Redressements de la première vague de l'enquête EpiCov : un exemple de correction des effets de sélection dans les enquêtes multimodes

Documents de travail

N° M2023-02 – Avril 2023

Laura CASTELL
Cyril FAVRE-MARTINOZ
Nicolas PALIOD
Patrick SILLARD

Insee
Mesurer pour comprendre

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
Série des documents de travail « Méthodologie Statistique »
de la Direction de la Méthodologie et de la Coordination Statistique et Internationale

M 2023/02

Redressements de la première vague de l'enquête EpiCov : un exemple de correction des effets de sélection dans les enquêtes multimodes

Laura CASTELL
Cyril FAVRE-MARTINOZ
Nicolas PALIOD
Patrick SILLARD

Insee

Avril 2023

Remerciements : Les auteurs remercient François Beck, Gwennaëlle Brilhault, Pauline Givord, Stéphane Legleye, Amandine Schreiber, Laurent Toulemon et l'équipe EpiCov pour leurs conseils et les nombreuses discussions qui ont nourri ce document. Ils tiennent également à remercier Hélène Chaput, Emmanuel Gros et Sylvie Lagarde pour leurs relectures attentives et leurs encouragements.

Direction de la méthodologie et de la coordination statistique et internationale
Département des Méthodes Statistiques -Timbre L001 -
88 Avenue Verdier - CS 70058 - 92541 Montrouge Cedex - France -
Tél. : 33 (1) 87 69 55 00 - E-mail : DG75-L001@insee.fr - Site Web Insee : http://www.insee.fr

Redressements de la première vague de l'enquête EpiCov : un exemple de correction des effets de sélection dans les enquêtes multimodes

Laura Castell[1], Cyril Favre-Martinoz[2], Nicolas Paliod[1], Patrick Sillard[1]

Résumé

La première vague de l'enquête Épidémiologie et Conditions de vie (EpiCov) a été collectée en mai 2020 dans le contexte de la pandémie de Covid-19 pour en mesurer l'impact sanitaire et social. Cette enquête est originale à plusieurs titres : thématique, objectif de diffusion départementale, mode de collecte essentiellement auto-administrée, réalisation d'auto-prélèvements. Ce document de travail décrit les redressements réalisés par l'Insee sur la première vague de l'enquête pour s'assurer de la qualité des résultats.

Ces redressements prennent en compte les spécificités de l'enquête pour corriger le biais de non-réponse, concernant d'une part le questionnaire et d'autre part les auto-prélèvements. Cependant, ces méthodes de correction usuelles ne s'avèrent pas suffisantes pour corriger certaines variables d'intérêt de l'enquête comme les symptômes déclarés dans l'enquête qui ont affecté les répondants. En théorie, le biais observé sur ces variables peut s'expliquer par une erreur de mesure, liée à l'utilisation de plusieurs modes de collecte, ou par l'existence d'une sélection liée, de manière résiduelle, aux variables d'intérêt, non corrigée par les méthodes de correction de la non-réponse sur variables observables. On montre dans ce document que le biais procède d'une sélection liée aux variables d'intérêt et non d'une erreur de mesure associée au mode. Une méthode de correction de la sélection liée aux variables d'intérêt, basée sur un modèle de sélection d'Heckman, est mise en œuvre pour estimer sans biais les valeurs moyennes des variables d'intérêt concernées. Pour les variables de symptômes, la correction peut représenter plus de 50 % du niveau de prévalence déclaré non-corrigé. Cette correction conduit à réduire le niveau de prévalence des symptômes, conformément à l'idée selon laquelle les répondants à l'enquête semblent, en moyenne, davantage affectés par les symptômes de la maladie que les non-répondants.

Mots-clés : EpiCov, enquête, multimode, sélection endogène, modèle d'Heckman, non-réponse

Codes JEL : C24, C34, C36, C93, C83

1 Insee-Département des méthodes statistiques
2 Insee- Pôle d'ingénierie pour les enquêtes ménages (PIMEN) de la Direction régionale de la Réunion

Table des matières

Introduction

L'enquête Épidémiologie et Conditions de vie (EpiCov) a été élaborée par l'Institut national de la santé et de la recherche médicale (Inserm) et la Direction de la recherche, des études, de l'évaluation et des statistiques (DREES) du ministère des solidarités et de la santé, en collaboration avec Santé Publique France (SPF) et l'Insee, dans le contexte de la pandémie de Covid-19 (Warszawski *et alii*, 2022). Elle a un double objectif de mesure de la prévalence de la Covid-19 et d'étude de l'impact de l'épidémie et du confinement sur les conditions de vie. Menée pour la première fois en mai 2020, en période de premier confinement lié à la pandémie, cette enquête est originale sur plusieurs points. Tout d'abord, par son ampleur : sur 371 000 personnes échantillonnées, 135 000 environ ont ainsi répondu à la première vague d'interrogation de l'enquête, ce qui représente un volume de répondants[3] nettement plus important que celui des enquêtes auprès des ménages habituellement réalisées par la statistique publique. Ensuite, par son protocole : les circonstances de la collecte et les objectifs fixés sur le nombre de répondants ont conduit à privilégier un mode de collecte principal par Internet et secondaire par téléphone, sur une période de collecte courte en pleine période de restrictions sanitaires. Enfin, par le type de collecte de données : le dispositif comprend à la fois un questionnaire et un auto-prélèvement sanguin pour la réalisation de tests sérologiques.

Dans ces circonstances, un enjeu important est de pouvoir s'assurer de la qualité des résultats de cette enquête. Pour cela, un certain nombre de travaux et de redressements ont été réalisés par l'Insee. Ce sont ces travaux qui sont présentés dans ce document de travail. Ce document ne revient donc pas sur l'ensemble du processus de l'enquête mais seulement sur les traitements aval permettant de redresser les résultats de l'enquête pour fournir des estimations fiables. Par ailleurs, ces travaux portent exclusivement sur la première vague d'interrogation de l'enquête, réalisée au second trimestre 2020.

Après une présentation de l'enquête, nous décrirons les redressements mis en œuvre pour corriger de la non-réponse sur variables observables. Cette correction ne permet cependant pas d'obtenir des estimations fiables pour certaines variables d'intérêt – en l'occurrence les symptômes déclarés. Des anomalies persistent en effet dans les statistiques descriptives associées à ces variables qui suggèrent une correction de non-réponse imparfaite. Nous détaillons les deux hypothèses pouvant expliquer ces anomalies : l'existence d'un effet de mesure et l'existence d'un biais de sélection sur inobservables. Nous présentons ensuite la méthode de correction qui a été proposée pour corriger le biais de sélection inobservable, qu'un faisceau d'éléments convergents conduisent à considérer, en l'occurrence, comme l'hypothèse la plus probable. Enfin, nous revenons également sur les redressements effectués spécifiquement sur les tests sérologiques réalisés pour un sous-échantillon de l'enquête.

Nous verrons au fil de ce document de travail que ces travaux, menés spécifiquement dans le cadre de l'enquête EpiCov, ont conduit à des évolutions méthodologiques substantielles pour mieux comprendre, appréhender et corriger certains biais potentiels des enquêtes auprès des ménages, notamment des enquêtes réalisées au moins pour partie par Internet. Dans un contexte d'évolution des enquêtes vers des protocoles multimodes et d'utilisation plus massive d'Internet comme mode de collecte, les enjeux soulevés par l'enquête EpiCov ont donc une portée plus générale sur le redressement des enquêtes auprès des ménages, dont il conviendra de tirer les enseignements.

I. L'enquête EpiCov

L'enquête Epidémiologie et Conditions de vie (EpiCov) a été mise en place par l'Inserm et la Drees, avec le concours de Santé Publique France et de l'Insee, dans le contexte de la pandémie

3 Bien que le taux de réponse soit plus faible... il est ici proche de 40 % alors que le taux, pour les enquêtes de la statistique publique, dépasse généralement 60 %.

de la Covid-19. L'objectif de cette enquête est de mesurer la dynamique de l'épidémie ainsi que les conséquences de l'épidémie et du confinement sur les conditions de vie des résidents de France, à un niveau national mais également départemental. L'enquête EpiCov prévoit plusieurs vagues d'interrogation.

La collecte de la première a eu lieu du 2 mai au 2 juin 2020, à la fin du premier confinement. Une deuxième vague a été réalisée en décembre 2020, à la fin du second confinement. Une troisième a été réalisée de mi-juin à début août 2021, après le troisième confinement. Une quatrième et dernière vague a été collectée à l'été 2022. Les travaux présentés ici portent uniquement sur la première vague d'enquête. La première vague d'interrogation comprend un questionnaire d'une durée comprise entre 25 et 35 minutes ainsi que des auto-prélèvements permettant de réaliser des tests sérologiques pour un sous-échantillon de répondants.

I.1. L'échantillonnage

L'enquête EpiCov est une enquête de grande envergure par rapport aux enquêtes auprès des ménages habituellement réalisées par la statistique publique. Un échantillon de 371 000 individus est sélectionné, dont 350 000 en métropole et 21 000 dans les départements et régions d'outre-mer (Drom). Le champ de l'enquête est celui des individus âgés de 15 ans ou plus au 1er janvier 2020, résidant en France métropolitaine, en Guadeloupe, en Martinique ou à la Réunion, et résidant en logements ordinaires comme en communautés, hors Établissement d'hébergement pour personnes âgées dépendantes (Ehpad ou maisons de retraite) et prisons. Les résidents en Ehpad et maisons de retraite sont écartés du champ du fait de la situation spécifique de cette population par rapport à l'épidémie et de la difficulté à les interroger à cette période (premier confinement en métropole) avec le protocole établi. La Guyane et Mayotte sont exclus du champ du fait d'un taux de renseignement des coordonnées de contact dans la base de sondage mobilisée insuffisant pour espérer contacter suffisamment d'individus avec le protocole établi qui s'appuie fortement sur les réponses par Internet.

L'échantillon est tiré dans la base des Fichiers démographiques sur les logements et les individus (Fidéli) 2018 de l'Insee (Chevalier *et al.*, 2022). En métropole, le tirage est stratifié d'une part par département, pour répondre à un objectif de diffuser des statistiques au niveau départemental et rendre compte des disparités locales, et d'autre part par une indicatrice de pauvreté du ménage de l'individu, pour sur-représenter les ménages en dessous du seuil de pauvreté, connus pour moins participer aux enquêtes, alors même que l'étude des conditions de vie de ces derniers pendant l'épidémie de Covid-19 constitue un des intérêts majeurs de l'enquête. Ainsi, un minimum de 1 834 individus sont sélectionnés par département afin d'obtenir au moins 700 répondants par département[4], y compris pour les départements les moins peuplés, pour assurer une diffusion départementale des résultats. Par ailleurs, l'échantillon est composé de 20 % d'individus en dessous du seuil de pauvreté, alors qu'ils représentent 13 % du champ. Au sein de chaque strate de tirage, les individus sont sélectionnés par tirage systématique et préalablement triés par :

- l'aire urbaine de la commune de l'individu ;
- la tranche d'unité urbaine de la commune de l'individu ;
- l'identifiant de la commune ;
- le revenu disponible du ménage ;
- le total des revenus du ménage ;
- l'identifiant du logement ;
- l'identifiant de l'individu.

Dans les Drom, le tirage est stratifié par zone d'emploi pour prendre en compte les disparités locales vis-à-vis de l'épidémie au sein de chaque Drom. Les individus sous le seuil de pauvreté ne sont pas sur-représentés, le taux de pauvreté étant plus élevé dans ces territoires. Au sein de chaque strate de tirage, les individus sont triés par :

4 Sous une hypothèse de taux de réponse de 40 %, conservatrice par rapport aux participations constatées lors des enquêtes de la statistique publique au protocole de collecte similaire pouvant ici servir de référence, comme l'enquête TIC.

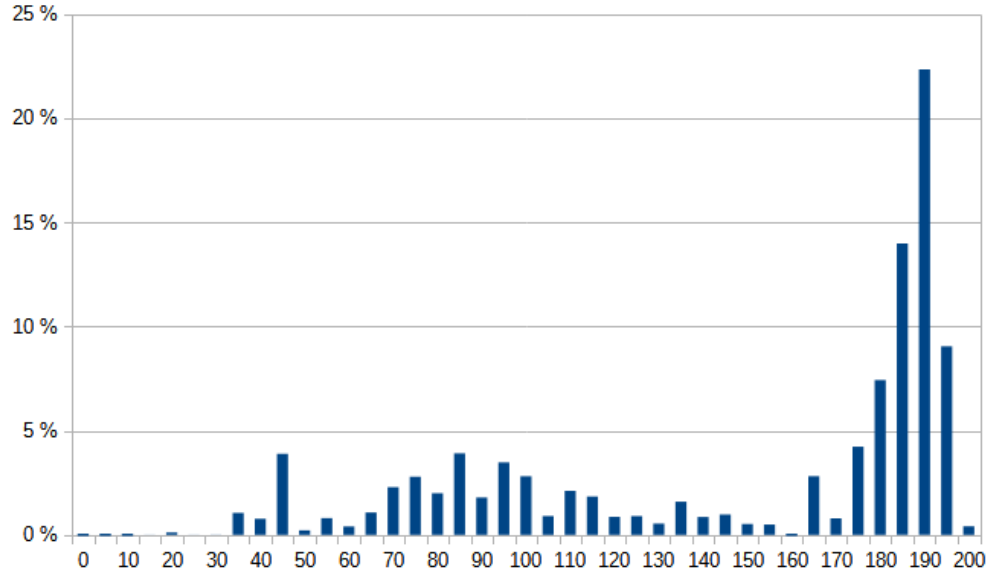
- le type de commune ;
- l'identifiant de la commune ;
- le revenu fiscal du ménage ;
- l'identifiant du logement.

Le poids de tirage de chaque individu sélectionné[5] est calculé de la façon suivante :

$$poidsTirage_{i \in S} = poidsFidéli_i * \frac{N_S}{n_S} \qquad (1)$$

avec *poidsFidéli*$_i$ le poids de l'individu *i* sélectionné dans la base Fidéli (soit 1 pour la plupart des individus et 0,5 pour les individus en résidence alternée) ; N_S le nombre d'individus dans la base de sondage appartenant à la strate *S* et n_S le nombre d'individus échantillonnés appartenant à la strate *S*.

Figure 1 : Histogramme des poids de tirage en métropole



Lecture : 2 % des individus échantillonnés en métropole ont un poids de tirage compris entre 70 et 75.
Champ : Ensemble des personnes sélectionnées dans l'échantillon métropolitain.
Source : Inserm-Drees, enquête EpiCov, vague 1.

En métropole, la dispersion des poids de tirage est de l'ordre d'un facteur 10, avec une forte concentration sur les poids élevés (figure 1). De fait, les sur-représentations qui conduisent à des poids plus faibles concernent essentiellement les individus en dessous du seuil de pauvreté résidant dans des départements peu peuplés. Or, ces individus sont minoritaires dans l'ensemble de l'échantillon. Au sein de chaque Drom, on est très proche de l'autopondération, *modulo* un facteur 2 (tableau 1) provenant des individus en résidence alternée.

5 i.e. le nombre d'individus de la population générale que représente un individu sélectionné dans l'échantillon d'enquête.

Tableau 1 : Statistiques descriptives sur les poids de tirage en métropole et dans les Drom

	Moyenne	Minimum	Maximum
Métropole	155,95	18,69	202,77
Guadeloupe	48,89	24,47	48,96
Martinique	45,36	22,69	45,42
Réunion	103,78	51,97	103,94

Lecture : un individu de l'échantillon métropolitain représente en moyenne 155,95 individus de la population générale.
Champ : Ensemble des personnes sélectionnées dans l'échantillon.
Source : Inserm-Drees, enquête EpiCov, vague 1.

L'échantillon est découpé en 20 sous-échantillons de 18 550 individus tirés aléatoirement[6]. Par la suite, on nommera ces sous-échantillons des lots. Ces lots permettent d'adapter au mieux le protocole de l'enquête en fonction des objectifs et de la charge pour les enquêtés et les enquêteurs. En particulier, ces lots permettent d'interroger une partie de l'échantillon seulement sur certains modules du questionnaire et de renforcer l'effort de collecte sur une partie de l'échantillon uniquement, le principe étant d'appliquer un protocole particulier à un lot entier (ou un groupe de lots) pour permettre une analyse d'impact statistique de ce protocole.

I.2. Le protocole

L'enquête EpiCov comprend plusieurs types de collecte de données : un questionnaire, portant sur les conditions de vie des enquêtés et de leur ménage, et un kit de prélèvement pour réaliser un test sérologique. Un appariement avec des données de santé est également prévu.

Le questionnaire principal dure environ 25 minutes. Tous les modules qui composent ce questionnaire sont proposés à l'ensemble des enquêtés[7]. Ils peuvent ainsi faire l'objet d'analyses départementales. Des modules supplémentaires[8] sont proposés à un dixième de l'échantillon. Ces modules sont considérés comme moins centraux et moins soumis à des disparités locales importantes. De fait, ces modules ne pourront faire l'objet que d'une diffusion au niveau national et d'analyses croisées moins fines que pour l'ensemble de l'échantillon. Le questionnaire comprenant ces modules supplémentaires – dit questionnaire long – dure environ 35 minutes. La collecte du questionnaire de première vague a été réalisée entre le 2 mai et le 2 juin 2020, soit durant un mois exactement.

Par ailleurs, un kit d'auto-prélèvement sanguin à réaliser à domicile sur un buvard et à renvoyer par la Poste est proposé à un sous-échantillon des répondants à l'enquête. L'objectif est de détecter la présence d'anticorps au SARS-Cov-2 pour mesurer un niveau de prévalence dans la population. Du fait des contraintes de charge des laboratoires à cette période, ce kit n'est proposé qu'à certains répondants tirés aléatoirement. Les sous-échantillons auxquels sont proposés les kits permettent une estimation de la prévalence au niveau national, ainsi qu'une estimation au niveau local pour les territoires suivants : l'Oise, le Bas-Rhin, le Haut-Rhin, Paris, la petite couronne parisienne (Hauts-de-Seine, Seine-Saint-Denis et Val-de-Marne), soit les départements les plus touchés par la pandémie lors de la première vague, pour lesquels il y avait un objectif de diffusion de la prévalence au niveau local (Warszawski *et alii*, 2020). L'échantillon permet également d'obtenir une estimation au niveau départemental pour le département des Bouches du Rhône, pour disposer de données relatives à un département peu touché par la pandémie lors de la première vague. Les enquêtés ayant accepté cet auto-prélèvement ont renvoyé leur prélèvement entre le 13 mai et le 1er juillet 2020, dont les trois quarts avant le 21 mai.

6 Cette affectation aléatoire à un sous-échantillon est réalisée en utilisant la strate et les variables de tri du tirage de l'échantillon complet. Cela permet de disposer de sous-échantillons aux caractéristiques similaires, notamment en termes de composition départementale ou de nombre de personnes sous le seuil de pauvreté.
7 Identification et lieux de vie, santé, emploi, enfants, sorties, tabac et alcool, ressenti sur la situation, opinion sur l'épidémie, parents, pratiques téléphone/internet
8 vie quotidienne, vie affective

Le mode de collecte principal choisi pour cette première interrogation est le mode Internet. Le mode téléphone est aussi utilisé mais de manière beaucoup plus restreinte que le mode internet. De fait, le mode de collecte internet est le plus adapté étant données les contraintes de coût sur un échantillon aussi important, les contraintes de collecte en période de confinement (impossibilité de se rendre sur place et disponibilité faible d'enquêteurs) et les contraintes de temps nécessitant une mise en œuvre et un traitement rapides des données.

Pour limiter les problèmes de couverture et de représentativité liés à Internet, un mode de collecte supplémentaire par entretien téléphonique avec un enquêteur est proposé à une partie de l'échantillon. Il n'est cependant pas proposé à l'ensemble de l'échantillon du fait de contraintes de coût et de disponibilité des enquêteurs. Par ailleurs, étant donnée la durée très courte fixée pour la collecte – un mois pour le questionnaire – le recours au téléphone est opéré très vite après l'invitation, voire en parallèle pour les personnes pour lesquelles on ne dispose pas d'adresse mail.

Ainsi, en métropole, trois lots sont exploités en multimode Internet-Téléphone quasiment dès le début de la collecte, c'est-à-dire en multimode de fait pratiquement concurrentiel tant le délai entre l'invitation postale à répondre sur Internet et l'appel téléphonique est bref. Un autre lot est proposé en multimode séquentiel Internet puis Téléphone, après quinze jours de collecte uniquement par Internet. À noter qu'il reste possible de répondre par Internet tout au long de la collecte et ce, dans tous les lots. Dans les Drom, davantage de lots sont proposés en multimode du fait du plus faible taux de réponse attendu par Internet dans ces territoires. Ce protocole est résumé dans le tableau 2.

Tableau 2 : Protocole de collecte selon les lots

	Type de questionnaire	Mode de collecte		
		Métropole	Martinique/ Guadeloupe	Réunion
Lot 1	Long	Internet+Téléphone concurrentiel	Internet+Téléphone	Internet+Téléphone
Lot 2	Court	Internet+Téléphone concurrentiel	Internet+Téléphone	Internet+Téléphone
Lot 3	Court	Internet+Téléphone concurrentiel	Internet+Téléphone	Internet+Téléphone
Lot 4	Court	Internet+Téléphone séquentiel(*)	Internet+Téléphone	Internet+Téléphone
Lot 5	Court	Internet	Internet+Téléphone	Internet+Téléphone
Lot 6	Court	Internet	Internet+Téléphone	Internet+Téléphone
Lot 7	Court	Internet	Internet+Téléphone	Internet+Téléphone
Lot 8	Court	Internet	Internet	Internet+Téléphone
Lot 9	Court	Internet	Internet	Internet+Téléphone
Lot 10	Court	Internet	Internet	Internet
Lot 11	Court	Internet	Internet	Internet
Lot 12	Court	Internet	Internet	Internet
Lot 13	Court	Internet	Internet	Internet
Lot 14	Court	Internet	Internet	Internet
Lot 15	Court	Internet	Internet	Internet
Lot 16	Court	Internet	Internet	Internet
Lot 17	Court	Internet	Internet	Internet
Lot 18	Court	Internet	Internet	Internet
Lot 19	Court	Internet	Internet	Internet
Lot 20	Long	Internet	Internet	Internet

(*) : la collecte téléphone débute 15 jours après le début de la collecte internet pour ce lot (4). Pour les autres lots (1-3) dans lequel le téléphone est utilisé, les débuts de collecte sont concomitants pour l'ensemble des modes.

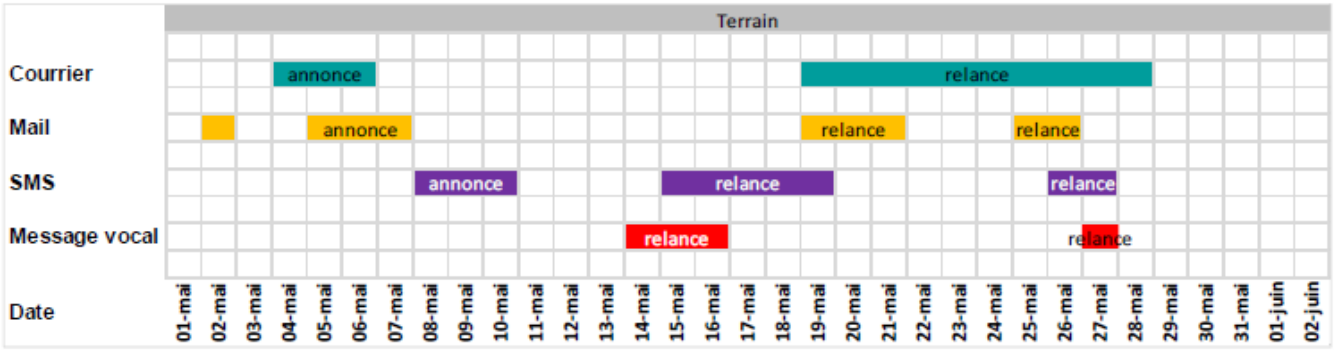
Quel que soit le protocole proposé, tous les individus reçoivent une lettre-avis par courrier avec les identifiants de connexion au questionnaire de l'enquête. Une plateforme dédiée ainsi qu'une hotline sont disponibles pour la prise de rendez-vous téléphonique pour les enquêtés des lots concernés.

Tous les moyens de communication disponibles sont utilisés pour maximiser les chances de contact (figure 2). Ainsi, en plus du courrier d'annonce, un mail-avis ainsi qu'un SMS[9] sont envoyés pour annoncer l'enquête lorsque les coordonnées nécessaires sont disponibles. Concernant les relances, des mails, des SMS ainsi que des messages vocaux déposés sur la messagerie sont envoyés lorsque les coordonnées nécessaires sont disponibles. Un courrier de relance est également envoyé à l'ensemble des non-répondants. Les dates des relances sont

9 Les coordonnées utilisées sont celles de la personne sélectionnée si elles sont disponibles dans les bases fiscales. A défaut, on utilise celles de la personne de référence du ménage. Il est possible qu'hormis l'adresse postale, il n'y ait pas de coordonnées mail ou téléphone. Dans ce cas, seul le courrier d'annonce est envoyé. Ce courrier contient des éléments pour répondre par internet.

extrêmement rapprochées par rapport aux écarts usuels entre les relances des enquêtes de la statistique publique, puisque les enquêtés n'ont qu'un mois pour répondre à l'enquête EpiCov.

Figure 2 : Calendrier des annonces et relances selon le support



Les chances de contact et l'intensité des relances sont donc fortement liées aux coordonnées disponibles. Dans Fidéli, des coordonnées téléphoniques et adresses mails sont renseignées, en plus des coordonnées postales disponibles pour l'ensemble de l'échantillon. Pour l'échantillon de l'enquête EpiCov, on dispose pour la majorité des enquêtés à la fois d'une adresse mail et d'au moins un numéro de téléphone (tableau 3). Pour 11 % des enquêtés, on dispose uniquement d'une adresse mail et pour 12 % des enquêtés, uniquement d'un numéro de téléphone. Au final, on ne dispose d'aucunes coordonnées – hors coordonnées postales – pour 18 % des enquêtés. Un enrichissement des numéros de téléphone est par ailleurs réalisé par Ipsos, le prestataire en charge de la collecte. Cet enrichissement permet d'obtenir un numéro de téléphone pour 10 % des individus, en plus d'enrichir les numéros de téléphone pour les individus pour lesquels on en dispose déjà d'au moins un. Après cet enrichissement, l'échantillon mis en exploitation ne contient pas de coordonnées (mail ou téléphone) pour seulement 11 % des enquêtés. Et 70 % des enquêtés disposent d'une adresse mail.

Tableau 3 : Coordonnées disponibles pour les individus sélectionnés dans l'échantillon

	Fidéli	Après enrichissement
Adresse mail uniquement	11 %	7 %
Téléphone uniquement	12 %	18 %
Adresse mail et téléphone	59 %	63 %
Ni adresse mail ni téléphone	18 %	11 %
Total	100 %	100 %

Lecture : 11 % des individus présents dans la base de sondage Fidéli ont une adresse mail et pas de coordonnées téléphoniques indiquées dans la base.
Champ : Ensemble des personnes sélectionnées dans l'échantillon.
Source : Inserm-Drees, enquête EpiCov, vague 1.

II. Mise en œuvre et résultats des redressements corrigeant de la non-réponse sur observables

Dans cette partie, nous allons présenter les résultats de la collecte et la méthode de redressement de la non-réponse totale au questionnaire de l'enquête, appliquée par défaut[10]. Les redressements spécifiques à l'analyse des tests sérologiques sont présentés à part dans la partie V.

Comme indiqué précédemment, l'enquête EpiCov comprend un questionnaire dit « court » et un questionnaire dit « long » proposé à un dixième de l'échantillon (voir tableau 2). L'existence de ces deux types de questionnement suppose de construire deux jeux de poids distincts : un poids pour l'usage des modules du questionnaire court, disponible pour l'ensemble des répondants ; un poids pour l'usage spécifique des modules du questionnaire long, disponible pour les seuls répondants au questionnaire « long ». La construction de deux modèles est liée à la différence de structure des deux types d'échantillon : la part des lots multimodes et donc des individus répondants par téléphone est plus importante pour le questionnaire long que pour l'ensemble des questionnaires. Par ailleurs, les niveaux de diffusion diffèrent : pour le questionnaire court, il est nécessaire d'avoir recours à une modélisation de la non-réponse très fine, car la diffusion au niveau départemental implique l'utilisation de variables de niveau département dans les modèles, ce qui est rendu possible par le volume important de l'échantillon.

Les méthodes de correction sont cependant similaires pour ces deux jeux de pondération. De fait, il n'y a pas vraiment de raison que la liste des déterminants de la réponse pour les lots 1 et 20 diffère de celle des autres lots[11]. Il a donc été décidé de fournir dans un premier temps des pondérations pour l'ensemble de l'échantillon afin d'être en mesure d'exploiter le questionnaire principal, puis de réaliser un nouveau modèle avec les mêmes variables pour les pondérations du questionnaire long[12].

II.1. Identification du statut de réponse

Dans un premier temps, il est nécessaire de définir le statut de réponse de l'ensemble des individus échantillonnés. Pour réaliser la correction de la non-réponse, il est important d'identifier les répondants, les non-répondants partiels (le questionnaire est globalement exploitable, mais toutes les questions ne sont pas renseignées), les non-répondants totaux (aucune information n'est collectée auprès de l'enquêté ou les informations collectées ne suffisent pas à rendre ses réponses exploitables) et les hors champ. Les individus avec un statut de réponse inconnu – c'est-à-dire qui sont soit hors champ soit non-répondants totaux, sans qu'on puisse le déterminer – sont classés ici comme non-répondants totaux, ce qui conduit à sous-estimer la part des hors champ dans l'échantillon.

Le mode de collecte essentiellement auto-administré a deux conséquences importantes :

- Dès que les individus ne répondent pas, il est complexe de faire la différence entre une sortie de champ (décès, départ à l'étranger par exemple) et une non-réponse (individu dans le champ mais injoignable, refus, etc.). La principale piste pour faire cette distinction est l'utilisation de sources administratives ;
- Il y a peu de contrôles sur le fait que la personne répondante à l'enquête est bien la personne du ménage qui a été échantillonnée. C'est tout particulièrement vrai pour les individus ayant moins de 20 ans dans la base de sondage (c'est-à-dire au 1er janvier 2018). Pour un nombre non négligeable d'entre eux, on dispose uniquement de leur année de

10 Cette méthode, qui correspond aux usages habituels dans la statistique publique, est ensuite discutée en partie III de ce document.

11 De part l'usage de modes de collecte différents, les déterminants de la réponse aux lots 1 à 4 peuvent différer de celle des autres lots. Ces lots étant utilisés, au moins en partie, pour le questionnaire court et pour le questionnaire long, la liste des déterminants n'a pas de raison de différer.

12 La seule raison qui aurait pu conduire à utiliser d'autres variables que celles mobilisées pour la correction de la non-réponse du questionnaire court serait que le nombre d'abandons dans le questionnaire long soit non négligeable par rapport au questionnaire court, ce qui n'a pas été observé.

naissance dans la base de sondage et non de leurs prénom et nom. Les lettres-avis et mails sont alors destinés à un des référents fiscaux (dans la grande majorité des cas, un parent) en précisant que la personne sélectionnée dans l'enquête est l'habitant du ménage fiscal qui est né une certaine année. Ce protocole génère évidemment un risque d'erreur : la personne répondante n'est pas nécessairement celle sélectionnée. Pour s'en prémunir, deux options sont possibles : inclure une vérification sur l'identité du répondant en début de questionnaire ; vérifier au moment des redressements que l'individu répondant est bien celui échantillonné. Pour mettre en œuvre cette dernière option, il convient de récupérer dans le questionnaire certaines informations sur l'individu échantillonné dont on dispose aussi dans la base de sondage (en l'occurrence, le sexe et la date de naissance). C'est ce qui a été fait (cf. *infra*).

Les hors champ doivent être exclus du modèle de correction de la non-réponse puisqu'ils ne font pas partie de la population cible sur laquelle l'enquête vise à être représentative. Plusieurs sources d'informations peuvent être utilisées pour identifier les hors champ. Un appariement avec Fidéli 2019 a été réalisé pour mettre à jour les informations de la base de sondage Fidéli 2018 dans laquelle l'échantillon a été sélectionné. Ainsi, les individus dans les situations suivantes en 2019 sont considérés comme hors champ :

- les individus décédés en 2018 ;
- les individus vivant à l'étranger, en Guyane ou à Mayotte dans Fidéli 2019 et n'ayant pas répondu à l'enquête ;
- les individus de 60 ans ou plus vivant dans une communauté de personnes âgées (EHPAD ou maison de retraite) ;
- les individus de 60 ans ou plus vivant à une adresse avec un complément d'adresse contenant « maison de retraite » ou « EHPAD » ;
- les individus vivant en prison.

Des informations remontées par la hotline durant la collecte permettent également d'identifier des individus décédés et des individus vivant en maison de retraite ou en EHPAD au moment de l'enquête. Enfin, dans l'enquête elle-même, les premières questions permettent d'identifier certaines situations de hors champ comme les individus résidant à l'étranger avant le confinement et les individus résidant en maison de retraite ou en EHPAD. Au total, 5 553 individus sont classés hors champ et mis de côté des analyses ultérieures.

Parmi les répondants à l'enquête, certains contrôles sont réalisés pour s'assurer de la qualité du questionnaire et éventuellement écarter des questionnaires inexploitables. Tout d'abord, il convient de vérifier l'identité du répondant pour s'assurer qu'il s'agit bien de l'individu échantillonné. Pour cela, la date de naissance renseignée dans l'enquête est comparée à celle indiquée dans la base de sondage. Si la date de naissance n'est pas celle de l'individu échantillonné mais correspond ou est proche de la date de naissance d'un des référents fiscaux du ménage, l'individu échantillonné est considéré comme non-répondant car ne correspondant pas à l'individu réellement enquêté[13]. Ce contrôle conduit à classer 915 individus non-répondants. Dans la plupart de ces cas (707 individus), il s'agit d'individus dont on ne dispose que de l'année de naissance comme information identifiante dans la base de sondage, alors même que ces cas sont minoritaires (5 % de l'échantillon). Par ailleurs, pour près de 1 400 autres individus, les informations de l'enquête (date de naissance et sexe) utilisées pour vérifier l'identité du répondant diffèrent de celles de l'individu échantillonné et des référents fiscaux renseignées dans la base de sondage. Dans la moitié des situations environ, le sexe renseigné dans l'enquête est différent de celui de la base de sondage. On estime donc également que ces quelques 1 400 individus sont non-répondants. En tout, 2 350 individus ont été classés non-répondants, car l'individu répondant n'est pas celui échantillonné.

13 Peu d'individus s'avèrent être dans cette situation. C'est probablement lié au fait qu'une vérification a été intégrée dès la passation du questionnaire pour s'assurer que la personne répondante à l'enquête avait bien la même année de naissance que l'individu échantillonné. Lorsque ce n'était pas le cas, la personne pouvait tout de même répondre à l'enquête mais en ayant confirmé au préalable qu'elle était bien la personne ciblée.

Certains questionnaires sont également considérés comme inexploitables non pas car l'enquêté n'est pas l'individu recherché mais car le contenu du questionnaire est insuffisamment renseigné, et donc insuffisamment informatif. Pour différencier les non-répondants partiels avec un questionnaire exploitable bien qu'incomplet de ceux avec un questionnaire inexploitable, il convient de définir des critères de choix. Pour cela, sept questions d'intérêt principal pour l'enquête ont été définies par la DREES et l'Inserm. Ainsi, tous les enquêtés ayant répondu à ces sept questions sont considérés comme répondants à l'enquête avec un questionnaire exploitable. Cela concerne 131 431 enquêtés. Par ailleurs, les enquêtés avec une ou deux non-réponses à ces sept questions mais qui sont au moins allés jusqu'à la fin du module B, qui constitue le module principal du questionnaire portant sur la santé, et ayant renseigné la moitié des questions portant sur les symptômes, sont également considérés comme répondants. Cela concerne 2 960 enquêtés. Les enquêtés ne répondant pas à l'une de ces deux situations sont considérés comme des non-répondants totaux ; leur questionnaire n'est pas du tout exploité.

Au final, 134 391 individus sont considérés comme dans le champ de l'enquête et répondants à la première vague de l'enquête EpiCov, soit un taux de réponse de 36,8 % en moyenne. Ce taux est significativement supérieur dans les lots multimodes – 46,0 % – que dans les lots monomodes – 34,5 % (figure 3).

Figure 3 : Taux de réponse par lots

Lecture : Le taux de réponse dans le lot 1 est de 45 %. La part des répondants au téléphone est de 20 %, celle des répondants par Internet est de 25 %.
Champ : Ensemble des personnes sélectionnées dans l'échantillon non identifiées comme hors champ.
Source : Inserm-Drees, enquête EpiCov, vague 1.

II.2. Méthode de correction

Après le plan de sondage, tous les individus de l'échantillon ont un poids de tirage, défini dans la partie précédente, qui est l'inverse de la probabilité d'inclusion de l'individu dans l'échantillon. Cependant, tous les individus de l'échantillon ne répondent pas à l'enquête. L'objectif est de trouver une pondération qui permette d'estimer sans biais les différentes variables de l'enquête sur l'ensemble de la population d'enquête à partir des réponses des seuls répondants. La correction

de la non-réponse suppose que tous les individus de l'échantillon ont une probabilité non nulle $p_{rép}$ de répondre à l'enquête qui constitue une étape d'auto-sélection, en plus de l'étape de sélection de l'échantillon. Or cette probabilité n'est pas connue et doit être estimée.

Pour estimer cette probabilité, les méthodes mises en œuvre ici sont relativement classiques en se basant sur une modélisation de la probabilité de répondre par régression logistique puis en appliquant des groupes de réponse homogène.

La modélisation de la probabilité de répondre à l'enquête est réalisée sur l'ensemble des individus de l'échantillon, après suppression des hors champ identifiés dans la partie précédente. Les variables à intégrer dans le modèle sont celles qui sont à la fois liées aux variables d'intérêt de l'enquête et au fait de répondre ou non à l'enquête. Du fait qu'on modélise la réponse pour l'ensemble des individus de l'échantillon, les variables disponibles sont les variables issues de la base de sondage Fidéli ou des variables géolocalisées. Dans la plupart des enquêtes auprès des ménages, notamment les enquêtes récurrentes, les variables prises en compte sont définies à partir des connaissances acquises des enquêtes précédentes ou de la littérature. Il s'agit en général de variables sociodémographiques classiques comme l'âge, le sexe, de variables caractérisant le lieu de résidence, le niveau de diplôme, le niveau de vie, la catégorie socio-professionnelle, le statut d'activité, le nombre d'individus dans le logement, etc. L'enjeu dans le choix de ces variables est de bien les calibrer : si on ne prend pas en compte toutes les variables corrélées à la fois aux variables d'intérêt et à la participation, les conditions sous lesquelles l'estimateur d'Horvitz-Thompson est convergent (indépendance conditionnelle de la variable d'intérêt et de la participation) ne sont pas vérifiées ; si on prend en compte des variables corrélées à la participation mais pas aux variables d'intérêt, le modèle sur-apprend et la variance des estimations est accrue sans gain en termes de correction de biais.

Dans le cas d'EpiCov, un certain nombre de variables sont envisagées *a priori* : des variables de contact, des variables de conditions de vie, des variables liées à la disponibilité et/ou l'activité de l'individu, des variables liées à la joignabilité des individus, des variables liées à la diffusion de l'enquête (département par exemple), des variables externes en lien avec le contexte épidémique local. Étant donnée la diversité des variables potentielles et la nouveauté de la thématique, le choix des variables à prendre en compte a été fait de manière agnostique, en étudiant au préalable la corrélation des variables disponibles avec les variables d'intérêt de l'enquête. Pour cela, plusieurs variables d'intérêt ont été définies comme essentielles par les concepteurs d'enquête :

- l'état de santé général ;
- les symptômes déclarés ;
- l'existence et le résultat éventuel d'un test lié à la Covid-19 ;
- le fait de télétravailler à temps complet les 7 derniers jours ;
- la fréquence de sortie du domicile dans les 7 derniers jours.

Une première étape consiste à conserver toutes les variables auxiliaires dont le V de Cramer[14] avec au moins une des variables d'intérêt principal et pour au moins un département[15] est supérieur à 0,1 pour les variables dichotomiques et 0,05 pour les autres. Environ 90 variables se trouvent dans ce cas sur plus de 200 variables disponibles dans Fidéli.

Une deuxième étape consiste à réaliser une sélection des variables à utiliser dans le modèle de non-réponse par validation croisée en mobilisant plusieurs critères de décision :

- la qualité de la prédiction par le jeu de variables mobilisées pour la correction de la non-réponse des variables d'intérêt principales à des niveaux départementaux.
- la qualité de la prédiction de la non-réponse mesurée par l'écart (en valeur absolue, quadratique, pondéré…) observé entre la probabilité de réponse prédite par le jeu de variables retenu et le comportement de réponse réel (répondant, non-répondant). De

14 Statistique d'association entre deux variables fondée sur la statistique du Chi-2 (Cramer, 1946 -chap. 21).

15 Dans le cas de la modélisation de la non-réponse au questionnaire « long », les corrélations sont étudiées au niveau national puisqu'aucune représentation départementale n'est souhaitée.

nombreux jeux de variables ont ainsi été testés pour la correction de la non-réponse, en retirant progressivement les variables les moins corrélées au fait de répondre ou non à l'enquête. On arrête la sélection de variables lorsque retirer des variables supplémentaires conduit à moins bien prédire à la fois une variable de l'enquête et le fait de répondre ou non à l'enquête. Les variables sélectionnées pour le modèle de correction de la non-réponse du questionnaire principal sont décrites en annexe (annexe A1) dans la colonne « modèle logistique ».

Le processus de validation croisée est réalisé en apprenant sur 4/5 de l'échantillon tiré aléatoirement et en testant le cinquième de l'échantillon restant. La probabilité de réponse est ensuite estimée par régression logistique à partir du modèle retenu.

Une fois ces probabilités de réponse estimées pour l'ensemble des individus de l'échantillon, des groupes de réponse homogène (GRH) sont identifiés en regroupant les individus qui ont une probabilité de réponse, et donc un comportement de réponse, proche. Ces regroupements sont effectués séparément pour chaque département, en utilisant la méthode d'Haziza et Beaumont (Haziza et Beaumont, 2007) en appliquant un algorithme de centres mobiles. Les individus appartenant à un groupe se voient appliquer comme probabilité de réponse le taux d'individus répondants au sein de ce groupe. L'application de GRH permet de diminuer la variance des estimations et de rendre la modélisation du comportement de réponse moins paramétrique. Le nombre de GRH dépend de la taille de l'échantillon dans le département et varie entre 4 (Corrèze, Corse du Sud, Loire, Nièvre, Sarthe, Tarn-et-Garonne, Vaucluse) et 12 (La Réunion) pour le questionnaire « court ». Ce nombre de groupes par département est fixé par plusieurs critères : chaque GRH doit contenir au moins 100 individus échantillonnés pour être suffisamment robuste ; conventionnellement, la constitution des GRH doit aboutir à au moins 5 GRH par département afin de ne pas sous-apprendre ; la part de variance de la probabilité de réponse expliquée par les GRH doit être la plus proche possible de 99 % pour ne pas sous-apprendre[16] ; les probabilités de réponse issues des GRH doivent être ordonnées de la même manière que les probabilités issues du modèle logit afin de ne pas sur-apprendre. Pour limiter le nombre d'individus avec un poids trop élevé, les groupes avec une probabilité moyenne de réponse inférieure à 10 % sont ensuite fusionnés avec les groupes ayant la probabilité la plus proche, jusqu'à ce que tous les GRH aient une probabilité moyenne de réponse supérieure à 10 %[17]. Ce choix est lié à un arbitrage biais-variance : pour éviter d'avoir des unités influentes (unités répondantes avec un poids trop élevé), on préfère effectuer ces regroupements manuels, quitte à réintroduire un peu de biais, en n'acceptant pas les probabilités de réponse trop basses.

Un des éléments de choix importants sur la méthode de correction de la non-réponse concerne le niveau de regroupement élémentaire sur lequel porte la correction. Des estimations des variables issues du questionnaire « court » sont souhaitées au niveau départemental[18]. La modélisation de la non-réponse comme l'application des GRH peut alors se faire au niveau départemental ou au niveau national. En ce qui concerne la modélisation de la non-réponse, réaliser des modèles distincts par département a un intérêt si les variables auxiliaires ont des effets différents sur la participation selon les départements et que ces effets sont corrélés aux variables d'intérêt. Si ce n'est pas le cas, les modèles vont conduire à un sur-apprentissage et une augmentation de la variance des estimations sans que ces différences ne reflètent des différences réelles entre départements. Le choix de modéliser la non-réponse au niveau départemental est écarté car il conduit à des probabilités de réponse très faibles et un sur-apprentissage très net. Le choix a donc été fait de réaliser un modèle logistique unique au niveau national mais en prenant en compte l'ensemble des variables corrélées aux variables d'intérêt dans au moins un département.

16 Elle est supérieure à 95 % pour tous les départements.

17 La seule exception concerne la Martinique et la Guadeloupe, dont les taux de réponse sont plus faibles qu'en France métropolitaine ou à La Réunion, ce qui conduit à accepter des probabilités de réponse plus faibles dans les GRH.

18 Pour le questionnaire « long », il n'y a pas d'objectif de diffusion départementale. Le modèle ainsi que l'application des GRH sont réalisés au niveau national.

Les GRH sont ensuite appliqués au niveau départemental. De fait, il est fort possible que les variables d'intérêt comme la propension à participer à l'enquête varient selon les départements. Dans ce cas, l'application de GRH par département permet de limiter les biais au niveau des estimations départementales. De fait, si les profils de répondants diffèrent entre départements, des regroupements départementaux d'individus permettent de tenir compte des spécificités de chaque département. À l'inverse, des regroupements au niveau national vont conduire à des estimations biaisées au niveau départemental. Dans le cas où les profils de répondants seraient similaires entre départements, le nombre d'individus enquêtés par département permet de constituer suffisamment de GRH par département pour ne pas sous-apprendre et d'inclure suffisamment d'individus dans chaque GRH pour éviter de sur-apprendre. On ne perd donc rien à constituer des GRH par départements dans l'enquête EpiCov plutôt qu'au niveau national.

Une fois ces probabilités de réponse obtenues, un poids intermédiaire est calculé pour l'ensemble des individus de l'échantillon en divisant le poids de tirage par la probabilité de réponse estimée. Une dernière étape de calage sur marges est alors réalisée pour l'ensemble des répondants à l'enquête. L'objectif du calage est quadruple : assurer la cohérence avec d'autres sources de référence ; gagner en précision pour les variables d'intérêt liées aux variables de calage ; réduire l'impact de défauts de couverture liés à l'impossibilité d'atteindre les entrées de champ (individus dans le champ de l'enquête mais qui ne sont pas présents dans la base de sondage) ; corriger de certains biais résiduels liés à la non-réponse.

Les marges de calage utilisées dans le cas d'EpiCov sont relativement classiques et répondent aux objectifs de diffusion départementale. En métropole, les marges utilisées sont la pyramide des âges[19] par sexe au niveau départemental et le niveau de diplôme[20] au niveau régional. Pour chacun des Drom (Réunion, Martinique, Guadeloupe), les marges utilisées sont la pyramide des âges par sexe, le niveau de diplôme et le lieu de naissance[21]. Ces marges sont issues des structures du dernier recensement de la population 2017[22] appliquées aux projections départementales de population au 1er janvier 2020, tout en appliquant un traitement pour retirer les individus en maison de retraite, en Ehpad ou en prison.

Pour réaliser ce calage, des redressements sur les variables déclarées dans l'enquête ont été nécessaires. Pour le département de résidence déclaré dans l'enquête, des redressements ont été réalisés pour corriger un certain nombre d'erreurs liées à l'auto-complétion dans le questionnaire et à des déclarations erronées de départ à l'étranger ou dans les COM. Pour le lieu de naissance des individus résidant dans les DOM, des corrections ont été faites lorsque les individus déclaraient par erreur être né dans une COM plutôt qu'un DOM. Pour les mois et année de naissance, les déclarations aberrantes ont été corrigées mais en cas de dissonance avec l'information disponible dans Fidéli, la variable déclarée a été privilégiée. Pour le diplôme, les valeurs manquantes ont dû être imputées par arbre de décision.

Une fois cette dernière étape de calage sur marges réalisée, un nouveau jeu de poids, cette fois définitif, est alors disponible pour l'ensemble des répondants à l'enquête. Pour toute la suite du document, nous nommerons cette pondération issue du modèle de correction de la non-réponse totale sur observables et du calage comme étant la « pondération sur observables ».

La variance des estimations à partir de cette pondération est estimée par la macro SAS %Everest développée par la division Sondages de l'Insee. La variance des estimations de moyennes de plusieurs variables d'enquête (voir tableau 4) est calculée sous les trois hypothèses suivantes :

(a) sondage aléatoire simple stratifié ;

19 L'âge est regroupé par tranche de 5 ou 10 ans en fonction de la taille des départements.

20 Le niveau de diplôme est regroupé en 4 modalités : non diplômés et CAP-BEP ; niveau Bac ; niveau Bac+2 ; Bac+3 ou plus. Les marges sont données pour les seuls individus âgés de 35 ans ou plus, car les risques de réponses divergentes entre le recensement de la population et l'enquête EpiCov ne sont pas négligeables lorsque les individus sont en cours de formation.

21 Le lieu de naissance est regroupé en 3 modalités : métropole ; DOM ; étranger.

22 Il s'agit du dernier recensement disponible au moment des redressements de la première vague d'EpiCov.

(b) correction de la non-réponse au sein de groupes de réponse homogène ;

(c) calage sur marges.

Les hypothèses (b) et (c) correspondent au cas d'EpiCov. L'hypothèse (a) en revanche n'est pas rigoureusement vérifiée. Cependant, le tirage de l'enquête EpiCov est proche d'un sondage aléatoire simple stratifié par département car le tirage mis en œuvre est un tirage systématique stratifié sur une base triée. Le tirage systématique réalisé introduit une pseudo-stratification complémentaire au sein de chaque strate départementale qui le rend *a priori* plus efficace en termes de précision que le sondage aléatoire simple départemental. On peut donc considérer que la variance présentée ici et calculée à partir de la macro %Everest surestime légèrement la vraie variance. Quelques exemples d'estimation de variance sont donnés dans le tableau 4.

Globalement, on observe un « design-effect » —ou effet de plan— variant entre 1,1 et 1,3 selon les variables d'intérêt considérées ici. Cela signifie que la variance de l'enquête EpiCov pour l'estimation d'une proportion au niveau national est entre 1,1 et 1,3 fois plus grande que la variance issue d'un plan de sondage dans lequel les répondants auraient été sélectionnés selon un sondage aléatoire simple au plan national. Ce « design-effect » qui porte sur les deux étapes de sélection successives (plan de sondage et participation) est supérieur à 1. Il résulte d'effets se rapportant à l'étape de sélection du plan et de correction de non-réponse qui se compensent en partie :

- Étape de sélection du plan : le plan de sondage de l'enquête EpiCov a été conçu pour permettre des exploitations départementales et donc il a fallu s'éloigner des allocations proportionnelles par strates qui auraient été plus efficaces pour l'estimation d'un total quelconque au niveau national. Cet effet se traduit par un « design-effect » supérieur à 1 par rapport à un sondage aléatoire simple.
- Étape de correction de non-réponse : par rapport à une correction uniforme de la non-réponse qui correspondrait à une sélection par sondage aléatoire simple des répondants parmi les enquêtés, les GRH et le calage sur marge améliorent la précision, ce qui réduit le « design-effect »

Tableau 4 : Estimation de variance

Variable d'intérêt	Type de question-naire	Estimation ponctuelle	Sondage aléatoire simple sans remise (SAS)		Macro %Everest[23]		Effet de plan Everest/SAS
			Ecart-type	Demi-longueur	Ecart-type	Demi-longueur	
Très bon état de santé général	court	0,3253	0,0013	0,0025	0,0014	0,0028	1,11
	long	0,3304	0,0039	0,0077	0,0043	0,0084	1,09
Toux	court	0,0817	0,0007	0,0015	0,0009	0,0017	1,17
	long	0,0824	0,0023	0,0045	0,0026	0,0051	1,14
Aucune sortie du domicile	court	0,0952	0,0008	0,0016	0,0009	0,0018	1,13
	long	0,1026	0,0025	0,0050	0,0032	0,0064	1,28

Lecture : 32,53 % des individus se déclarent en très bon état de santé général. L'intervalle de confiance à 95 % de cette estimation est [32,28 ; 32,78]. La colonne « Demi-longueur » désigne la demi-longueur de l'intervalle de confiance à 95 %.
Note : moyennes et écarts-types en niveau pour des variables binaires (dont les modalités individuelles possibles sont 0 ou 1).
Source : Inserm-Drees, enquête EpiCov, vague 1.

23 Les variables d'intérêt étant des proportions, il est nécessaire d'estimer les résidus de calage mobilisés dans le calcul de la variance. Ici, sont présentés les résultats utilisant une estimation basée sur des résidus issus d'un modèle linéaire. Une autre estimation faisant intervenir les résidus d'un modèle GLM de type logit donne des résultats similaires.

II.3. Caractéristiques des poids obtenus après correction de la non-réponse

Dans cette partie, nous revenons sur quelques caractéristiques des pondérations obtenues pour évaluer leur qualité.

Tout d'abord, on étudie le rapport entre le poids calé final et le poids de tirage pour évaluer la déformation de la distribution liée à la correction de la non-réponse. On constate que 90 % des rapports de poids calés sur les poids de tirage restent contenus, car inférieurs à un facteur 8 (tableau 5)[24]. Néanmoins, on observe des rapports de poids maximum élevés. En moyenne, sur l'ensemble des départements pour le questionnaire « court », le rapport de poids maximal est de 13,2, avec un maximum de 24,9 pour le département de l'Aude. Pour le questionnaire « long », le rapport de poids maximal sur l'ensemble de l'échantillon est de 11.

Tableau 5 : Rapport poids calés sur poids de tirage

	Min	0,10	0,25	0,5	0,75	0,9	Max
Questionnaire court (national)	0,7	1,3	1,5	2,0	2,9	5,4	24,9
Questionnaire court (moyenne départementale)	1,0	1,3	1,6	2,1	3,0	5,6	13,2
Questionnaire long (national)	1,0	1,3	1,5	1,9	2,7	4,6	11,0

Champ : ensemble des personnes considérées comme répondantes à l'enquête.
Source : Inserm-Drees, enquête EpiCov, vague 1.

Un indicateur permettant d'alerter sur la présence de valeurs influentes dans les estimations est également fourni. Une valeur est considérée comme influente si elle a un impact significatif sur l'erreur quadratique moyenne de l'estimateur considéré. Une façon classique de déterminer l'influence d'une unité de l'échantillon est de calculer son biais conditionnel, qui est la mesure d'influence mobilisée en théorie des sondages. Ici, comme les variables de l'enquête sont très majoritairement des variables qualitatives, on présente un proxy de ce biais conditionnel en regardant la contribution des unités à l'estimation selon leur place dans la distribution des poids. Plus précisément, on va étudier la contribution[25] des X % des individus avec les poids les plus élevés dans les estimations (tableau 6). On constate que 25 % des individus ayant les poids les plus élevés pèsent pour environ la moitié des estimations, ce qui semble tout à fait raisonnable. Mais on constate par ailleurs que 5 % des répondants avec les poids les plus élevés pèsent pour 17,2 % des estimations. Ce résultat incite à une certaine prudence dans les interprétations et les comparaisons départementales notamment.

24 Ces rapports de poids élevés, pour une enquête de la statistique publique, s'expliquent par un taux de réponse plus faible que les enquêtes usuelles de la statistique publique. Cela a impliqué la constitution de GRH dont la probabilité de réponse s'élève à 10 % environ, ce qui conduit à multiplier par 10 les poids de tirage à l'issue de l'étape de correction de la non-réponse.

25 On appelle ici contribution le ratio exprimé en pourcentage entre la somme des poids des X % des individus avec les poids les plus élevés sur la somme de l'ensemble des poids des individus de l'échantillon des répondants.

Tableau 6 : Pourcentage de l'estimation porté par les individus en fonction de leur quantile de poids

Quantile →	0 %	10 %	25 %	50 %	75 %	90 %	95 %	100 %
Questionnaire court	100 %	96,7 %	88,7 %	71,2 %	47,9 %	27,6 %	17,2 %	0 %
Questionnaire long	100 %	96,7 %	88,3 %	69,9 %	45,9 %	25,9 %	16 %	0 %

Lecture : Tous les individus ayant un poids supérieur ou égale au quantile à 25 % des poids jouent pour 88,7 % de l'estimation dans le questionnaire court. Autrement dit, 75 % des individus influent sur 88,7 % des totaux du questionnaire court.
Champ : ensemble des personnes considérées comme répondantes à l'enquête.
Source : Inserm-Drees, enquête EpiCov, vague 1.

L'utilisation de la pondération du questionnaire « court » pour des estimations départementales nécessite une certaine vigilance. De fait, certains départements ont peu de répondants. C'est par exemple le cas des deux départements de Corse, qu'il convient de regrouper pour gagner en robustesse dans les estimations. Par ailleurs, les taux de réponse dans les deux départements des Antilles ont été plus faibles qu'ailleurs. Par conséquent, des seuils plus faibles dans les taux de réponse des GRH ont été autorisés pour éviter de trop biaiser les estimations. Cependant, cela conduit à des rapports élevés de poids calés sur poids de tirage. Une vigilance accrue doit être également portée à ces départements. Sur ces quatre départements, et sur d'autres départements avec peu de répondants, les croisements et les estimations sur domaines doivent être réalisées avec prudence.

Quelques contrôles de cohérences externes ont été réalisés sur les pondérations finales. Pour cela, on compare l'estimation obtenue dans l'enquête avec la pondération sur observables avec des marges connues par ailleurs *via* le recensement ou des estimations de population, pour des questions posées de la même manière dans le recensement de la population et dans l'enquête EpiCov. Deux variables ont été retenues : la nationalité et le type d'activité (tableau 7).

Les résultats obtenus sont relativement proches entre le questionnaire court et le questionnaire long sur la variable de type d'activité sauf sur la modalité autre situation. Dans l'enquête EpiCov, les individus se sont davantage déclarés dans une autre situation que dans le recensement. Il s'agit probablement d'un biais lié à la compréhension de la question. Les personnes étant en chômage partiel ont possiblement renseignées cette modalité au détriment de la modalité en emploi ou au chômage, ce qui expliquerait les différences observées avec les marges issues du recensement.

L'écart est un peu plus marqué pour l'estimation de la nationalité. La différence est significative, mais d'assez peu (tableau 7). Or on a introduit un léger biais dans les estimations du questionnaire court en appliquant un mécanisme de seuillage des probabilités de réponse petites, dans le calcul des groupes de réponse homogènes (GRH). Les premières estimations sans seuillage des probabilités de réponse minimum au sein des GRH fournissaient des estimations de la part d'étrangers proche de 7,2 % mais la dispersion des poids résultante était trop importante pour être conservée en l'état. En conséquence, même si les estimations de la proportion de nationalités étrangères diffèrent entre les deux types de questionnaires, il est possible que cette différence ne soit qu'un artefact numérique et non réellement significative. On revient sur cette question en annexe A8.

Tableau 7 : Estimation du type d'activité et de la nationalité selon la source (en %)

	Type d'activité						Nationalité	
	En emploi	Chômage	Retraite	En étude	Au foyer	Autre	Française	Étrangère
Quest. court	47,6	6,1	27,0	9,2	3,7	6,4	93,1	6,9
Quest. long	47,8	6,0	27,4	9,5	3,4	5,9	92,4	7,6
Référence	49,9	8,0	26,5	8,2	3,0	4,4	92,7	7,3

Lecture : on estime la part d'individus de nationalité étrangère avec la pondération du questionnaire court dans EpiCov à 6,9 %, alors qu'elle est de 7,3 % dans le recensement de la population 2017. La ligne « Référence » correspond aux statistiques issues du recensement ou des estimations de population sur le champ de l'enquête EpiCov.
Champ : personnes âgées de 15 ans ou plus résidant hors Ehpad, maisons de retraite et prisons, en France métropolitaine, en Martinique, en Guadeloupe et à la Réunion.
Source : Inserm-Drees, enquête EpiCov, vague 1 ; recensement de la population 2017.

II.4. Analyse des résultats par protocole

La pondération mise en œuvre et présentée ci-dessus assure la représentativité des répondants des 20 lots par rapport à l'échantillon global. Il est possible d'aller plus avant dans l'analyse en regardant ce qui se passe pour les différents lots de l'enquête. De fait, un redressement similaire peut être réalisé sur chacun des lots, sans avoir l'objectif d'exploiter les résultats de cette analyse à des fins de diffusion, mais simplement pour vérifier si les redressements décrits précédemment sont fiables pour l'enquête EpiCov. Dans ce cas, la pondération assure la représentativité des répondants de chaque lot par rapport à la même population d'enquête puisque chaque lot sélectionné est censément représentatif de cette population, chacun d'entre eux étant tiré aléatoirement au sein de l'échantillon global. Si la méthode de correction de la non-réponse conduit à des estimateurs sans biais, alors l'utilisation de pondérations spécifiques à chaque lot devrait donner des estimations similaires pour chacun des lots repondérés séparément.

La plupart des lots ayant des protocoles identiques, nous proposons ici de les regrouper de façon à redresser séparément d'une part les répondants des lots multimodes et d'autre part les répondants des lots monomodes. De fait, les taux de réponse et les caractéristiques des répondants sont parfaitement similaires au sein des lots multimodes d'une part[26] et des lots monomode d'autre part (annexe A2).

La modélisation de la probabilité de réponse pour chaque type de lots est semblable à celle réalisée précédemment pour l'ensemble des lots. En revanche, les étapes de GRH et de calage ne sont pas appliquées ici[27].

Le tableau 8 présente quelques exemples d'estimations réalisées sur les lots monomodes et les lots multimodes avec leur pondération associée. Nous pouvons noter que les estimations de différentes variables socio-démographiques sont significativement identiques au regard des intervalles de confiance de ces estimations.

En ce qui concerne les variables d'intérêt principales de l'enquête, l'estimation de la part des personnes en télétravail est également significativement identique. Les estimations de la part de personnes en bonne santé et de la part de personnes n'étant pas sorties du domicile dans les 7 derniers jours sont également très proches, bien que leurs intervalles de confiance à 95 % soient

26 Le lot 4 en métropole se différencie des lots 1 à 3 par un protocole séquentiel avec l'entrée du mode téléphone seulement 15 jours après le début de la collecte. Le taux de participation est légèrement moindre (0 à 2 points selon l'autre lot multimode pris en référence). Mais les caractéristiques générales des répondants sont très proches.
27 Ces étapes sont néanmoins faisables ; elles n'ont pas été réalisées car non essentielles dans les résultats présentés ici. À noter que sur l'ensemble des répondants, les poids avant GRH et calage donnent des résultats très semblables en moyenne à ceux obtenus avec le poids final, après GRH et calage. Quelques résultats sont présentés en annexe A3.

disjoints : [9,1 ; 9,5] dans les lots monomodes et [9,8 ; 10,4] dans les lots multimodes, pour la part de personnes n'étant pas sorties du domicile. Il est possible qu'un phénomène du type de celui observé sur la variable de nationalité (cf. § II.3) entre le questionnaire court et le questionnaire long soit à l'œuvre. Cet écart ténu ne paraît donc pas gênant à ce stade.

En revanche, les estimations obtenues sur la part de personnes ayant déclaré des symptômes sont très nettement différentes selon que l'estimation est réalisée à partir des lots monomodes ou à partir des lots multimodes. Par exemple, l'estimation de la prévalence de la toux se situe entre 8,5 et 8,8 % dans les lots monomodes alors qu'elle se situe entre 6,9 et 7,5 % dans les lots multimodes. Ce constat est vrai pour l'ensemble des symptômes, hormis pour la perte de goût ou d'odorat pour laquelle la différence est la moins importante. Nous pouvons également noter que l'estimation de la prévalence des symptômes est systématiquement plus élevée dans les lots monomodes que dans les lots multimodes.

Ces différences d'estimation, particulièrement importantes pour les variables de symptômes, révèlent une différence entre les protocoles qui n'est pas corrigée par la méthode de correction de la non-réponse sur observables précédemment présentée. Cette différence peut s'expliquer de deux façons :

- Il existe un effet de mesure, c'est-à-dire que les personnes qui répondent par Internet déclarent, toutes choses égales par ailleurs, davantage de symptômes que si elles avaient répondu par téléphone. Il s'agit de l'effet causal direct du mode de collecte. D'où une estimation moindre dans les lots multimodes, composés à la fois de répondants Internet et de répondants téléphone.
- Le modèle de correction de la non-réponse ne conduit pas à une estimation sans biais pour ces variables. Cela peut venir du fait que l'hypothèse, sous-jacente au modèle de correction de non-réponse, d'indépendance conditionnelle entre la participation et les variables de symptôme, conditionnellement aux variables observables intégrées dans la modélisation de la non-réponse, n'est pas vérifiée. Il reste alors des variables corrélées à la fois à la participation et aux symptômes qui ne sont pas prises en compte (i.e. des variables omises) dans le modèle de correction présenté ci-dessus ce qui engendre un biais de sélection résiduel. Cette situation est connue sous le nom de MNAR (Missing-Not-At-Random) (Little and Rubin, 2020).

Ces deux hypothèses ne sont pas exclusives l'une de l'autre et peuvent se révéler compliquées voire impossibles à tester si aucun protocole n'a été mis en œuvre dans ce but. Dans EpiCov, toutefois, il est possible, grâce au protocole, d'aller assez loin dans l'identification de ces deux phénomènes. Leurs effets ne se corrigent pas de la même manière :

- la première suppose un travail particulier sur les réponses apportées par les répondants au mode alternatif par rapport au mode retenu comme référence, et donc une action sur les réponses des répondants au mode alternatif, au niveau individuel ou en moyenne ;
- la seconde suppose d'adapter en conséquence le modèle de correction de la non-réponse, et donc une action sur la pondération des réponses dans les estimateurs de moyennes issus de l'enquête.

Il est donc important d'établir l'origine des différences d'estimation précédentes en distinguant bien ces deux effets (Hox *et alii*, 2015).

Tableau 8 : Estimations avec des pondérations distinctes pour les lots monomodes et les lots multimodes (en%)

	Lots monomodes	Lots multimodes	*Ensemble*
Variables socio-démographiques			
Âge moyen	48,3	48,6	*48,4*
	(0,19)	(0,32)	*(0,16)*
En situation de pauvreté	14,1	14,1	*14,1*
	(0,13)	(0,23)	*(0,11)*
Au chômage	6,1	6,2	*6,1*
	(0,10)	(0,16)	*(0,08)*
Plus haut diplôme obtenu : Bac ou plus	42,4	41,7	*42,3*
	(0,19)	(0,31)	*(0,16)*
Variables d'intérêt			
Télétravail	11,2	10,7	*11,1*
	(0,12)	(0,20)	*(0,10)*
Aucune sortie du domicile **^**	9,3	10,1	*9,5*
	(0,11)	(0,19)	*(0,10)*
Bon ou très bon état de santé **^**	76,4	77,3	*76,6*
	(0,16)	(0,26)	*(0,14)*
Fièvre **^**	8,0	7,0	*7,6*
	(0,10)	(0,16)	*(0,09)*
Toux **^**	8,7	7,2	*8,2*
	(0,10)	(0,16)	*(0,09)*
Perte de goût ou d'odorat	2,8	2,5	*2,6*
	(0,06)	(0,10)	*(0,05)*
Au moins 1 symptôme **^**	27,2	23,5	*26,0*
	(0,17)	(0,27)	*(0,15)*

Note 1 : **^** indique que l'estimation fondée sur les lots monomodes diffère significativement (à 95%) de celle fondée sur les lots multimodes.
Note 2 : entre parenthèses, sont indiquées les écart-types estimés via la macro Everest de l'Insee, modélisant d'une part l'effet du plan de sondage (qui s'écarte d'un sondage stratifié par département à allocation proportionnelle) et, d'autre part la correction de non-réponse par GRH et le calage.
Lecture : la prévalence de la fièvre estimée avec une pondération spécifique dans les lots monomodes est de 8,0 %. L'écart-type de cette estimation est de 0,10 points de pourcentage, soit un intervalle de confiance à 95 % de [7,8 % ; 8,2 %].
Champ : personnes âgées de 15 ans ou plus résidant hors Ehpad, maisons de retraite et prisons, en France métropolitaine, en Martinique, en Guadeloupe et à la Réunion.
Source : Inserm-Drees, enquête EpiCov, vague 1.

III. Biais de mesure ou biais de sélection ?

Dans un premier temps, nous allons donc évaluer l'hypothèse d'existence d'un effet de mesure. Pour cela, on analyse plus spécifiquement les variables de symptômes, qui sont les variables d'intérêt principales pour lesquelles la différence des estimations fondées sur les lots monomodes et les lots multimodes est la plus nette. Pour une analyse plus lisible, seuls certains symptômes seront parfois présentés : la fièvre, la toux, la perte de goût ou d'odorat et l'indicatrice d'au moins un symptôme. Les principaux résultats portent néanmoins sur l'ensemble des variables de

symptôme. Une première sous-partie développe quelques considérations générales sur la notion d'effet de mesure. Puis la partie III.2 présente une classe de méthodes d'estimation des effets des mesures fondées sur des méthodes de score de propension dont les résultats sont donnés au paragraphe III.3. Le paragraphe III.4 présente une autre méthode d'estimation, dite « Local average treatment effect-LATE », vraisemblablement plus incertaine mais aussi plus robuste car fondée sur un instrument du mode, rendue possible par le protocole très particulier mis en œuvre dans EpiCov. Enfin, un dernier paragraphe (III.5) conclut de l'ensemble de ces résultats sur l'effet de mesure que l'origine probable des constats réalisés dans la partie précédente est principalement un effet de sélection endogène, dont la correction est étudiée dans la partie suivante (IV).

III.1. Les effets de mesure dans les enquêtes multimodes

L'enquête EpiCov est une enquête multimode, en tout cas pour un cinquième de l'échantillon. Les répondants Internet diffèrent des répondants Téléphone en raison de leurs caractéristiques personnelles. Dans EpiCov, les répondants Internet sont plus jeunes, plus aisés, plus souvent en couple avec enfant et disposent plus souvent d'une adresse mail dans Fidéli. Et ces différences de caractéristiques sont susceptibles de jouer dans les réponses que les répondants apportent à l'enquête, indépendamment du mode choisi pour répondre. Il s'agit dans ce cas d'un effet de composition qui en lui-même ne porte pas à conséquence[28], si ce n'est qu'il est susceptible de perturber l'identification d'un effet de mesure proprement dit (i.e. toutes choses égales par ailleurs).

Les écarts de réponses sur les variables d'intérêt peuvent également s'expliquer par un effet de mesure, ou biais de mesure[29]. Il s'agit du fait qu'un même individu – ou deux individus parfaitement identiques, c'est-à-dire toutes caractéristiques étant égales, observables ou non – réponde différemment sur Internet et au téléphone. Dans la littérature, on distingue principalement deux types d'effets cognitifs pouvant causer un biais de mesure :

- un effet de désirabilité sociale (De Leeuw, 1992). Cet effet correspond à la formulation de réponses insincères dans le but de donner une image de soi en accord avec les attendus normatifs. Il est considéré plus important dans le cas d'enquêtes intermédiées. C'est pourquoi on recommande en général de privilégier un protocole auto-administré pour des questions sensibles ou d'opinion, davantage sujettes à ce type de biais.
- un effet de *satisficing* (Krosnick, 1991). Cet effet correspond à une faible implication de l'enquêté qui se contente d'une réponse approximative suite à un arbitrage entre l'effort à fournir pour répondre et les intérêts à le faire. Il peut se manifester par de la non-réponse partielle, des réponses systématiques, des effets d'ordre, des arrondis ou encore une mauvaise compréhension des consignes liée à une lecture hâtive. Il est considéré plus important dans le cas d'enquêtes auto-administrées.

Concernant les symptômes, qui sont demandés sur une période précise, l'effet de mesure attendu est que la déclaration soit plus importante par Internet que par téléphone. D'une part, du fait d'une moindre attention accordée à la période de référence sur Internet et de l'absence du contrôle informel des déclarations qui s'opère naturellement en présence de l'enquêteur au téléphone. D'autre part, du fait d'une potentielle réserve à déclarer des problèmes de santé à un enquêteur plutôt qu'en auto-administré.

En théorie, le biais de mesure est donc défini au niveau individuel. Toutefois, cet effet individuel ne peut jamais (ou presque) être mis en évidence car seule la réponse observée sur le mode de collecte effectif est connue. Par conséquent, la réponse alternative est inconnue ; elle est dite

28 Ce ne sont pas les mêmes individus qui répondent à chacun des deux modes. Mais l'ensemble de ces individus forme la population d'intérêt donc c'est l'ensemble de leurs réponses qui importe, et non de savoir que les uns et les autres ne répondent pas de la même chose.

29 Dans ce texte, on utilise indifféremment les expressions « biais de mesure », « effet de mesure » ou « erreur de mesure » pour désigner cette composante du biais.

contrefactuelle. Le biais de mesure ne peut donc jamais[30] être mesuré avec certitude car la réponse sur le mode contrefactuel est inconnue et donc estimée. Le cadre usuellement retenu pour l'estimation du contrefactuel est celui du modèle causal de Rubin (Imbens et Rubin, 2015). La stratégie d'estimation consiste, dans ce cadre, à évaluer l'effet du mode sur des individus comparables, du point de vue de caractéristiques observables ou que le protocole appliqué permet de rendre comparables. Ces méthodes reposent principalement sur deux hypothèses importantes :

i. La première est l'hypothèse d'indépendance conditionnelle. La recherche d'un contrefactuel se fait en rendant comparables les répondants de chaque mode de collecte sur un ensemble de caractéristiques observables. Toutes les caractéristiques expliquant à la fois la sélection à un mode plutôt qu'un autre et les variables d'intérêt doivent être prises en compte pour une bonne comparaison. Si une ou plusieurs de ces caractéristiques sont inobservables, alors le contrefactuel sur observables ne sera pas parfaitement comparable. L'effet de mode mesuré, neutralisé des différences observables, peut alors comprendre un effet de mesure mais également un effet de composition inobservable. L'objectif des méthodes d'estimation des effets de mesure consiste ainsi à utiliser le maximum d'informations disponibles pertinentes pour limiter le risque qu'il subsiste un biais de composition inobservable mais il n'est pas toujours possible de l'exclure totalement.
ii. La seconde hypothèse repose sur l'existence d'un support commun, c'est-à-dire qu'il existe des individus comparables dans les deux modes de collecte. Si ce n'est pas le cas, il ne sera pas possible de trouver un contrefactuel pour les individus en dehors du support commun. L'estimation de l'effet de mesure ne pourra alors se faire que pour les individus appartenant à ce support commun. D'où l'intérêt de disposer du support commun le plus large possible car le reste de la population d'enquête n'appartenant pas à ce support commun peut aussi générer son propre biais de mesure qui devient *de facto* non mesurable.

III.2. Méthodes d'estimation des effets de mesure sur fondement du score de propension

La définition de l'effet de mesure conduit à chercher à comparer des groupes de répondants similaires en tout point sauf le mode de collecte. Plusieurs méthodes peuvent être mises en œuvre pour évaluer l'effet du mode de collecte après avoir rendu comparables les répondants Internet et les répondants téléphone. Nous en avons testé plusieurs d'entre elles pour assurer la robustesse des résultats. Quatre méthodes, fondées sur une méthode de score de propension, sont présentées dans ce paragraphe et les résultats correspondants sont donnés au paragraphe III.3.

La première méthode appliquée repose sur la propriété de score équilibrant. Elle consiste d'abord à modéliser la probabilité de répondre sur un mode plutôt que l'autre, ici le téléphone plutôt qu'Internet. Si le mécanisme d'allocation du mode est complet et bien modélisé, alors conditionnellement à la probabilité d'être affectés à un mode plutôt que l'autre, les individus d'un mode et de l'autre sont parfaitement comparables. Cette méthode permet ainsi de réduire le nombre de dimensions du problème de la comparaison des groupes de répondants à une seule dimension : si les hypothèses d'indépendance conditionnelle déjà évoquées sont vérifiées, il suffit de comparer des répondants Internet et Téléphone ayant la même probabilité de répondre sur Internet plutôt que de comparer des répondants Internet et téléphone ayant les mêmes caractéristiques (socio-démographiques ou autres) dont la liste peut être très longue. Cette probabilité de répondre sur un mode plutôt que l'autre est connue sous le terme de score de

30 On suppose ici un schéma classique dans lequel l'individu ne répond qu'une seule fois à l'enquête. A titre expérimental, il est possible de réaliser deux enquêtes analogues sur deux modes différents et donc d'interroger deux fois un même individu sous deux modes différents (Biemer, 2001 ; Klausch et al, 2017). Cependant, dans ce cas, il est possible que l'individu apporte deux réponses différentes car dans le temps séparant les deux enquêtes, sa situation a pu évoluer. Et par ailleurs, le fait de ne plus découvrir l'enquête lors de la deuxième interrogation peut aussi, par effet d'apprentissage, modifier la réponse apportée par l'individu. Donc cette approche, guère généralisable, n'aboutit pas non plus à des résultats « définitifs » de ce point de vue.

propension (Rosenbaum et Rubin, 1983). En pratique, la méthode repose sur une repondération mobilisant l'inverse du score de propension[31] :

$$poids_i^{ps} = \begin{cases} w_i / ps_i(X_i) & pour\ les\ répondants\ Téléphone \\ w_i / (1 - ps_i(X_i)) & pour\ les\ répondants\ Internet \end{cases} \quad (2)$$

avec *w* la pondération initiale, *ps* le score de propension et *X* le vecteur des variables auxiliaires prises en compte dans le score. À noter qu'il est également possible de réaliser la repondération sans utiliser le poids initial *w* (à remplacer par 1 dans les expressions précédentes, dans ce cas).

La validité de cette approche repose sur celle du score de propension : le modèle doit être sans biais, c'est-à-dire, en l'espèce, qu'aucune variable expliquant la sélection à un mode de collecte plutôt qu'un autre et qui jouerait en même temps sur la variable d'intérêt ne doit être omise. Sous cette hypothèse, s'il persiste une différence significative dans les réponses données par les répondants Internet et les répondants téléphone après repondération, cette différence est alors assimilable à un effet de mesure.

Il faut noter que cette repondération a pour seul objectif de rendre les répondants Internet et les répondants Téléphone comparables, c'est-à-dire en leur donnant la même structure marginale qui n'est ni celle des répondants Internet, ni celle des répondants Téléphone, ni celle de la population cible. Elle ne vise pas nécessairement la représentativité de la population cible mais plutôt celle du support commun (cf. *supra*). Elle quantifie le biais de mesure par différence des valeurs moyennes pondérées des deux groupes de répondants, sous les hypothèses (i) et (ii) précédentes et pour la population du support commun aux populations répondantes à chacun des modes comparés.

La seconde méthode consiste à modéliser non pas le mode mais la variable d'intérêt, en contrôlant par le mode de collecte et tous les facteurs associés à la fois au mode de collecte et à la variable d'intérêt. Si la modélisation est correcte, alors l'effet du mode est une estimation non biaisée de l'effet de mesure. En pratique, la méthode repose sur une régression logistique des variables d'intérêt conditionnellement au mode de collecte (avoir répondu par téléphone plutôt qu'Internet ici) et à un certain nombre de variables de contrôle. De manière similaire à ce qui vaut pour le score de propension, la validité de cette méthode repose sur celle de sa modélisation sous-jacente : tous les facteurs de confusion doivent être mesurés et inclus dans le modèle qui prédit alors efficacement et sans biais la variable d'intérêt. L'hypothèse (i) d'indépendance conditionnelle est également nécessaire, comme précédemment. À cela, s'ajoute une hypothèse (ii') de dépendance linéaire de la variable d'intérêt au biais de mesure et aux covariables, qui n'existe pas dans l'approche par le score de propension. En conséquence, cette méthode est légèrement plus restrictive que la précédente sur le score de propension pour la qualification du biais de mesure.

La troisième méthode consiste à coupler les deux précédentes, c'est-à-dire de réaliser une régression de la variable d'intérêt pondérée par l'inverse du score de propension. Il est prouvé que cette méthode présente une double robustesse : l'estimation de l'effet est sans biais si l'une au moins des modélisations (celle du score ou celle de la variable d'intérêt) l'est (Quantin, 2018).

Enfin, une quatrième méthode consiste à utiliser le score de propension évoqué précédemment pour réaliser un appariement entre des répondants Internet et des répondants Téléphone ayant un score proche. Cette méthode consiste à considérer comme contrefactuel l'individu le plus proche par rapport au score de propension ayant répondu sur l'autre mode de collecte. Contrairement à la régression, cette méthode a l'avantage d'être non paramétrique et ne nécessite pas de faire des hypothèses sur la linéarité de la distribution conditionnelle des variables d'intérêt. L'appariement permet d'évaluer l'effet du mode de collecte pour les seuls répondants Téléphone (average treatment effect on treated ou ATT) ou pour l'ensemble des répondants (average treatment effect

31 Ici, on cherche à estimer l'effet moyen de la réponse par Internet plutôt qu'au téléphone pour tous les répondants, quel que soit le mode de collecte effectif. On parle alors de pondération ATE (average treatment effect).

ou ATE). Ici l'appariement est réalisé parmi les répondants appartenant au support commun sur les voisins les plus proches en termes de score de propension. Là encore, les hypothèses (i) et (ii) seront des conditions nécessaires d'une estimation sans biais.

Quelle que soit la méthode, il est important de contrôler l'ensemble des facteurs de confusion, c'est-à-dire des variables corrélées à la fois à la sélection au mode et aux variables d'intérêt. L'objectif est de purger toutes les différences entre répondants Internet et répondants téléphone pour que la différence restante ne soit attribuable qu'au mode de collecte, et donc à un effet de mesure. Si des facteurs de confusion sont omis, l'effet du mode résiduel qui sera estimé se composera d'un effet de mesure et d'un effet de composition. Et dans ce cas, les deux effets ne sont pas séparables.

Les variables de conditionnement peuvent être classiquement des variables issues de la base de sondage ou des variables externes. Des variables issues de l'enquête peuvent également être utilisées, à condition qu'elles ne soient pas elles-mêmes biaisées par un éventuel effet de mesure. De fait, les méthodes d'estimation d'effet de mesure se fondent sur les seuls répondants à l'enquête. Il est donc possible d'utiliser davantage de variables que ce qui peut être fait dans le cas d'une méthode de correction de la non-réponse totale. Enfin, d'autres variables de l'enquête susceptibles d'expliquer une partie de la sélection au mode (c'est-à-dire le choix d'un mode plutôt que l'autre), et éventuellement des variables d'intérêt, peuvent également être utilisées. L'objectif de ces variables est de contrôler la sélection non expliquée par les variables socio-démographiques classiques.

Les variables retenues pour la modélisation du score et les régressions sont les suivantes[32] :

(a) Variables auxiliaires issues de la base de sondage Fidéli et variables externes : âge, lien avec le référent fiscal, indicatrice de pauvreté du ménage, naissance à l'étranger, décile de niveau de vie, statut d'occupation, type de ménage, taille d'unité urbaine, quartiers prioritaires de la ville (QPV), densité, taux d'hospitalisation, indicateur communal d'accessibilité aux médecins généralistes (APL), grille communale de densité et type de coordonnées disponibles (mail, téléphone portable, au moins une coordonnée pour l'individu et pour les référents fiscaux).

(b) Variables auxiliaires issues de l'enquête : sexe, type de logement pendant le confinement, présence d'enfants selon l'âge, présence d'enfant en garde alternée pendant le confinement, diplôme, situation principale, profession, a travaillé dans les 7 derniers jours, usage du téléphone fixe, du téléphone portable et d'Internet. Ces variables sur l'usage du téléphone et d'Internet ont été introduites dans le questionnaire dans le but de contrôler les différences entre les répondants Internet et les répondants Téléphone et ainsi de mesurer au mieux les effets de mesure dans l'enquête. Ces variables, de même que celles indiquant les coordonnées renseignées dans les fichiers fiscaux, sont particulièrement explicatives du fait de répondre à un mode plutôt qu'un autre. Il peut donc être très judicieux d'introduire de telles variables dans le questionnaire.

(c) Variables additionnelles issues de l'enquête : l'état de santé général, les limitations fonctionnelles, le fait d'attribuer ses symptômes à la Covid-19, le fait qu'un autre membre du ménage ait été malade, la peur d'être contaminé en allant se faire soigner ou en allant travailler, l'évolution de la situation financière, le nombre de relance et le délai de réponse en tranches.

III.3. Résultats d'estimation de l'effet de mesure par méthodes du score de propension

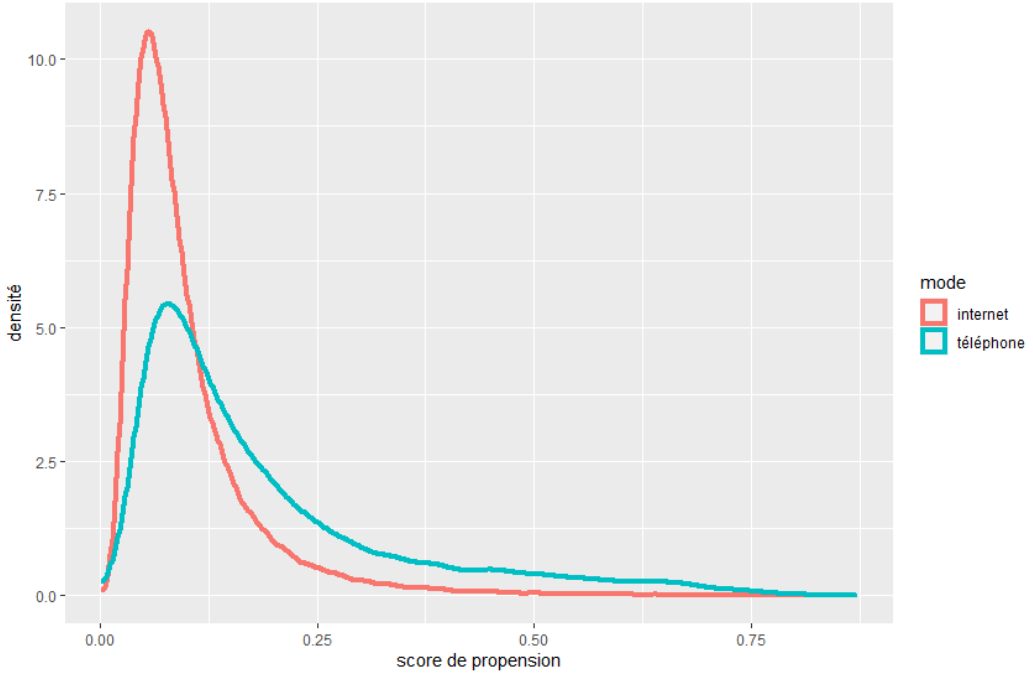
32 Pour la modélisation du score de propension, des interactions ont également été ajoutées : les variables continues au carré ; les interactions deux à deux des variables suivantes : âge, sexe, diplôme, situation principale, profession, indicatrice de pauvreté, type de ménage, être né à l'étranger, QPV, densité, lien avec le référent fiscal ; les interactions entre l'âge, le sexe, le diplôme, le niveau de vie et les types de coordonnées et l'usage du téléphone et d'Internet.

Ici, nous présentons les résultats en comparant les répondants Téléphone des lots multimodes et les répondants Internet des lots monomodes. De fait, cette comparaison limite l'existence d'éventuels biais de composition inobservables et d'un support commun petit par rapport à une comparaison au sein des lots multimodes. Lorsque le protocole est séquentiel, comparer les répondants au premier mode et les répondants au mode de relance conduit à comparer des répondants de première intention avec des répondants relancés. Or ces comportements peuvent être liés à des caractéristiques difficiles à contrôler pouvant conduire à des biais de composition inobservables. Dans le cas d'EpiCov, cette difficulté est moindre puisque les lots multimodes se rapprochent davantage de protocoles concurrentiels, hormis le lot 4. Les résultats comparant les répondants Téléphone et les répondants Internet des lots multimodes sont d'ailleurs similaires à ceux présentés ici (annexe A4). Par ailleurs, les résultats sont présentés pour les départements de métropole uniquement, les variables disponibles n'étant pas exactement les mêmes dans les Drom. L'ajout des répondants des Drom ne modifie cependant pas les résultats.

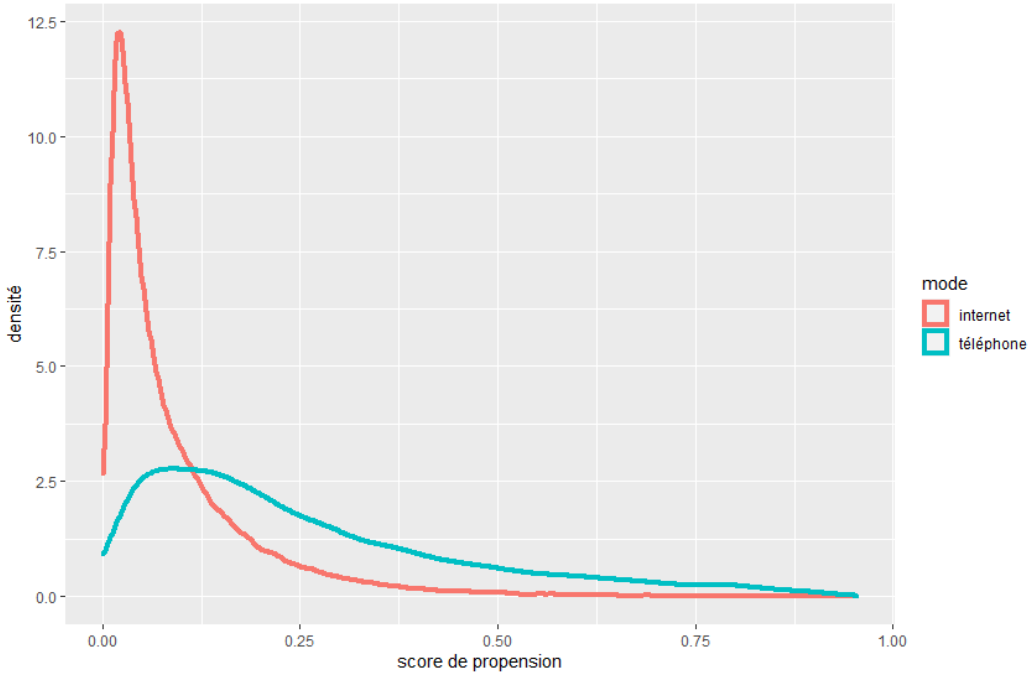
La figure 4a présente la distribution du score de propension modélisé avec les variables auxiliaires, qu'elles soient issues de la base de sondage ou de l'enquête. La figure 4b présente la distribution du score de propension modélisé avec l'ensemble des variables décrites ci-dessus, y compris les variables additionnelles. On modélise ici la probabilité d'avoir répondu par téléphone plutôt que par Internet. On observe qu'une grande partie des répondants Internet (courbe rouge) ont une propension à répondre par téléphone plutôt que par Internet très concentrée vers le bas. Les répondants Téléphone (courbe bleue) ont une probabilité de répondre par ce mode plutôt que par Internet qui reste faible mais beaucoup plus étalée. Les variables additionnelles jouent particulièrement sur la propension des répondants Téléphone, à répondre par téléphone plutôt que par Internet. L'ajout de ces variables modifie moins la propension des répondants Internet, bien que celle-ci s'en trouve resserrée. La prise en compte des variables additionnelles dans le score améliore sensiblement la comparabilité des répondants aux différents modes et conduit à des distributions davantage différenciées que celles issues d'un score sur variables auxiliaires uniquement.

Figure 4 : Distribution du score de propension modélisant la propension à répondre par téléphone plutôt que par Internet

a. Modélisation du score à partir des variables auxiliaires (a) et (b)



b. Modélisation du score à partir des variables auxiliaires et additionnelles (a), (b) et (c)



Note : la courbe rouge représente la distribution du score de propension, modélisant la propension à répondre par téléphone plutôt que par Internet, pour les répondants Internet; la courbe bleue représente la distribution du même score pour les répondants Téléphone.
Champ : personnes âgées de 15 ans ou plus résidant hors Ehpad, maisons de retraite et prisons, en France métropolitaine, ayant répondu par Internet dans les lots monomodes ou par Téléphone.

Source : Inserm-Drees, enquête EpiCov, vague 1.

Ces distributions permettent de définir le support commun, c'est-à-dire la plage de variation du score de propension commune aux répondants Téléphone et Internet. Ce support commun peut être défini de façon plus ou moins restrictive. D'un côté, il est important de disposer d'un support commun le plus large possible puisque l'effet de mesure ne pourra être estimé que sur ce support commun ; en dehors, il n'existe pas d'individus comparables (du point de vue de la propension à répondre par téléphone) présents dans les deux modes de collecte. D'un autre côté, pour disposer d'une meilleure comparabilité des répondants à chacun des modes, on peut choisir de restreindre davantage le support commun.

Avec une définition extensive du support commun – on supprime uniquement les répondants Téléphone avec un score supérieur au score maximum des répondants Internet[33] – seuls quelques répondants seraient mis de côté dans les analyses (respectivement 8 et 9 répondants Téléphone). Or, il apparaît dans les figures 4a et 4b des queues de distribution très différentes pour les répondants Internet et les répondants Téléphone. Ici, nous choisissons donc de restreindre le support commun aux scores de propension inférieurs à 0,40 pour le modèle avec variables auxiliaires uniquement (figure 4a) et aux scores inférieurs à 0,50 pour le modèle avec variables auxiliaires et additionnelles (figure 4b). Cette restriction conduit à supprimer 2 % des observations (respectivement 2 387 et 2 422 sur 109 469 répondants). Les observations supprimées sont essentiellement des répondants Téléphone ; plus de 10 % d'entre eux sont écartés de l'analyse (respectivement 1 324 et 1 568 sur 11 945 répondants Téléphone). À l'inverse, seulement 1 % des répondants Internet sont écartés (respectivement 1 063 et 854 sur 97 524 répondants Internet).

Le tableau 9 présente les résultats obtenus à partir de la méthode de pondération par l'inverse du score de propension sur les supports communs définis ci-dessus. Comme indiqué plus haut, le niveau mesuré sur chaque mode ainsi pondéré n'est pas signifiant, la pondération par l'inverse du score de propension n'ayant pas de visée représentative, contrairement à la pondération sur observables. Ce sont les écarts d'estimation entre les répondants Internet et les répondants Téléphone qu'il convient d'analyser. On observe alors que, pour la fièvre et le fait de déclarer au moins un symptôme, ces écarts sont très légèrement diminués par la pondération par l'inverse du score de propension modélisé par les variables auxiliaires (tableau 9.a). Ce résultat n'est pas très surprenant puisque la pondération sur observables prend déjà en compte la plupart des variables socio-démographiques utilisées dans le modèle du score. Lorsque les variables additionnelles, susceptibles d'être porteuses d'un effet de mode lié aux variables d'intérêt, sont ajoutées comme variables de contrôle au modèle du score (comparer les tableaux 9.a et 9.b), on observe que des écarts persistent mais ils diminuent nettement. Ainsi, la différence de prévalence d'au moins un symptôme passe de 8,2 points (soit un écart relatif de 29 %) à 3,9 points (soit un écart relatif de 14 %). La différence devient même non significative pour la fièvre et la perte de goût et d'odorat.

33 On peut faire la même chose pour le bas de la distribution. Dans notre cas, les scores les plus bas sont très proches pour les répondants Internet et les répondants Téléphone.

Tableau 9 : Effet du mode estimé par pondération par l'inverse du score de propension (en %)

a. Modélisation du score à partir des variables auxiliaires

En %	Pondération sur observables		Pondération par l'inverse du score de propension	
	Internet	Téléphone	Internet	Téléphone
Fièvre	8,2	5,3	8,5	6,0
	(0,09)	(0,21)	(0,09)	(0,23)
Toux	8,8	5,5	9,2	5,8
	(0,09)	(0,23)	(0,09)	(0,23)
Perte de goût ou d'odorat	2,8	2,2	2,8	2,2
	(0,05)	(0,15)	(0,05)	(0,14)
Au moins 1 symptôme	27,5	18,3	28,4	20,2
	(0,15)	(0,39)	(0,14)	(0,38)

b. Modélisation du score à partir des variables auxiliaires et additionnelles

En %	Pondération sur observables		Pondération par l'inverse du score de propension	
	Internet	Téléphone	Internet	Téléphone
Fièvre	8,2	5,4	8,4	7,8
	(0,09)	(0,21)	(0,09)	(0,27)
Toux	8,8	5,9	9,0	7,3
	(0,09)	(0,24)	(0,09)	(0,27)
Perte de goût ou d'odorat	2,7	2,4	2,7	2,6
	(0,05)	(0,15)	(0,05)	(0,16)
Au moins 1 symptôme	27,5	19,4	28,1	24,2
	(0,14)	(0,38)	(0,14)	(0,43)

Note : sont indiqués entre parenthèses les écarts-types calculés par Bootstrap.
Lecture : la prévalence de la fièvre pour les répondants Internet des lots monomodes est de 8,2 % avec la pondération sur observables. A noter que les niveaux indiqués ici sont sans signification particulière puisqu'ils se rapportent aux populations comparées. Ce sont les différences entre les estimations (internet-téléphone) qui sont signifiantes. Pour savoir si ces différences sont signifiantes, on rapporte la différence (par exemple, sur la première ligne : 8,2-5,3=2,9) à l'écart-type de la différence (ici $(0,09^2+0,21^2)^{0,5}=0,23$). La valeur de ce rapport est ici de 12,6, soit très supérieure, en valeur absolue, à 2 (quantile de la loi normale centrée-réduite à 97,5%). La différence est donc significative.
Champ : personnes âgées de 15 ans ou plus résidant hors Ehpad, maisons de retraite et prisons, en France métropolitaine, ayant répondu par Internet dans les lots monomodes ou par Téléphone et faisant partie du support commun.
Source : Inserm-Drees, enquête EpiCov, vague 1.

Le tableau 10 présente les résultats obtenus à partir des méthodes de régression et de régression doublement robuste (deuxième et troisièmes méthodes exposées au paragraphe III.2). On observe que l'effet du mode de collecte (ici, avoir répondu par Internet plutôt que par téléphone) augmente, de plus de la moitié, la prévalence aux différents symptômes. Ceci à l'exception des symptômes de perte de goût et d'odorat pour lesquels l'ampleur est moindre. Comme pour la méthode de pondération par l'inverse du score de propension, l'effet du mode diminue très peu en contrôlant par les variables auxiliaires (comparaison des modèles 1 et 2). En revanche, l'ajout des variables additionnelles diminue nettement l'effet du mode pour les différents symptômes, même si

l'effet reste significatif[34] – hormis pour la perte de goût ou d'odorat (modèle 3). Pour la fièvre par exemple, l'effet du mode diminue de plus de la moitié.

Tableau 10 : Effet du mode estimé par régression (odds ratios)

	Régression non pondérée			Régression doublement robuste
	Sans contrôles (1)	Variables auxiliaires (2)	Variables auxiliaires et additionnelles (3)	(modèle 3)
Fièvre	1,59 ***	1,50 ***	1,24 ***	1,20 ***
Toux	1,63 ***	1,58 ***	1,37 ***	1,42 ***
Perte de goût ou d'odorat	1,28 ***	1,27 ***	ns	1,16 ***
Au moins 1 symptôme	1,72 ***	1,67 ***	1,63 ***	1,51 ***

Note : les résultats présentés correspondent aux odds ratios de l'effet du mode de collecte. Cet effet est significatif (écart à l'unité) au seuil de *** 1 %; ** 5 %; * 10 %. Le modèle 3 est restreint au support commun correspondant à la modélisation du score de propension avec variables auxiliaires et additionnelles.
Lecture : Le fait de répondre par Internet plutôt que par téléphone augmente de 72 % les chances de déclarer avoir eu au moins un symptôme.
Champ : personnes âgées de 15 ans ou plus résidant hors Ehpad, maisons de retraite et prisons, en France métropolitaine, ayant répondu par Internet dans les lots monomodes ou par Téléphone et faisant partie du support commun.
Source : Inserm-Drees, enquête EpiCov, vague 1.

Le tableau 11 présente les résultats obtenus par appariement sur le score de propension. On observe que la différence du fait de répondre par Internet plutôt que par téléphone pour l'ensemble des répondants (estimation ATE) est très proche de ce qu'on observe par repondération par l'inverse du score de propension. De la même manière, la prise en compte des variables additionnelles diminue nettement ces différences, même si ces dernières restent significatives, hormis pour la perte de goût et d'odorat. Les résultats sont similaires lorsqu'on réalise un appariement sur les seuls répondants téléphone (estimation ATT).

Les résultats présentés ici correspondent à un appariement avec remise avec au plus 2 voisins dont la distance (caliper) est inférieure à 15 points de propension. Aucune observation n'est supprimée pour cause d'inexistence de voisins. La restriction initiale au support commun au préalable ainsi que les effectifs conséquents expliquent ce résultat. Ces résultats sont robustes à ces choix de paramètres puisqu'ils ne sont pas modifiés avec un nombre de voisins différent, une distance plus faible encore ou avec un appariement sans remise.

34 Rappelons ici que les effectifs disponibles dans l'enquête EpiCov sont particulièrement importants, rendant plus aisée la significativité des résultats.

Tableau 11 : Effet du mode estimé par appariement sur le score de propension (en points de %)

	Effet moyen sur les répondants Téléphone (ATT)		Effet moyen sur l'ensemble des répondants (ATE)	
	variables auxiliaires	variables auxiliaires et additionnelles	variables auxiliaires	variables auxiliaires et additionnelles
Fièvre	2,55 ***	0,86 ***	2,62 ***	1,04 **
	(0,26)	(0,27)	(0,32)	(0,44)
Toux	2,98 ***	1,55 ***	3,26 ***	1,96 ***
	(0,26)	(0,28)	(0,33)	(0,45)
Perte de goût et d'odorat	0,63 ***	0,06	0,51 ***	0,16
	(0,16)	(0,17)	(0,20)	(0,27)
Au moins 1 symptôme	8,43 ***	5,43 ***	8,34 ***	4,19 ***
	(0,43)	(0,46)	(0,53)	(0,72)

Note : Entre parenthèses, sont indiqués les écarts-types. L'effet moyen de répondre par Internet plutôt que par téléphone est significatif au seuil de *** 1 %; ** 5 %; * 10 %.
Lecture : La différence estimée liée au mode de collecte sur le fait de déclarer au moins un symptôme pour les répondants téléphone (ATT) est de 5,43 points de pourcentage quand on apparie sur le score modélisé à partir de l'ensemble des variables auxiliaires et additionnelles.
Champ : personnes âgées de 15 ans ou plus résidant hors Ehpad, maisons de retraite et prisons, en France métropolitaine, ayant répondu par Internet dans les lots monomodes ou par Téléphone et faisant partie du support commun.
Source : Inserm-Drees, enquête EpiCov, vague 1.

L'utilisation de ces différentes méthodes conduit à des résultats très semblables, assurant leur robustesse. D'autres tests de robustesse de ces résultats sont réalisés. Ils consistent à restreindre encore le support de comparaison pour s'assurer de contrôler au maximum les différences de composition inobservables éventuelles. Pour cela, on restreint l'analyse aux individus disposant d'un numéro de téléphone portable et d'une adresse mail dans la base de sondage[35]. On peut effectivement faire l'hypothèse que ces individus sont susceptibles de répondre aux deux modes de collecte indifféremment. Ces individus représentent plus de la moitié des répondants. De la même manière que précédemment, on modélise un score de propension à partir des variables auxiliaires et des variables additionnelles. Les distributions de ce score sont relativement similaires à celles observées sur l'ensemble des répondants (annexe A5). Nous proposons d'évaluer l'effet de mode à partir d'un support commun similaire au précédent – en supprimant les scores de propension supérieurs à 0,50 – puis d'un support commun plus restreint, consistant à supprimer tous les répondants avec un score supérieur à 0,25. Les résultats présentés au tableau 12 sont fondés sur la méthode par appariement sur le score de propension modélisé à partir des variables auxiliaires et additionnelles. Les autres méthodes donnent des résultats semblables ; la méthode issue de la pondération par l'inverse du score de propension est présentée en annexe (annexe A6).

35 Un enrichissement des numéros de téléphone a été réalisé par le prestataire en charge de la collecte. Ces informations ne sont pas prises en compte car nous n'en disposons pas. Cependant, l'enrichissement conduit surtout à récupérer des numéros de téléphone fixes.

Tableau 12 : Effet moyen du mode (ATE) estimé par appariement sur le score de propension selon le support (en points de%)

		Ensemble des répondants		Répondants disposant d'un numéro de téléphone et d'un mail	
				Score sur variables auxiliaires et additionnelles	
	Prévalence pondérée (1)	Score sur variables auxiliaires	Score sur variables auxiliaires et additionnelles	Support commun large	Support commun restreint (2)
Fièvre	7,72	2,62 *** (0,32)	1,04 ** (0,44)	0,50 (0,58)	0,23 (0,66)
Toux	8,25	3,26 *** (0,33)	1,96 *** (0,45)	1,72 *** (0,58)	1,73 *** (0,65)
Perte de goût et d'odorat	2,65	0,51 *** (0,20)	0,16 (0,27)	0,10 (0,35)	0,06 (0,39)
Mal de tête	13,70	5,38 *** (0,41)	3,62 *** (0,56)	4,46 *** (0,72)	4,28 *** (0,80)
Fatigue	11,06	3,86 *** (0,37)	1,51 *** (0,52)	1,76 *** (0,67)	1,54 ** (0,75)
Courbatures	8,82	3,49 *** (0,33)	2,09 *** (0,45)	2,17 *** (0,57)	2,01 *** (0,65)
Difficultés respiratoires	4,51	1,74 *** (0,24)	0,93 *** (0,34)	0,77 * (0,44)	0,63 (0,50)
Douleurs thoraciques	4,99	1,88 *** (0,26)	0,87 ** (0,36)	1,08 ** (0,46)	0,87 * (0,53)
Au moins 1 symptôme	26,23	8,34 *** (0,53)	4,19 *** (0,72)	4,35 *** (0,91)	3,61 *** (1,02)

Note : Entre parenthèses, sont indiqués les écarts-types. L'effet moyen de répondre par Internet plutôt que par téléphone est significatif au seuil de *** 1 %; ** 5 %; * 10 %.
Lecture : 26,23 % des individus déclarent au moins un symptôme. La différence estimée liée au mode de collecte est de 8,34 points de pourcentage quand on apparie sur le score modélisé à partir des variables auxiliaires.
Champ : personnes âgées de 15 ans ou plus résidant hors Ehpad, maisons de retraite et prisons, en France métropolitaine, ayant répondu par Internet dans les lots monomodes ou par Téléphone et faisant partie du support commun. La prévalence est calculée pour l'ensemble des répondants en métropole.
Source : Inserm-Drees, enquête EpiCov, vague 1.

On observe que l'effet du mode de collecte sur la plupart des variables de symptômes est encore réduit par rapport à l'effet sur l'ensemble des répondants, d'autant plus lorsqu'on se restreint à un support commun très réduit, et donc à une population *a priori* plus comparable. L'effet du mode n'est plus significatif pour la fièvre, la perte de goût et d'odorat et les difficultés respiratoires, alors qu'il était significatif pour l'ensemble des symptômes sur l'ensemble des répondants et sans prise en compte des variables additionnelles. Il devient également peu significatif pour la fatigue et les douleurs thoraciques. Même pour les autres symptômes, pour lesquels on observe toujours un effet de mode significatif, celui est fortement réduit. Ainsi, l'effet du mode sur le fait de déclarer au moins un symptôme passe de plus de 8 points à moins de 4.

III.4. Étude complémentaire sur la robustesse de l'effet de mesure avec un estimateur LATE

Dans ce paragraphe, nous présentons un autre estimateur fondé sur un instrument du mode que nous permet de construire le protocole de collecte EpiCov. Les méthodes présentées au paragraphe III.3. reposent sur l'hypothèse d'indépendance conditionnelle des variables d'intérêt à la variable de sélection du mode de collecte (en l'occurrence le téléphone plutôt qu'internet), conditionnellement aux variables qui entrent dans l'appariement ou le score de propension. C'est évidemment une hypothèse forte qui, dans le cas où le choix du mode est *in fine* laissé à l'enquêté, peut être discutée et surtout impossible à tester. On est donc obligé de s'en remettre à cette hypothèse, sans véritable garde-fou. Un moyen de sortir de cette difficulté est de construire, par le protocole, un instrument du choix du mode, c'est-à-dire une variable qui joue dans le choix du mode mais qui est indépendante des variables d'intérêt. Un tel instrument ne peut en réalité exister que dans le cas d'une expérience aléatoire. Par exemple, un protocole qui scinde les répondants en deux lots sélectionnés de manière indépendante et qui donne lieu à une composition différente des modes au sein des lots génère naturellement un instrument du mode : l'indicatrice d'appartenance au lot. EpiCov offre une situation de ce type, en tirant parti des différences de protocoles entre les lots multimodes, sous des hypothèses que nous allons expliciter.

Tableau 13 : Quelques statistiques descriptives des lots multimodes

Lot	Taux de réponse (%)	Proportion de répondants téléphone parmi les répondants (%)
Lot 1	45,4	45,4
Lot 2	47,6	40,7
Lot 3	46,5	39,5
Lot 4	45,0	29,3

Champ : personnes âgées de 15 ans ou plus résidant hors Ehpad, maisons de retraite et prisons, en France métropolitaine, en Martinique, en Guadeloupe et à la Réunion.
Source : Inserm-Drees, enquête EpiCov, vague 1.

Dans les lots 1 à 3, le téléphone et Internet ont été proposés aux enquêtés dès le début de la collecte (protocole concurrentiel) avec cependant de petites différences entre le lot 1 et les deux autres lots concurrentiels (2-3), liées à la logistique d'enquête, le lot 1 faisant l'objet du questionnaire long de l'enquête, les deux autres étant soumis au questionnaire court. La proportion de répondants téléphone est donc identique pour les lots 2 et 3 et sensiblement inférieure à celle observée pour le lot 1 (tableau 13). Le lot 4 a fait, quant à lui, l'objet d'un protocole séquentiel au bout de deux semaines ; avant cela, seul le mode internet était proposé, à l'instar des protocoles appliqués dans tous les autres lots d'enquête (lots 5 et suivants). Ceci se traduit par une proportion de répondants téléphone plus faible dans ce lot 4 que dans les trois autres lots multimodes.

Ainsi, on dispose de trois groupes (le lot 1, les lots 2 et 3, et le lot 4) de sous-échantillons sélectionnés aléatoirement dont la différence principale consiste en ces différences de protocoles multimodes.

Le taux de réponse, en revanche, est très semblable entre ces quatre lots, de l'ordre de 46 %, et nettement supérieur à celui des lots monomodes, s'échelonnant d'un minimum de 33,7 % à un maximum de 35,5 % pour l'ensemble des 16 lots concernés (sur le même champ). Cette proximité des taux de réponse des lots multimodes est conforme à l'idée selon laquelle, la population répondante, dans ces quatre lots, est la même. Considérés dans leur ensemble, les individus qui ont participé à l'enquête, au sein de chacun des quatre lots multimodes, sont ceux qui auraient participé si on leur avait proposé de répondre par téléphone **ou** par internet. Certains sont indifférents : ils ont répondu au premier mode qui leur a été proposé. Les autres ont participé au

téléphone ou à internet selon les circonstances de contact et leurs propres préférences. Quoi qu'il en soit, la population répondante est la même : si la proportion de répondants téléphone au sein de ces lots diffère, c'est parce que le protocole, c'est-à-dire le portefeuille de modes, proposé à l'enquêté, diffère lors du premier contact, et non parce que les enquêtés répondants sont différents.

Ainsi, l'appartenance[36] au lot 1 ou aux lots 2 et 3 d'un côté, plutôt qu'au lot 4 de l'autre, constituent de bons instruments pour évaluer l'effet du mode de collecte sur les réponses.

Cette situation nous permet de mettre en œuvre une méthode d'évaluation par Local Average Treatment Effect (LATE - Imbens and Angrist, 1994) de l'erreur de mesure. Si on utilise le vocabulaire de l'économétrie de l'évaluation :

- l'impact recherché est l'effet propre du mode téléphone (M=1) par rapport au mode internet (M=0) ;
- l'incitation à participer au traitement est donnée pour le groupe traité (lots 1, 2 et 3 ou une combinaison de ces lots ; Z=1), tandis que le lot témoin est le lot 4 (Z=0).

Dans ce contexte, l'appartenance au lot 1 (ou aux lots 1 à 3 selon le cas), indicatrice notée Z, est un instrument du traitement (le fait de répondre par téléphone), indicatrice notée M par la suite. En effet, ce couple de variables (Z,M) remplit les deux conditions essentielles suivantes :

- la variable instrumentale Z est corrélée au mode de collecte M. Cela se manifeste, entre autres, ici par le fait que la part des modes de collecte diffère entre les lots traités (1, 2 et 3 ou une combinaison de ces lots) et le lot témoin (4).
- L'affectation au lot (Z) n'a pas d'incidence, par elle-même, sur la réponse (Y — aussi appelée « variable d'intérêt »), conditionnellement au mode de réponse (M — i.e. Y(M,Z) = Y(M)).

Une troisième condition dite de « monotonie » est également vérifiée : elle correspond à la situation dans laquelle un répondant téléphone du groupe témoin (Z=0) aurait participé sous le même mode s'il avait été sélectionné dans les lots traités (Z=1). Le protocole retenu assure en effet qu'un répondant téléphone du lot 4 se serait vu proposer le mode téléphone s'il avait été sélectionné dans l'un des lots traités, s'il n'avait pas répondu par internet avant d'être sollicité pour une réponse téléphonique.

Sous ces trois conditions, on peut estimer sans biais l'effet propre du mode de collecte avec l'estimateur de Wald :

$$\hat{\delta} = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[M|Z=1] - E[M|Z=0]} \qquad (3)$$

Le tableau 14 donne les résultats de l'estimateur de Wald pour les principales variables de symptômes, pour différentes configurations d'estimation. A une exception près, aucune estimation n'est significative (à 95 %). On note également qu'en niveau, les estimations d'effet de mesure sont, en général, nettement plus faibles que celles, significatives, obtenues avec des méthodes de conditionnement sur observables par score de propension (comparer le tableau 14 aux tableaux 11 et 12). Ce résultat est conforme à l'idée selon laquelle les hypothèses d'indépendance conditionnelle de la sélection au mode et des variables d'intérêt, requises avec les méthodes de score de propension, ne sont probablement pas parfaitement assurées, ou du moins, la variance estimée est sous-évaluée.

36 Compte tenu des petites différences que l'on observe sur la composition en modes de chacun des lots, il s'avère en effet cohérent de distinguer les lots 1 et les lots 2 et 3. Un test de robustesse peut consister à rassembler les lots 1, 2 et 3.

La variable de prévalence de la toux pourrait être affectée d'un effet de mesure, au sens du test de comparaison du protocole du lot 1 et du lot témoin 4. Cependant, cet effet n'est pas confirmé à l'aide des lots 2-3 ainsi que pour l'ensemble des lots traités.

Ces nouveaux résultats suggèrent que si un effet de mesure existe, il n'est pas décelable car trop faible par rapport aux écarts-types de l'estimateur de Wald[37].

Tableau 14 : Estimateurs LATE de l'effet de mesure « téléphone (M=1) par rapport à internet (M=0) » sur différentes variables de symptôme (en points de%)

Variable	Groupe traité		
	Lot 1	Lots 2-3	Lots 1-2-3
Fièvre	-0,04	-4,19	-2,56
	(2,25)	(3,27)	(2,63)
Toux	5,04 ***	-2,19	0,84
	(2,49)	(3,33)	(2,49)
Perte de goût ou d'odorat	0,68	0,47	0,56
	(1,49)	(1,91)	(1,66)
Au moins 1 symptôme	0,30	-4,37	-2,42
	(4,02)	(5,31)	(4,47)

Champ : personnes âgées de 15 ans ou plus résidant hors Ehpad, maisons de retraite et prisons, en France métropolitaine, en Martinique, en Guadeloupe et à la Réunion.
Lecture : l'effet de mesure estimé pour le téléphone par rapport à internet (à l'aide d'un estimateur de Wald) est 0,04 points si on compare le lot1 (groupe traité) au lot 4 (groupe non traité).
Note : Ecarts-types entre parenthèses, estimés par Bootstrap (1 000 réplications). Les variables M et Z font référence à la relation (3). Dans tous les cas, le groupe traité est le lot 4. Significativité au seuil de *** 1 %; ** 5 %; * 10 %.
Source : Inserm-Drees, enquête EpiCov, vague 1.

III.5. Hypothèse d'une prédominance de l'effet de sélection

L'hypothèse d'un effet de mesure sur les déclarations de symptôme n'est pas complètement écartée au vu des résultats précédents. Pour autant, la comparaison de populations davantage comparables – par exemple selon les coordonnées disponibles et en restreignant le support commun – conduit à diminuer nettement l'ampleur estimée des effets de mesure, voire à rendre ces effets non significatifs pour certains d'entre eux.

Les écarts observés de prévalence d'au moins un symptôme, entre les lots multimodes et monomodes corrigés de la non-réponse sur observable s'élèvent à 3,7 points (cf. tableau 8). Dans les lots multimodes, le téléphone représente environ 40 % des répondants (tableau 13). Si une erreur de mesure de 3 à 4 points sur la prévalence était associée au téléphone par rapport à internet, celle-ci expliquerait entre un tiers[38] et 40 % des écarts de prévalence observés entre les lots multimodes et les lots monomodes. Un raisonnement analogue sur la fréquence du symptôme de toux (écart multimode - monomode de 1,5 points — cf. tableau 8 — et erreur de mesure de 2 points — cf. tableau 12) montre que l'erreur de mesure n'expliquerait, au plus, que 50 % de l'anomalie constatée entre les lots multimodes et les lots monomodes (tableau 8). Ce constat est plus généralement valable pour les écarts observés sur les variables de symptômes. Par

37 Rappelons que les intervalles de confiance les plus petits obtenus avec l'estimateur de Wald proposé sont de 8 points de demi-amplitude. Donc si erreur de mesure il y a, elle est inférieure, en valeur absolue, à 8 points. Ce résultat n'est évidemment pas extrêmement informatif puisque l'intervalle est grand. Mais l'ensemble des résultats de cette partie IV, y compris le niveau estimé de l'estimateur de Wald, inférieur en valeur absolue à 2 points, vont dans le même sens d'un effet de mesure très limité voire inexistant.
38 précisément 0.4x3.0/3.7=0,32 ; 0.4x4.0/3.7=0,43.

conséquent, une fraction majoritaire de cet écart ne peut s'expliquer par un autre phénomène qu'un effet de sélection non purgé par la correction de non-réponse sur observables[39].

Par ailleurs, on observe une très nette diminution de l'effet de mode résiduel lorsqu'on contrôle par les variables additionnelles issues de l'enquête. Or ces variables constituent un proxy d'un intérêt pour la thématique, corrélé aux variables d'intérêt mais qui pourrait n'être que partiellement corrélé aux variables sociodémographiques plus classiques[40]. Ce résultat montre que les variables sociodémographiques ne suffisent pas à expliquer les différences de composition des répondants Internet et des répondants Téléphone, et que de tels facteurs de confusion sont aussi vraisemblablement omis dans la correction de la non-réponse sur observables.

Enfin, l'estimateur de Wald, fondé sur les différences de composition des lots multimodes en termes de réponse par mode, indique que l'erreur de mesure n'est pas décelable et *a priori* de faible ampleur si elle existe.

Ces différents éléments conduisent à formuler l'hypothèse d'un biais de sélection inobservable sur les variables de symptômes dans l'enquête EpiCov. L'existence d'un tel biais est d'ailleurs fortement soupçonné dans le cas des enquêtes en lien avec la Covid-19 (Schaurer et Weiss, 2020 ; Class et Kohler, 2020).

Tout d'abord, la plupart de ces enquêtes sont réalisées par Internet. Or participer à ce mode de collecte auto-administré résulte — probablement — davantage d'une action volontaire de l'enquêté que dans le cas d'enquêtes intermédiées par un enquêteur, et donc mobilise davantage son intérêt pour l'enquête et sa thématique. Il est à craindre que l'empressement à répondre soit plus marqué sur Internet qu'au téléphone. Par ailleurs, on peut aussi souligner que les délais très courts de collecte ont sans doute limité les efforts pour joindre les plus réticents et les convaincre de répondre.

De plus, la thématique de la Covid-19 est un sujet ayant un impact très important sur la vie sociale, économique et individuelle d'une très grande partie de la population, qui a suscité beaucoup d'inquiétude et dont la médiatisation fut très importante. Il est ainsi très probable que les personnes les plus concernées par la maladie, qu'elles furent inquiètes, affectées personnellement ou bien que leurs proches aient eu à en souffrir, ont pu se décider à participer davantage que les autres. Les circonstances mêmes de l'enquête, dont la collecte a débuté le 2 mai 2020 tandis que le premier confinement en France se terminait (il a débuté le 17 mars 2020 et s'est achevé le 11 mai 2020) ont certainement renforcé cette tendance. Rappelons qu'à cette époque, la population ne disposait ni de masque ni de test, de sorte qu'une enquête dont la lettre-avis indiquait qu'elle avait « pour objectif de renseigner sur la diffusion du virus dans la population et sur les conséquences de l'épidémie sur la vie des personnes », ne pouvait que susciter l'intérêt des personnes ayant développé une sensibilité particulière à la maladie. Enfin, une des promesses d'EpiCov était de procéder à des tests sérologiques. Cette information avait été communiquée aux sélectionnés pour ces tests (soit un sous-échantillon d'environ 17 000 personnes —voir partie V) et était affichée dans la FAQ de l'enquête. Cette particularité, dans ces circonstances, ne pouvait que

39 *Stricto sensu*, trois phénomènes de collecte peuvent causer un biais sur des estimateurs de Hajek : (i) une erreur de mesure, (ii) un effet de sélection (i.e. lien entre variable d'intérêt et participation) non purgé par les variables de conditionnement utilisées pour le calcul du modèle de participation et (iii) une probabilité nulle de participer de certaines sous-populations. Dans le cas présent, l'anomalie constatée entre les lots multimodes et les lots monomodes porte, *a priori*, sur la sous-population dont la propension à participer à l'enquête n'est pas nulle dans au moins un des lots. Cette question fait l'objet d'une discussion complémentaire en annexe A8.

40 Les corrélations avec l'outcome « au moins un symptôme » sont en moyenne plus importantes pour les variables additionnelles que pour les variables auxiliaires socio-démographiques. C'est notamment le cas du fait d'attribuer ses symptômes à la Covid et de la présence d'un membre du ménage malade. Ces corrélations diminuent peu lorsqu'on conditionne par des variables socio-démographiques.

renforcer l'attrait de l'enquête aux yeux de personnes sensibles. Ce biais de sélection en fonction de la thématique de l'enquête est bien connu (Groves, Singer et Corning, 2000). Toutes les enquêtes sont *a priori* victimes de cette auto-sélection des répondants. Cependant, cette auto-sélection peut être corrigée si cet intérêt pour la thématique de l'enquête est bien corrélé aux variables de contrôle mobilisées dans le modèle de correction de la non-réponse. Dans le cas de la Covid-19, cette hypothèse est discutable. En effet, l'intérêt pour cette thématique a des chances d'être moins lié à des caractéristiques socio-démographiques connues qu'à des traits de personnalité ou des attitudes politiques, qui ne sont en général pas observables et ne peuvent donc pas être contrôlées (Schaurer et Weiss, 2020).

Un certain nombre d'autres résultats viennent étayer cette hypothèse de l'existence d'un biais de sélection inobservé. Ces résultats ne permettent pas de tester statistiquement cette hypothèse. Cependant, ils constituent un faisceau d'éléments allant tous dans le même sens :

- *Le lien entre le contexte épidémique et la participation à l'enquête.* On observe que les enquêtés résidant dans les territoires les plus touchés par l'épidémie sont ceux qui répondent le plus à l'enquête et le plus rapidement. Cet effet est particulièrement visible quand seul le mode de collecte par Internet leur est proposé. Ainsi, la corrélation entre le taux de réponse dans l'ensemble de l'échantillon et le taux d'hospitalisation ou de décès du département est de 0,21. Cette corrélation est surtout valable en début de collecte : la corrélation entre le taux de réponse et le taux d'hospitalisation est de 0,22 en première semaine de collecte contre 0,07 en dernière semaine. Cette corrélation avec le contexte épidémique est essentiellement visible dans les lots monomodes : la corrélation entre le taux de réponse et le taux d'hospitalisation est de 0,24 dans les lots monomodes alors qu'elle est de 0,06 dans les lots multimodes[41]. On peut en conclure que la participation, notamment par Internet, est très liée à la proximité et l'intérêt pour la thématique de l'enquête.
- *Le lien entre l'empressement à répondre et la réponse à l'enquête*. On observe que les répondants les plus prompts à répondre, comme ceux ayant été moins relancés, déclarent davantage de symptômes. Cette corrélation est observée sur Internet comme au téléphone mais elle est plus marquée sur Internet. Ainsi, la corrélation entre la déclaration d'au moins un symptôme et le nombre de relances est de -0,06 ; cette corrélation est plus importante dans les lots monomodes (-0,07) que dans les lots multimodes (-0,05). Les coefficients de corrélation sont similaires lorsqu'on étudie le délai de réponse[42]. Ce lien entre promptitude à répondre et déclaration de symptômes reste significatif quand on contrôle par d'autres caractéristiques socio-démographiques. On peut en conclure que les répondants les plus prompts à participer sont les plus concernés par l'enquête, et on peut faire *a fortiori* l'hypothèse que les répondants sont davantage concernés que les non-répondants, hypothèse confortée par le résultat suivant.
- *Le lien entre la participation à l'enquête et la réponse à l'enquête*. On observe que les personnes ayant une plus forte probabilité de participer à l'enquête déclarent davantage de symptômes. Ainsi, la corrélation entre la probabilité prédite de répondre à l'enquête et le fait de déclarer au moins un symptôme est de 0,020. Elle est semblable dans les lots multimodes (0,021) et les lots monomodes (0,018).

Si l'hypothèse d'un effet de mesure n'est pas écartée, il semble que l'existence d'un effet de sélection inobservé soit prédominante — ou *a minima* nécessaire — pour expliquer le biais observé sur l'estimation des variables de symptômes déclarés. Pour pallier ce problème et permettre des estimations plus fiables pour les variables de symptômes, une méthode de correction est proposée dans la partie suivante. À noter que pour les autres variables d'intérêt non

41 On observe également ce résultat au niveau individuel : la corrélation entre la probabilité de répondre et le taux d'hospitalisation (respectivement le taux de décès) est de 0,038 (resp. 0,093) dans les lots monomodes contre 0,000 (resp. 0,050) dans les lots multimodes.

42 La variable de délai de réponse est proposée en 4 tranches : réponse en moins de 5 jours ; de 5 à 10 jours ; de 10 à 20 jours ; plus de 20 jours.

sujettes à ce biais, il convient d'utiliser la pondération associée à la méthode de correction sur observables, présentée à la partie II. Cependant, l'existence ou non d'un biais de sélection endogène n'a pas été systématiquement testé pour l'ensemble des variables d'intérêt de l'enquête. Si un tel risque existe du fait du lien avec la thématique de l'enquête, il convient de tester l'existence d'une éventuelle sélection endogène avant d'utiliser la pondération usuelle sur observables. L'annexe A8 apporte aussi quelques compléments sur ce point.

IV. Corriger l'effet de sélection endogène

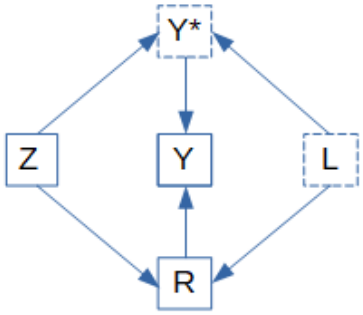
IV.1. Le problème de la sélection endogène

Les méthodes de redressement classiques sur observables, telles que présentées dans la partie II, conduisent à des estimations sans biais sous l'hypothèse d'un mécanisme de non-réponse ignorable. Sous cette hypothèse, appelée aussi Missing-At-Random (MAR), la participation est indépendante de la variable d'intérêt, conditionnellement aux caractéristiques observables prises en compte dans le modèle de correction de la non-réponse.

Pour simplifier, plaçons-nous dans le cadre d'une enquête monomode et supposons qu'il existe une (ou plusieurs) variable(s) corrélée(s) à la fois à la participation et à la variable d'intérêt non prise(s) en compte dans le modèle de participation. Usuellement, on considère que toutes les variables de ce type sont fortement corrélées aux caractéristiques observables classiques (âge, diplôme, revenu, etc.). Si ce n'est pas le cas, alors le mécanisme de non-réponse est non-ignorable (ou Missing-Not-At-Random —MNAR— cf. Little et Rubin, 2020). Dans ce cas, l'hypothèse sur laquelle reposent les méthodes de redressement sur observables n'est pas vérifiée et ces méthodes conduisent à des estimateurs de moyenne ou de sommes sur la population d'enquête (estimateurs d'Horvitz-Thompson) biaisés. Ce biais est d'autant plus grand que le taux de réponse est faible et que les variables omises sont corrélées à la variable d'intérêt.

La figure 5 illustre cette situation : la variable d'intérêt observée pour les seuls répondants Y est expliquée par sa variable latente inobservée Y* —qu'on peut assimiler ici à la vraie valeur que l'on souhaite estimer pour l'ensemble de la population — et par la participation R. Les caractéristiques observables Z expliquent à la fois la participation et la variable d'intérêt ; elles peuvent être contrôlées. Cependant, s'il existe une ou plusieurs autres variables L inobservées mais explicatives à la fois de la participation et de la variable d'intérêt, un autre chemin s'ouvre entre la participation et la variable d'intérêt. Or, on ne peut pas contrôler ce chemin, ce qui biaise la relation estimée entre la participation R et la variable d'intérêt Y*.

Figure 5 : Schéma causal dans le cas d'une sélection endogène



Note : chaque flèche représente un lien causal entre la variable origine et la variable destination. Les carrés en pointillés correspondent à des variables non observées, contrairement aux carrés pleins.

IV.2. Application du modèle d'Heckman

Ici, nous proposons d'appliquer un modèle d'Heckman pour redresser ce biais de sélection. Pour ce faire, nous faisons l'hypothèse qu'il n'existe pas d'effet de mesure, hypothèse sur laquelle nous reviendrons plus tard. Le modèle de Heckman (Heckman, 1979) consiste à modéliser simultanément la variable d'intérêt *y* et la participation *r* sous la forme suivante :

$$\begin{cases} \text{(i)} & y_i & = & c^1 + \mathbf{z}_i\chi + \epsilon_i^1 \\ \text{(ii)} & r_i^* & = & c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi + \epsilon_i^0 \\ \text{(iii)} & r_i & = & \mathbf{1}(r_i^* \geqslant 0) \end{cases} \tag{4}$$

avec ***z*** le vecteur des variables de contrôle et ***w*** un instrument expliquant la participation mais pas la variable d'intérêt. Formellement, les conditions d'identification du modèle sont les suivantes[43] :

$$\begin{cases} \mathbf{E}\left(\begin{pmatrix}\epsilon_i^0\\ \epsilon_i^1\end{pmatrix}\middle|\, \mathbf{z}_i, \mathbf{w}_i\right) = 0 \\ \begin{pmatrix}\epsilon_i^0\\ \epsilon_i^1\end{pmatrix} \hookrightarrow \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma\right) \\ \Sigma = \begin{pmatrix}1 & \varrho\sigma\\ \varrho\sigma & \sigma^2\end{pmatrix} \end{cases} \tag{5}$$

C'est à travers la matrice de variance des aléas Σ, et le facteur de corrélation *ρ,* qu'est modélisée la formation simultanée de la variable d'intérêt et de la participation.

La résolution du modèle de Heckman permet de construire deux estimateurs :

43 **E** désigne l'espérance mathématique ; **P** la probabilité.

- l'un par imputation de la variable d'intérêt pour les non-répondants. Le modèle permet en effet de prédire tous les paramètres nécessaires pour estimer $\mathbf{E}(y_i|z_i,w_i,r_i=0)$.
- l'autre par repondération, en utilisant la probabilité d'inclusion conditionnelle $\mathbf{E}(r_i|z_i,w_i,y_i)$ calculée avec les mêmes paramètres.

Usuellement, en économétrie, c'est le premier estimateur par imputation, basé sur l'espérance conditionnelle de y, qui est mis en œuvre à partir du modèle de Heckman. Ici, nous proposons d'utiliser l'estimateur par repondération (Castell & Sillard, 2021). Cette méthode, non utilisée à notre connaissance dans la littérature, correspond davantage aux usages en matière d'exploitation d'enquêtes, dans lequel il est habituel de disposer et diffuser un jeu de pondération, plutôt que de procéder par imputation des non-répondants.

Le modèle de Heckman présenté ci-dessus peut être adapté aux variables d'intérêt binaires (Galimard *et allii*, 2018). L'équation (i) correspond alors à la modélisation de la variable latente y_i^* de la variable d'intérêt :

$$\begin{cases} (0) & y_i &= \mathbb{1}(y_i^* \geqslant 0) \\ (\mathrm{i}) & y_i^* &= c^1 + \mathbf{z}_i\chi + \epsilon_i^1 \\ (\mathrm{ii}) & r_i^* &= c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi + \epsilon_i^0 \\ (\mathrm{iii}) & r_i &= \mathbb{1}(r_i^* \geqslant 0) \end{cases} \quad (6)$$

Les conditions d'identification restent identiques, à ceci près que l'on peut poser $\sigma=1$ dans les conditions d'identification. Nous obtenons alors les estimateurs suivants :

$$\mathbf{P}(r_i=1|z_i,w_i,y_i=1)=\frac{\Phi_2(c^0+z_i\beta+w_i\psi,c^1+z_i\chi;\rho)}{\Phi(c^1+z_i\chi)} \quad (7)$$

$$\mathbf{P}(r_i=1|z_i,w_i,y_i=0)=\frac{\Phi_2(c^0+z_i\beta+w_i\psi,-(c^1+z_i\chi);-\rho)}{\Phi(-(c^1+z_i\chi))} \quad (8)$$

avec Φ la fonction de répartition de la loi normale centrée réduite et Φ_2 la fonction de répartition de la loi normale bivariée centrée réduite de corrélation ρ ou $-\rho$. Le modèle est estimé par maximum de vraisemblance (voir par exemple Lollivier, 2002).

Pour construire un jeu de poids qui corrige de la sélection endogène pour la variable *y*, on utilise la même méthode que pour la pondération sur observables, présentée dans la partie II, mais en remplaçant la probabilité de participer calculée par modèle logistique par ces probabilités issues du modèle de Heckman. On divise ainsi le poids de tirage par cette probabilité de participer puis on applique les GRH puis un calage sur marges, de la même manière que pour la pondération sur observables.

L'existence d'instruments *w*, qui expliquent la participation mais pas la variable d'intérêt, est nécessaire à la bonne convergence du modèle. Cet instrument implique que le principe de

monotonie (voir par exemple Imbens et Angrist, 1994) soit vérifié par la variable latente de participation à l'enquête. Supposons que l'instrument *w* est binaire. Le modèle implique que $r_i*(w_i=1)-r_i*(w_i=0)=\psi$ pour tout *i*. ψ étant une constante, la participation est soit croissante, soit décroissante avec l'instrument. En d'autres termes, si ψ est positif, tout individu participant à l'enquête en l'absence d'instrument ($w_i=0$) aurait également participé en présence de l'instrument ($w_i=1$) ; et réciproquement si ψ est négatif.

Cette propriété doit être vérifiée par l'instrument pour la cohérence du modèle appliqué. C'est ensuite grâce au protocole d'enquête qu'on peut justifier ou non que l'instrument utilisé vérifie effectivement cette propriété. Dans le cas de l'enquête EpiCov, le protocole pour une partie monomode et pour l'autre partie multimode constitue un instrument indiscutable pour appliquer le modèle d'Heckman : l'instrument envisagé est donc la variable indicatrice d'appartenance aux lots multimodes. De fait, appartenir aux lots multimodes plutôt qu'aux lots monomodes a un effet significatif sur la participation : la différence est de plus de 10 points (45,9 % contre 34,4 %). Et dans le même temps, l'appartenance aux lots est indépendante des variables d'intérêt de l'enquête, les lots étant tirés aléatoirement. Il s'avère enfin que cet instrument respecte l'hypothèse de monotonie puisque le mode de collecte par Internet est proposé dans les deux cas : les enquêtés qui ont répondu par Internet dans les lots monomodes auraient pu répondre de la même manière s'ils avaient été sélectionnés dans les lots multimodes.

Le modèle s'avère contraint par la simultanéité de l'estimation. Ainsi, il n'est pas possible de prendre en compte l'ensemble des variables de contrôle utilisées dans la modélisation logistique de la participation présentée dans la partie II, le modèle ne disposant pas d'assez de degrés de liberté. Par exemple, la variable de stratification (croisement département*indicatrice de pauvreté) n'est pas intégrée au modèle. Cependant, les variables les plus explicatives sont prises en compte ainsi que des croisements avec une indicatrice de taux d'hospitalisation pour rendre compte des disparités locales liées à l'épidémie, du fait de l'absence de la variable de stratification. Les variables[44] prises en compte sont présentées en annexe (annexe A1).

IV.3. Résultats

Pour l'ensemble des symptômes déclarés, la correction par le modèle d'Heckman conduit à diminuer nettement la prévalence par rapport à la correction sur observables (tableau 15). Par exemple, pour la fièvre, la prévalence, qui est de 7,6 % avec la pondération sur observables, diminue à 5,2 % avec la pondération corrigée du biais de sélection endogène, soit une baisse d'un tiers environ. Cette diminution est du même ordre pour l'ensemble des symptômes étudiés.

Le modèle d'Heckman conduit à considérer qu'il existe une sélection endogène importante. Le paramètre ϱ qui correspond à la corrélation des aléas des deux équations d'*outcome* et de participation est toujours positif et nettement significatif : il varie entre 0,14 pour la perte de goût ou d'odorat et 0,47 pour le mal à la tête. Ce qui montre qu'il y a bien une sélection inobservable avec des répondants qui déclarent davantage de symptômes que ce que n'auraient fait les non-répondants.

44 L'annexe A8 propose une discussion sur les variables à intégrer dans le modèle de Heckman en examinant le comportement du modèle sur la variable indicatrice de naissance à l'étranger.

Tableau 15 : Prévalence des symptômes selon la pondération utilisée

	Pondération sur observables (%) (1)	**Pondération Heckman (sans calage) associé à chaque symptôme (%)** (2)	***Pondération Heckman (après calage) associé à « au moins 1 symptôme » (%)***	ρ ***(modèle d'Heckman)*** (3)
Fièvre	7,6	5,2	*4,7*	*0,26 ****
	(0,09)	(0,38)		*(0,05)*
Toux	8,2	4,8	*5,0*	*0,38 ****
	(0,09)	(0,28)		*(0,05)*
Perte de goût ou d'odorat	2,6	2,0	*1,6*	*0,14 **
	(0,05)	(0,28)		*(0,08)*
Mal à la tête	13,6	7,6	*8,3*	*0,47 ****
	(0,11)	(0,30)		*(0,04)*
Fatigue inhabituelle	10,9	6,8	*6,6*	*0,35 ****
	(0,10)	(0,38)		*(0,05)*
Courbatures / douleurs musculaires	8,7	4,7	*5,2*	*0,43 ****
	(0,09)	(0,26)		*(0,05)*
Difficultés respiratoires, essoufflement inhabituel	4,5	2,8	*2,7*	*0,29 ****
	(0,07)	(0,22)		*(0,07)*
Douleurs thoraciques, oppression	4,9	2,9	*3,0*	*0,34 ****
	(0,07)	(0,22)		*(0,07)*
Au moins 1 symptôme	26,0	15,6	*15,6*	*0,44 ****
	(0,15)	(0,54)		*(0,03)*

Champ : personnes âgées de 15 ans ou plus résidant hors Ehpad, maisons de retraite et prisons, en France métropolitaine, en Martinique, en Guadeloupe et à la Réunion.

Lecture : la proportion de personnes ayant souffert de Fièvre au sens de l'enquête s'élève à 7,6 % si on estime cette proportion par un estimateur d'Horvitz-Thompson pondéré avec un modèle de probabilité d'inclusion fondé sur les observables (variables de la base de sondage). Cette proportion est de 5,2 % avec des pondérations issues d'un modèle d'Heckman. Elle est de 4,7 % si on utilise les pondérations du modèle d'Heckman obtenues sur la variable « Au moins 1 symptôme ». La corrélation estimée par modèle d'Heckman entre les aléas de la variable latente d'intérêt et celle de participation est de 0,26.

Notes (1) et (2) : les écarts-types indiqués sont issus, pour (1), du calcul de variance tenant compte du plan de sondage et d'une non-réponse corrigées par GRH et calage (cf. § II.2) ; pour (2), issus d'un Bootstrap sur les échantillonnés, avec 1000 réplications. Théoriquement, ces derniers n'intègrent pas (à tort) l'effet du plan de sondage (estimé à un facteur multiplicatif de 1,2 sur l'écart-type) et le calage éventuel (négligeable en l'espèce puisque les variables de calage ne sont pas corrélées aux variables de symptômes). En même temps, les écarts-types de (1) n'intègrent pas l'incertitude du modèle de non-réponse, estimée là-encore à un facteur multiplicatif de 1,2 pour un modèle Probit non calé (cas de la correction de non-réponse sur observable) mais réellement dominant dans le cas du modèle d'Heckman. C'est d'ailleurs la raison pour laquelle les écarts-types des estimations fondées sur une pondération de Heckman sont aussi élevés par rapport à ceux des estimations fondées sur une pondération sur observables. (pour plus de détails sur les calculs de précision, se reporter à l'annexe A7 : « Composantes de l'incertitude des estimateurs de prévalence des symptômes »).

Note (3) : estimation du coefficient de corrélation du modèle d'Heckman par maximum de vraisemblance (voir § IV.2) et écarts-types associés entre parenthèses. Significativité au seuil de *** 1 %; ** 5 %; * 10 %.

Source : Inserm-Drees, enquête EpiCov, vague 1.

La pondération étant issue d'une modélisation jointe de la participation *r* et d'une variable d'intérêt *y* (modèle ci-dessus), chaque pondération n'est applicable qu'à la variable d'intérêt ayant servi au modèle. Pour une estimation corrigée du biais de sélection endogène d'une variable d'intérêt, un jeu de pondération spécifique à cette variable doit donc être utilisé.

Cependant, pour limiter le nombre de jeux de poids et faciliter l'utilisation de ces pondérations, on peut chercher un jeu de poids unique qui permet de corriger une grande partie des biais de sélection pour chacun des symptômes. Ici, on peut voir que le jeu de pondération associé à la variable indicatrice d'au moins un symptôme permet de corriger une grande partie du biais pour l'ensemble des symptômes (tableau 15). Pour faciliter l'usage de ces méthodes, nous préconisons donc d'utiliser ce jeu de pondération pour l'analyse des différentes variables de symptômes, ainsi que pour des regroupements éventuels de symptômes, par exemple pour calculer des indicateurs de suspicion de Covid à partir d'un certain nombre de symptômes.

La prévalence d'au moins un symptôme déclaré diminue de 26,0 % avec la pondération sur observables à 15,6 % avec la pondération issue du modèle de Heckman, soit une baisse de 40 % (tableau 15). On peut voir que les différences d'estimations entre lots monomodes et lots multimodes sont nettement atténuées par rapport à ce qui est observé avec la pondération sur observables (comparer tableaux 8 et 16) : la prévalence estimée par le modèle d'Heckman est de 15,9 % dans les lots monomodes et de 14,6 % dans les lots multimodes. La différence n'est d'ailleurs pas significative si on intègre l'incertitude du modèle de correction de la non-réponse dans les écarts-types (voir aussi annexe A7). L'écart résiduel entre lots multimode et lots monomode est porteur, si elle existe, d'une éventuelle erreur de mesure du mode téléphone par rapport au mode internet. On a vu plus haut que cette erreur, bornée à 3 ou 4 points dans nos analyses d'erreur de mesure, expliquerait un surcroît d'environ un point entre les moyennes observées sur les lots monomodes et celles des lots multimodes, pour la prévalence d'au moins un symptôme. Notablement, on retrouve, une fois corrigé de la sélection endogène, un écart de cet ordre entre les moyennes calculées sur les lots monomodes et les lots multimodes. Ceci suggère, à nouveau, qu'il peut exister une erreur de mesure dans le cas considéré, mais qu'elle est d'ampleur beaucoup plus faible (impact de l'ordre de 1 point sur la valeur moyenne de la prévalence d'au moins un symptôme) que le biais de sélection endogène (impact de plus de 10 points sur la moyenne de la variable).

Tableau 16 : Estimations de la prévalence d'au moins un symptôme selon la pondération utilisée

Pondération sur observables	**Pondération Heckman (après calage)**		
	Ensemble	Lots monomodes	Lots multimodes
26,0 % (0,15)	15,6 % (0,09)	15,9 % (0,11)	14,6 % (0,18)

Note : entre parenthèses, sont présentés les écarts-type des estimations, calculés à partir de la macro SAS %Everest. Ils prennent en compte la variance induite par le sondage stratifié, les GRH et le calage. La variance propre au modèle de la probabilité de réponse n'est pas prise en compte ici. C'est la raison pour laquelle les écarts-types sont beaucoup plus petits que ceux indiqués dans le tableau 15. Mais naturellement, ces écarts-type sont, ici, fictivement petits car il y a lieu, en général, de tenir compte de l'incertitude du modèle de non-réponse. Elle est souvent négligeable, mais ne l'est pas dans le cas présent avec un modèle d'Heckman (voir à ce propos l'annexe A7).
Champ : personnes âgées de 15 ans ou plus résidant hors Ehpad, maisons de retraite et prisons, en France métropolitaine, en Martinique, en Guadeloupe et à la Réunion.
Source : Inserm-Drees, enquête EpiCov, vague 1.

La dispersion des poids issus du redressement par le modèle d'Heckman est accentuée par rapport à celle du modèle sur observables (tableau 17). Si 90 % des rapports de poids entre le poids final et le poids de tirage restent contenus, inférieurs à 6, les rapports de poids maximaux peuvent atteindre des niveaux relativement élevés (près de 30 pour la pondération issue du modèle d'Heckman).

Tableau 17 : Rapport poids calé sur poids de tirage

	min	0,10	0,25	0,5	0,75	0,9	max
Pondération sur observables	0,86	1,31	1,54	1,98	2,88	5,40	24,92
Pondération Heckman	0,61	1,13	1,43	1,93	2,91	5,72	29,45

Champ : Champ : ensemble des personnes considérées comme répondantes à l'enquête.
Source : Inserm-Drees, enquête EpiCov, vague 1.

Pour l'analyse mobilisant une ou plusieurs de ces variables de symptômes, il est donc recommandé d'utiliser la pondération issue du modèle de Heckman calculée à partir de l'indicatrice d'au moins un symptôme, plutôt que la pondération sur observables. En revanche, pour les autres variables d'intérêt principales de l'enquête ne semblant pas soumises à un biais de sélection endogène, il est recommandé d'utiliser la pondération sur observables, détaillée dans la partie II, qui est à la fois plus précise et soumise à moins d'hypothèses qu'une pondération issue du modèle de Heckman. Cependant, un éventuel biais de sélection endogène n'a pas été testé pour l'ensemble des variables du questionnaire. Si certaines variables sont susceptibles d'être sujettes à un tel biais du point de vue de leur lien avec la thématique de l'enquête, il peut être utile de vérifier l'absence de sélection endogène par la méthode proposée avant d'utiliser la pondération sur observables, qui reste à privilégier en l'absence de sélection endogène.

IV.4. Discussion et compléments

Si la correction permise par le modèle de Heckman peut être très importante, comme c'est le cas ici de la prévalence des symptômes, l'utilisation de cette méthode comporte plusieurs limites.

Tout d'abord, la variance des estimateurs de moyenne calculés grâce aux probabilités d'inclusion d'Heckman est nettement plus élevée que celle d'estimateurs fondés sur une correction de non-réponse sur observables. L'écart-type associé à l'estimation de la prévalence d'au moins un symptôme calculée avec la pondération Heckman (prévalence de 15,6%) est de 0,54, alors que l'écart-type équivalent est de[45] 0,17 pour une correction de non-réponse sur observables (avec GRH et calage). Cette augmentation de variance peut être un frein à l'utilisation de cette méthode dans le cas d'échantillons plus petits ou d'objectifs élevés de précision. Cependant, cette méthode permet *a minima* de calculer un intervalle de ce que serait une estimation corrigée du biais de sélection endogène. Or ce type de biais peut être très important pour des enquêtes avec un faible taux de réponse et une thématique peu corrélée aux caractéristiques socio-démographiques classiques, comme c'est le cas de l'enquête EpiCov.

Une deuxième limite tient au cadre d'application du modèle et aux hypothèses paramétriques sous-jacentes. De fait, le modèle d'Heckman fait la double hypothèse de (i) normalité des aléas des variables latentes de participation et d'intérêt (quand cette dernière est binaire) et (ii) de dépendance linéaire de ces variables latentes aux variables explicatives. Par ailleurs, l'instrument mobilisé étant binaire, cette relation linéaire est estimée *de facto* à partir de seulement deux valeurs de l'instrument : le lien observé dans les lots monomodes et le lien observé dans les lots multimodes (voir Castell et Sillard, 2021). Le coefficient directeur donné par le modèle est estimé

45 L'écart-type indiqué ici est légèrement plus élevé que celui indiqué dans le tableau 16, conséquence de la prise en compte, ici, de l'incertitude du modèle de non-réponse alors que celle-ci est négligée dans le tableau 16.

localement, de sorte que, si la relation entre r et y n'est pas linéaire, l'estimation issue du modèle d'Heckman n'est valable qu'au voisinage du taux de participation (ici autour de 35-45 %). Rien ne garantit que cette relation puisse être extrapolée à d'autres niveaux de propension à participer. Ainsi, si le lien entre participation et symptômes est encore plus fort parmi les non-répondants que pour les répondants au protocole multimode et non au protocole monomode, alors on sous-estime le biais de sélection et on corrige moins que nécessaire. À l'inverse, si le lien est moins fort parmi les non-répondants, alors on sur-estime le biais de sélection et on corrige plus que nécessaire.

Il est utile de rappeler que le modèle d'Heckman ne fournit pas une estimation des prévalences des symptômes corrigées de la sélection sur Internet plutôt qu'au téléphone, mais une estimation reposant sur la sélection liée au protocole monomode Internet relativement au protocole multimode Internet-téléphone.

Nous pouvons vérifier que la pondération issue du modèle de Heckman ne distord pas les estimations sur d'autres caractéristiques. Pour cela, nous comparons les estimations obtenues à partir du modèle de correction sur observables à celles obtenues à partir du modèle de Heckman pour des caractéristiques socio-démographiques déclarées dans l'enquête et non prises en compte dans les modèles de correction. Les résultats montrent que les estimations sont sensiblement identiques (tableau 18). Ce résultat ne prouve pas que l'hypothèse de linéarité est justifiée. Cependant, il montre que si distorsion il y a, elle ne s'opère pas sur les caractéristiques socio-démographiques analysées, ce qui est un élément rassurant.

Tableau 18 : Estimations de variables socio-démographiques selon la pondération utilisée

	Pondération issue du modèle sur observables	Pondération issue du modèle de Heckman
En emploi	46,0 %	46,3 %
Au chômage	6,1 %	6,0 %
Retraité	27,0 %	27,0 %
Mère née à l'étranger	20,3 %	20,2 %
Père né à l'étranger	21,1 %	20,9 %

Champ : personnes âgées de 15 ans ou plus résidant hors Ehpad, maisons de retraite et prisons, en France métropolitaine, en Martinique, en Guadeloupe et à la Réunion.
Source : Inserm-Drees, enquête EpiCov, vague 1.

Enfin, dans le cas d'EpiCov où l'on utilise comme instrument la différence entre le protocole monomode et le protocole multimode, une hypothèse supplémentaire importante doit être faite : l'absence d'effet de mesure sur les variables d'intérêt modélisées dans le modèle d'Heckman. De fait, l'effet de mesure se traduit par l'intervention du mode de collecte dans l'équation d'*outcome*. Les paramètres du modèle sont identifiables si le mode de collecte n'est pas colinéaire à l'instrument $\boldsymbol{w}$. Or, dans l'enquête EpiCov, l'instrument est intrinsèquement lié au mode de collecte. Il est donc nécessaire de faire l'hypothèse qu'il n'y a pas d'effet de mesure sur les variables d'intérêt modélisées pour mettre en œuvre le modèle d'Heckman.

L'analyse séparée des résultats obtenus par lots, avec les pondérations issues du modèle d'Heckman permet néanmoins de revenir sur d'éventuelles erreurs de mesure, même si cette approche n'est pas parfaitement cohérente au plan du modèle puisque, pour les raisons évoquées ci-dessus, le modèle d'Heckman est, lui, estimé sous l'hypothèse d'absence d'erreur de mesure.

Des travaux d'économètres visent à proposer un cadre plus directement cohérent, comme la méthode proposée par Lee (2009). Cette méthode vise à estimer l'effet d'un traitement dans le cas d'une sélection endogène. Elle repose sur le modèle suivant :

$$\begin{cases} \text{(i)} & y_i = c^1 + \mathbf{z}_i\chi + \mathbf{D}_i\gamma + \epsilon_i^1 \\ \text{(ii)} & r_i^* = c^0 + \mathbf{z}_i\beta + \mathbf{D}_i\psi + \epsilon_i^0 \\ \text{(iii)} & r_i = \mathbb{1}(r_i^* \geqslant 0) \end{cases} \quad (9)$$

Contrairement au modèle d'Heckman, il n'est pas nécessaire de disposer d'un instrument : la variable de traitement D peut expliquer à la fois la participation et la variable d'intérêt. Par ailleurs, l'estimation ne repose pas sur des hypothèses paramétriques, supposant l'invariance de l'effet mesuré dans la population, comme c'est le cas du modèle d'Heckman. Cette méthode ne conduit pas à une estimation moyenne de l'effet de mesure mais à l'estimation d'une borne inférieure et d'une borne supérieure.

La méthode repose sur deux hypothèses. La première est une hypothèse d'indépendance des aléas par rapport à la variable de traitement et aux variables de contrôle ***z***, ce qui est assuré dans le cas d'une affectation aléatoire au traitement (variable D). La seconde est une hypothèse de monotonie de la participation en fonction de la variable de traitement : les enquêtés qui participent sans s'être vu proposé le traitement (D=0) auraient participé si on le leur avait proposé (D=1). Cette hypothèse est identique à l'hypothèse de monotonie de l'instrument dans le modèle de Heckman. Comme pour le modèle d'Heckman, l'application de cette méthode à l'enquête EpiCov est rendue possible par le protocole, d'une part monomode et d'autre part multimode, qui répond à ces deux hypothèses (affectation aléatoire et monotonicité liée à l'adjonction du mode téléphone dans le protocole multimode). L'effet modélisé ici correspond donc à l'effet du protocole, et non à un effet de mesure qui comparerait le fait de répondre par Internet au fait de répondre par téléphone. De fait, le mode de collecte ne répond pas aux hypothèses du modèle, le mode n'étant pas affecté aléatoirement et n'assurant pas la monotonicité.

Le tableau 19 présente les estimations de l'effet du protocole (appartenir aux lots multimodes plutôt qu'aux lots monomodes) en situation de sélection endogène issues du modèle de Lee. Les résultats montrent que les intervalles ainsi calculés sont très larges. Par exemple, l'effet du protocole sur la déclaration d'au moins un symptôme est compris entre -0,20 et 0,04. Cet intervalle ne permet pas de conclure à la véracité ou non de l'hypothèse d'absence d'effet de mesure sur les symptômes. Ces résultats montrent que, lorsqu'on ne fait pas d'hypothèse ni sur l'existence d'un effet de sélection endogène ni sur l'existence d'un effet de mesure, on n'est pas capable de conclure et de distinguer ces deux effets dans le cas des symptômes déclarés avec un protocole tel que celui de l'enquête EpiCov.

Tableau 19 : Effet du protocole sur les variables de symptômes d'après le modèle de Lee

	Borne inférieure	Borne supérieure
Fièvre	-0,076	0,013
Toux	-0,083	0,008
Perte de goût ou d'odorat	-0,022	0,007
Au moins 1 symptôme	-0,197	0,041

Note : Les coefficients représentent l'effet d'avoir participé au protocole multimode plutôt qu'au protocole monomode sur les variables d'intérêt, contrôlés de l'âge, du type de ménage, de l'appartenance à un ménage pauvre, du taux d'hospitalisation du département, de la densité en tranche de la commune et du type de coordonnées disponibles dans Fidéli.
Champ : personnes âgées de 15 ans ou plus résidant hors Ehpad, maisons de retraite et prisons, en France métropolitaine, en Martinique, en Guadeloupe et à la Réunion.
Source : Inserm-Drees, enquête EpiCov, vague 1.

Ces résultats, qui n'invalident pas ceux obtenus grâce à la modélisation d'Heckman puisque la nullité du coefficient γ, assimilable à l'erreur de mesure pour le modèle (9), est dans l'intervalle de confiance, invitent à une certaine prudence. Il est très probable que les données soient ici caractérisées par un comportement de sélection endogène. Néanmoins, l'ampleur de l'impact de cette sélection sur les estimateurs de moyenne est forte, puisque les prévalences de symptômes, sous l'hypothèse d'absence d'erreur de mesure, sont généralement divisées par 2, après correction de la sélection endogène par modèle d'Heckman.

Pour réaliser l'ampleur de la correction, il est intéressant de calculer les différences de taux moyen de participation entre les personnes qui ont souffert d'au moins un symptôme et les autres, si la prévalence réelle s'élevait effectivement à 15,6 % (niveau obtenu après correction par modèle d'Heckman — cf. tableau 16). Étant donnés le taux de participation observée dans les lots monomodes (34,5 %) et la prévalence observée dans ces lots (27,2 %), cela signifie que le taux de participation des personnes qui ont souffert d'au moins un symptôme serait de 60 %, tandis que celles qui n'en ont pas souffert auraient participé avec un taux de 30 %[46]. Ces taux de participations, allant du simple au double, ne sont pas irréalistes compte-tenu des circonstances de l'enquête (voir la discussion du paragraphe III.5). Néanmoins, ils traduisent, pour être plausibles, un comportement de participation pour les personnes qui ont souffert de symptômes très différent de celui des autres.

L'hypothèse d'absence d'erreur de mesure est donc forte, comme on le voit dans cette discussion. Il est donc possible que erreur de mesure et sélection endogène soient mêlées dans le cas présent. Néanmoins, le protocole de la vague 1 d'EpiCov ne permet pas de trancher définitivement entre une hypothèse où la sélection endogène se combine à une erreur de mesure et une hypothèse de sélection endogène expliquant seule les anomalies constatées sur les déclarations de symptômes. Les travaux présentés aux paragraphes III.3 et III.4 montrent cependant que l'erreur de mesure seule ne peut pas expliquer l'ampleur des anomalies constatées.

V. Redressement de la séro-prévalence

Les redressements liés à la non-réponse des résultats des auto-prélèvements sont distincts des redressements du questionnaire car il faut prendre en compte une nouvelle étape de sélection qui consiste à accepter ou refuser l'auto-prélèvement une fois que l'individu a répondu au questionnaire. Par ailleurs, les auto-prélèvements n'ont pas été proposés à l'ensemble des individus de l'échantillon en raison des contraintes, à cette période, sur les capacités des laboratoires.

Pour l'estimation nationale de la séro-prévalence, les auto-prélèvements ont été proposés à l'ensemble des répondants du lot 2 résidant en métropole, soit un vingtième de l'échantillon. Pour les estimations locales dans les départements ciblés, le nombre de lots a été choisi pour pouvoir

46 Précisément, soit τ la participation des affectés (en%) et $\bar{\tau}$, celle des non-affectés. Alors : 0,156 τ+ (1-0,156) $\bar{\tau}$=34,5 et 15,6 τ / 34,5=27,2. Il en découle que τ=60 % et $\bar{\tau}$=30 %.

disposer d'un minimum de 1 000 tests sérologiques pour chaque zone. Ainsi, les répondants des lots suivants se sont vus proposés un auto-prélèvement :

- les lots 1 à 16 pour le département du Haut-Rhin ;
- les lots 1 à 14 pour le département du Bas-Rhin ;
- les lots 1 à 8 pour le département de l'Oise ;
- les lots 1 à 6 pour le département de Paris ;
- les lots 1 à 3 pour les départements de la petite couronne (92, 93 et 94) ;
- les lots 1 à 12 pour le département des Bouches-du-Rhône.

Une pondération spécifique a donc été réalisée pour chacune de ces zones ainsi que pour l'estimation au niveau national.

Le principe de la correction de la non-réponse pour les auto-prélèvements consiste à recalculer la probabilité de réponse au questionnaire sur les sous-échantillons concernés, puis à modéliser l'étape supplémentaire de non-réponse spécifique à ces tests, sachant que l'individu a répondu au questionnaire.

Sont considérés comme répondants aux auto-prélèvements, les répondants au questionnaire qui font partie des lots sélectionnés ci-dessus, qui acceptent de recevoir le kit d'auto-prélèvement, qui le retournent ensuite par la Poste et pour lesquels le prélèvement est exploitable. Pour l'ensemble des répondants tirés au sort pour réaliser l'auto-prélèvement, 87,6 % ont accepté de recevoir le kit au moment du questionnaire (tableau 20). Parmi eux, 82,8 % ont finalement retourné ce kit, dont une très faible partie s'est révélée inexploitable. Au final, si on tient compte uniquement des prélèvements exploitables, on dispose de 12 114 observations, soit un taux de réponse de 71 % parmi les répondants échantillonnés, ce qui est tout à fait satisfaisant. Pour chaque zone, on dispose finalement de 1 740 observations pour le Haut-Rhin, 1 191 pour le Bas-Rhin, 1 087 pour l'Oise, 1 061 pour Paris, 957 pour la petite couronne et 1 454 pour les Bouches-du-Rhône.

Tableau 20 : Les différentes étapes de sélection pour les tests sérologiques

	Tirés au sort pour l'auto-prélèvement	Réception du kit acceptée	Envoi du kit	Prélèvements exploitables
Effectifs	17 142	15 012	12 433	12 114
Pourcentage	100 %	87,6 %	82,8 %	97,4 %
Pourcentage cumulé	100 %	87,6 %	72,5 %	70,7 %

Lecture : Parmi les 12 433 kits envoyés, 12 114 sont exploitables, soit 97,4 % des kits envoyés et 70,7 % de l'ensemble des répondants tirés au sort pour l'auto-prélèvement.
Champ : ensemble des personnes considérées comme répondantes à l'enquête.
Source : Inserm-Drees, enquête EpiCov, vague 1.

L'objectif de la correction de la non-réponse aux auto-prélèvements est de donner des estimations robustes et non biaisées de la prévalence de la Covid 19 dans chacune des zones sélectionnées ainsi qu'au niveau national. On construit donc autant de pondérations que de zones d'intérêt. Pour cela, on commence par construire un nouveau poids corrigé de la non-réponse au questionnaire en modélisant la non-réponse dans chacune de ces zones prises séparément et en appliquant les GRH par zones également. Prendre directement la pondération sur observable calculée précédemment sur l'ensemble de l'échantillon conduirait à une estimation biaisée puisque le nombre de lots monomodes et de lots multimodes échantillonnés pour les auto-prélèvements ne correspond pas à la répartition de l'ensemble de l'échantillon, d'où la nécessité de ré-estimer une probabilité de répondre au questionnaire pour ces sous-populations. Prenons par exemple le cas de l'estimation de la séro-prévalence au niveau national, qui se fait à partir du lot 2 uniquement. Un

individu du lot 2 a en moyenne plus de chances de répondre qu'un individu pris au hasard dans un des 20 lots. Cela provient du fait que le lot 2 fait l'objet d'une collecte multimode, avec un taux de réponse plus élevé, alors que la majorité des 20 lots sont monomodes. Lorsqu'on s'intéresse uniquement au lot 2, les déterminants du fait de répondre ou non à l'enquête ne sont pas forcément les mêmes que pour l'ensemble de l'échantillon. Ainsi, pour chaque zone d'intérêt, une probabilité de réponse au questionnaire a été recalculée en se restreignant aux lots concernés, selon une méthodologie de correction de la non-réponse (modélisation logit et application de GRH) identique à celle décrite en partie II. À noter que cette probabilité est calculée à chaque fois pour l'ensemble des départements, alors même qu'elle ne servira que pour certains d'entre eux. Cela vient du fait que les individus peuvent déménager entre la base de sondage et la date de la collecte, et donc changer de département. Pour diffuser sur un département, il convient donc de calculer la probabilité de réponse au questionnaire de tous les individus qui y vivent au moment de l'enquête, quel que soit leur département dans la base de sondage. C'est cette dernière information qui est utile à la correction de la non-réponse au questionnaire puisque le département de résidence au moment de l'enquête est inconnu pour les non-répondants au questionnaire.

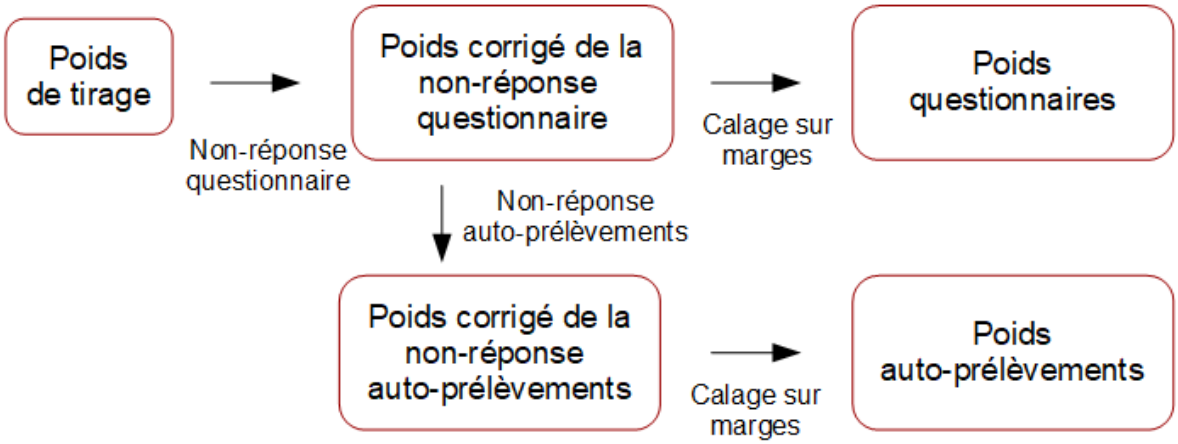
Une fois ces poids calculés pour l'ensemble des répondants au questionnaire faisant partie de l'échantillon auto-prélèvement, on modélise cette fois la non-réponse aux auto-prélèvements, c'est-à-dire le fait d'avoir retourné un prélèvement exploitable. Cette modélisation est réalisée par régression logistique puis application de GRH, de la même façon que pour la correction de la non-réponse au questionnaire détaillé dans la partie II. Les modèles logit sont calculés séparément pour le lot servant aux estimations nationales d'une part, et pour l'ensemble des participants aux estimations départementales d'autre part. Cette séparation tient au fait que les estimations nationales concernent l'ensemble du territoire, tandis que les estimations départementales concernent des territoires particulièrement touchés durant la première vague de l'épidémie de Covid-19. La non-réponse dans ces départements peut donc avoir des caractéristiques particulières.

La principale spécificité de la modélisation de la non-réponse aux auto-prélèvements est que la sélection porte sur les seuls répondants à l'enquête. Ainsi, des variables issues de l'enquête peuvent être prises en compte dans le modèle. Ces variables, qui portent directement sur la thématique de l'enquête, comme les symptômes déclarés par exemple, sont potentiellement fortement corrélées à la séro-prévalence. Le risque de sélection endogène liée à l'omission d'une variable expliquant à la fois la participation et les variables d'intérêt est donc nettement plus faible que pour les variables d'intérêt du questionnaire[47]. L'ensemble des variables retenues pour la correction de la non-réponse aux auto-prélèvements pour le lot servant aux estimations nationales et pour les lots servant aux estimations départementales sont respectivement détaillées en annexe (annexe A1).

En divisant le poids corrigé de la non-réponse au questionnaire par la probabilité de répondre aux auto-prélèvements, on obtient un nouveau poids corrigé de la non-réponse aux auto-prélèvements. Comme pour la pondération du questionnaire, on réalise ensuite un calage sur marges pour obtenir les poids finaux applicables à chacune des zones d'intérêt.

47 Toutefois, s'il y avait de la sélection endogène issue de la non-réponse au questionnaire qui n'a pu être corrigée, alors elle perdurera au moment de l'estimation de la séro-prévalence. L'étape de correction de la non-réponse aux auto-prélèvements ne peut pas corriger les erreurs d'estimation résiduelles liées à la non-réponse au questionnaire.

Figure 6 : Schéma des différentes étapes de correction de la non-réponse



Du fait de la taille des échantillons mobilisés et de la succession des étapes de sélection (réponse au questionnaire puis réponse aux auto-prélèvements), l'estimation de la séro-prévalence pour chacune des zones d'intérêt est encore plus sujette à des problèmes de rapports de poids élevés et de valeurs influentes que la pondération du questionnaire. En ce qui concerne le rapport des poids avec le poids de tirage, on observe que 90 % des rapports de poids restent contenus en deçà d'un facteur 8 (tableau 21). En revanche, les rapports de poids maximaux sont élevés, variant entre 20 et 30 selon les zones. En ce qui concerne les valeurs influentes, on remarque que dans toutes les zones 25 % des individus ayant les poids les plus élevés jouent sur 50 % des estimations (tableau 22). Par ailleurs, 5 % des répondants avec les poids les plus élevés jouent sur 16,4 à 20,7 % des estimations selon les zones. Ces résultats contribuent nécessairement à une certaine volatilité des estimations issues des répondants aux auto-prélèvements et à être vigilant quant à la comparaison des résultats par zones.

Tableau 21 : Rapport poids calés sur poids de tirage

	Min	0,10	0,25	0,5	0,75	0,9	Max
National	0,93	1,41	1,67	2,17	3,13	5,32	28,19
Bouches-du-Rhône	1,22	1,81	2,19	2,90	4,16	7,97	29,97
Oise	1,09	1,62	1,88	2,47	3,76	6,99	22,17
Bas-Rhin	1,25	1,55	1,89	2,45	3,62	5,80	23,81
Haut-Rhin	1,32	1,71	2,02	2,60	3,80	7,70	26,01
Paris	1,16	1,67	1,89	2,32	3,20	6,82	20,50
Petite couronne	1,24	1,54	1,88	2,40	3,76	6,91	32,90

Champ : ensemble des personnes considérées comme répondantes à l'enquête, ayant un prélèvement exploitable.
Source : Inserm-Drees, enquête EpiCov, vague 1.

Tableau 22 : Pourcentage de l'estimation porté par les individus en fonction de leur quantile de poids

	0 %	10 %	25 %	50 %	75 %	90 %	95 %	100 %
National	100 %	96,7 %	88,9 %	71,8 %	49,0 %	28,7 %	18,5 %	0 %
Bouches-du-Rhône	100 %	96,1 %	88,6 %	72,8 %	51,9 %	32,1 %	20,7 %	0 %
Oise	100 %	95,7 %	87,8 %	71,5 %	49,4 %	28,3 %	18,2 %	0 %
Bas-Rhin	100 %	96,2 %	87,9 %	70,5 %	47,9 %	28,2 %	18,3 %	0 %
Haut-Rhin	100 %	95,6 %	97,5 %	71,4 %	49,5 %	30,3 %	19,2 %	0 %
Paris	100 %	95,3 %	86,5 %	68,8 %	46,9 %	27,7 %	16,4 %	0 %
Petite couronne	100 %	95,8 %	88,0 %	72,0 %	49,8 %	29,1 %	18,0 %	0 %

Lecture : Tous les individus ayant un poids supérieur ou égale au quantile à 25 % des poids des individus participant à l'estimation de la séro-prévalence nationale jouent pour 88,9 % de l'estimation de la séro-prévalence au niveau national. Autrement dit, 75 % des individus influent sur 88,9 % des totaux.

Champ : ensemble des personnes considérées comme répondantes à l'enquête, ayant un prélèvement exploitable.

Source : Inserm-Drees, enquête EpiCov, vague 1.

Conclusion

De par son ampleur, sa thématique particulière et son protocole visant une collecte courte et essentiellement réalisée par Internet, l'enquête EpiCov est spécifique dans le paysage de la statistique publique. Pour assurer la qualité des estimations issues de la première vague de l'enquête, des redressements approfondis ont été mis en œuvre. La présentation de ces redressements fait l'objet de ce document de travail. De fait, les enseignements de ces travaux ont une portée plus générale, pouvant être utiles à de nouvelles enquêtes auprès des ménages.

Si la problématique de la sélection endogène n'est pas nouvelle, elle est peu traitée du fait que la plupart des enquêtes de la statistique publique portent sur des thématiques connues, supposées bien corrélées à des caractéristiques socio-démographiques classiques, et sur une collecte intermédiée avec de bons taux de réponse. L'enquête EpiCov montre que, dans certains cas, cette problématique s'avère centrale et nécessite un protocole et une correction adaptés pour assurer la qualité de l'enquête. Dans un contexte où les enquêtes rapides, sur des thématiques nouvelles et avec des modes de collecte auto-administrés (essentiellement par Internet) se développent, cette problématique de la sélection pourrait devenir un enjeu réel pour certaines enquêtes. Ce document de travail montre que des méthodes de correction existent. Pour cela, il est cependant nécessaire d'adapter le protocole en amont de la collecte, en prévoyant notamment un sous-échantillon aléatoire permettant une participation renforcée. Le multimode apparaît alors comme une solution efficace pour assurer cette participation renforcée. En revanche, la mobilisation de plusieurs modes de collecte oblige à s'interroger sur la présence, en complément de la sélection endogène, d'éventuels effets de mesure. Ce document montre que bien discerner les deux effets peut se révéler compliqué ou déboucher sur des niveaux d'incertitude tels qu'il n'est pas possible de conclure.

Dans le cas de l'enquête EpiCov, d'autres éléments, qui n'ont pas été analysés à ce stade, pourraient permettre de distinguer encore davantage ces deux effets. L'appariement envisagé avec les données du système national des données de santé (SNDS) fournira par exemple de nouvelles informations sur les enquêtés, informations en lien fort avec la thématique sanitaire de l'enquête. Une analyse des vagues suivantes de l'enquête pourrait également permettre d'aller plus loin dans l'identification d'éventuels effets de mesure.

En ce qui concerne les méthodes de redressement sur variables observables, plusieurs enseignements peuvent être tirés des spécificités de l'enquête EpiCov pour d'autres enquêtes ménages. Tout d'abord, lors d'une collecte auto-administrée, il convient de s'assurer que l'individu répondant est bien celui échantillonné. Cette vérification, effectuée spontanément par l'enquêteur dans le cas d'une collecte intermédiée, nécessite l'ajout de quelques questions pour être réalisée pendant la passation du questionnaire et/ou à l'aval par comparaison avec les données disponibles dans la base de sondage.

Lorsqu'une enquête vise une représentativité départementale, une réflexion spécifique doit être menée. L'expérience de l'enquête EpiCov montre qu'il est utile, dans ce cas, d'intégrer une dimension départementale aux redressements. Il est apparu pertinent d'intégrer cette dimension en constituant des groupes de réponse homogène départementaux, à partir d'une probabilité de réponse modélisée au niveau national. La représentativité départementale conduit également à réaliser un calage départemental (voir à ce propos Vihta et *al.*, 2022).

Bibliographie

BAJOS N. *et alii* (2020) : Les inégalités sociales au temps du COVID-19, *Questions de santé publique,* IreSP, n°40.

BECK F., CASTELL L., LEGLEYE S. et SCHREIBER A. (2022) : Le multimode dans les enquêtes auprès des ménages : une collecte modernisée, un processus complexifié, *Courrier des Statistiques, Insee*, n°7.

BIEMER P. (2001) : Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing, *Journal of Official Statistics,* 17 pp. 295-320.

CASTELL L. et SILLARD P. (2021) : Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman, *Document de travail « Méthodologie Statistique », Insee*, n°M2021/02.

CHEVALIER M. *et alii* (2022) : *Le renouvellement de l'échantillon-maître des enquêtes auprès des ménages et de l'échantillon de l'enquête Emploi de l'Insee*, Insee Méthodes n°141.

Cramer H. (1946) : Mathematical Methods of Statistics. Princeton University Press.

GALIMARD J.E., CHEVRET S., CURIS E., RESCHE-RIGON M. (2018): Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors, *BMC medical research methodology,* 18(1).

GIVORD P. et SILHOL J. (2020) : Confinement : des conséquences économiques inégales selon les ménages, *Insee Première*, n°1822.

GROVES R.M., SINGER E. & CORNING A. (2000) : Leverage-Saliency Theory of Survey Participation : Description and Illustration, *The Public Opinion Quaterly*, 64(3), pp. 299-308.

HAZIZA D. & BEAUMONT J.F. (2007) : On the Construction of Imputation Classes in Surveys, *International Statistical Review*, 75(1) pp. 25-43.

HECKMAN J.J. (1979) : Sample Selection Bias as a Specification Error, *Econometrica*, 47(1), pp. 153-161.

HOX J., DE LEEUW E. & KLAUSCH T. (2017) : Mixed Mode Research: Issues in Design and Analysis, in BIEMER *et alii,* Total Survey Error in Practice, *John Wiley & Sons*, p. 511-530.

IMBENS G.W. & ANGRIST J.D. (1994) : Identification and Estimation of Local Average Treatment Effects, *Econometrica*, 62(1), pp. 467-475.

IMBENS G.W. & RUBIN D.B. (2015) : Causal Inference for Statistics, Social, and Biomedical Sciences : An Introduction, *Cambridge University Press.*

KLAUSCH T., HOX J.J. & SCHOUTEN B. (2013) : Measurement effects of survey mode on the equivalence of attitudinal rating scale questions, *Sociological Methods & Research*, 42(3), pp. 227-263.

KROSNICK J.A. (1991) : Training strategies for coping with the cognitive demands of attitude measures in surveys, *Applied Cognitive Psychology*, 5(3), pp. 213-236.

LOLLIVIER S. (2002) : *Endogénéité dans un système d'équations normal bivarié avec variables qualitatives*, Journées de Méthodologie Statistique 2002, Insee.

LITTLE R.J.A. & RUBIN D.B. (2020) : *Statistical Analysis with Missing Data*, 3rd edition, Wiley.

LEE D.S. (2009) : Training, wages, and sample selection : Estimating sharp bounds on treatment effects, *The Review of Economic Studies*, 76(3), p. 1071-1102.

DE LEEUW E. (1992) : *Data Quality in Mail, Telephone and Face to Face Surveys,* PhD Thesis of the Amsterdam University.

POST J.C., CLASS F. & KOHLER U. (2020) : Unit Nonresponse Biases in Estimates of SARS-CoV-2 prevalence, *Survey Research Methods*, 14(2), pp. 115-121.

ROSENBAUM P.R. & RUBIN D.B. (1983) : The central rôle of propensity scores in observational studies for causal effects, *Biometrika*, 70(1), pp. 45-55.

SCHAURER I. & WEISS B. (2020) : Investigating selection bias of online surveys on coronavirus-related behavioral outcomes, *Survey Research Methods*, 14(2), pp. 103-108.

TCHENGEN TCHENGEN E. & WIRTH K. (2018) : A general instrumental variable framework for regression analysis with outcome missing not at random, *Biometrics*, 73(4), pp. 1123-1131.

TILLÉ, Y. (2019) : Théorie des sondages-2e éd.: Échantillonnage et estimation en populations finies. Cours et exercices corrigés. Dunod.

VANNIEUVENHUYZE J., LOOSVELDT G. & MOLENBERGHS G. (2014) : Evaluating Mode Effects in Mixed-Mode Survey Data Using Covariate Adjustment Models, *Journal of Official Statistics,* 30(1), pp. 1-21.

VIHTA K.D. *et alii*. (2022): COVID-19 Infection Survey, Symptoms and Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Positivity in the General Population in the United Kingdom, *Clinical Infectious Diseases*, Vol. 75, Issue 1, 1 July 2022, pp. e329–e337, https://doi.org/10.1093/cid/ciab945.

WARSZAWSKI J. *et alii* (2020) : En mai 2020, 4,5 % des la population vivant en France métropolitaine a développé des anticoprs contre le SARS-CoV-2. Premiers résultats de l'enquête nationale EpiCov, *Etudes et Résultats*, n°1167.

WARSZAWSKI J., BEAUMONT AL., SENG R. *et alii* (2022) : Prevalence of SARS-Cov-2 antibodies and living conditions: the French national random population-based EPICOV cohort, *BMC Infect Dis*, 22(41).

Annexes

Annexe A1 : Liste des variables auxiliaires prises en compte dans les modèles d'estimation de la probabilité de réponse

	Questionnaire		Auto-prélèvements	
	Modèle logistique	**Modèle Heckman**	**Modèle national**	**Modèles départe-mentaux**
Strate d'échantillonnage	x			x
Variables issues de Fidéli de niveau individu				
tranche d'âge quinquennal	x	x	x	x
sexe	x	x		x
indicatrice de garde alternée	x			
indicatrice de bilocalisation de l'individu	x		x	
zone géographique de naissance	x	x		
lien de l'individu avec le référent 1 du logement	x	x	x	x
année du dernier événement matrimonial de l'individu	x		x	x
mobilité résidentielle de l'individu l'année précédente	x		x	
indicatrice de perception d'allocations chômage ou préretraite	x	x	x	x
indicatrice de perception d'allocation chômage ou préretraite deux années successives	x			
indicatrice de perception de salaire deux années successives	x		x	
indicatrice de perception de salaire	x	x	x	x
indicatrice de perception de pensions, retraites et rentes	x	x		x
indicatrice de perception de bénéfices agricoles	x		x	
indicatrice de perception de bénéfices non commerciaux	x			
indicatrice de perception de pensions alimentaires	x			
indicatrice « l'individu vit dans un logement connu de la taxe d'habitation »	x			
Variables issues de Fidéli de niveau ménage				
type de ménage	x	x		x
nombre de personnes dans le logement	x			x
indicatrice de surpeuplement du logement	x		x	
présence d'enfants de moins de 15 ans dans le logement	x	x		
présence d'enfants de moins de 3 ans dans le logement	x			
présence d'enfants de moins de 3 à 5 ans dans le logement	x			
présence d'enfants de 6 à 10 ans dans le logement	x			
présence d'enfants de 11 à 14 ans dans le logement	x		x	
statut d'occupation du logement (propriétaire, locataire…)	x	x	x	x

indicatrice de perception d'allocation chômage ou préretraite par au moins une personne du logement	x		x	
indicatrice de perception de salaire par au moins une personne du logement	x		x	
indicatrice de perception de salaire par au moins une personne du logement durant deux années consécutives	x		x	
indicatrice de perception de bénéfices agricoles par au moins une personne du logement	x			
indicatrice de perception de pensions alimentaires par au moins une personne du logement	x			x
indicatrice de perception de pensions, retraites ou rentes par tous les individus du logement	x		x	
indicatrice de propriété d'une ou plusieurs résidences secondaires par au moins une personne dans le logement	x			
sexe du référent 1 du logement	x		x	
sexe du référent[48] 2 du logement	x			
Variables issues de Fidéli de niveau logement				
type de logement (appartement, maison, information non renseignée)	x		x	
surface du logement en tranches	x		x	
indicatrice de logement social	x		x	
date de construction du logement en tranches	x		x	
nombre de pièces du logement en tranches	x		x	x
présence d'ascenseur dans le bâtiment	x			
Variables externes géolocalisées				
accessibilité potentielle localisée en tranches (indicateur développé par la Drees et l'Irdes pour mesurer l'adéquation spatiale entre l'offre et la demande de soins de premier recours)	x	x	x	
indicatrice d'appartenance du logement à un quartier prioritaire de la ville	x	x		x
croisement de la tranche d'aire urbaine avec la catégorie de commune dans les aires urbaines calculées par l'Insee	x			x
taille d'unité urbaine		x		
grille de densité communale	x	x	x	
taux départemental d'hospitalisation pour 10 000 habitants		x		
taux départemental de décès pour 10 000 habitants		x		
taux de logements sociaux dans le carreau	x			
taux de ménages vivant en logement collectif dans le carreau	x			
taux de ménages propriétaires dans le carreau	x		x	
taux de ménages pauvres dans le carreau	x		x	

48 Conjoint du référent 1. Le référent 1 est la personne de référence du foyer fiscal. Quand plusieurs foyers fiscaux cohabitent au sein d'un même logement, le référent 1 est conventionnellement celui qui reçoit le revenu fiscal le plus élevé parmi les différentes personnes de référence des foyers fiscaux.

taux de ménages de 5 individus ou plus dans le carreau[49]	x			
taux de ménages monoparentaux dans le carreau	x		x	
taux d'individus de 65 ans ou plus dans le carreau	x			
Informations de contact issues de Fidéli				
indicatrice de présence de mail pour l'individu sélectionné	x	x	x	
indicatrice de présence de numéro de téléphone portable pour l'individu sélectionné	x	x	x	
indicatrice de présence d'au moins un numéro de téléphone portable ou d'au moins une adresse mail pour l'individu tiré	x	x	x	
indicatrice de présence d'au moins un numéro de téléphone portable ou d'au moins une adresse mail ou d'au moins un numéro de téléphone fixe pour l'individu sélectionné	x			
indicatrice de présence de mail pour un des deux référents fiscaux	x	x	x	x
indicatrice de présence de numéro de téléphone portable pour un des deux référents fiscaux	x	x	x	x
indicatrice de présence d'au moins un numéro de téléphone portable ou d'au moins une adresse mail pour un des deux référents fiscaux	x	x	x	
indicatrice de présence d'au moins un numéro de téléphone portable ou d'au moins une adresse mail ou d'au moins un numéro de téléphone fixe pour un des deux référents fiscaux	x			
Variables issues de Filosofi au niveau logement ou individu				
indicatrice de perception de l'allocation logement à caractère familial (ALF)	x		x	
indicatrice de perception de l'allocation logement à caractère social (ALS)	x			
indicatrice de perception de l'aide personnalisée au logement (APL)	x		x	
indicatrice de perception de minima sociaux	x		x	x
indicatrice de perception d'allocations familiales	x		x	
indicatrice de ménage sous le seuil de pauvreté	x		x	
indicatrice de versement de pensions alimentaires	x			
indicatrice de perception de revenus de valeurs mobilières	x		x	
indicatrice de perception de revenus du foncier	x		x	
décile de niveau de vie	x	x	x	x
décile de revenu disponible	x			
Variables issues de l'enquête relatives à l'individu enquêté				
sexe			x	
diplôme			x	x
nationalité			x	

49 Voir https://www.insee.fr/fr/statistiques/4176290?sommaire=4176305

Département habituel de résidence, avec le début de la pandémie			x	
État de santé général de l'individu			x	x
Indicatrice de problème de santé chronique ou durable			x	x
Indicatrice de perte d'autonomie en raison d'un problème de santé			x	
A été testé pour la Covid-19, et résultat du test le cas échéant			x	x
Symptômes de la Covid-19 durant le confinement			x	x
Pense avoir été contaminé par la Covid-19			x	x
Le cas échéant, absence de contact avec un professionnel de santé au moment de l'apparition des symptômes			x	x
Le cas échéant, absence de sortie à partir de l'apparition des symptômes			x	x
Situation vis-à-vis de l'emploi avant le confinement			x	x
Chômage partiel durant le confinement			x	x
Exercice d'une activité professionnelle la semaine précédente			x	x
Télétravail à temps complet la semaine précédente			x	
Profession dans le milieu médical			x	
Fréquence de sortie la semaine précédente			x	x
Le cas échéant, port du masque lors de ces sorties			x	x
Le cas échéant, absence de rencontre extérieure durant ces sorties			x	x
Tabagisme actuel			x	
A déjà fumé quotidiennement pendant plus de 6 mois			x	x
Fréquence de consommation d'alcool			x	x
Variables issues de l'enquête relatives au logement occupé durant le confinement				
Nombre d'occupants			x	x
Lien de l'individu avec les autres occupants			x	
Un autre occupant a été testé positif à la Covid-19				x
Un autre occupant a eu des symptômes de la Covid-19 durant le confinement			x	x
Sortie d'un occupant la veille			x	x
Le logement n'est pas le logement usuel de l'individu				x
Nombre de pièces			x	
Type de logement * aménagements extérieurs			x	x
Détérioration de la situation financière du ménage depuis le début du confinement			x	

À noter : certaines modalités ont parfois été regroupées afin de disposer de suffisamment d'individus par regroupement de modalité dans l'échantillon.

Annexe A2 : Caractéristiques des répondants en métropole selon le type de lot

	Lots monomodes	Lots multimodes	Lot 1	Lot 2	Lot 3	Lot 4
Moins de 30 ans	21%	21%	21%	21%	21%	21%
30-44 ans	24%	24%	24%	24%	24%	24%
45-59 ans	28%	27%	27%	26%	27%	27%
60-74 ans	21%	21%	21%	21%	21%	22%
75 ans ou plus	5%	7%	7%	8%	7%	7%
Ménage en dessous du seuil de pauvreté	13%	14%	14%	14%	14%	14%
décile de niveau de vie moyen	6,16	5,94	5,97	5,92	5,92	5,95
célibataire	14%	16%	16%	16%	16%	15%
en couple sans enfant	25%	25%	24%	24%	26%	25%
en couple avec enfant	45%	44%	44%	44%	43%	44%
famille monoparentale	10%	10%	10%	10%	10%	10%
ménage complexe	5%	6%	6%	5%	6%	6%
vit en maison	66%	65%	65%	65%	65%	66%
vit dans un quartier prioritaire de la ville	5%	5%	5%	5%	5%	5%
commune rurale	25%	25%	25%	25%	26%	25%
moins de 20000 habitants	18%	18%	19%	18%	18%	18%
de 20000 à 100000 habitants	14%	14%	13%	14%	13%	14%
plus de 100000 habitants	28%	28%	28%	28%	28%	27%
agglomération parisienne	15%	15%	15%	15%	15%	15%
mail disponible	75%	70%	70%	70%	70%	71%
téléphone portable disponible	39%	39%	40%	39%	38%	40%
au moins 1 coordonnée disponible	78%	77%	77%	77%	77%	77%
Taux de réponse	*34,9*	*45,7*	*45,2*	*47,5*	*46,2*	*44,7*
Taux d'hospitalisation du département	*15,13*	*14,94*	*14,95*	*14,91*	*14,97*	*14,99*

Annexe A3 : Estimations pondérées avant et après application des GRH et du calage sur marges

En %	Poids avant GRH et calage	Poids finaux
Âge moyen	48,44	48,35
En situation de pauvreté	14,02	13,90
Au chômage	5,49	6,11
Né à l'étranger	7,23	6,89
Plus haut diplôme obtenu : Bac ou plus	42,21	40,05

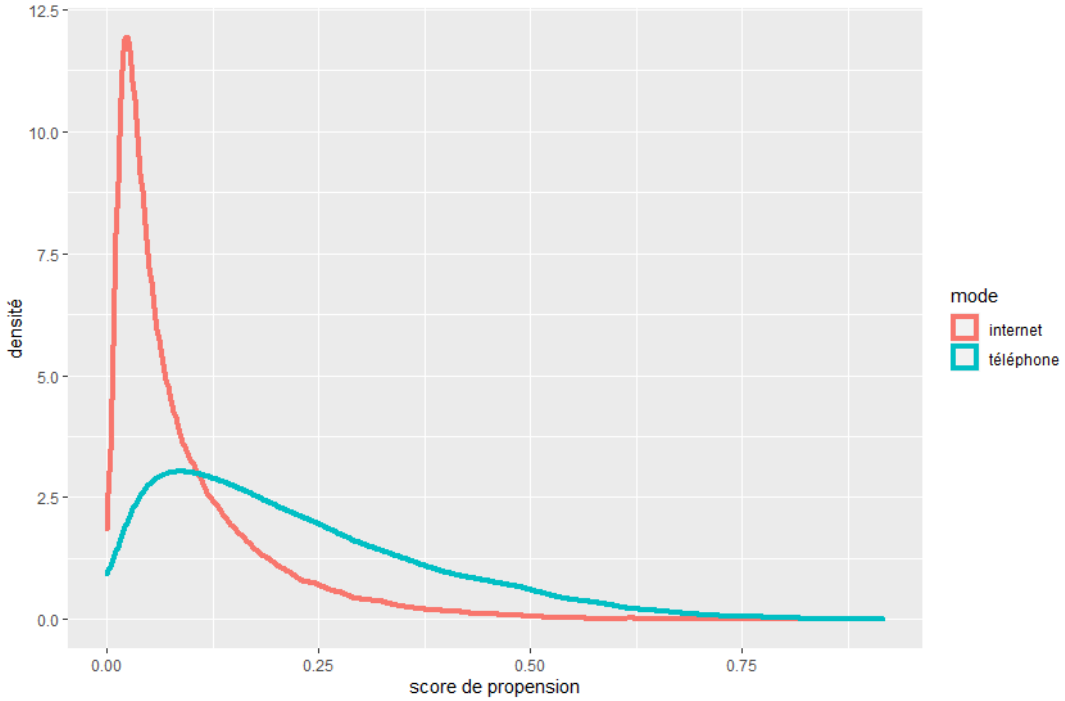
En %	Poids avant GRH et calage	Poids finaux
Télétravail	11,05	10,37
Aucune sortie du domicile	9,50	9,52
Bon ou très bon état de santé	76,67	76,49
Fièvre	7,71	7,64
Toux	8,25	8,17
Perte de goût ou d'odorat	2,67	2,62
Au moins 1 symptôme	26,15	26,00

Annexe A4 : Estimations pour les répondants Internet et Téléphone des lots multimodes selon la pondération

En %	Pondération sur observables		Repondération par l'inverse du score de propension			
			Variables auxiliaires		Variables auxiliaires et additionnelles	
	Internet	Téléphone	Internet	Téléphone	Internet	Téléphone
Fièvre	8,2 (0,19)	4,7 (0,20)	8,5 (0,21)	5,8 (0,21)	7,9 (0,19)	6,8 (0,24)
Toux	8,4 (0,19)	5,1 (0,19)	8,8 (0,20)	5,7 (0,21)	8,4 (0,19)	6,5 (0,22)
Perte de goût ou d'odorat	2,7 (0,12)	2,0 (0,13)	2,8 (0,12)	2,1 (0,13)	2,6 (0,11)	2,3 (0,14)
Au moins 1 symptôme	27,2 (0,31)	17,5 (0,36)	27,6 (0,32)	19,7 (0,35)	26,7 (0,32)	21,9 (0,40)

Note : entre parenthèses, sont présentés les écarts-types calculés par Bootstrap.
Lecture : l'estimation de la prévalence de la fièvre avec la pondération sur observables est de 8,2 % pour les répondants Internet et 4,7 % pour les répondants Téléphone.
Source : Inserm-Drees, enquête EpiCov, vague 1.

Annexe A5 : Distribution du score de propension pour les répondants disposant d'un numéro de téléphone portable et d'une adresse mail



Note : la courbe rouge représente la distribution du score de propension, modélisant la propension à répondre par téléphone plutôt que par Internet, pour les répondants Internet; la courbe bleue représente la distribution du même score pour les répondants Téléphone.
Champ : individus résidant en métropole disposant d'un numéro de téléphone portable et d'une adresse mail ayant répondu soit par téléphone soit par Internet dans les lots monomodes.
Source : Inserm-Drees, enquête EpiCov, vague 1.

Annexe A6 : Effet du mode estimé par repondération par l'inverse du score de propension parmi les répondants disposant d'un téléphone et une adresse mail

a. Support commun large

En %	Pondération sur observables		repondération par l'inverse du score de propension	
	Internet	Téléphone	Internet	Téléphone
Fièvre	9,4	7,7	9,3	9,5
	(0,12)	(0,34)	(0,12)	(0,39)
Toux	9,8	7,5	9,9	8,1
	(0,13)	(0,34)	(0,13)	(0,35)
Perte de goût ou d'odorat	3,1	2,8	3,0	3,1
	(0,07)	(0,22)	(0,07)	(0,22)
Au moins 1 symptôme	30,3	23,6	30,3	26,7
	(0,19)	(0,54)	(0,20)	(0,56)

b. Support commun restreint

En %	Pondération sur observables		repondération par l'inverse du score de propension	
	Internet	Téléphone	Internet	Téléphone
Fièvre	9,5	8,6	9,5	9,9
	(0,12)	(0,44)	(0,12)	(0,48)
Toux	9,9	7,8	10,0	8,3
	(0,14)	(0,41)	(0,14)	(0,46)
Perte de goût ou d'odorat	3,1	3,0	3,1	3,2
	(0,08)	(0,25)	(0,07)	(0,28)
Au moins 1 symptôme	30,4	25,4	30,4	27,3
	(0,20)	(0,71)	(0,20)	(0,66)

Note : sont indiqués entre parenthèses les écarts-types calculés par Bootstrap.
Lecture : la prévalence de la fièvre pour les répondants Internet des lots monomodes est de 9,5 % avec la pondération sur observables sur le support commun restreint des répondants avec un téléphone et un mail.
Champ : support commun du score calculé sur les variables auxiliaires et additionnelles pour les répondants résidant en métropole, disposant d'un numéro de téléphone portable et d'une adresse mail, ayant répondu soit par téléphone soit par Internet dans les lots monomodes.
Source : Inserm-Drees, enquête EpiCov, vague 1.

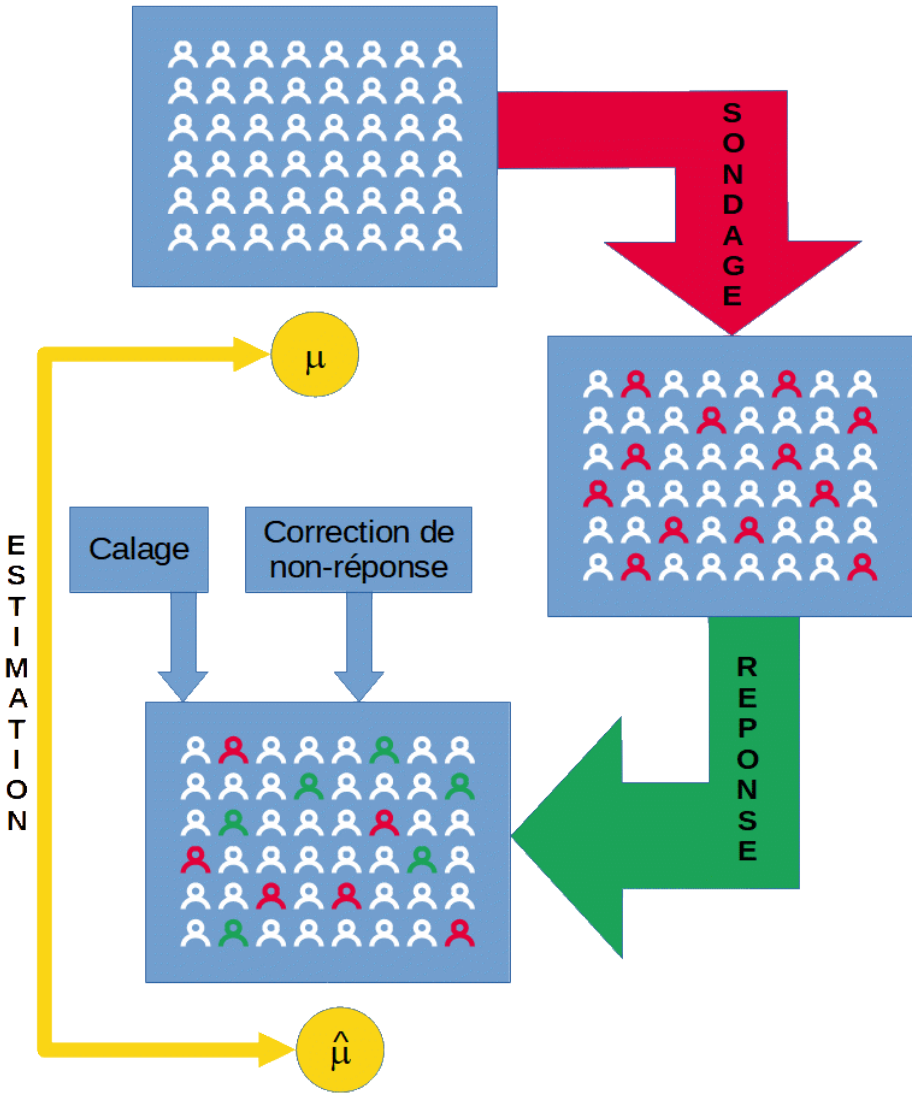
Annexe A7 : Composantes de l'incertitude des estimateurs de prévalence des symptômes

Cette annexe vise à présenter des ordres de grandeur sur les sources d'incertitude des estimateurs de moyenne fondés sur les échantillons observés dans EpiCov. La population répondante résulte des deux étapes de sélection (figure ci-après) : 1) le sondage qui sélectionne les échantillons parmi la population générale et 2) le processus de réponse des enquêtés, certains d'entre eux décidant de ne pas répondre à l'enquête.

Il s'agit, à partir de la population des répondants de calculer un estimateur $\hat{\mu}$ d'une quantité μ qui serait — théoriquement — observée si on était en mesure d'interroger tous les individus composant la population d'intérêt de l'enquête.

La variable aléatoire $\hat{\mu}$ est un estimateur d'Horvitz-Thompson (ou de Hajek) dont le principe théorique est de combiner linéairement les observations réalisées sur les répondants dans le cadre de l'enquête en les pondérant par l'inverse de la probabilité d'inclusion des individus concernés dans la population des répondants. Sous l'hypothèse où l'on connaît cette probabilité, alors $\hat{\mu}$ estime la quantité μ sans biais. Cette estimation est affectée d'une incertitude (variance) qui découle des aléas suivants :

A) l'aléa de sondage
B) l'aléa de la variable de réponse à l'enquête
C) l'aléa associé à l'estimation du modèle de correction de non-réponse et à son utilisation en prédiction des probabilités d'inclusion



Lecture de la figure : On veut connaître la moyenne μ d'une variable sur la population générale (ensemble en haut). On sélectionne aléatoirement, selon un plan de sondage donné, des individus dans la population générale. Connaissant ce plan de sondage, on peut, si tous les individus sélectionnés répondent (personnages en rouge dans l'ensemble du milieu), construire un estimateur sans biais de μ (estimateur d'Horvitz-Thompson, noté $\hat{\mu}$). Cependant, seules certaines personnes sélectionnées répondent effectivement (ensemble en bas, personnages en vert). Comme précédemment, si on connaît la règle de sélection des répondants parmi la population sélectionnée dans le plan de sondage, on peut calculer un estimateur sans biais. Cette nouvelle étape de sélection, contrairement à la précédente, n'est pas aléatoire et doit être inférée grâce aux caractéristiques connues des individus sélectionnés dans le plan. L'analyste construit donc un modèle de non-réponse pour estimer la règle de sélection. Il dispose d'informations supplémentaires sur la population générale qu'il peut aussi utiliser par calage. Si cette information est corrélée avec la variable d'intérêt, alors l'estimateur $\hat{\mu}$ s'en trouvera amélioré (plus précis).

La portée de ces aléas est réduite par les informations externes qui sont mobilisées pour le calcul des poids de l'estimateur, notamment le calage qui consiste à calculer un système de poids, proche des poids initiaux, qui annule la différence entre l'estimateur modifié et des marges connues sur la population générale, notamment grâce aux enquêtes de recensement ou à la base de sondage. La réduction de variance qui découle de l'usage de cette information complémentaire est directement liée à la corrélation qui existe entre la variable à estimer μ et les variables de calage. En l'espèce, les variables de calage sont très peu corrélées aux variables de symptômes collectées dans l'enquête.

Le tableau A7 indique la valeur des écarts-types des estimateurs d'Horvitz-Thompson (ou de Hajek) obtenus sous différentes hypothèses concernant les aléas évoqués ci-dessus.

On note, dans les résultats présentés au tableau A7, le faible écart entre les précisions calculées sous l'hypothèse d'un sondage aléatoire simple (SAS) et sous l'hypothèse d'un sondage stratifié (comparer les lignes 1 et 2 du tableau). Du point de vue de la précision, le plan de sondage est donc un quasi-SAS. Moyennant quoi, l'estimation des précisions, quel que soit l'estimateur fondé sur les données d'enquête, peut être réalisée simplement par Boostrap sur les individus sélectionnés.

La contribution de l'incertitude de modèle à la variance des estimateurs est, pour le modèle Probit de correction de non-réponse, faible et du même ordre de grandeur que l'écart de précision entre une hypothèse de plan de sondage stratifié et l'hypothèse de SAS (comparer les lignes 3 et 2, puis 3 et 1). En revanche, cette contribution du modèle de non-réponse à l'incertitude est dominante dans le cas du modèle d'Heckman de correction de non-réponse : la prise en compte de cette incertitude multiplie par 4 les écarts-types n'en tenant pas compte (comparer les lignes 1 et 4), tandis que l'augmentation n'est que d'environ 30 % dans le cas d'un modèle de correction de non-réponse par Probit (comparer les lignes 1 et 3). La modélisation d'Heckman a donc un coût en termes de précision, mais évidemment, elle permet ici de construire des estimateurs sans biais, tandis qu'ils sont fortement biaisés lorsque la correction de non-réponse est fondée sur un modèle Probit, ce dernier ne permettant pas de prendre en compte la sélection endogène.

Tableau A7 : écarts-types estimés selon les hypothèses d'aléas (en points de %)

	Modélisation *(i.e. aléas pris en compte et forme supposée de ceux-ci)*			Variable	
	A *aléa de sondage*	B *aléa de participation*	C *aléa de modèle de participation utilisé en prédiction des poids*	Toux	Au moins 1 symptôme
(1)	SAS	SAS	Ø	0,07	0,12
(2)	Sondage stratifié et SAS au sein des strates	CNR	Ø	0,09	0,15
(3)	SAS	CNR	Oui (Probit)	0,09	0,14
(4)	SAS	CNR	Oui (Heckman)	0,28	0,54

Lecture : les lettres A, B, C se réfèrent aux trois formes d'aléas cités dans le texte. L'écart-type de l'estimateur de prévalence de la Toux en population générale fondé sur les répondants à l'enquête et sous l'hypothèse d'une sélection en double sondage aléatoire simple (SAS) et pas d'aléa associé à l'estimation de la probabilité de réponse vaut 0,07 points de pourcentage.

Notes : **(1)** : on suppose par ailleurs que la sélection des répondants résulte d'un double SAS pour les deux étapes de sélection (donc globalement d'un SAS). On n'intègre pas ici d'aléa lié au modèle de participation (colonne C) donc les poids de sondages sont uniformes et certains. Le calcul résulte de l'application de la formule de variance pour un SAS appliqué aux effectifs de répondants (voir Tillé, 2019 ; p. 31). On obtient les mêmes résultats en réalisant un calcul de Bootstrap sur l'estimateur de moyenne, avec 1000 boucles sur les répondants. **(2)** : la correction de non-réponse appliquée ici est fondée sur un modèle logistique accompagné d'un traitement par groupe de réponses homogènes et d'un calage sur marge. La variance du plan est cohérent avec le véritable plan choisi (stratifié —voir partie I). Les mêmes résultats sont obtenus par application de la formule de variance de Sen-Yates-Grundy (Tillé, 2019 ; p. 23) avec les poids de sondages normalisés sur les répondants, en application de la formule de Hajek (Tillé, 2019 ; p. 31) et des probabilités d'inclusion doubles égales au produit des probabilités simples (pas de covariances issues du plan). **(3)** estimation par boucle Bootstrap sur l'estimateur de moyenne, avec 1000 boucles sur les sélectionnés dans le plan de sondage et calcul du modèle de correction de non-réponse par Probit, les poids obtenus via le modèle étant d'abord winsorisés pour ceux supérieurs à 10 et ensuite multipliés par les poids de sondage. **(4)** estimation par boucle Bootstrap sur l'estimateur de moyenne, avec 1000 boucles sur les sélectionnés dans le plan de sondage et calcul du modèle de correction de non-réponse par modèle d'Heckman dans chaque boucle, les poids obtenus étant d'abord winsorisés pour ceux supérieurs à 10 et ensuite multipliés par les poids de sondage.

Source : Inserm-Drees, enquête EpiCov, vague 1.

Annexe A8 : Étude de la variable indicatrice de naissance à l'étranger et recommandations dans la mise en œuvre des modèles d'Heckman

La variable indicatrice de naissance à l'étranger est intéressante, dans cette enquête, à plusieurs titres. D'abord elle est disponible dans la base de sondage. Elle est d'ailleurs, pour cette raison, utilisée comme variable de calage[50] dans les Drom (sous la forme d'une variable à trois modalités :

50 Pour être précis, la variable indicatrice est utilisée sous sa forme présente dans la base de sondage dans les DROM. En revanche, en métropole, c'est la réponse à l'enquête, redressée à l'aide de la variable de la base de sondage, qui est

métropole/DOM/étranger —voir paragraphe II.2). Il est donc possible de calculer l'estimateur de proportion de nés à l'étranger pour la population sélectionnée dans l'enquête. Cette valeur s'établit à **14,0 %**, pour un écart-type de sondage de 0,05 points de pourcentage (écart-type pour un sondage aléatoire simple sans remise —voir annexe A7 pour une discussion à ce propos). Par ailleurs, une estimation de la moyenne de cette variable fondée sur les seuls répondants pondérés grâce aux poids de sondage conduit (tableau A8.1) à une estimation fortement biaisée de **10,2 %** (± 0,09). Dans la même configuration, elle est aussi significativement différente selon qu'on la calcule sur les lots **internet** (**10,1** % ± 0,09) ou sur les lots **multimodes** (**10,5 %** ± 0,17). Tout se passe comme si, les nés à l'étranger avaient globalement une propension à répondre plus faible que les autres individus. On peut observer la signature de ce comportement à deux niveaux : d'une part, le protocole qui génère un surcroît de réponse (le protocole multimode) par rapport au protocole alternatif conduit à une estimation de la proportion de nés à l'étranger plus élevée ; d'autre part, l'estimation fondée sur les répondants à l'enquête, sous une hypothèse de non-réponse MAR (*missing-at-random*), est globalement sous-évaluée par rapport à la valeur vraie.

Il est intéressant, dans ces circonstances (biais de sélection apparent et connaissance de la vraie valeur de la moyenne de la variable sur la population d'enquête), d'examiner le comportement des estimateurs de Heckman et des estimateurs alternatifs d'enquête. C'est l'objet de cette annexe.

Le tableau A8.1 montre différentes estimations de la proportion de nés à l'étranger fondées sur les répondants à l'enquête[51], avec différents systèmes de pondérations. Naturellement, le système (ligne 1 en italique du tableau) qui utilise la connaissance *in extenso*[52] de cette variable est spontanément convergents. Ceci est conforme à la théorie qui indique que si la probabilité de réponse est conditionnée par la variable d'intérêt, tandis que celle-ci est connue sur l'ensemble de la population échantillonnée, alors l'estimateur d'Horvitz-Thompson qui en découle est sans biais[53]. Le calage est ici une autre étape de convergence, c'est-à-dire que si cette variable n'était pas utilisée dans le modèle de correction de non-réponse, son utilisation dans le calage sur marge conduirait également à un estimateur non-biaisé[53].

Bien sûr, en général, la variable d'intérêt n'est pas connue sur les non-répondants et ne peut donc pas être utilisée pour la correction de non-réponse ou le calage. C'est la raison pour laquelle, la situation particulière permet ici de simuler ce qu'auraient été les estimateurs, en l'absence de la variable indicatrice de naissance à l'étranger dans la base de sondage, et de comparer ces estimateurs à la valeur vraie.

On sait, dans le cas présent, que les nés à l'étranger ont tendance à sous-participer à l'enquête. Donc, la participation décroît avec la propension à être né à l'étranger. La corrélation de la propension à participer avec celle d'être né à l'étranger devrait donc être négative. C'est ce qu'on observe lorsqu'on estime un modèle de Heckman avec comme explicatives, des variables caractéristiques de l'environnement socio-économique du ménage, mais sans intégrer de caractéristiques individuelles (ligne 5 du tableau A8.1). En revanche, la même régression intégrant des variables individuelles comme explicatives conduit à estimer une corrélation positive (ligne 4 du tableau A8.1). Ces deux coefficients de corrélation estimés sont significativement différents. On observe, en outre, que les estimateurs de moyenne qui découlent de la correction de non-réponse de Heckman intégrant des variables individuelles sont biaisés[54]. A l'inverse, quand on n'intègre pas

utilisée, confrontée aux marges du recensement. Il convient de préciser que la variable présente dans la base de sondage (source fiscale) comporte une modalité « lieu de naissance non renseigné » pour 6% des individus qui sont, pour l'analyse présente, considérés être nés en France. On peut noter que ceci n'est nullement de nature à fausser les résultats de l'analyse conduite dans cette annexe puisque celle-ci est conduite en intégralité sur cette variable binaire, disponible pour **tous** les sélectionnés dans l'enquête.

51 On utilise directement la variable de la base de sondage ; il n'y a donc pas d'erreur de mesure possible dans ce cas.

52 i.e. pour les répondants et pour les non-répondants.

53 asymptotiquement

54 i.e. un test d'égalité de l'estimateur à la vraie valeur est rejeté.

de variables individuelles dans le modèle d'Heckman, les estimateurs ne diffèrent pas significativement de la vraie valeur[55].

Ces observations nous permettent de préciser le cadre d'application des différentes stratégies d'estimation, en les illustrant :

- Dans le cas où le design d'enquête combine des sous-échantillons constitués aléatoirement, certains monomodes et d'autres multimodes, ces derniers donnant lieu à une participation accrue par rapport aux premiers, il est très informatif d'examiner les moyennes de variables, obtenues sur les deux types de sous-échantillons. Différentes hypothèses de correction de non réponse doivent être examinées dans ce cas. Si ces moyennes diffèrent, alors il y a vraisemblablement un phénomène de participation endogène et/ou une erreur modale de mesure. Il faut alors être vigilant et réaliser différents tests, dont des corrections de non réponse par modèles d'Heckman, pour rendre plus robustes les hypothèses sous-jacentes au calcul final. Dans le cas présent, la ligne 2 du tableau A8.1 doit nous alerter (différence des moyennes monomode-multimode quasi-significative).
- Les modèles usuels de correction de non-réponse sur observables posent comme seules hypothèses (i) celle de l'indépendance conditionnelle de la variable d'intérêt et de la participation, conditionnellement aux variables explicatives et (ii) celle de la forme fonctionnelle de la dépendance conditionnelle (Probit, Logit...). En particulier, ces modèles n'exigent pas de postuler l'existence d'une variable latente, donc ils n'imposent pas d'indépendance entre explicatives et aléa qui seraient associés à cette variable latente. Une autre façon de présenter les choses est de noter que ces modèles ne sont utilisés que pour prédire la probabilité d'inclusion. Par conséquent, les coefficients du modèle sont sans importance autre que leur faculté à bien prédire cette probabilité. C'est ainsi qu'à l'extrême, si on dispose de la variable d'intérêt pour l'ensemble de la population sélectionnée, on a intérêt à intégrer cette variable dans le modèle de probabilité d'inclusion et l'estimateur d'Horvitz-Thompson qui en découle est sans biais. On observe bien ce résultat à la ligne (1) du tableau A8.1.
- Le modèle d'Heckman pose l'hypothèse de l'existence de variables latentes et comme hypothèse d'exclusion que les explicatives sont indépendantes des aléas des variables latentes (voir la relation 5 —§ IV.2). Dans la correction de non-réponse, le paramètre de corrélation du modèle joue un rôle essentiel, comme on peut l'observer grâce aux formules de calcul des probabilités d'inclusion (relations 7 et 8 —§ IV.2). Or l'estimation des paramètres du modèle, réalisée par maximum de vraisemblance, est sans biais si les conditions d'exclusion sont effectivement vérifiées, c'est-à-dire que les explicatives sont « exogènes ». En général, et en particulier ici, il est douteux que des variables de caractéristiques individuelles soient réellement indépendantes d'aléas, eux-mêmes propres à l'individu : rien ne permet de le justifier au plan théorique. Au contraire, si des variables sont endogènes dans les explicatives potentielles, c'est évidemment d'abord dans les caractéristiques individuelles qu'on les trouve. Le risque d'endogénéité est beaucoup moins fort pour des caractéristiques décrivant le contexte dans lequel évolue l'individu. Ce contexte, souvent mesuré à l'aide de moyennes sur un ensemble d'habitants du voisinage, est certainement explicatif des variables d'intérêt voire de la participation. En revanche, il est peu probable que l'individu lui-même ait une influence significative sur ces variables. Par conséquent, ces variables sont probablement exogènes, en ce sens qu'elles vérifient très probablement, en tant qu'explicatives du modèle d'Heckman, les conditions d'exclusion (5). C'est précisément ce qu'illustre le calcul du modèle d'Heckman intégrant des caractéristiques individuelles : dans ce calcul, le signe du coefficient de corrélation est manifestement inversé avec la réalité (il devrait être négatif comme indiqué *supra* ; il est ici positif) et la conséquence est que l'estimateur d'Horvitz-Thompson résultant est biaisé

55 Bien que l'estimateur soit relativement éloigné numériquement de la vraie valeur, celle-ci figure bien dans l'intervalle de confiance à 95 % (2 écarts-types). C'est d'ailleurs une des vertus importantes de l'approche par correction de non-réponse fondée sur un modèle de Heckman que de générer des intervalles de confiance réalistes.

(ligne 4, tableau A8.1). La suppression, dans le modèle d'Heckman, de l'ensemble des variables individuelles comme explicatives permet de restaurer une estimation sans biais des paramètres, et en particulier du signe de la corrélation. *In fine*, l'estimateur d'Horvitz-Thompson qui en découle est non biaisé[56, 53] (ligne 5, tableau A8.1).

- S'agissant du cas particulier de la variable indicatrice de naissance à l'étranger, le modèle de correction de non-réponse usuel fondé sur les variables individuelles et de contexte disponibles dans la base de sondage permet d'approcher une estimation sans biais et surtout de corriger de la sélection endogène. En effet, après application de ce modèle, la différence des estimateurs sur les lots monomodes et multimodes est nulle (ligne 3 du tableau A8.1). On note que la correction d'Heckman (ligne 5 du tableau A8.1) produit des estimations de niveaux comparables. On en déduit que la sélection endogène mise en évidence par la différence des estimateurs monomodes et multimodes sur une pondération par les poids de sondages (ligne 2 du tableau A8.1) est, en l'espèce, convenablement prise en compte et corrigée par les différences de caractéristiques individuelles figurant dans la base de sondage. Il n'y a donc pas lieu, ici, d'utiliser un modèle d'Heckman pour corriger de la sélection endogène. En revanche, l'utilisation du modèle d'Heckman permet de valider la bonne prise en compte de la sélection endogène par le modèle usuel et de calculer des intervalles de confiance qui sont davantage conformes avec la vraie valeur : on note en effet que la vraie valeur n'est pas dans l'intervalle de confiance calculé avec le modèle usuel (ligne 3 du tableau A8.1), tandis qu'elle l'est avec celui du modèle d'Heckman (ligne 5 du tableau A8.1). Le gain en précision qu'on pourrait avancer pour justifier d'utiliser le modèle usuel plutôt qu'un modèle d'Heckman n'épuise pas le sujet : cet argument de meilleure précision du modèle usuel peut même fausser l'inférence puisqu'*in fine*, la vraie valeur n'est pas dans l'intervalle associé et donc la précision estimée est en partie illusoire. L'utilisation du modèle d'Heckman permet donc ici d'étudier la précision effective, même si elle n'est pas recommandable pour le calcul de l'estimateur proprement dit.
- Comme évoqué à la note de bas de page n°39, en pratique, les estimateurs portent sur la fraction de la population d'enquête dont la probabilité de participer n'est pas nulle. En l'absence de sélection endogène, l'écart observé entre la vraie valeur et l'estimateur usuel (ligne 3 du tableau A8.1) pourrait s'expliquer par l'existence d'une sous-population ayant une probabilité nulle de participer à l'enquête. Dans l'exemple proposé, cela reviendrait à considérer que certains nés à l'étranger ont une probabilité nulle de participer à l'enquête, ce qui n'est pas, en soi, impossible. *De facto*, l'estimateur de Heckman est, théoriquement exposé au même biais : si une fraction de la population a une probabilité nulle de participer, alors le modèle de Heckman ne peut pas redresser l'information pour les individus dont la probabilité de participer serait nulle. Le mal est même plus profond dans ce cas puisque la forme fonctionnelle du modèle ne permet pas non plus de traiter le cas où la probabilité nulle de participer serait liée au mode. Cela étant, dans le cas qui nous intéresse ici, le fait que la valeur vraie de la proportion de nés à l'étranger est dans l'intervalle de confiance de l'estimateur fondé sur le modèle d'Heckman, tandis qu'elle est hors de l'intervalle de confiance de l'estimateur usuel suggère plutôt :
  i. l'absence de biais de couverture (pas de sous-population ayant une probabilité nulle de participer à l'enquête) ;
  ii. une sélection endogène dont la variance de modèle pour la corriger n'est pas estimée convenablement par la formule de variance de l'estimateur usuel.

56 bien qu'incertain.

Tableau A8.1 : Estimation de la proportion de nés à l'étranger avec différents systèmes de pondération, pour les lots monomodes et multimodes et pour l'ensemble

	Pondération	Lots multimodes	Lots monomodes	Ensemble	ρ *(modèle d'Heckman)*
(1)	*Poids calés*	*13,9 (0,26)*	*13,9 (0,15)*	*13,9 (0,13)*	
(2)	Poids de sondage	10,5 (0,17)	10,1 (0,09)	10,2 (0,09)	
(3)	Poids CNR Probit non calés	12,2 (0,22)	12,3 (0,13)	12,2 (0,11)	
(4)	Poids CNR Heckman incluant des explicatives individuelles	10,9 (2,10)	10,9 (1,21)	10,9 (1,16)	0,080 (0,051)
(5)	Poids CNR Heckman sans explicatives individuelles	12,2 (1,89)	12,2 (1,00)	12,2 (0,97)	-0,082 (0,048)
(6)	Poids CNR Probit non calés et sans explicatives individuelles	10,9 (0,18)	10,6 (0,10)	10,6 (0,09)	

Lecture : CNR signifie « correction de non-réponse ».

Notes : **(1)** les poids calés sont issus du modèle de non-réponse (logistique), des GRH et du calage sur marge tel que décrit au paragraphe II.2. Ce modèle intègre, dès la CNR logistique la variable indicatrice de naissance à l'étranger comme explicative. Donc la probabilité d'inclusion est sans biais, ce qu'on observe ici. Le calage (la variable de naissance à l'étranger est, elle aussi, intégrée dans les variables de calage —la liste complète des variables de calage est indiquée au paragraphe II.2) rajoute une étape supplémentaire de convergence vers la valeur exacte. En général, ce calcul est *impossible* pour les variables d'enquêtes puisqu'on ne les observe que pour les répondants. Donc la vraie valeur n'est pas connue et la variable n'intervient pas dans les explicatives du modèle ou dans les marges. **(3)** La variable indicatrice de naissance à l'étranger est ici retirée des explicatives (voir le (a) du paragraphe III.2 pour une liste complète des autres variables explicatives intégrées dans le modèle de non-réponse). Pas de GRH ni de calage sur marge. **(4)** Hors la variable indicatrice de naissance à l'étranger (voir le (a) du paragraphe III.2 pour la liste complète). Le coefficient de corrélation estimé est significativement non nul à 10 %. **(5)** Les variables intégrées dans le modèle sont : la taille d'unité urbaine, l'indicatrice de quartier prioritaire de la ville (QPV), le taux local d'hospitalisation (en tranches), l'indicateur communal d'accessibilité aux médecins généralistes (APL), la densité en tranches issue de la grille communale de densité. Le coefficient de corrélation estimé est significativement non nul à 10 %. **(6)** Les variables explicatives utilisées sont les mêmes que celles utilisées pour la régression (5).

Les poids des lignes 3 à 6 sont ceux issus des modèles de non-réponse multipliés par les poids de sondages, conformément aux notes 3 et 4 du tableau A7.

Source : Inserm-Drees, enquête EpiCov, vague 1.

En complément, le tableau A8.2 montre les résultats du modèle d'Heckman sur quelques symptômes avec un modèle, tel que présenté au paragraphe IV.2, utilisant des explicatives individuelles et contextuelles puis, un second modèle n'utilisant que des explicatives contextuelles. Les estimateurs ne diffèrent pas significativement. Cependant, on note que le niveau de corrélation entre la participation et la variable d'intérêt est réduit dans le cas où le modèle d'Heckman intègre des variables individuelles. C'était aussi le cas, en valeur absolue, pour la proportion de nés à l'étranger. *In fine*, les estimateurs de prévalence de symptômes sont plus faibles encore après correction d'Heckman sans variables individuelles qu'avec des variables individuelles. Comme on l'a vu précédemment, ce sont plutôt ces derniers résultats qu'il conviendrait de privilégier. Les résultats du texte principal ont néanmoins été calculés avec des explicatives individuelles, les écarts n'étant, en l'espèce, pas significatifs au regard des incertitudes calculées.

Tableau A8.2 : Estimation de la prévalence de certains symptômes selon qu'on intègre ou pas les explicatives individuelles dans le modèle d'Heckman de correction de la non-réponse

Symptôme	calcul	Prévalence (en%)	ρ *(modèle d'Heckman)*
Perte de goût ou d'odorat	(a)	2,0 (0,28)	0,14* (0,08)
	(b)	1,8 (0,30)	0,18* (0.08)
Toux	(a)	4,8 (0,28)	0,38*** (0,05)
	(b)	4,0 (0,31)	0,50*** (0,06)
Au moins 1 symptôme	(a)	15,6 (0,54)	0,44*** (0,03)
	(b)	14,2 (0,55)	0,53*** (0,03)

Lecture : Significativité au seuil de *** 1 %; ** 5 %; * 10 %. Les écarts-type sont indiqués entre parenthèses.

Notes : **(a)** modèle d'Heckman avec variables explicatives individuelles et contextuelles (la liste est indiquée au (a) du paragraphe III.2) ; **(b)** modèle d'Heckman sans variables explicatives individuelles.

Source : Inserm-Drees, enquête EpiCov, vague 1.

Série des Documents de Travail « Méthodologie Statistique »

9601 : Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population.
G. DECAUDIN, J.-C. LABAT

9602 : Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises.
N. CARON, P. RAVALET, O. SAUTORY

9603 : La procédure FREQ de SAS - Tests d'indépendance et mesures d'association dans un tableau de contingence.
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : Les principales techniques de correction de la non-réponse et les modèles associés.
N. CARON

9605 : L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration.
P. RAVALET

9606 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT).
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes.
N. CARON, D. LE BLANC

9701 : Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?
J.-C. DEVILLE

9702 : Modèles univariés et modèles de durée sur données individuelles.
S. LOLLIVIER

9703 : Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises.
N. CARON, J.-C. DEVILLE

9704 : La faisabilité d'une enquête auprès des ménages.
1. au mois d'août.
2. à un rythme hebdomadaire

C. LAGARENNE, C. THIESSET

9705 : Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine.
P. GIRARD.

9801 : Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS.
K. ATTAL-TOUBERT, D. LADIRAY

9802 : Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation.
J.-C. DEVILLE

9803 : Pour essayer d'en finir avec l'individu Kish.
J.-C. DEVILLE

9804 : Une nouvelle (encore une !) méthode de tirage à probabilités inégales.
J.-C. DEVILLE

9805 : Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish.
J.-C. DEVILLE

9806 : Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE.
N. CARON, J.-C. DEVILLE, O. SAUTORY

9807 : Estimation de données régionales à l'aide de techniques d'analyse multidimentionnelle.
K. ATTAL-TOUBERT, O. SAUTORY

9808 : Matrices de mobilité et calcul de la précision associée.
N. CARON, C. CHAMBAZ

9809 : Échantillonnage et stratification : une étude empirique des gains de précision.
J. LE GUENNEC

9810 : Le Kish : les problèmes de réalisation du tirage et de son extrapolation.
C. BERTHIER, N. CARON, B. NEROS

9901 : Perte de précision liée au tirage d'un ou plusieurs individus Kish.
N. CARON

9902 : Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.
N. CARON

0001 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT) (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER

0002 : Modèles structurels et variables explicatives endogènes.
J.-M. ROBIN

0003 : L'enquête 1997-1998 sur le devenir des personnes sorties du RMI - Une présentation de son déroulement.
D. ENEAU, D. GUILLEMOT

0004 : Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables.
O. GODECHOT

0005 : Estimation dans les enquêtes répétées : application à l'Enquête Emploi en Continu.
N. CARON, P. RAVALET

0006 : Non-parametric approach to the cost-of-living index.
F. MAGNIEN, J. POUGNARD

0101 : Diverses macros SAS : Analyse exploratoire des données, Analyse des séries temporelles.
D. LADIRAY

0102 : Économétrie linéaire des panels : une introduction.
T. MAGNAC

0201 : Application des méthodes de calages à l'enquête EAE-Commerce.
N. CARON

C 0201 : Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation.
L. ARRONDEL, A. MASSON, D. VERGER

C 0202 : Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002.
J.-A. VALLET, G. BONNET, J.-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA

0203 : General principles for data editing in business surveys and how to optimise it.
P. RIVIERE

0301 : Les modèles logit polytomiques non ordonnés : théories et applications.
C. AFSA ESSAFI

0401 : Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques.
V. COHEN, C. DEMMER

0402 : La macro SAS CUBE d'échantillonnage équilibré
S. ROUSSEAU, F. TARDIEU

0501 : Correction de la non-réponse et calage de l'enquêtes Santé 2002
N. CARON, S. ROUSSEAU

0502 : Correction de la non-réponse par répondération et par imputation
N. CARON

0503 : Introduction à la pratique des indices statistiques - notes de cours
J-P BERTHIER

0601 : La difficile mesure des pratiques dans le domaine du sport et de la culture - bilan d'une opération méthodologique
C. LANDRE, D. VERGER

0801 : Rapport du groupe de réflexion sur la qualité des enquêtes auprès des ménages
D. VERGER

M2013/01 : La régression quantile en pratique
P. GIVORD,
X. D'HAULTFOEUILLE

M2014/01 : La microsimulation dynamique : principes généraux et exemples en langage R
D. BLANCHET

M2015/01 : la collecte multimode et le paradigme de l'erreur d'enquête totale
T. RAZAFINDROVONA

M2015/02 : Les méthodes de Pseudo-Panel
M. GUILLERM

M2015/03 : Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse
E. GROS
K. MOUSSALAM

M2016/01 : Le modèle Logit Théorie et application.
C. AFSA

M2016/02 : Les méthodes d'estimation de la précision de l'Enquête Emploi en Continu
E. GROS
K.MOUSSALAM

M2016/03 : Exploitation de l'enquête expérimentale Vols, violence et sécurité.
T. RAZAFINDROVONA

M2016/04 : Savoir compter, savoir coder. Bonnes pratiques du statisticien en programmation.
E. L'HOUR
R. LE SAOUT
B. ROUPPERT

M2016/05 : Les modèles multiniveaux
P. GIVORD
M. GUILLERM

M2016/06 : Econométrie spatiale : une introduction pratique
P. GIVORD
R. LE SAOUT

M2016/07 : La gestion de la confidentialité pour les données individuelles
M. BERGEAT

M2016/08 : Exploitation de l'enquête expérimentale Logement internet-papier
T. RAZAFINDROVONA

M2017/01: Exploitation de l'enquête expérimentale Qualité de vie au travail
T. RAZAFINDROVONA

M2018/01: Estimation avec le score de propension sous R
S. QUANTIN

M2018/02: Modèles semi-paramétriques de survie en temps continu sous R
S. QUANTIN

M2019/01 : Les méthodes de décomposition appliquées à l'analyse des inégalités
B. BOUTCHENIK
E. COUDIN
S. MAILLARD

M2020/01 : L'économétrie en grande dimension
J. L'HOUR

M2021/01 : R Tools for JDemetra+ - Seasonal adjustment made easier
A. SMYK
A. TCHANG

M2021/02 : Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman
L. CASTELL
P. SILLARD

M2021/03 : Conception de questionnaires auto-administrés
H. KOUMARIANOS
A. SCHREIBER

M2022/01 : Introduction à la géomatique pour le statisticien : quelques concepts et outils innovants de gestion, traitement et diffusion de l'information spatiale
F. SEMECURBE
E. COUDIN

M2022/02 : Le zonage en unites urbaines 2020
V. COSTEMALLE
S. OUJIA
C. GUILLO
A. CHAUVET

M2023/01 : Les réseaux de neurones appliqués à la statistique publique : méthodes et cas d'usages
D. BABET
Q. DELTOUR
T. FARIA
S. HIMPENS

M2023/02 : Redressements de la première vague de l'enquête epicov : un exemple de correction des effets de sélection dans les enquêtes multimodes
L.CASTELL
C. FAVRE-MARTINOZ
N. PALIOD
P. SILLARD