgogne-Franche-Comté	88,1%	10,8%	0,6%	0,4%	0,0%	88,1%	7,2%	2,8%	1,5%	0,4%	0,0%	0,0%
	Bretagne	83,7%	15,4%	0,8%	0,2%	0,0%	83,7%	10,7%	3,0%	1,8%	0,8%	0,0%	0,0%
	Corse	75,2%	9,9%	4,0%	8,9%	2,0%	75,2%	0,0%	4,0%	4,0%	16,8%	0,0%	0,0%
	Centre-Val de Loire	88,6%	10,6%	0,3%	0,3%	0,2%	88,6%	5,2%	3,9%	1,3%	1,0%	0,0%	0,0%
	Grand Est	87,7%	12,3%	0,0%	0,0%	0,0%	87,7%	11,2%	0,8%	0,1%	0,2%	0,0%	0,0%
	Guadeloupe	48,1%	10,4%	6,5%	24,7%	10,4%	48,1%	0,0%	0,0%	2,6%	49,4%	0,0%	0,0%
	Guyane	69,6%	0,0%	4,3%	26,1%	0,0%	69,6%	0,0%	0,0%	0,0%	13,0%	17,4%	0,0%
	Hauts-de-France	91,4%	8,5%	0,1%	0,1%	0,0%	91,4%	7,4%	0,9%	0,2%	0,1%	0,0%	0,0%
	Île-de-France	82,6%	16,7%	0,4%	0,2%	0,0%	82,6%	13,6%	2,2%	1,3%	0,3%	0,0%	0,0%
	La Réunion	56,2%	31,0%	3,5%	9,3%	0,0%	56,2%	1,6%	7,0%	20,9%	14,3%	0,0%	0,0%
	Martinique	22,7%	9,1%	3,0%	28,8%	36,4%	22,7%	0,0%	0,0%	0,0%	77,3%	0,0%	0,0%
	Mayotte	Aucun cas identifié dans les données					Aucun cas identifié dans les données						
	Nouvelle-Aquitaine	82,0%	17,3%	0,6%	0,1%	0,0%	82,0%	15,4%	1,7%	0,9%	0,1%	0,0%	0,0%
	Normandie	90,2%	8,9%	0,6%	0,3%	0,0%	90,2%	6,1%	1,9%	0,9%	0,9%	0,0%	0,0%
	Occitanie	84,8%	14,9%	0,3%	0,1%	0,0%	84,8%	12,8%	1,8%	0,5%	0,1%	0,0%	0,0%
	Provence-Alpes-Côte d'Azur	85,3%	14,4%	0,1%	0,3%	0,0%	85,3%	11,4%	2,1%	0,6%	0,6%	0,0%	0,0%
	Pays de la Loire	83,4%	16,0%	0,5%	0,1%	0,0%	83,4%	11,9%	2,7%	1,0%	1,0%	0,0%	0,0%

Lecture : 48,1% des personnes hospitalisées en hospitalisation conventionnelle avec test RT-PCR positif en Guadeloupe sont appariées avec Vacsi. Dans la méthode antérieure au 7 avril 2022, 24,7% étaient imputés à partir des observations de la même région et de la même semaine sans critère d'âge.
Source : appariements Sidep Sivic Vacsi, traitements Drees.

Dès lors, plus le pourcentage d'observations appariées ou imputées sur la base d'un nombre important de variables est élevé, plus les données associées sont fiables. Afin d'apporter cette information aux utilisateurs, deux colonnes supplémentaires sont désormais ajoutées aux données régionales mises en ligne pour préciser le taux d'appariement des personnes testées et celui des personnes entrées en hospitalisation conventionnelle. Par exemple, lorsque les effectifs sont faibles (c'est notamment le cas en général dans les régions d'Outre-Mer et la Corse, pour les personnes décédées, ou encore en période de faible circulation épidémique), l'imputation concerne une grande part de la population hospitalisée, et elle repose sur l'observation du statut vaccinal de personnes peu nombreuses ; l'interprétation des données doit donc être faite avec précaution.