jms Journées de méthodologie statistique de l'Insee 2022

EPICOV: A LARGE NATIONAL POPULATION-BASED COHORT IN COVID-19 TIMES - METHODOLOGICAL ISSUES

Josiane WARSZAWSKI (*), Guillaume BAGEIN (**), Thomas DEROYON (**), Nathalie BAJOS (***) and the EpiCov study group[1]

(*) INSERM CESP U1018, Université Paris-Saclay, Le Kremlin-Bicêtre, France

(**) DREES, bureau Etat de santé de la population

(***) IRIS, INSERM, EHESS, CNRS Aubervilliers, France

guillaume.bagein@sante.gouv.fr

Key words: cohort survey, Covid-19, data collection protocol, mixed-mode survey, home sampling kits

Mots-clés : enquête de cohorte, Covid-19, protocole de collecte, enquête multimode, kits d'autoprélèvement

Domaines concernés : 4 – Collecte ; 14 – Mesures et impact de la pandémie de Covid-19

Abstract:

This article aims to describe the protocol of the three first waves of the EpiCov cohort, initiated in the context of the French national lockdown in March 2020. The three first rounds occurred in May 2020, November 2020 and June 2021. EpiCov had general objectives to study diffusion of Covid-19 virus in the population and to study relations between the epidemic, health and living conditions. This population-based cohort combines a questionnaire survey with Covid-19 serological tests performed on home blood self-samples of respondents. The initial sample of 371 000 individuals was divided into 20 sub-samples called "batches", to allow more flexibility in the protocol design and for the monitoring of the fieldwork phases. Data collection is still ongoing, as a fourth round and further enrichments of the collected data are planned. Some of EpiCov's main strengths and limitations are discussed.

Résumé en français :

La cohorte EpiCov (Épidémiologie et conditions de vie au temps du Covid-19) est une enquête nationale menée par l'Inserm et la Drees avec une forte collaboration de l'Insee et de Santé publique France, s'appuyant sur un échantillon initial de 371 000 personnes. Trois vagues d'enquête ont déjà eu lieu (mai 2020, novembre 2020 et juin 2021), et une quatrième est actuellement en préparation. Ce rythme

[1] EpiCov study group: J. Warszawski and N. Bajos (co-principal investigators), G. Bagein, F. Beck, E. Counil, F. Jusot, Nathalie Lydie, C. Martin, L. Meyer, P. Raynaud, A. Rouquette, A. Pailhé, D. Rahib, P. Sillard, R. Slama, A. Spire

intense de préparation et d'exploitation d'une enquête de cette ampleur a nécessité un travail très important de la part de toutes les personnes impliquées (recherche publique, statistique publique, prestataire de collecte), et n'a été possible que grâce à des procédures accélérées mises en place par les autorités de contrôle compétentes (CNIS, CPP, CNIL, …).

Les données collectées sur les participants proviennent de questionnaires administrés par des enquêteurs téléphoniques ou par internet, ainsi que des analyses sérologiques visant à attester la présence d'anticorps dirigés contre le virus SARS-CoV-2 à partir d'échantillons sanguins prélevés par les participants eux-mêmes. Une division de l'échantillon en 20 "lots" dès la phase de tirage permet de moduler le protocole de collecte de façon fine, tout en apportant des garanties sur la comparabilité des lots entre eux ; la taille de l'échantillon total permet par ailleurs même au sein de chaque lot de disposer d'un nombre de répondants suffisamment conséquent pour permettre des analyses assez détaillées.

Le taux de réponse en vague 1 a été de 36 %, avec des différences significatives notamment selon les modes de collecte employés (multimode CATI-CAWI ou monomode CAWI), selon les caractéristiques socio-démographiques des individus (sexe, niveau de revenu…), ou encore entre la métropole et les trois Drom (Martinique, Guadeloupe, La Réunion) inclus dans le champ de l'enquête. Les taux de réponse des vagues suivantes se situent à des taux plus élevés, de l'ordre de 80%, mais les mêmes effets de sélection semblent se dessiner.

Les kits permettant de réaliser des analyses sérologiques ont été proposés à une partie des répondants en vague 1, à l'ensemble des répondants de vague 2, ainsi qu'aux cohabitants âgés de 6 ans ou plus d'une partie des répondants de vague 2. Ces analyses ne portent pas sur les infections au SARS-CoV-2 au moment du test, mais plutôt dans les mois passés. Elles ont permis notamment de fournir les premières estimations de prévalence de la maladie à partir d'un échantillon représentatif de la population à la fin de la première vague de l'épidémie.

Avec la vaste palette de sujets abordés par l'enquête, la vague 4 en préparation, et grâce à plusieurs enrichissements prévus pour la cohorte (données de profession, de revenus, du SNDS), nombre de sujets, EpiCov constitue une source extrêmement féconde pour la recherche et la statistique publique, autant d'un point de vue méthodologique que pour les thématiques abordées elles-mêmes.

Introduction:

The Covid-19 pandemic began in Europe at the start of 2020. It immediately generated enormous pressure on the health system, and demands for continual information updates from public authorities and other stakeholders. In the absence of vaccination and specific treatment options, public health interventions to curb the spread of the pandemic were launched worldwide [1, 2]. Each country reacted in its own way, at different speeds and with *ad hoc* solutions, based, in particular, on stay-at-home orders, rules for social distancing, the use of personal protective equipment, the isolation of individuals with confirmed infection, the quarantine of their contacts, border restrictions and total or partial lockdowns. The need to follow the evolution of the pandemic, and its influence on living conditions, with sufficient precision and on a fine geographical scale, was common to all countries.

Surveillance systems were set up to estimate the temporal dynamics of the SARS-CoV-2 virus, with a monitoring of the number of hospitalizations, deaths, and positive virology and serology tests in the population, mostly based on data from medical structures, or from repeated cross-sectional studies in various selected populations, such as blood donors or healthcare professionals [3, 4, 5]. Studies on seroprevalence[2], based on SARS-CoV-2 antibody tests, have been recommended as a means of estimating the cumulative incidence[3] of COVID-19, the disease caused by SARS-CoV-2 [6]. Random sampling of the target population remains the gold standard for achieving representativeness [7]. However, this approach also requires extensive resources and commitment from the community to be implemented in the general population, and few SARS-Cov2 seroprevalence surveys based on probability samples have been conducted in the general population at national or territorial level [8, 9, 10, 11, 12, 13].

Discussions about a national representative population-based cohort designed to monitor the consequences of the pandemic were initiated at the start of the first French national lockdown in March 2020. Its first objective was to provide an initial point estimate of the seroprevalence of SARS-CoV-2 and a precise description of the effects of living conditions on health and of the epidemic on living conditions, at both national and local (*département*, equivalent to a county and forming the NUTS-3 level in Eurostat nomenclature for metropolitan France) levels [14]. The study was also designed to provide reliable data for the various social groups, including hard-to-reach subgroups, such as people in precarious situations, and to be repeated at different time points.

Time between the first ideas and the start of the first round of the population-based EpiCov cohort, on May 2, 2020 was less than two months. The study was conducted by the National Institute for Health and Medical Research (Inserm) and the French Ministry of Health and Solidarities statistics centre (DREES), in collaboration with the French national institute of statistics (Institut national des statistiques et des études économiques) and the French national public health agency (Santé publique France). It was based on a large random sample of people living in France (371 000 individuals, among whom 350 000 in metropolitan France and 7 000 in three of the French overseas *départements*). Initially, the main objectives were to estimate the immunity status both nationally and locally and in various subpopulations, including populations with socioeconomic deprivation, to study the intra-household circulation of the virus, to describe the situation of the population in the context of the national lockdown. In addition to a detailed questionnaire, self-administered home blood collection kits (including a prick, a blotting paper intended to collect blood drops, and a return envelope) were offered to the participants to perform Covid-19 serological tests.

The perspective of this cohort also was to describe changes in health and living conditions in relation to the Covid-19 pandemic in France. The follow-up rounds took place in November 2020 and June 2021, and one last round is under preparation for May 2022.

[2] *Seroprevalence* is the proportion of a population to have antibodies against Covid-19

[3] *Cumulative incidence* is the proportion of a population at risk to have developed in a certain period of time the studied outcome; in this context, the proportion of a population to have been infected in a given period.

This paper aims to provide an extensive description of the methodology of this cohort, including the context of the elaboration of the sample design and data collection, the corrections for nonresponse bias, and the fieldwork, and to set out both the strengths and the limitations of its design. It is a continuation of a first article describing solely the first round of the survey [15]. We will then describe the sampling design for each round, as well as the protocol for the fieldwork and discuss its results. Finally we will present the perspectives for the fourth round, two years after the starting pandemy, also including possibilities for further use of the data by matching the survey data to several administrative databases. Round 4 sets out to assess and correct selection and information biases by questioning not only the respondents from round 3 but all the individuals who were initially included in the sample.

1. Elaboration of EpiCov according to the initial context and subsequent changes

In early 2020, the rapid spread of the coronavirus pandemic led to a need for information systems to monitor the epidemic and guide strategies of control. Two main approaches were combined: adapting existing epidemiological surveillance systems, and designing new data sources or ad hoc studies.

In France, the SiVic database, originally developed to monitor the health follow-up of terrorism victims, was switched to monitor Covid-related hospitalizations, while two other registries were set up to monitor Covid-19 virology tests (Si-Dep) and later Covid-19 vaccinations (Vac-Si)[4].

Several studies have suggested geographical and social heterogeneity in Covid-19 mortality and hospitalization. As most people had little access to virology diagnostic tests, and as those tested were likely to be at higher risk of being infected, two population-based cohorts were developed during the first epidemic wave, including systematic serological testing on a large number of respondents, with two socio-epidemiological objectives: the SAPRIS and EpiCov cohorts.

SAPRIS was designed in March 2020 by pooling four pre existing national cohorts (Constances, Elfe-EpiPage, E3N-E4N and Nutrinet santé) resulting in a survey base of 600 000 individuals, who were already used to answering questionnaires about their health, and therefore more likely to participate in new surveys. Sapris had the particular advantage of including participants for whom a wealth of data on medical history and biological biobanks was already available [16].

The EpiCov Cohort was designed to provide national indicators during the crisis, with repeated rounds to study the evolutions, with sufficient power to produce local estimations and maintain representativeness over time. The first-round questionnaire was adapted from the Sapris questionnaire.

EpiCov is one of the largest samples among French surveys with 371 000 people selected at random, and among them 134 000 participated in the first-round of the survey. This sample size makes it complementary to many other smaller and more flexible panel surveys that have been conducted since March 2020. An example of this is CoviPrev[5], which was also launched during the March 2020 lockdown. It conducted 31 study rounds between March 2020 and January 2022, providing highly frequent updates but at the cost of lower statistical quality both in terms of numbers of participants and representativeness of the sample.

1.1. Regulatory aspects and quality assessments of the EpiCov cohort

Although the time between the first discussions about the EpiCov survey and the start of the first round was very short, authorization was obtained from all the appropriate regulatory committees for public

[4] See for instance: https://drees.solidarites-sante.gouv.fr/sources-outils-et-enquetes/les-appariements-si-vic-si-dep-et-vac-si

[5] See for instance: https://www.santepubliquefrance.fr/etudes-et-enquetes/coviprev-une-enquete-pour-suivre-l-evolution-des-comportements-et-de-la-sante-mentale-pendant-l-epidemie-de-covid-19

statistics and biomedical research, which greatly accelerated their procedures to ensure rapid processing and implementation of the study[6]. Even with the accelerated process, the requirements for obtaining ethics and data protection authorizations were maintained. However, the Comité du Label could not provide a notification of statistical quality according to their usual standards, as an appropriate test of the questionnaire could not be conducted before the committee examined the protocol.

1.2. Themes studied over time

As many unexpected pandemic changes occurred with time, including new COVID-19 waves, new variants and availability of vaccines, the relative importance of the different themes explored in EpiCov evolved and new issues emerged.

In round 1 (May 2021), the questionnaire focused on a variety of themes. For one tenth of the sample (see below for a more precise description of the selection criterion), the questionnaire also included additional modules with no local objective in the first round, forming what was called the EpiCov long questionnaire. First demographic and socioeconomic characteristics were reported. Health status was described from general questions on health status, self-perceived health, Covid-19 like symptoms, and access to healthcare during lockdown. The employment situation and the working conditions were then studied, with specific questions to determine whether the worker actually went to work, worked from home, or stopped working during lockdown. Questions on household organization, on children in the household, and about relationships with partners followed, only in the long questionnaire. All respondents were asked questions about their compliance with health regulations and about their consumption of alcohol and smoking behaviors. Mental health was a topic that was restricted to the long questionnaire, as were questions regarding trust in the institutions. Finally, questions about the possible migratory background of the respondent's parents and about the respondent's use of Internet and telephone (for methodological purposes) ended the questionnaire.

In round 2 (November 2021), the overall questionnaire structure and most questions remained the same. Three major updates were included. First, a description of all household members living with the respondent was conducted, starting with a list of these people and detailing a few questions for each of them (gender, age, relationship with the respondent, and other questions for the subjects belonging to the "household subsample"). Second, an extensive description of the professions was added, allowing to replace the respondents in the standardized occupational coding nomenclature (PCS 2020), thanks to the French national institute of statistics. Third, questions on vaccination reluctance were asked.

In round 3 (July 2022), a new theme concerned the respondents' vaccination status , their willingness to get vaccinated, their motives to receive or refuse vaccination or oppose it, and their representations towards some controversies regarding vaccines. Mental health was also broached more fully. Only the PHQ-9 depression scale was used in the previous rounds, as well as suicidal thoughts and attempts in round 2. Here there were other questions, such as questions about anxiety, eating disorders, and social and psychological support. Questionnaires on psychosocial strengths and difficulties among children were also included, providing insights into children's mental state more than one year after the pandemic outbreak. On the other hand, parts of the long questionnaire in round 1 and 2 were not repeated, and physical health was less detailed.

[6] A notice of opportuneness from the National Council for Statistical Information (CNIS) was obtained on April 17, 2020, and a notice of review from the Comité du label de la statistique publique on April 21, 2020. The EpiCov cohort protocol was also approved by the CPP ("Comité de Protection des Personnes", the French equivalent of the Research Ethics Committee) on April 24, 2020. The CNIL (Commission nationale de l'informatique et des libertés, the French independent administrative authority responsible for data protection) authorized the study on April 25, 2020.

Finally, round 4, planned for May 2022, will explore the themes common to rounds 2 and 3, with more extensive data to explore "long" Covid-19.

1.3. Covid-19 serological tests

Serological tests are not intended for the detection of current infection, which is assessed by virology tests (a RT-PCR test or less frequently by an antigen test), with a short period of positivity. Serology tests detect the presence of SARS-Cov2 antibodies, which appear in the blood within 3 weeks after infection, and reflect a past contact with the virus.

The choice for a centralized analysis of blood samples obtained thanks to home self-sample kits in order to detect SARS CoV-2 antibodies was made after the experience of a 2016 study on seroprevalence for HIV and hepatitis B and C based on a random telephone survey in general population in France [17]. This was the only feasible solution for such a research in the context induced by national lockdown, as sending participants to medical analysis laboratories to perform tests was not an option.

Serological tests were limited to a national metropolitan subsample in the first round of the EpiCov survey, conducted at the end of the first lockdown (May 2020), when the supply of materials for kits and laboratory tests was limited.

All respondents to the second round of EpiCov, which took place in November 2020, were eligible for a serological test. In addition, in order to evaluate intra-household transmission and estimate the seroprevalence among children and adolescents in autumn 2020, serological tests were offered for all members aged 6 years or more in a subsample of index participants.

2. Sampling protocol

2.1. Sampling frame

The sampling frame was the French database called Fidéli (Housing and individual demographic files) [18], a comprehensive database obtained by merging several administrative tax databases. Administrative tax data have been used since 2009 as a sampling frame for the Labor Force Survey. Fidéli was developed to enhance the quality of this taxation data for statistical purposes and is now used by the French National statistical Institute Insee as its sampling frame for all household surveys and its Labor Force surveys starting from 2021 [19]. These databases are updated yearly from annual tax returns, and are monitored by the National Institute of Economics and Statistics (Insee), to eliminate duplicates and to identify community housing facilities (nursing homes, prisons, military barracks, etc.) and residential hotels for separate treatment. The Fidéli database compiled in 2018 was used for EpiCov.

All dwellings included in Fidéli are associated with a postal address, which can be used to contact the individuals sampled. Additional modes of contact are possible, because at least an email address, a landline or a mobile phone number was available for 83% of the dwellings in 2018 (one telephone number for at least 69%, one mobile phone number for at least 45%, and one e-mail address for at least 71% in mainland France).

Fidéli includes a wide range of relevant auxiliary information at individual and household level. This information is useful for stratification purposes and for the subsequent correction of non-response bias after data collection.

2.2. Target population

The target population consisted of all individuals aged 15 years or older on January 1, 2020, living in mainland France or one of three overseas *départements* (Martinique, Guadeloupe and Réunion Island).

Because of the poor quality of the sampling frame, poor internet access and the need to translate the questionnaire into numerous languages to ensure comprehension by all potential respondents, two other overseas departments, French Guiana and Mayotte, were excluded from the study.

We also excluded individuals living in prisons at the time of the study, and people living in residential institutions for the dependent elderly, as caregivers were not available during the epidemic period to help them with Internet access or phone calls.

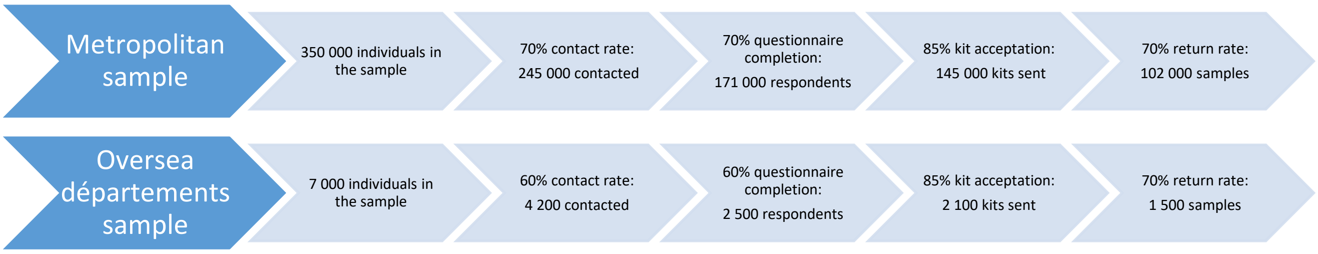
2.3. Sample size

The sample size was calculated to ensure sufficient precision for the seroprevalence estimate, with the objective of obtaining a 95% confidence interval of 2 points for a prevalence of 5% in administrative subdivisions of 600 000 inhabitants. This required the collection of at least 500 blood samples from individuals in each of the 96 *départements* in mainland France.

The parameters used for the calculation of the sample size were the following. The expected rate of contact through the Internet or by telephone was 70% (60% in overseas territories) of the initial sample from Fidéli. Among the individuals contacted, 70% were expected to complete the questionnaire (60% in overseas territories), resulting in an expected complete response rate of 49% in metropolitan France and 36% in the overseas *départements*. 85% of the respondents were expected to agree to receive the home sampling kit, with 70% of these individuals effectively posting a dried blood sample to the biobank. This resulted in an overall expected return rate of blood samples of 60% among the respondents selected.

Based on these parameters, 350 000 individuals were randomly selected from mainland France, in order to obtain 170 000 respondents for the questionnaire, and 100 000 participants tested. Concerning the overseas departments, 7 000 individuals from each of the three overseas departments were included in the sample, with the expectation of retrieving 2 500 questionnaires and 1 500 blood samples. These hypotheses are presented in Figure 1.

Figure 1: Hypotheses used for EpiCov sample size calculation


Metropolitan sample
350 000 individuals in the sample
70% contact rate: 245 000 contacted
70% questionnaire completion: 171 000 respondents
85% kit acceptation: 145 000 kits sent
70% return rate: 102 000 samples
Oversea départements sample
7 000 individuals in the sample
60% contact rate: 4 200 contacted
60% questionnaire completion: 2 500 respondents
85% kit acceptation: 2 100 kits sent
70% return rate: 1 500 samples

2.4. Sampling design

Eligible individuals were selected with a stratified systematic sampling design, with stratification according to two criteria: administrative area (*départements* in mainland France and three overseas), and a binary indicator of poverty, defined as living over or under a threshold of 60% of the median national per capita household income.

The *département* allocation was linked to the population size of the *département* itself, but included an overrepresentation of the least populated *départements*, to ensure that there were at least 900 respondents in each, assuming a response rate of 50% among those selected. Individuals living in a household below the poverty threshold were overrepresented in mainland France[7], constituting 20%

[7] In the three overseas *départements*, the proportion of the population living below the poverty threshold was considered to be high enough to ensure that this population would be represented even without an overrepresentation in the sample.

of the sample rather than the 13% of the Fidéli sampling frame, as a lower response rate was expected for this subpopulation.

Inside each sampling stratum, the sampling frame was also sorted by urban subdivisions, municipality, household income level, and the identification numbers of the dwelling and the individuals. This systematic sampling process ensured an implicit stratification for these variables, and prevented the selection of two individuals from the same household.

The overall sample was divided into 20 subsamples of 18 550 individuals (17 500 individuals in metropolitan France, 350 in Martinique, Guadeloupe and La Réunion), according to the same sampling design used to select the whole sample, called “batches”, separately for French mainland and overseas *départements*, each with the stratum allocation equal to that of the overall sample divided by 20. This ensures that each stratum represents the same share of the whole sample in the complete sample and in the batch samples. This method was chosen as a flexible means of selecting subsamples for specific purposes.

2.5. Differences across batches: collection modes, questionnaires and serology sampling

In EpiCov rounds 1 and 2, the overall response burden and the cost of telephone calls and response collection were limited by using two versions of the questionnaire. A short version (mean duration: 26 minutes) was proposed for 90% of the sample (18 of the 20 batches). A longer version (mean duration: 34 minutes, including all the questions in the short version) was administered to 10% of the EpiCov participants (the two remaining batches). Local representativeness was not an objective for the longer questionnaire. In round 2, the questionnaire duration for the short questionnaire remained at 26 minutes and the long questionnaire reached a mean duration of 36 minutes.

In round 3, there were no distinctions between the short and long questionnaires. However, parts of the long questionnaire in round 1 and 2 were not repeated, and physical health was less detailed.

Both the long and short versions were implemented as self-completed questionnaires, through the computer-assisted web interview system (CAWI), or were administered by qualified and supervised professional interviewers via a computer-assisted telephone interview system (CATI). As the context of national lockdown during round 1 limited the number of investigators that could be mobilized at the same time, the batches also enabled the control of how the investigators' efforts were allocated: only batches 1 to 4 were called during this round, and round 4 only was called after two weeks, in order to make sure that the maximum effort would be made for every individual called.

In rounds 1 and 2, the batches also helped to design the sampling process for serological tests. In round 1, the medical labs working with EpiCov did not have enough resources to analyze the planned number of tests, so a selection process had to be used. One batch (batch 2) was selected to serve as a national sample in which all respondents were invited to carry out a serological analysis. In several specific *départements*, where the disease was spreading rapidly, other batches were added, so that local estimations could be produced. Overseas *départements* could not be included in the serology sample in wave 1 due to logistical constraints.

In round 2, all respondents were offered a serological test. 4 out of the 20 batches were selected to form the “household sample”, in which not only the respondents but also the other members of the household had the possibility of performing this test.

The different characteristics of these batches are summarized in table 1.

Table 1 : Distinctive characteristics of EpiCov batches :

	Data collection protocol	Serological sample

Batch	Short / Long questionnaire	Round 1	Round 2	Round 1						
				Metro	92-93-94	75	60	13	67	68
1	Long	Competitive		No	Yes	Yes	Yes	Yes	Yes	Yes
2	Short	Competitive		Yes	Yes	Yes	Yes	Yes	Yes	Yes
3	Short	Competitive	Household	No	Yes	Yes	Yes	Yes	Yes	Yes
4	Short	Sequential	Household	No	No	Yes	Yes	Yes	Yes	Yes
5	Short	Internet		No	No	Yes	Yes	Yes	Yes	Yes
6	Short	Internet		No	No	Yes	Yes	Yes	Yes	Yes
7	Short	Internet		No	No	No	Yes	Yes	Yes	Yes
8	Short	Internet		No	No	No	Yes	Yes	Yes	Yes
9	Short	Internet		No	No	No	No	Yes	Yes	Yes
10	Short	Internet		No	No	No	No	Yes	Yes	Yes
11	Short	Internet		No	No	No	No	Yes	Yes	Yes
12	Short	Internet		No	No	No	No	Yes	Yes	Yes
13	Short	Internet		No	No	No	No	No	Yes	Yes
14	Short	Internet		No	No	No	No	No	Yes	Yes
15	Short	Internet		No	No	No	No	No	No	Yes
16	Short	Internet		No	No	No	No	No	No	Yes
17	Short	Internet		No	No	No	No	No	No	No
18	Short	Internet	Household	No	No	No	No	No	No	No
19	Short	Internet	Household	No	No	No	No	No	No	No
20	Long	Internet		No	No	No	No	No	No	No
Total	**2**	**3**	**4**	**1**	**3**	**6**	**8**	**12**	**14**	**16**

2.6. Children sub-sample

Specific questions were asked to the respondents concerning one of their children (aged between 3 and 17) selected at random while running the questionnaire. In rounds 1 and 2, these questions concerned schoolwork, sleeping difficulties and screen exposure in the long version of the questionnaire. In round 3, questions on the children's psychosocial strengths and weaknesses were asked, as well as other questions on possible explanatory factors.

The interest and the statistical relevance of exploiting the panel dimension of this child sample was not judged to be a goal in itself, so that even if a respondent's child had already been sampled in the previous survey waves, another child was sampled in round 3 (possibly, but not necessarily the same as before).

3. Covid-19 Serology

3.1. Inclusion criteria for serological tests in rounds 1 and 2

In the first round, self-testing was offered to a metropolitan random subsample in order to perform the expected 6 000 tests, and to obtain local indicators for five areas with overrepresentation, with at least 1 000 expected serology tests each. Three of these five areas had the highest Covid-19 risk indicators (Haut-Rhin, Paris and the inner suburbs, Oise) and two were considered to be at lower risk (Bas-Rhin and Bouches-du-Rhône) during the first epidemic wave.

During the second round of the EpiCov survey, home blood self-samples were offered to all study participants. In addition, 20% of the respondents (4 batches among the 20) were selected for testing of all members of the household aged six years and over, including the index[8]. Information on cohabitants was obtained from the questionnaire completed by the index on people cohabiting with them. The main questionnaire included for every respondent questions about gender, age, and family relationships for each of the respondent's household members. Individuals selected for household testing in batches 3, 4, 18 and 19 answered a few more questions, since they were also asked about the educational level of cohabiting individuals, whether they had already contracted Covid-19, or if they have been hospitalized since the beginning of the coronavirus crisis. The question about the age of each household member acted as a filter to trigger the dispatch of a blood sampling kit, and also helped to distinguish between the several models of sampling kits and instruction notices according to the children's age.

3.2. Self-administered tests

Serological tests were based on capillary blood samples collected by the participants themselves at home by pricking their finger tip and laying a few drops of blood on a dried blood spot card (903 Whatman paper kit). The participants then sent them to a centralized virology laboratory in order to have them analyzed.

Self-administered tests were offered during the telephone or online questionnaires, with appropriate explanations about how the test should be performed and how it would be used for subsequent analysis.

Sampling kits were delivered by express mail to each participant who had agreed to undergo testing. The kits included all the necessary material and printed instructions on how to perform the sampling, together with a prepaid addressed envelope for the return of the dried-blood sample to one of the EpiCov biobanks (located in Bordeaux since the first round, Amiens, Montpellier, Saint-Pierre, Fort-de-France and Pointe-à-Pitre in the subsequent rounds). A hotline telephone number was provided to allow participants to ask any questions they might have.

At the biobanks, DBS cards were first stored in 2D FluidX 96- Format 0.5 mL tubes (Brooks) at -30°C. Once most of the cards had arrived, up to four 4.7-mm discs were punched from the spots on the Whatman paper, using a PantheraTM machine (PerkinElmer).

3.3. Serological tests

The tubes were sent to the virology laboratory (Unité des virus Emergents, Inserm/IRD, Marseille, France) for elution of the dried blood and serological analysis. Eluates were processed with a

[8] It was also possible for the respondent to only ask for a single kit for themself.

commercial ELISA (Euroimmun®, Lübeck, Germany) to detect anti-SARS-CoV-2 antibodies (IgG) directed against the S1 domain of the spike protein of the virus (ELISA-S), according to the manufacturer's instructions. All samples with an ELISA-S test optical density ratio ≥ 0.7 were also tested with an in-house microneutralization assay to detect neutralizing anti-SARS-CoV-2 antibodies (SN) [20]. For these tests, VeroE6 cells cultured in 96-well microplates, 100 TCID50 of SARS-CoV-2 strain BavPat1 (courtesy of Prof. Drosten, Berlin, Germany) and serial dilutions of serum (1/20–1/160) were used. Dilutions associated with the presence or absence of a cytopathic effect on day 4.5 post-infection, were considered to be negative and positive, respectively.

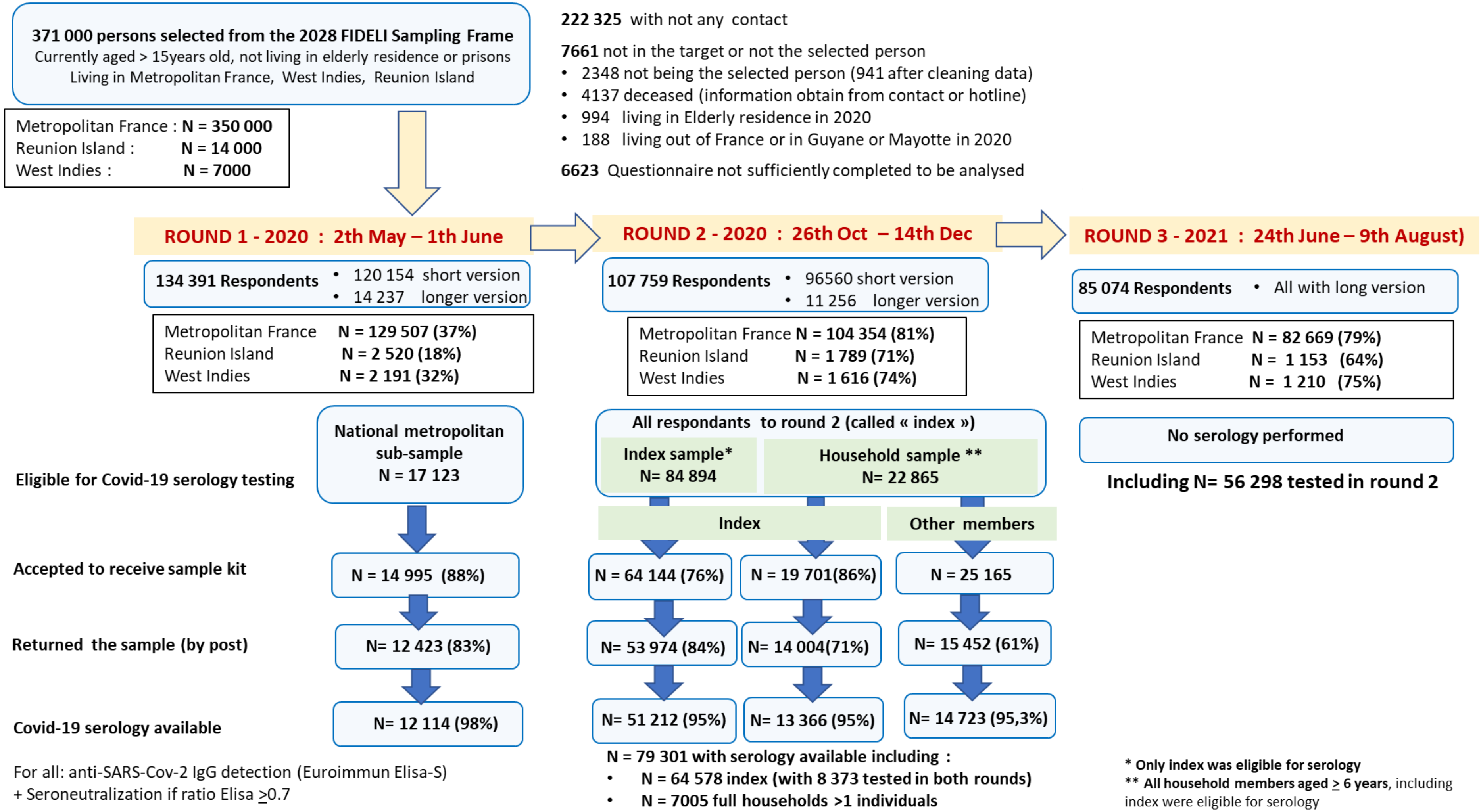
The serological results were sent to the participants by postal or secured email, at the end of the study, with information concerning the lack of scientific knowledge about individual protection against future re-infection for those who had tested positive for antibodies.

4. Fieldwork of rounds 1 to 3

Round-1 questionnaires were collected from May 2 to June 1, 2020, and the last blood samples were received until June 26, 2020. Round-2 questionnaires were collected from October 26 to December 14, 2020, and the blood samples were accepted until January 19, 2021. Round-3 questionnaires were collected from June 24 to August 9, 2021.

A flowchart showing the participation rates across the three rounds of the survey is to be found in Figure 2.

Figure 2: Flowchart of the EpiCov participation from round 1 to round 3


371 000 persons selected from the 2028 FIDELI Sampling Frame
Currently aged > 15years old, not living in elderly residence or prisons
Living in Metropolitan France, West Indies, Reunion Island
Metropolitan France : N = 350 000
Reunion Island : N = 14 000
West Indies : N = 7000
222 325 with not any contact
7661 not in the target or not the selected person
• 2348 not being the selected person (941 after cleaning data)
• 4137 deceased (information obtain from contact or hotline)
• 994 living in Elderly residence in 2020
• 188 living out of France or in Guyane or Mayotte in 2020
6623 Questionnaire not sufficiently completed to be analysed
ROUND 1 - 2020 : 2th May – 1th June
134 391 Respondents
• 120 154 short version
• 14 237 longer version
Metropolitan France N = 129 507 (37%)
Reunion Island N = 2 520 (18%)
West Indies N = 2 191 (32%)
ROUND 2 - 2020 : 26th Oct – 14th Dec
107 759 Respondents
• 96560 short version
• 11 256 longer version
Metropolitan France N = 104 354 (81%)
Reunion Island N = 1 789 (71%)
West Indies N = 1 616 (74%)
ROUND 3 - 2021 : 24th June – 9th August)
85 074 Respondents
• All with long version
Metropolitan France N = 82 669 (79%)
Reunion Island N = 1 153 (64%)
West Indies N = 1 210 (75%)
Eligible for Covid-19 serology testing
National metropolitan sub-sample
N = 17 123
All respondants to round 2 (called « index »)
Index sample*
N= 84 894
Household sample **
N= 22 865
No serology performed
Including N= 56 298 tested in round 2
Index
Other members
Accepted to receive sample kit
N = 14 995 (88%)
N = 64 144 (76%)
N = 19 701(86%)
N = 25 165
Returned the sample (by post)
N= 12 423 (83%)
N= 53 974 (84%)
N= 14 004(71%)
N= 15 452 (61%)
Covid-19 serology available
N= 12 114 (98%)
N= 51 212 (95%)
N= 13 366 (95%)
N= 14 723 (95,3%)
For all: anti-SARS-Cov-2 IgG detection (Euroimmun Elisa-S)
+ Seroneutralization if ratio Elisa >0.7
N = 79 301 with serology available including :
• N = 64 578 index (with 8 373 tested in both rounds)
• N = 7005 full households >1 individuals
* Only index was eligible for serology
** All household members aged > 6 years, including index were eligible for serology

4.1. Collection modes

The 370 000 randomly selected individuals were initially sent a personalized contact letter including a presentation of the survey, together with access codes for a web link to the questionnaire. Whenever possible, the information was also provided by e-mail, text message, and phone call. In all communications, the first name and surname of the person selected from the household were indicated. As the telephone contact details from tax files are not necessarily those of the person selected in the household, the person contacted was asked to forward the letter and information sheet with the internet link to the intended recipient if necessary.

In round 1, the availability of telephone interviewers was reduced because of lockdown, making it impossible to use CATI for all the subsamples. A concurrent mixed mode was then assigned to three of the 20 "batches" in mainland France (that is, from the start of the study, interviewers tried to contact selected individuals who had also received the web link to answer online). A sequential mixed-mode procedure was assigned to one batch, in which interviewers tried to reach respondents only two weeks after the start of the study, i.e. half way through the fieldwork. An exclusive CAWI mode was assigned to the 16 other batches. The number of CATI batches was higher for the overseas departments (7 in Guadeloupe and Martinique, 9 in La Réunion).

For later rounds, both data collection modes were opened for all participants. The large number of calls that this protocol implied and the strong pressure it put on the interviewers led to the need to define a priority to monitor the way the individuals were called. Therefore, individuals who responded to the questionnaires of earlier rounds by phone were called first. The justification for this choice was that respondents were more likely to answer by Internet if they had already done so before. In order to obtain as many respondents as possible at the same time and the largest share of Internet respondents, it was deemed more efficient to try and first call all those who were less likely to respond by Internet anyway. Secondly, the batches were used to operate another level of prioritization, which facilitated day-to-day monitoring of the fieldwork.

4.2. Contacts and reminders

The telephone numbers available in the Fidéli sampling frame were supplemented by a telephone directory search, which increased the proportion of available numbers from 71% to 81%. Letters were sent to 370 928 (349,936 participants in mainland France and 20 992 in the overseas *départements*. As postal services were not fully functional in France during this period, emails and text messages were also sent, at the same time whenever possible: 258 867 e-mails (246 019 in mainland France and 12 848 in the overseas departments) and 165 028 text messages . Only 4.2% of the letters were undelivered, 7.4% of the e-mails and 17% of the text messages were bounced back as spam. The interviewers reported that some respondents contacted by telephone had not received the initial letter in time, almost certainly due to poor postal deliveries during the pandemic period or because they had left their usual place of residence during the lockdown period.

Sequential reminders were sent with different wordings and modalities: 253 801 letters, 163 434 and 148 820 e-mails at two different times, 116 600 and 97 505 text messages, 112 578 voice messages on mobile phones and 26 856 on landline telephones. Each reminder led to a new peak in questionnaire completion. Additional reminders were also carried out after the questionnaire collection period, to increase the number of blood samples returned.

In follow-up rounds, similar types of contacts and reminders were mobilized. As the duration of the fieldwork was longer in rounds 2 and 3, the number of emails and text reminders per individual was higher than in round 1.

4.3. Response rate to questionnaires

No contact was established for 222 325 of the individuals selected. For the others, a specific assessment was made to check that the respondent was indeed the individual selected, by comparing gender and birth date recorded on the questionnaire to those available on the Fidéli sampling frame. This assessment led to the exclusion of 2 348 individuals who were not the selected individual. Another 5487 individuals were considered to be outside the sampling frame (deceased, living in a care home for the elderly, no longer living in France, or living in French Guyana or Mayotte), and 6 23 completed too few items to be retained in the final database.

For the first round, there were 134 391 respondents in all (Fig. Flow chart): 120 154 completed the short form of the questionnaire and 14 237 completed the long questionnaire. The response rate was 37% in mainland France, 32% for Reunion Island and 18% in the French West Indies. For the mixed-mode subsample batches contacted by Internet and telephone, the response rate was 46% in mainland France, 41% for Reunion Island and 33% in the French West Indies. For those contacted solely on Internet, the response rate was 35% in mainland France, 24% in Reunion Island and 23% in the French West Indies. For those responding via the Internet, 20% did so on a smartphone and 7% on a computer tablet. In mainland France, the response rate for people living below the poverty threshold was lower than that of the population as a whole, by a factor of 1.7.

The response rates in the two subsequent rounds were 80% (107 759) and 79% (85 032), respectively. As in round 1, there was a significant difference in response rates between mainland France (81% and 79%) and overseas *départements* (71% and 64% in the West Indies and 74% and 75% in Reunion Island). Although all batches had the possibility of answering through a phone interview, the share of CATI interviews remained significantly higher in the four mixed-mode batches of round 1: 29% versus 12% in round 2, and 24% versus 12% in round 3.

4.4. Return rate for serological tests

In the first round, 17 123 participants were eligible to receive a sampling kit, among them 88% (14 995) agreed to receive it, 83% (12 423) returned it, conducting to available serology for 98% (12 114).

In the second round, 107 759 respondents were eligible to receive a sampling kit, out of whom 22 865 were also included in the household batches.

Among all index respondents, 83 845 (77%) agreed to receive the home sample kit with available mailing addresses, 67 978 (82%) returned the dried blood spot sample to biobank, among them 64 924 (96%) had sufficient quality to be punched for serology testing process at the virological laboratory. Finally, a serologic result was obtained for 64 578 (99,5%), whereas not enough blood was available in the remaining tubes to obtain an interpretable result.

Considering separately metropolitan and overseas areas, the response cascade was respectively : 85 350 (82%) and 2 495 (73%) index who accepted to receive the tests, among them 66 826 (78%) and 1152 (46%) returned the sample to the biobank, which conducted to an available serology for 63 858 (96%) and 1 066 (93%).

Concerning the other 25 165 eligible household members of the index participants, 61% returned the sample and 95% were tested. Overall, serological tests were performed for all members aged 6 years or more in 7005 non-single households.

The median date and interquartile range for serological tests in mainland France were for the first and second rounds in 2020 : May 21st 2020; (18th – 28th May) for 12 114 participants and November 24th 2020; IQR: (18th November– 4th December) for 63 524 participants, respectively.

Overall, a serological result was available at both round 1 and 2 for 8373 metropolitan index participants.

The proportion of kits returned to the biobank with insufficient quality was higher in Epicov round 2 than round 1, likely due to changes in lancing devices to prick the finger included in the home self-sample kit, less easy to use than in the second round.

5. Post-field phase: correction for non-response and weighting calculations

The corrections of selection biases induced by non-response were treated by a production of weights from non-response models and calibration on margins derived from the census. Partial non-response was not addressed and it was left to researchers to proceed to imputations concerning variables they intended to use according to their needs. Multiple imputations were for instance performed for papers on mental health based on rounds 1 and 2 of the EpiCov survey [21, 22].

5.1. Definition of the samples retained for the calculation of weights

The information gathered through the Epicov data collection protocol is very diverse: answers to the standard questionnaire items in rounds 1 to 3, answers to the long questionnaire items in rounds 1 and 2, answers to the children's questionnaire in round 3, test results for the respondent in rounds 1 and 2, and test results for the members of the respondent's household in round 2. Due this great variety of information, available for different parts of the original sample, numerous subsamples had to be produced, each associated with a specific estimation weight elaborated to account for the sampling design according to which it had been selected, the several non-response steps that were passed through for and the auxiliary information available to enhance the precision of its estimates.

The main samples that could be produced with this information were the samples of respondents to the main questionnaire in rounds 1 to 3. These samples were selected according to a multi-phase sampling design: the first phase describes the stratified systematic sampling design according to which the original sample of Epicov was selected, the following phases describe the successive non-response steps this original sample had to undergo. In rounds 2 and 3, the samples on which data collection was conducted were limited to the main questionnaire respondents in the preceding rounds, so that non response to each round could be described as a succession of nested non-response steps, each of them treated using a Poisson sampling design. In rounds 1 and 3, total non-response was defined based on a list of the main variables of interest, whereas in round 2, partial non-response was limited enough to allow individuals whose questionnaire has been validated by the data collection process to be defined as respondents.

For each of these steps, the same method was used to estimate the non-response probabilities, described in more details in the section "General method used for weight computation". The non-response adjusted estimation weight taking into account the whole sampling design used the classic formula for the estimation weight for multi-phase sampling designs [23, 24]: the initial sampling weight is divided by the estimated response probabilities of the successive non-response steps. This non-response adjusted weight is then calibrated on known margins to reduce its variance. This calibration is also described in more detail in the section "General method used for weight computation".

Other samples of interest are formed of respondents to the long questionnaire in round 1, in both rounds 1 and 2, and respondents in round 3 questionnaire who also responded to the long questionnaires in rounds 1 and 2. These samples are in particular used to analyze cross-sectional and longitudinal results for mental health. The sampling design of these samples is very similar to that of the main samples described earlier; the only difference lies in the sampling design of the first round. For the long questionnaire samples, it is a stratified systematic sampling of one tenth of the main sample, according to the same strata.

Another group of subsamples of interest is formed with the subsample of sampled individuals who returned their blood samples in rounds 1 and 2.

Non-response for the serologies sample was defined as refusing to receive the serological kit, or not sending a blood sample back after receiving one. Cases where a sampling kit was sent back and a blood sample was indeed collected afterwards were counted as valid. In round 1, two different weights were calibrated according to the geographical area analyzed: a specific weight was calculated to estimate local indicators from the serological data collected in the oversampled *départements*, and another weight was designed to enable the use of all serological samples to estimate national prevalence. In round 2, a first set of weights was computed for the sample of all index serologies, and a second set of weights for the sample of individuals tested in rounds 1 and 2.

For all these samples, the computation of weights follows the same steps. For instance, the weights of round 2 samples of all index tests are computed from the weights of the sample of respondents to the main questionnaire. Probabilities of sending back a blood sample are estimated with the same method used for the non-response processing in the main sample. The non-response adjusted weights of the index test samples are then calibrated on the same margins used to calibrate round 2 main sample. Lastly, specific weighting was needed for the analysis of the "household" and of the "children" samples in round 2.

5.2. General method used for weight calculations

The method was similar for the non-response adjustments of most samples and is the classic two-step approach involving reweighting with homogeneous response groups and calibration [25, 26].

In the first step, the survey weight (inverse of the inclusion probability) was divided by an estimate of the probability of response. These response probabilities were estimated using a scoring method. A first version of the response probabilities was estimated using logit models or statistical learning algorithms such as random forests, taking into account auxiliary variables linked to both the response mechanism and the main variables of interest in the EpiCov survey. For the first round, the sampling frame provided numerous auxiliary demographic and socioeconomic variables, 90 of which were correlated with at least one variable of interest in at least one *département*. The quality of contact information, and variables describing the respondent's residential neighborhood, such as population density, proportion of people aged over 65 or below the poverty threshold, obtained from geo-referencing information included in the Fidéli sampling frame, were also used. For rounds 2 and 3, answers to the preceding round questionnaires were also used.

Response homogeneity groups were derived from these estimated probabilities using k-means and Haziza-Beaumont algorithms [27]. For the main samples, homogeneous response groups were constituted inside each *département* in rounds 1 and 2 and inside each region in round 3 (corresponding to the NUTS-2 level in the European nomenclature of local units), whereas for other samples they were computed directly on the national sample. The response probability was then estimated from the percentage of respondents in each homogeneity group, yielding first-step weights. The lower the response weight in the homogeneous response group of a respondent, the higher its non-response adjusted weight. This results in a correction of the effects of differences in response rates between subgroups.

In the second step, these weights were calibrated [28] on the margins of the population census data and population projections for several variables. For the main samples, these margins are the structure of the population per aggregated diploma and region, by gender, 10-year age categories and *département* (region in round 3) and by *département* and place of birth in the three French overseas *départements*. Weights in the long questionnaire samples were calibrated on national margins: the structure of the population by region, by gender and 10-year age groups and by diploma. Weights for the serological subsample were calibrated at national and local level for the five overrepresented areas in round 1. In round 2, the sample of all index tests was calibrated on the same margins as round 2 main sample, whereas the longitudinal samples of individuals tested in rounds 1 and 2 was calibrated on national margins (the same margins on which the long questionnaire sample was calibrated) and margins describing the five oversampled areas in round 1. This calculation was designed to decrease the variance and the residual bias for variables correlated with margins.

The whole list of weights that were calculated is as follows: cross-sectional weights (for rounds 1, 2 and 3), two weights for the serologies in round 1 (for national and local analyses), one weight for round 2 index serologies, and one additional weight to analyze the serologies of individuals who were tested in rounds 1 and 2. Two specific sets of weights are detailed below: the "household sample" in round 2 household member serologies, and the "children samples" in the different rounds.

5.3. Design of “children” weighting

In each round, part of the questionnaire was designed to ask the respondent about one randomly selected child in the household. Children taken into account for this selection are the son or daughter of the respondent or their partner, or children placed with the respondent or their partner by the child protection administration. The child concerned by the questions was selected for each respondent according to a simple random sampling design of one observation among all the children in the respondent’s household. In rounds 1 and 2, time and resource constraints as well as the fact that these questions were more likely to be analyzed as a qualification of the responding parent themself (for instance in order to analyze how housework and more specifically school coaching was taken on by one or other parent during a year affected by lockdowns) did not lead to the calculation of specific weights for this specific subsample.

In round 3, specific questions about children’s psychosocial strengths and difficulties, as a way to assess their mental health, required this calculation. The sampling probability for a given child was different from that of their responding parent, since the number of children of their parent directly affected this sampling probability: for children without any sisters or brothers living with them the questions asked to their respondent parent would concern them alone, but children living with one sister or one brother aged between 3 and 17 only had a 50% chance to be the subject of these questions.

In order to deal with this, the weights calculated for the analysis of round 3 results regarding children were obtained through several steps of adjustment. First, non-response was accounted for. Starting with the non-response corrected questionnaire weights described earlier, two additional models for subsequent non-response were calculated: non-response regarding the composition of the household, and after that non-response for the questions regarding the child[9]. These two models made it possible to adjust twice to these non-response behaviors, finally obtaining respondent weights adjusted for these three non-response steps. Using these as entry weights, a simultaneous calibration [29] on respondent and child margins was performed. The final weights were then obtained by multiplying these weights by the number of children who could potentially have been selected in the household, so that children living in large families would not be under-represented.

5.4. Adjustment for the “household” sample

The specific collection of serologic data on respondents and their household members needed a specific weight, since only 4 out of 20 batches were concerned and a different selection was induced by this: the respondents in these batches had the choice to receive blood sampling material for them alone or for their whole household, or solely for them if they reported living alone in the household.

The probability for a given individual to be included in the sample as a household member is twofold. On the one hand, if this individual is aged 15 years or more, they can be selected as a direct respondent. On the other hand, every person aged 6 years or more living with at least one person who matches the previous criteria also had a chance to be indirectly selected. The probability of this indirect selection is linked to the number of persons they live with, as well as to the inclusion probabilities of these cohabiting individuals, and even to the probability for this directly selected cohabitant to engage their household in this part of the survey. For children under 15 in May 2020, the probability of direct selection is zero, so that this indirect selection is the only way they can be taken into the sample. This kind of situation leads to the fact that the sampling probabilities of each person in this household are different in nature, which prevents the use of regular adjustment methods.

The classic method to weight samples obtained with indirect sampling is the generalized weight sharing method [30, 31]. We first formed the sample of index respondents who detailed the composition of the household[10] and whose household members all returned their blood samples. This

[9] In practice, respondents who did not answer enough questions to calculate the SDQ scores were considered as non-responding in this step.

[10] due to partial non-response, the complete composition was not available for all households

sample was obtained from the main sample in round 2 with two additional phases of non-response for the correction of which we applied the method described earlier. The sample was also calibrated on the same margins as the main sample in round 2[11].

The generalized weight share method (GWSM) was applied to the non-response adjusted and calibrated weights of the sample of index respondents who detailed the composition of their household and whose household members returned their tests. We then obtained a sample of households, the households of the index respondent, weighted by the GWSM weights. This sample describes the population of households at the time round 2 was collected, at least one member of which could have been selected in the original EpiCov sample. The GWSM weights can also be used to weight the sample of the index respondent household members, describing the population living in households where at least one member could have been an index respondent.

6. Enrichment and matching of the data collected

6.1. Occupation coding: between PCS 2003 and PCS 2020 nomenclatures

From round 2 the EpiCov survey used the new protocol constructed by a working group under supervision of the National council of statistical information (CNIS) in 2018 and 2019 to collect and classify the reported professions of the respondents. This protocol consists in the collection of a main description, possible supplementary descriptions for disambiguation, and several ancillary questions to help encode the results into the nomenclature. The nomenclature is structured in four nested layers, and the best possible precision was expected for the survey. While all respondents were asked to complete this module in round 2, a filter question was added in the following rounds of questioning to avoid asking the respondent once more to describe a profession that was already known.

A self-completion field was included, helping to standardize the descriptions collected, with the possibility of drafting an open description if needed. Self-completed descriptions mostly involved an automatic recognition of the profession in the PCS 2020 nomenclature, but the use of an non-definitive self-completion list for the survey (the stabilized list was then not available) led to some coding rejections. Nonetheless, the fields collected that were not recognized by the PCS 2020 environment were also submitted to the PCS 2003 environment, which is the former version of the PCS nomenclature. The use of a transfer matrix between the two versions of the nomenclature made it possible to raise the automatic codification rate. Cases that were recognized by neither version, or for which the PCS 2003 classification did not lead to a single connection in the PCS 2020 framework were then analyzed by hand by INSEE experts.

The initial PCS 2020 environment allowed the direct encoding of 84.2% of the job descriptions (76,276 among 90,582 collected). Among the rest, 62.9% (8,868) were coded in PCS 2003, and 45.2% (6,372) could be directly translated into the PCS 2020 nomenclature at the best precision level. Overall, this original protocol made it possible to divide by almost two the volume of descriptions that needed to be analyzed manually, even though the material used for data collection was not entirely stabilized.

6.2. Income data from social and taxation sources

With the objective to follow, among other things, the economic consequences of the coronavirus crisis for the French population, income data is of prime importance for some exploitations of the EpiCov cohort. The sampling database Fidéli already contained some information concerning this dimension, but the information dates back to 2018, the year of the Fidéli issue used to construct the sample.

Due to the methodological difficulties arising when direct questions are asked about individual income (difficulties distinguishing between individual and household income, unclear perimeter of the concept

[11] The calibration was applied before application of the weight sharing, because the margins that were available to calibrate the sample represented the respondents and not their dwellings.

for some respondents, memory bias, tendencies to round off amounts etc.), no direct question about it was included in the questionnaires, apart from questions about perceived financial situation. Instead, a matching with social and taxation data was considered a better way to deal with this dimension.
Using personal data from the initial sampling database Fidéli, from repertory searches, or from direct information given by the respondents when answering the survey, matching on personal characteristics such as name, surname, age or address was performed. A specialized INSEE team, using the method described in [32], performed this operation. This allowed the survey to be augmented by precise income data with distribution by type of income (allowances, wages…) for the year 2020.

6.3. National Health Data System (SNDS)

The National Health Data System, known as SNDS, contains a wealth of data on health, notably every health-care act that is associated with a health insurance reimbursement. This includes in particular data managed by the French health insurance agency, hospital data, or medical causes of death. A 19-year conservation duration for this data makes studies on medium and long-term health trajectories possible.
For a cohort such as EpiCov, being matched to the SNDS can be useful in many ways. First, the examination of the factors associated with the severity of Covid-19 infection or their longer term consequences count among the main objectives of this cohort. Other study themes such as healthcare access during the crisis can also be explored with this design, and the range of questions included in EpiCov makes it possible to precisely link different medical data to social causes and consequences. Another interesting way to make use of this matched data could be to provide more insight on the non-response observed in the different questionnaire rounds. In particular, it will make it possible to control for non-response behavior directly linked to severe forms of Covid-19.
In order to achieve this matching, direct matching will be performed on the basis of the national registration number (NIR). This number is a single identifier associated with each person born in France or living in France, and is used by the health insurance agency to identify patients. It can be reconstructed using the information in the sampling database Fidéli (name, surname, sex, date of birth, place of birth) via the national directory for the identification of natural persons (RNIPP). These NIRs are then sent to the National health insurance agency in order to perform a selected extraction from the SNDS. This extracted data will then be linked to the cohort data itself, making it possible to perform the aforementioned analyses.

7. Strengths and limitations

7.1. A broad range of possible exploitations

EpiCov undoubtedly provides one of the richest sources of information about virus diffusion and living conditions since the beginning of the epidemic in France. Few such national population-based studies, aiming to achieve national and local representativeness, have been set up as quickly in other countries. Each EpiCov round enables the study of numerous aspects. In October 2020, the first published papers described social heterogeneity in the probability of having been in contact with the coronavirus before the end of the first lockdown, and the economic consequences of the lockdown. Later, analyses were published on the dynamics of the epidemic throughout the year 2020, on mental health during and after the first lockdown, on representations and practices regarding Covid-19 vaccination, or on specific populations (young people, disabled persons…). These results have been published or are under review, in both public statistics reports and in peer-reviewed journals. Other analyses concern the risk of prolonged Covid-19 symptoms, the relationship between Covid-19 infection and occupation, adult and child mental health, home-care workers, or health professionals.
Several collaborations between the EpiCov team and other research teams were initiated, before the EpiCov database was made accessible for research projects through the center for secured access to

data (CASD). For each round, this access is made possible 9 months after data collection is completed, including weighting and serological results, in each round, according to a specific regulatory process (CNIL, committee for statistical confidentiality). Selected utilizations of the biological samples will also be possible, although the sparseness of the material makes a strict selection process essential.

7.2. The pace of survey preparation, fieldwork and publications

During the follow-up questioning process, the calendar was very dense: although the EpiCov team tried to take more time than initially planned, the mobility of the epidemic and the subsequent health regulations made the preparation of the study arduous. On the one hand, the objectives of the study and the high expectations from its results put pressure on the preparation work and maintained that pressure on timing. On the other hand, these evolutions in the general context led to considerable changes between the conception of the questionnaires and the fieldwork, and even in the course of the fieldwork. For instance, the fieldwork of round 1 (May 2 to June 4, 2020) began during the first national lockdown (March 17 to May 11, 2020) and ended after the end of the lockdown; this change in the health context had obvious consequences on the responses to some parts of the questionnaire. Even though the regulatory procedures were more accurately followed in the later rounds, there was not enough time to assess the statistical quality of the survey protocol according to gold standards (especially to carry out sufficient tests) and the aforementioned committees were asked again to perform their assessments in a less stringent mode. Overall, the three survey rounds which each lasted between one and two months were on the field in one year and a quarter.

7.3. Interpretation of serological tests

The serological tests served to detect antibodies that reflect past contact with the infection, whereas virological tests remain positive for only a short period after the infection.
However the date of the infection, especially in case of non-symptomatic infection, cannot be derived from positive serology. The Euroimmun ELISA-S test has a sensitivity of 94.4% according to the manufacturer's cutoff. It has been evaluated in various studies, which reported specificity ranging from 96.2% to 100% and sensitivity ranging from 86.4% to 100% [33, 34, 35]. Anti-Sars-Cov2 IgG antibody levels have been reported to decline more or less rapidly, particularly among the elderly and subjects with mild or asymptomatic forms. Changes in pattern of factors associated with seropositivity remain of major interest to understand changes in exposure risks between May and December 2020. The persistence of IgG antibodies partly reflects the level of protection. In the fourth round, the quantity of antibodies will be measured from self-administered tests and compared according to vaccination status, history of symptoms, self-reported positive virological test, and previous serological results.

7.4. The effects of the mode of data collection

EpiCov offered mixed-mode data collection to limit selection bias, especially due to poorer access or ability to internet self-questionnaire for part of the population, which increased global cost of the study. While surveys using several data collection modes have the advantage of increasing response rates and representativeness, it can also induce biases. Two kinds of effects can be distinguished. Firstly, selection effects, since Internet respondents could have different characteristics from phone respondents which could therefore lead to discrepancies between online and phone results. Secondly, measurement effects, as a given person can give different answers depending on whether they answer online or on the phone. This effect can notably be related to the direct interaction with the investigator on the phone, as response may be driven for instance by social desirability[12] bias, whereas the lack of

[12] Social desirability bias occurs when respondents tend to give the answers they imagine will make them more likeable for the surveyor.

a direct interaction during an online interview can lead to other biases like satisficing[13]. A more extensive description of these issues in the general context of mixed-mode surveys can be found in [36].

An endogenous selection effect was observed concerning the report of COVID-19-associated symptoms : in round 1, the response rates in CAWI/CATI mixed-mode batches were higher than in CAWI single-mode batches, but the respondents in mixed-mode batches were less likely to report Covid-19 symptoms than in single mode batches. Among mixed-mode batches, there were no significant differences in response rates (between 44% and 46%), but differences between the proportions of answers obtained online (46% of batch 1 respondents were phone respondents, as were 29% of batch 4 respondents). Nevertheless these differences were not associated with differences in Covid-19 symptom prevalence between mixed-mode batches. This was interpreted as a sign that the differences observed were more likely to reflect a selection effect than a measurement effect.

We accounted for this, by using a Heckman model [37, 38, 39] to generate specific weights to correct for this bias in the estimation of symptom prevalence. This estimate was based on the simultaneous modeling of participation and output variables (collected by the survey). For identification purposes, the model needs an instrumental variable, explaining survey participation but not playing any role in the output variables. In the case of EpiCov, this instrument was the binary variable distinguishing between people selected from the mixed CAWI/ CATI subsample batches and those in the single-mode CAWI subsample [40]. This binary variable was a reliable instrument. Indeed the division of the whole sample into 20 batches was random, so that the differences of response rates between single-mode and mixed-mode batches are also randomly assigned to the individuals: the participation rate was about 10 points higher in the four CAWI/CATI batches than in the 16 CAWI single-mode batches. In the general weight estimation framework presented above, the Heckman step replaced the logit model step, allowing to control for potential endogenous selection effects, while everything else remained unchanged.

In order to further investigate, evaluate and potentially correct the respective part of selection and measurement biases, round 4 will include two fully single-mode batches (one online and one phone).

7.5. Social structure of EpiCov respondents

Despite considerable efforts to adjust for non-response bias, some issues were identified without the possibility to correct them by appropriate adjustment mechanisms. This is specifically the case to estimate the prevalence of disabled persons as defined by the general activity limitation indicator (GALI)[14], which is one of the most commonly used indicators for disability in France and in Europe [41]. EpiCov data, after all weightings were calculated to adjust as far as possible for non-response biases, estimated this proportion at around 5% of the French population. Meanwhile, sources of reference tend to provide estimations between 8% and 10% [42]. The slight difference in survey scope (EpiCov includes communities while most surveys do not) fails to explain such a significant difference.

In the same way, some discrepancy remained when comparing the professional distribution in the survey to the French Labor Force Survey (LFS survey, enquête Emploi [43]), as shown by Table X.

Table X: Social structure by main professional groups of the active population in 2020, according to the LFS 2020 and EpiCov (round 2)

[13] Satisficing refers to the tendency of respondents to answer as quickly as possible to the questions they are asked, which can lead them to select the first possible answer that approximately matches their situation without examining all of them before selecting the most fitting.

[14] This indicator is constructed from responses to the following question: “Over at least the past six months, to what extent have you been limited because of a health problem in activities people usually do? Would you say you have been: severely limited? limited but not severely? not limited at all?”. Respondents answering “severely limited” to this question are considered here as “disabled”.

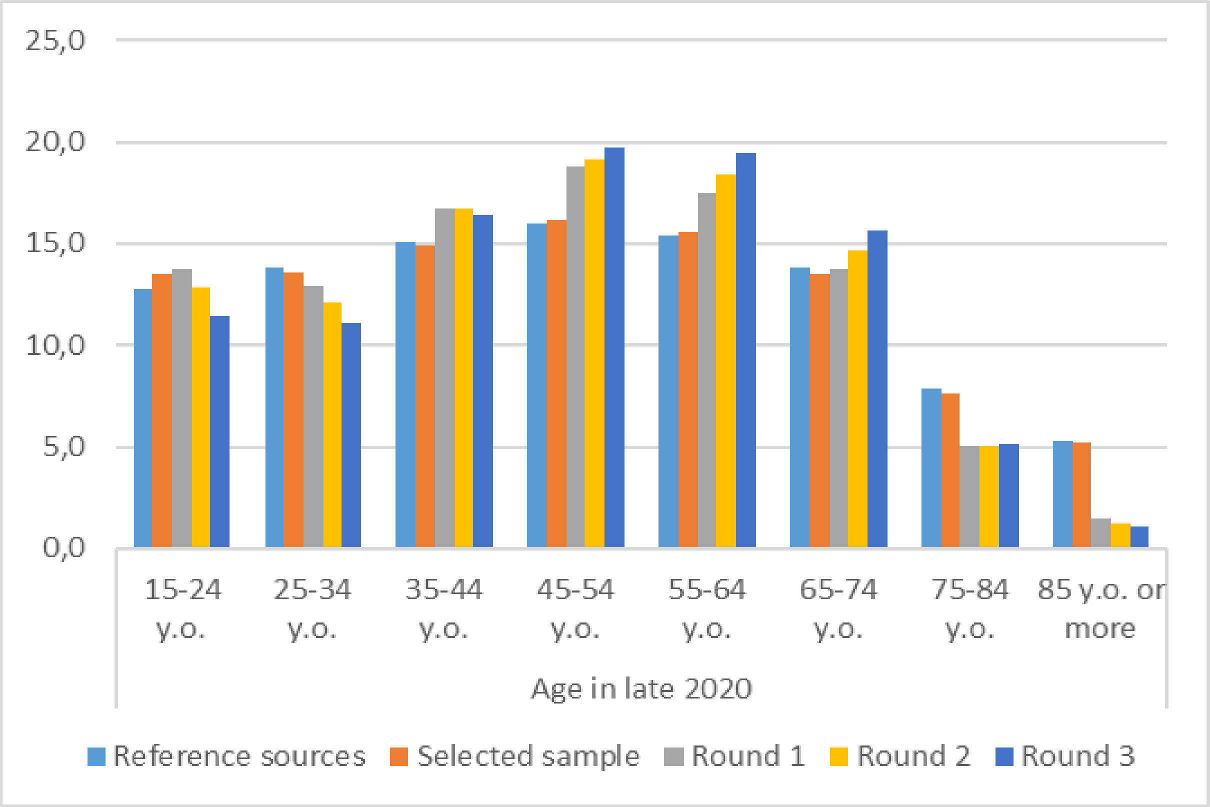
Socio-professional group according to the PCS nomenclature (1 position)	LFS 2020 (reference)	LFS 2020 (web pilot)	EpiCov, unweighted	EpiCov, weighted
Farmers *(agriculteurs exploitants)*	1.2	1.9	1.4	1,6
Craftspersons, salespersons, company heads *(artisans, commerçants, chefs d'entreprise)*	6.4	6.4	5.3	5,8
Executives and intellectual occupations *(cadres, professions intellectuelles supérieures)*	19.0	20.0	27.7	22,3
Technicians and associated professionals *(professions intermédiaires)*	25.5	24.2	30.0	27,7
Office workers *(employés)*	27.1	27.7	23.7	25,6
Factory and manual workers (*ouvriers)*	20.9	19.8	11.9	17,0

Sources: French Labor Force Survey 2020 and pilot for web LFS, EpiCov round 2
Findings: according to the LFS 2020 survey, 19.0% of the employed population in France belongs to the "executives and intellectual occupations" group, while this proportion is estimated at 22.3% in EpiCov. 27.7% of the respondents to the survey were classified as such, before the application of weightings accounting notably for non-response biases.
Scope: Working population aged over 15. LFS: population living in ordinary housing, EpiCov: excluding EHPADs, retirement homes and prisons.

The discrepancies in the occupation structure between EpiCov and the LFS survey reflects that the selection bias, made stronger by the overall low response rates in round 1 also affects representativeness according to socio-professional status. The calculated weights mitigated this issue, but not sufficiently to totally correct the bias and reach the same structure as in the reference survey. Two main arguments can explain that these differences persist despite the adjustments for non-response. The first is that EpiCov round 1 took place in the very specific context of the March 2020 national lockdown, which may have induced specific participation behaviors. The second is that adjustments need to be based on variables available for both respondents and non-respondents linked to the indicators of interest. The variables included in the Fidéli sampling frame did not provide enough information to properly model the mechanism of non-response among disabled persons. In later rounds, GALI was included as an auxiliary variable in non-response models in order to maintain a constant bias in estimates over time.

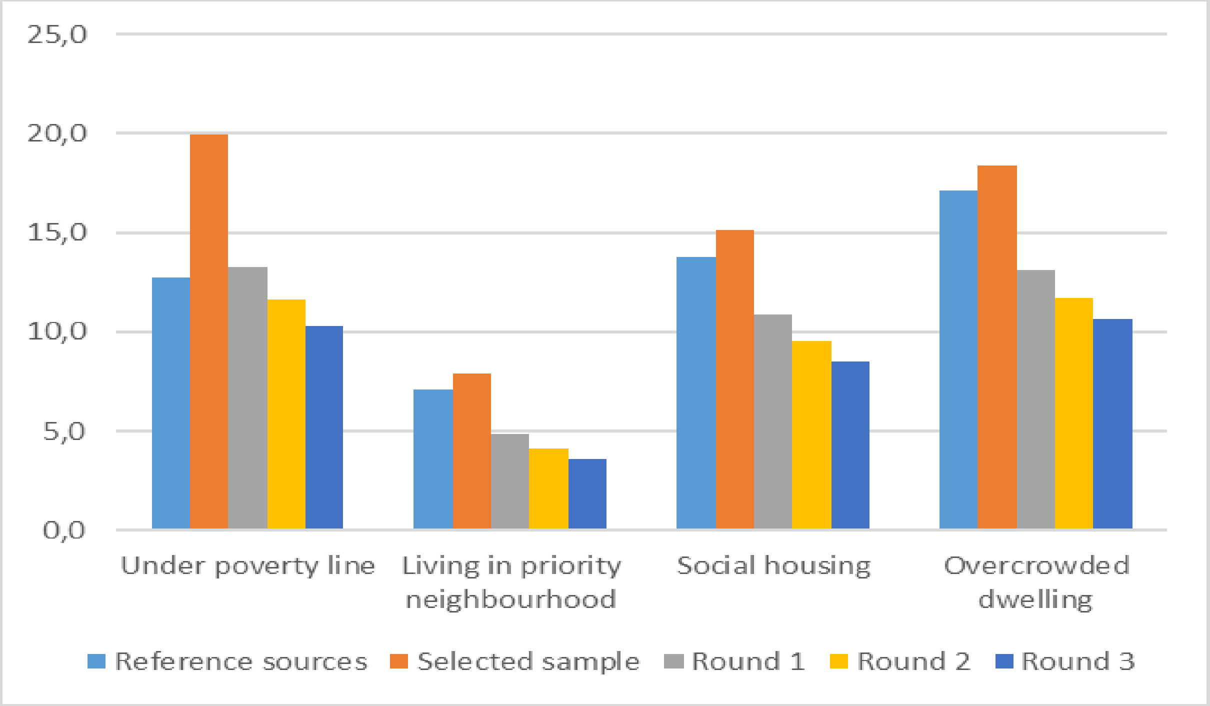
With several rounds of questioning and the use of a rich sampling frame for initial sample selection, EpiCov can provide a good source for the study of participation mechanisms over time. Figures 3-1 and 3-2 show for instance the shifts in age structure and in housing characteristics across the rounds of questioning. While round 1 response rates induced by far the most significant selection effect compared to additional selections in rounds 2 and 3, populations who were less likely to participate in the first round are also less likely to continue participating in further rounds. This increased therefore the selection biases with time. The very biases presented in these figures can easily be dealt with by the use of appropriate weighting processes. Other biases are yet likely not to be entirely corrected by the implemented post-field processes, as soon as the appropriate information does not stem from the sampling frame.

Figure 3-1: Age of EpiCov respondents across questioning rounds



Source: EpiCov, Fidéli database (reference)
Scope: Individuals aged 15 years or over, residing in metropolitan France, excluding EHPADs, retirement homes and prisons.
Findings: While people over 85 represent 5.3% of the population in the Fidéli database, they accounted for only 1.5% of EpiCov round 1 respondents and 1.0% of round 3 respondents.

Figure 3-2: Housing characteristics of EpiCov respondents across questioning rounds[15]



Source: EpiCov, Fidéli database (reference)
Scope: Individuals aged 15 years or over, residing in metropolitan France, excluding EHPADs, retirement homes and prisons.
Findings: People living under the poverty threshold, accounting for 12.8% of the population in the Fidéli database, were oversampled in the initial sample to reach 19.9% of the sample, and 13.1% of EpiCov round 1 respondents.

A more in-depth analysis of the response probabilities in round 1 is to be found in [44], with special emphasis on metadata influence on response probability.
As a means to increase representativeness by reducing non response and attrition biases and loss of statistical power, round 4 will include a "replenishment" of the cohort, by re-contacting some of the individuals initially selected for round 1.

Conclusion :

The EpiCov survey provides an extreme example of a large national statistical and epidemiological survey, conducted under troubled conditions (a national lockdown) and in a great emergency context by multiple institutions. This led to an extreme haste in the methodological conception and regulatory assessment procedures, in order to collect a large variety of epidemiologic, health, behavioral and socio-economic indicators from the end of the first wave of the Covid-19 pandemic, at local and national level.
This context demanded rapid adaptation to multiple and unforeseeable changes affecting some of the indicators at the core of the survey, both related to rapid epidemiological changes and governmental responses.

[15] Priority neighborhoods are a priority geography of urban policy, established by the planning law for urban affairs and urban cohesion of 21 February 2014 (check https://www.insee.fr/en/metadonnees/definition/c2114 for more details). A household is considered as overcrowded if the area is less than 18 m² per person for dwellings of more than one person and less than 25 m² for dwellings occupied by only one person.

The goals of the EpiCov cohort were very ambitious: to provide a panoramic overview of the consequences of the health crisis, on a national and infra-national scale. This survey provides an exceptional source of data on the first two years of the Covid-19 pandemic in France, including serologic evaluation at two time points before vaccines availability and one point in mid-2022. Data enrichments through matchings with for instance the SNDS database will also permit to study subsequent health of the population.

EpiCov can not only provide a wealth of information to document the social determinants of Covid-19 health issues and consequences, but it can also serve as a large playground for methodological research on survey design, data collection methods, and non-response correction methods. These developments have already started, but there is still much work to be done on these topics. Furthermore, the fourth and probable last round of EpiCov, which should begin in early May 2022, will provide even more material for later methodological investigations.

Bibliography

[1] Wilder-Smith A, Chiew CJ, Lee VJ. “Can we contain the COVID-19 outbreak with the same measures as for SARS?” *Lancet Infect Dis*. 2020 May; 20(5):e102–7.

[2] OECD “First lessons from government evaluations of COVID-19 responses: A synthesis”, January 4 2022

[3] Arora RK, Joseph A, Van Wyk J, Rocco S, Atmaja A, May E, et al. SeroTracker: a global SARS-CoV-2 seroprevalence dashboard. Lancet Infect Dis. 2020 Aug; S1473309920306319.

[4] Johns Hopkins Coronavirus Resource Center. Home [Internet]. Johns Hopkins Coronavirus Resource Center. [cited 2021 Feb 23]. Available from: https://coronavirus.jhu.edu/

[5] ECDPC. Homepage | European Centre for Disease Prevention and Control [Internet]. [cited 2021 Feb 23]. Available from: https://www.ecdc.europa.eu/en

[6] Koopmans M, Haagmans B. “Assessing the extent of SARS-CoV-2 circulation through serological studies.” *Nat Med*. 2020 Aug;26(8):1171–2

[7] Cornesse C, Blom AG, Dutwin D, Krosnick JA, De Leeuw ED, Legleye S, et al. “A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research”. *J Surv Stat Methodol*. 2020 Feb 1;8(1):4–36.

[8] Bobrovitz N, Arora RK, Cao C, Boucher E, Liu M, Rahim H, et al. « Global seroprevalence of SARS-CoV-2 antibodies: a systematic review and meta-analysis” [Internet]. *Public and Global Health*; 2020 Nov [cited 2021 Feb 2]. Available from: http://medrxiv.org/lookup/doi/10.1101/2020.11.17.20233460

[9] Pollán M, Pérez-Gómez B, Pastor-Barriuso R, Oteo J, Hernán MA, Pérez-Olmeda M, et al. “Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study.” *Lancet Lond Engl*. 2020 Aug 22;396(10250):535–44

[10] Stringhini S, Wisniak A, Piumatti G, Azman AS, Lauer SA, Baysson H, et al. “Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-POP): a population-based study.” *The Lancet*. 2020 Aug;396(10247):313–9

[11] Ward H, Cooke G, Atchison C, Whitaker M, Elliott J, Moshe M, et al. “Declining prevalence of antibody positivity to SARS-CoV-2: a community study of 365,000 adults” [Internet]. Infectious Diseases (except HIV/AIDS); 2020 Oct [cited 2021 Feb 2]. Available from: http://medrxiv.org/lookup/doi/10.1101/2020.10.26.20219725

[12] Sood N, Simon P, Ebner P, Eichner D, Reynolds J, Bendavid E, et al. “Seroprevalence of SARS-CoV-2–Specific Antibodies Among Adults in Los Angeles County, California, on April 10-11”, 2020. *JAMA*. 2020 Jun 16;323(23):2425–7

[13] Silveira MF, Barros AJD, Horta BL, Pellanda LC, Victora GD, Dellagostin OA, et al. « Population-based surveys of antibodies against SARS-CoV-2 in Southern Brazil.” *Nat Med*. 2020 Aug;26(8):1196–9.

[14] Warszawski J, and al. « 4 % de la population a développé des anticorps contre le SARS-CoV-2 entre mai et novembre 2020 ». *Études et Résultats*, 2021, n°1202

[15] Warszawski, J. and al. “A national mixed-mode seroprevalence random population-based cohort on SARS-CoV-2 epidemic in France: the socio-epidemiological EpiCov study”, medRxiv 2021.02.24.21252316; doi: https://doi.org/10.1101/2021.02.24.21252316

[16] Carrat F, and al., “Seroprevalence of SARS-CoV-2 among adults in three regions of France following the lockdown and associated risk factors: a multicohort study”, available at medRxiv 2020.09.16.20195693; doi: https://doi.org/10.1101/2020.09.16.20195693

[17] Lydié N, Saboni L, Gautier A, Brouard C, Chevaliez S, Barin F, et al. “Innovative Approach for Enhancing Testing of HIV, Hepatitis B, and Hepatitis C in the General Population: Protocol for an Acceptability and Feasibility Study (BaroTest 2016).” *JMIR Res Protoc*. 2018 Oct 12;7(10):e180.

[18] Merly-Alpa T, Sillard P. “French admin dataset Fidéli” [Internet]. CROS - European Commission. 2019 [cited 2021 Feb 4]. Available from: https://ec.europa.eu/eurostat/cros/content/french-admin-dataset-Fidéli_en

[19] Loonis V. La construction du nouvel échantillon de l'enquête emploi en continu à partir des fichiers de la taxe d'habitation. 2009

[20] Gallian P, Pastorino B, Morel P, Chiaroni J, Ninove L, de Lamballerie X. Lower prevalence of antibodies neutralizing SARS-CoV-2 in group O French blood donors. *Antiviral Res*. 2020 Sep;181:104880

[21] Hazo J.B., Costemalle V. et al. "Confinement du printemps 2020 : une hausse des syndromes dépressifs, surtout chez les 15-24 ans Résultats issus de la 1re vague de l'enquête EpiCov et comparaison avec les enquêtes de santé européennes (EHIS) de 2014 et 2019", *Études et Résultats*, 2021, n°1185

[22] Hazo J.B, Costemalle V., Rouquette A., Bajos N. et al. "Une dégradation de la santé mentale chez les jeunes en 2020 – Résultats issus de la 2e vague de l'enquête EpiCov", *Études et Résultats*, 2021, n°1210

[23] Särndal C.-E., Swensson B., Wretman J. (1992), Model-assisted survey sampling, Springer, chapter 4

[24] Fuller W. (2009), Sampling Statistics, Wiley, chapter 1

[25] Haziza D., Lesage E. (2016), A discussion of weighting procedures for unit nonresponse, *Journal of Official Statistics*, vol. 32 n°1, p.129-145

[26] Haziza D., Beaumont J.-F. (2017), Construction of weights in surveys: a review, *Statistical Science*, vol. 32 p.206-226

[27] Haziza D., Beaumont J.F., "On the construction of imputation classes in surveys", *International Statistical Review*, vol.75 p.25-43

[28] Deville J.C., Särndal C.E. (1992), Calibration estimators in survey sampling, *Journal of the American Statistical Association*, vol.87 n°418, p.376-382

[29] Estevao V., Särndal C.E. (2002), The ten cases of auxiliary information for calibration in two-phase sampling, *Journal of Official Statistics*, vol.18 n°2, p.233-255.

[30] Deville J.C., Lavallée P. (2006), "Indirect sampling: foundations of the generalised weight share method", *Survey Methodology*, vol. 32 n°2, p.185-196

[31] P. Lavallée (2007), *Indirect Sampling*, Springer

[32] Jabot P., Treyens P-E., « Appariement de l'enquête Care par identification du plus proche écho », *Journées de méthodolgoie statistique de* l'Insee, 2018 available at http://www.jms-insee.fr/2018/S20_1_ACTEv2_TREYENS_JMS2018.pdf

[33] Beavis KG, Matushek SM, Abeleda APF, Bethel C, Hunt C, Gillen S, et al. "Evaluation of the EUROIMMUN Anti-SARS-CoV-2 ELISA Assay for detection of IgA and IgG antibodies." J Clin Virol Off Publ Pan Am Soc Clin Virol. 2020 Aug;129:104468.

[34] Krüttgen A, Cornelissen CG, Dreher M, Hornef M, Imöhl M, Kleines M. "Comparison of four new commercial serologic assays for determination of SARS-CoV-2 IgG." J Clin Virol. 2020 Jul;128:104394.

[35] Kohmer N, Westhaus S, Rühl C, Ciesek S, Rabenau HF. "Clinical performance of different SARS-CoV-2 IgG antibody tests" J Med Virol. 2020 Oct;92(10):2243–7.

[36] Razafindranovona T. : « La collecte multimode et le paradigme de l'erreur d'enquête totale », *Document de travail INSEE*, 2015, available at https://www.insee.fr/fr/statistiques/1381054

[37] Heckman J. "Sample Selection Bias as a Specification Error" *Econometrica*. 1979;47(1):153–61

[38] Morrissey K, Kinderman P, Pontin E, Tai S, Schwannauer M. "Web based health surveys: Using a Two Step Heckman model to examine their potential for population health analysis." *Soc Sci Med* 1982. 2016 Aug;163:45–53

[39] Galimard J-E, Chevret S, Curis E, Resche-Rigon M. Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors. BMC Med Res Methodol. 2018 Aug 31;18(1):90.

[40] Sillard P, Beck F, Castell L, Favre-Martinoz C, Givord P, Legleye S, et al. Development of online and mixed-mode national household surveys: the missing-not-at-random (MNAR) challenge. JRSS. 2021 submitted

[41] Berger, N., Van Oyen, H., Cambois, E. et al. Assessing the validity of the Global Activity Limitation Indicator in fourteen European countries. *BMC Med Res Methodol* 15, 1 (2015). https://doi.org/10.1186/1471-2288-15-

[42] Dauphin and Eideliman 2021, « Élargir les sources d'étude quantitative de la population handicapée : Que vaut l'indicateur « GALI »? », *Dossiers de la DREES*, n°74, 2021

[43] Jauneau Y., Vidalenc J., Une photographie du marché du travail en 2020 – L'emploi résiste, le halo autour du chômage augmente. *Insee Résultats*, 2021

[44] Charrance G. et al., Apprendre des paradonnées pour améliorer les protocoles de collecte : l'exemple d'EpiCov, *Journées de méthodologie statistique*, 2022