- *Full view* -

EHIS_NESQRS_5_FR_2019_0000

National Reference Metadata in ESS Standard for Quality Reports Structure (ESQRS)

Compiling agency: Direction de la recherche, des études, de l'évaluation et des statistiques (Drees) Institut de Recherche et Documentation en Economie de la Santé (Irdes)

Eurostat metadata

Reference metadata

For any question on data and metadata, please contact: EUROPEAN STATISTICAL DATA SUPPORT

1. Contact

Top

1.1. Contact organisation	Direction de la recherche, des études, de l'évaluation et des statistiques (Drees) Institut de Recherche et Documentation en Economie de la Santé (Irdes)
1.2. Contact organisation unit	Sous-direction Observation de l'Assurance Maladie, Bureau Etat de Santé de la population Observation of health insurance departement, Health State of Population Unit
1.3. Contact name	Thomas Deroyon (Drees) / Thierry Rochereau (Irdes)
1.4. Contact person function	Methodologist in charge of surveys on population health / Researcher
1.5. Contact mail address	Drees : 10 Place des Cinq Martyrs du Lycée Buffon, 75015 Paris, France / Irdes : 117 bis rue Manin, 75019 Paris, France
1.6. Contact email address	thomas.deroyon@sante.gouv.fr / rochereau@irdes.fr
1.7. Contact phone number	Thomas Deroyon : +33140568785 Thierry Rochereau : +33153934332
1.8. Contact fax number	[not requested]

2. Statistical presentation

Top

2.1. Data description

- Survey name(s) in the national language(s):

Enquête santé européenne 2019

- Survey name in English:

European health interview survey 2019

- Link to the survey website:

https://drees.solidarites-sante.gouv.fr/sources-outils-et-enquetes/enquete-sante-europeenne-ehis

and

https://www.irdes.fr/recherche/enquetes/ehis-enquete-sante-europeenne/actualites.html

2.1.1. Combination of EHIS with another survey/questionnaire

Type	Name of the survey that hosted the EHIS questionnaire
Health Interview Survey	
Health Examination Survey	
Disability Survey	
Labour Force Survey	
Living Conditions Survey	
Multipurpose Survey	
Other	

2.1.2. Indication of the type of survey if 'Multipurpose Survey' or 'Other' are marked

2.2. Classification system

2.2.1. Versions and breakdowns (level) of the classifications used for the data collection

Acronym	Version	Level
NACE	2008	1
ISCO	2008	2
ISCED	2011	A
ICD	11	

IPC		
ICF		
NUTS	2016	2
LAU		2
DEGURBA		2

2.2.2. Deviations from ESS or international standards

2.3. Coverage - sector

[not requested]

2.4. Statistical concepts and definitions

No deviation

2.5. Statistical unit

The sampled, collection and dissemination units are individuals

2.6. Statistical population

Population of 15 years old or more residing in an ordinary household as principal residence in mainland France

2.6.1. Main characteristics of the survey population

See **ANNEX 1** Survey population characteristics.

Annexes:
Annex 1

2.6.2. Participation and non-participation in the survey

See **ANNEX 2** Summary table on participation and non-participation.

Annexes:
Annex 2

2.6.3. Structure of the target population, of the sample population, and of response and non-response

See **ANNEX 3** Structure of target, sample, response & non-response population.

Annexes:
Annex 3

2.7. Reference area

Mainland (Metropolitan) France : including Corsica and excluding oversea departments (Guyane, Guadeloupe, Martinique, Reunion, Mayotte).

2.8. Coverage - Time

Year 2019. Data collected during 9 months, with a first wave between April and July (more precisely, between 1st April 2019 and 1st August 2019) and a second wave between September 2019 and January 2020 (more precisely between 27th august 2019 and 1st february 2020).

2.9. Base period

[not requested]

3. Statistical processing

Top

3.1. Source data

3.1.1. Sampling frame

Fideli : fichier démographique pour les logements et les invidus (demographic file for dwellings and individuals)

3.1.1.1. Type and name of data source used for building the sampling frame

Type	Name of data source used for building the sampling frame
Population register	
Household register	
Dwelling register	
List of phone numbers	
Postcode address file	
Another survey sample	
Other	Fideli : administrative file built with tax files from the fiscal administration, on dwellings, tax households and individuals

3.1.1.2. Description of data source used for building the sampling frame

Fideli is a file built by Insee, the French National Statistical Institute, with administrative files produced by the fiscal administration. Fideli uses files on :
- fiscal compulsory declarations on incomes that fiscal households have to complete everey year ;
- tax on dwellings property (taxe foncière) and occupation (taxe d'habitation) ;
- dwellings maintained by the fiscal administration for its own purposes.

These data are completed with information linked to the precise localisation of the dwellings : geographic coordinates, caracteristics of the dwelling's neighbourhood especially its administratrive status (does this neighbourhood belong to the category of porr neighbourhodd benefiting from special policies).

More information on Insee's website : https://www.insee.fr/fr/metadonnees/source/serie/s1019

3.1.1.3. Frequency of the updates of data source used for building the sampling frame

Yearly

3.1.1.4. Details if 'Every … years' or 'Irregular' are marked

3.1.1.5. Date(s) of the data source used for the selection of the sampling units

2018 : it was the most recent version of Fideli available at the time the sampled was selected

3.1.2. Sampling design of the survey

3.1.2.1. Ultimate sampling unit(s)

Individuals

3.1.2.1.1. Number of households belonging to a selected dwelling interviewed if 'Dwellings' is marked

3.1.2.1.2. If 'More than 1 household' is marked, specification of the number

3.1.2.1.3. Number of individuals belonging to a selected household interviewed if 'Households' is marked

3.1.2.1.4. If 'More than 1 individual' is marked, specification of the number

3.1.2.2. Sampling design(s)

Multi-stage sampling

3.1.2.2.1. Specification if ''Combination of designs' is marked

3.1.2.2.2. Stratification variables if 'Stratified Sampling' is marked

3.1.2.2.3. List of the different stages and the probabilities for every stage if 'Multiple Stage Sampling' is marked

As EHIS 2019's data collection involved a part of face-to-face surveys, and on the suggestions of the statisticial methodological unit of Insee, a multistage sampling was used to select EHIS' sample. This multistage sampling was a two stage sample. At the fist stage, a sample of primary units consisting of gathering of at least neighbouring 2500 dwellings was selected ; then a sample of individuals was selected in each of the first stage primary units.

Sampling design at the first stage :

At the time EHIS's sample was selected, Insee was carrying out the important task of renovating its master sample used for the data collection of its main household surveys. The first stage of EHIS's sampling design used a preliminary version of the primary units Insee used for its new master sample. More details on this new master sample and the methodology involved in http://jms-insee.fr/jms2018s09_1 and https://www.insee.fr/fr/information/4497081?sommaire=4497095

Primary units are composed to contain at least 2500 dwellings (more precisely, 2500 ordinary dwellings used as principal residence) as close as possible of each other so as to limit the costs and burdens of data collection for face to face interviews. They are based on towns or gathering of towns respecting the frontiers of each French department (departments in France correspond to NUTS3 division of the country). Numerous applications of a salesman problem resolution algorithm on each department were launched and for each departement, the most compact primary units constituted according this algorithm were chosen. The primary units finally chosen by Insee were marginally modified and validated by Insee's regional services after this algorithmis step. This step was not integrated in the primary units used for EHIS's selection, which were only the produce of the algorithmic construction.

Primary units were built based on the French official geographical code available at the 1st of January 2016, defining the lists and compositions of towns. 35 722 towns were gathered in 5 155 primary units.

The sample of primary units selected at the first stage of EHIS consists of 250 primary units. The sampling design is stratified according to the 13 NUTS1 regions of mainland France. Probabilities of selection of each primary units are computed according to the following steps :

- the number of primary units to be selected in each stratum is proportional to the number of dwellings (ordinary dwellings used as princial residence) in the stratum ;
- in each stratum, the probability of selection of each primary unit is proportional to its size in terms of number of dwellings ;
- after this step, some primary units have a selection probability greater than 1. They are automatically selected in the sample and called exhaustive primary units. The calculation of selection probabilities in each stratum is repeated until all probabilities of selection are inferior or equal to 1.

Due to roundings in the computation of selection probabilities, the size of the sampled is finally equal to 249 primary units.

Once probabilities of selection are computed, the sample of primary units is selected according to a doubly balanced spatial sampling, corresponding to the algorithms described in the following scientific publications :

- Grafström, A. & Tillé, Y. (2013), Doubly balanced spatial sampling with spreading and restitution of auxiliary totals, Environmetrics, 2013, 24(2), 120-131.
- Chauvet, G. (2009), Stratified balanced sampling, Survey Methodology, 35(1), 15-119.
- Deville, J.-C. & Tillé, Y. (2004), Efficient balanced sampling: the cube method, Biometrika, 91(4), 893-912.

To implement these algorithms, the R package BalancedSampling (http://www.antongrafstrom.se/balancedsampling/) was used.

The idea of the algorithm of doubly balanced spatial sampling is to enrich the traditional approach of balanced sampling, introduced by Deville and Tillé, whose goal is to select a sample in which the estimates of some balancing variables totals are equal to the totals known in the population, with a spatial dimension : the algorithm avoids selecting units spatially close to each other, to avoid the loss of precision due to spatial correlation of socioeconomic variables.

The balancing variables used in the selection of the sample are the following ones :

- probability of selection (balancing variable used to guarantee the fixed size of the sample) ;
- fiscal total income of the primary unit in 2014 ;
- total amount of old age pensions of the primary unit in 2014 ;
- population aged between 15 and 29 years old in the primary unit ;
- population aged between 30 and 44 years old in the primary unit ;
- population aged between 45 and 59 years old in the primary unit ;
- population aged between 60 and 74 years old in the primary unit ;
- population aged between 75 and 89 years old in the primary unit ;
- population aged of 90 years old or more in the primary unit ;
- number of dwellings belonging to a city in the primary unit ;
- number of dwellings belonging to the direct surroundings of a city in the primary unit ;
- number of dwellings belonging to the periurban space in the primary unit ;
- number of dwellings belonging to rural areas in the primary unit.

The selection of the sample is made according to the following steps :

- initial fly phase in each stratum ;
- second fly phase at the national level, for primary units whose selection in the sample has not yet been decide
- final landing phase carried out in each stratum, without spatial spreading.

Finally, 249 primary units were selected :

- 57 in NUTS1 FR1, including 20 exhaustive primary units ;
- 11 in NUTS1 FRB, none of which are exhaustive ;
- 12 in NUTS1 FRC, none of which are exhaustive ;
- 14 in NUTS1 FRD, none of which are exhaustive ;
- 24 in NUTS1 FRE, 1 being exhaustive ;
- 23 in NUTS1 FRF, 1 being exhaustive ;
- 15 in NUTS1 FRG, 1 being exhaustive ;
- 14 in NUTS1 FRH, none of which are exhaustive ;
- 25 in NUTS1 FRI, one being exhaustive ;
- 23 in NUTS1 FRJ, including two exhaustive primary units ;
- 39 in NUTS1 FRK, including 9 exhaustive primary units ;
- 33 in NUTS1 FRL, including 17 exhaustive primary units ;
- 1 non exhaustive primary unit in NUTS1 FRM.

Sampling design at the second stage :

The second stage is a classical self-weighted stage of individual selection, whose goal is to limit the variablity of estimation weights. However, due to the specificity of the population and its higher non response rate, at the second stage, individuals were selected according to a systematic stratified sampling. In each primary unit, population was divided between two strata, according to whether persons live in a priority neighbourhood for urban policy (QPV: quartiers prioritaires de la politique de la ville, priority neighbourhood for urban policy, are poor neighbourhoods defined by the ministry of urban policy and benefiting from specific policy measures). Probabilities of selection at the second stage were twice higher in the strata of individuals living in QPV.

In each of the two strata, the sample is selected according to a systematice sample algorithm, units being sorted by the following variables :

- a variable equal to one if a phone number was available in the sampling frame in the fiscal household of the individual ;
- age of the individual, divided according to the following age groups : 15-24 years old, 25-34 y.o., 35-44 y.o., 45-54 y.o, 55-64 y.o., 65-74 y.o. and 75 years old or more ;
- revenue of the household, defined as the total fiscal revenue of the household divided by the number of consumption units in the household.

In order to limit the statistical burden suffered by each household, in each household one person was selected according to a simple random sampling. The systematic sampling was applied to a file in which the observations corresponding to the non selected members of the household were replaced by the characteristics of the selected person. If the household contains m persons, the observation corresponding to the selected person was repeated m times. This technique enabled us to maintain a non-biased sampling selection whereas preventing from selecting more than one person per household.

The final sample contains 27 600 individuals, 3752 living in a QPV.

3.1.2.3. Oversampling of specific populations

Persons living in poor neighbourhood benefiting from specific policies (priority neighbourhoods for urban policy) are oversampled. Their selection probability is twice higher than that of persons not living in priority neighbourhoods for urban policy.

3.1.2.4. Stratified oversampling methods

3.1.2.5. Methods used for drawing up the sample

At the first stage, the sample of primary units was selected according to the method of doubly balanced spatial sampling developed by A. Grafström and Y. Tillé. At the second stage, individuals were selected according to a classic systematic sampling applied on sorted file.

3.1.2.6. Assumptions used for determining the sample size

A the first stage, the number of primary units selected was determined according to the needs of Insee, the French National Statistical Institute, for the constitution of its new master sample, destined to be used for 10 years. The characteristics of primary units and the number of primary units selected were chosen according to Insee's needs, especially the number of interviewers available.

The number of respondents needed to abide by the precision requirements set in annex II of commission regulation n°2018/255 implementing regulation n°1338/2008 of the European Parliament and of the Council of 16 December 2008 on Community statistics on public health and health and safety at work, if the sample is selected according to a simple random sampling without replacement is around 11 500. The sampling design used to select EHIS's sample is however not a simple random sampling, but a two-stage sampling design entailing a loss of precision due to the cluster effect. The sampling design used to select primary units at the first stage of the sample is however new and we lack elements to judge on real data the real design effect it implies. The simulations realised by Insee during the construction of its new master sample showed however that the design effect could be small. We therefore made the assumption that a number of 12 000 respondent was sufficient to fulfill the precision requirements set by the regulation.

This expected number of respondent was used to compute the needed sample size with the following hypotheses on non response rates. As described in the section on data collection, data are collected using cati and capi surveys. More precisely, the individuals for which we have no phone number in the sampling frame are directly collected face to face, whereas individuals with phone numbers are first contacted by phone. A random half of individuals contacted by phone but that our interviewers were not able to reach are then also contacted by face to face interviewers.

Our hypotheses on non response rates for these populations were the following :

- we will have a phone number in the sampling frame for 70 % of the sample ;
- 50 % of individuals contacted by phone will respond ;
- 60 % of individuals initially contacted face to face will respond ;
- 30 % of persons initially contacted by phone and then sent to the face to face survey will respond.

These response rates are a little lower than the ones observed on households surveys on the field in 2017, at the time the sampling design or EHIS was determined.

With these hypotheses on non response rates, the sample size should be of 21 600 individuals. We nevertheless selected a larger sample of 27 600 individuals in which we randomly selected a main sample of 21 600 individuals and a reserve sample to be used if response rates were lower than expected. In fact, this reserve sample was used, so that the final size of EHIS's sample is 27 600 individuals.

3.2. Frequency of data collection

There is currently no other national health survey than EHIS. The frequency of data collection is thus that of EHIS waves.

3.3. Data collection

3.3.1. Data collection method used

Mixed mode

3.3.2. Mode(s) for data collection

Face-to-face, electronic version
Telephone, electronic version

Annexes:

elements on data collection modes for EHIS 2019

3.3.2.1. Specifications if 'Other' is marked

3.3.3. Topics (submodules/ variables) administered via a self-completion questionnaire

For CAPI-mode, the following submodules were administered via a self-completion questionnaire on tablets (to persons having accepted):

- MH
- SK
- AL

3.3.4. Variables completed from an external source

HHINCOME from national tax files. The original information stems from the sampling frame, its reference year is 2018.

3.4. Data validation

The first data validation step stands in the construction and validation of the questionnaire, by implementing the filters as required in the EHIS manual, so as to ensure that every question was asked to the relevant individuals. The correctness of the programming of the questionnaire for both CAPI and CATI modes have been validated carefully.

We also ensured that every question was mandatory to be answered, letting always the possibility to answer 'don't know' or 'refuse to answer' - but without citing these response categories - in order to reduce item non-response as much as possible.

As regards open questions, we implemented lower and upper limits to responses according to the EHIS manual, so as to avoid outliers or comprehension mistakes.

For example, we implemented a control to take into account the response given in PE6 on the number of days per week when the respondent practise leisure physical activity in the response given to PE7 : the response to PE6 is used to define the range of values possible in PE7.

Therefore, the data validation process after receiving the datasets was relatively limited. Some variables had to be constructed using several variables from the French questionnaire, such as HHTYPE, constructed using the number of people living in the household and the links between them. For many variables, no changes were made as regard response categories. For all variables we implemented the codes -1 for not stated, -2 for not applicable and -3 for proxy. we however checked the coherence of the responses given to the survey.

For instance, we had a lot of difficulties with the questions on unmet needs, who were hard to understand for interviewees and interviewers as well. These elements were already mentioned after the survey's tests and were confirmed after the survey field. Interviewers in their final feedback on the survey field mentioned that these questions were not properly

phrased and were very hard to understand for interviewees. The analyses we realised on the responses to these questions show a high number of incoherences we finally chose to let in the data as they are.

For instance persons declaring in questions UN2 that they were not concerned by the medical cares mentioned also mentioned in previous questions they resorted to health care during the preceeding 12 months. It is the case of 87 % of persons declaring they are not concerned in question UN1A, 88 % for question UN1B and 87 % for question UN1C. These answers are incoherent, and the high prevalence of these incoherences show a major difficulty in understanding the questions, whose formulation should be reviewed.

We also realised controls ex-post to check whether the respondent was indeed the sampled person. We compared for each respondent the information on its sex, date of birth and first name with the same informations available in the sampling frame (the comparisons on first names were implemented by the firm in charge of data collection, as we could not have access to these data). A few cases corresponded to proxy situations that were not correctly identified by the interviewer and were corrected. Other situations could however not correspond to proxies. This enabled to identify a limited number of cases where the person ho gave the answers was not the sampled person. These cases were dismissed and treated as non response. 50 persons were in that situation.

3.5. Data compilation

Data compilation steps will be detailed in the following sections.

The first step was to identify respondents, nonrespondents and out-of-scope units in the sample. For this, we analysed the results codes of the different contact attempts for CATI and CAPI data collection. We also analysed the patterns of partial nonreponse to part respondents with partial nonresponse from questionnaires to seldomly filled to be used.

Once this first step was realised, data editing and imputation on the one side and weights computation and the other side, were carried out in parallel. Standard methods were used for both processes : imputations have for instance been implemented with standard hotdeck techniques, and total nonresponse treatments with usual homogeneous nonresponse methods.

Data editing and imputation was focused on questions necessary to produce eurostat microdata files. Variables specific to the French questionnaire are still being processed.

3.5.1. Method applied to correct for 'item non-response'

Simple imputation (stochastic) method

3.5.1.1. Details of the method if ‘Multiple imputation approach‘ or ‘Other’ are marked

3.5.1.2. Auxiliary information used for stratification

Item nonresponse was treated with random hotdeck imputation in imputation cells. Groups of correlated variables were imputed together with the informations coming from the same donor to prevent imputation from altering relationships between variables. Imputation cells were based on values of the following variables : age, taken into accouts with five years groups, household income per consumption unit, type of household and sex. In case a variable is filtered by the responses to a previous variable, the values of the filtering variables are also taken into account to build the imputation cells.

3.5.2. Calculation of weighting factors and weight adjustments

3.5.2.1. Method for calculation of weighting factors

Weighting factors take into account the following elements:

- sampling design ;
- total non response treatments ;
- calibration.

More precisely, the final estimation weights are obtained by applying the calibration described in the following sections to the pre-calibration weighting factors equal to the weights taking into account the sampling design divided by the estimated response probabilities.

Sampling Design :

As described earlier, the sample was selected with a two-stage sampling design. Estimation weights taking into account this design are obtained as the inverse of the product of :

- the inclusion probability of the primary unit each sampled individual belong to. These probabilities are proportional to the number of ordinary dwellings used as primary residence at the time the sampling frame was built in the primary units ;
- the inclusion probability of each individual used at the second stage, for the selection of the sample of individuals in each primary unit, conditional on the sample of primary units selected at the first stage. These probabilities are determined according to a self-weighted sampling scheme, taking into account a stratification on whether the individual's dwelling in the sampling frame belongs to a priority neighbourhoud for urban policy. The inclusion probability of individuals inhabiting a priority neighbourhood for urban policy are twice higher.

Total nonresponse treatment :

Total nonresponse is corrected through classical reweighting methods. Nevertheless, due to the data collection process used for EHIS 2019, total nonresponse treatment is implemented in two steps.

Indeed, the initial sample is first divided into two parts : individuals for which no phone numbers are available in the sampling frame are directly collected by face-to-face CAPI interviaws. Individuals for which phone numbers are available in the sampling frame are first collected by phone. Phone interviewers will however be unable to have a direct contact with a part of them. For half of this part, data collection is resumed with face-to-face interviews.

Hence :

- for individuals directly contacted with face-to-face surveys, we select and estimate a logit model explaining their probability to answer to the survey based on information available for respondents and non respondents (these variables will be detailed below). This model gives us an estimation of the response probability we use as an input to create homogeneous response groups with the algorithm suggested by J.F. Beaumont and D. Haziza (see Haziza, D., Beaumont, J.F., On the construction of imputation classes in surveys, International Statistical Review, 2007, 75(1)). We then estimate response probabilities in these homogeneous response groups as the weighted mean of response indicator (the weights used in this computation are the estimation weights taking into account the two phases of the sample selection). The weights after nonresponse treatments are computed by dividing the estimation weight taking into account the initial sample selection by the estimated response probabilities obtained with the homogeneous response groups.
- for individuals firstly contacted by phone, total nonresponse is treated in two steps, as the data collection. First, we select and estimate a model to predict the fact an individual has been either respondent by phone, or not contacted by an interviewer. The variable we apply the model to is therefore equal to zero for individuals contacted by interviewers who did not answer to the survey.

We use this model to estimate a probability of being either respondent or non contacted with homogeneous response groups, as described for direct CAPI interviews. Then :

- for respondents, their nonresponse treated weight is obtained by dividing their estimation weight taking into account the two step of initial sample selection by their estimated probability of being either respondent or non contacted ;
- for persons not contacted, we identify the ones that have been selected to participate to the CAPI data collection, and build a model to estimate their probability of responding in face to face interviews. This model is used to obtain an estimation of their response probability, with homogeneous response groups as for direct CAPI interviews. Their final nonresponse treated weight is equal to the initial estimation weight taking into account the two step of sample selection, divided by the estimated probabilities of being respondent or non contacted by phone and being respondent in face to face data collection, and multiplied by 2.

The weights obtained with the nonresponse treatment process we described are then submitted to a calibration on margins to obtain the final weights.

3.5.2.2. Adjustments applied to mitigate non-response (weight adjustments)

The total nonresponse adjustment process, as we mentioned earlier, uses multiple models :

- a first model describes the probability of responding to the survey versus nonresponse for individuals directly submitted to a face to face data collection ;
- another model describes the probabity of responding or not being contacted, versus being contacted and not responding for individuals submitted to a phone data collection ;
- a final model describes the probability of respondinf versus non response for individuals firstly collected by phone that CATI interviewers were not able to contact and that were randomly selected to face another step of CAPI data collection.

For each of these models, the process of model selection, the initial list of auxiliary variables and the final use of the model are the same. We then describe these three elements:

Model selection

The initial sample participating in the estimation of the model is randomly divided into two parts, a learning sample representing 66 % of the initial sample and a test sample made of the rest of the sample. Different models ares bult on the learning sampled and applied to the test sample which is used to assess their accuracy.

For each model, parameters are estimated on the learning sample, and applied without modification on the test sample data, so that these data do not take part in the estimation of the parameters. The accuracy of the different models are compared on a set of metrics typical of problems of classification in whci the goal is not to predict the class but to estimate the probability according to which a person belongs to a class : Brier score, average quadratic and absolute errors, log-loss, F-statistics, sensitivity, specificity, precision, negative predictive value.

The model which performs best for the majority of the metrics is chosen.
The different algorithms compared, for each model, are the following :
- a classic CART algorithm ;
- a random forest on all auxiliary variables ;
- a logit model on all auxiliary variables ;
- a logit model on all auxiliary variables on which a stepwise selection process is applied.

Auxiliary variables

The auxiliary variables we use for our nonresponse models come from two sources :
- the sampling frame, that is fiscal data files ;
- the geographic informations we can link to the adress od the sampled individuals in the sampling frame.

The auxiliary variables using geographic information are :
- relative to the urban unit (https://www.insee.fr/fr/metadonnees/definition/c1501) the adress belongs to : size and nature of the urban unit ;
- relative to the urban area (https://www.insee.fr/fr/metadonnees/definition/c2070#:~:text=Une%20aire%20urbaine%20ou%20%C2%AB%20grande,r%C3%A9sidente%20ayant%20un%20emploi%20travaille) the adress belongs to : size and nature of the urban area, nature of the town of residence inside the urban are ;
- relative to the town the adress belongs to : number of dwellings, share of principal residence among dwellings, share of social housing among dwellings, share of primary residence owned by their inhabitants among primary residences of the town ; number of births in the town divided by the size of the population for the last year census data were available (which was 2017 at the time of nonresponse treatments), number of deaths divided the the size of the population, growth rate of the town population between 2012 and 2017 ; activity and unemployment rate in the town, share of the population and active age (between 15 and 64 years old) ; share of the active population working into the different economic sectors at the nace selection level.

The auxiliary information coming from the sampling frame are :
- sex and age of the sampled person ;
- disposable income ;
- variables indicating the presence of certain types of incomes in the sampled person income : pensions, rents, salaries, unemployement benefits, income from independant activities, social security benefits, family allowances ;
- variables describing the household of the sampled person in the sampling frame : size of the household, presence of at least a person of a certain age, with 10-years age groups, presence of at least one man, one woman in the household.
- variables describing the dwelling the sampled person lives in : is this dwelling a house or an appartment, is the sampled individual the owner of the dwelling or a tenant,...

Final use of the model

Once the final model is estimated, it is used to obtain estimates of response probabilities for each individual (respondent and non respondent). We then apply to all the individual taking part in the estimation of the model the algorithm of Haziza and Beaumont. The goal of this algorithm is to divide the sample into groups of units whose estimated response probabilites are close. To do so, the algorithme uses the following steps :
- first, we compute a distance between each pair of individuals taking part in the process and the square difference between their estimated response probabilities ;
- we start by dividing the sample into two groups, with a k-means algorithm. We randomly select a 1000 different starting points for the algorithm and select the two groups whose within sum of square is the lowest.
- we then estimate a model linking the initial estimated response probability coming from the selected logit model and the two variables indicating to which of the two groups each individual belongs. If the R2 of the model is higher than a given threshold we set à 99 %, we stop the algorithm and use the two groups obtained. Otherwise, we resume the process with three groups. The algorithme carries on until we reach a number of groups sufficient to be able to explain more than 99 % of the information contained in the initial estimated response probabilities.

We then estimate the final response probabilities as the sum of the weights of respondents in each group, divided by the sum of the weights of respondents and nonrespondents in this group.

3.5.2.3. Adjustments (calibration techniques) applied and list of the external data sources

We apply a calibration to the sample after total nonresponse treatments.
Margins are relative to the target population at the time of the survey.
They are estimated with the sample of the Labor Force Survey for year 2019, as is usual for household surveys in France. LFS is indeed a survey with a large sample, whose target population matches the one of EHIS. Moreover, contrary to the sample of other surveys such as EHIS, the LFS sample contains also dwellings that are out scope at the time the sampling frame is built (because the dwellings are not occupied or are occupied as secondary dwellings). These dwellings may change their status at the time of data collection and become ordinary dwellings occupied as primary residence. Therefore, LFS gives a complete image of the target population at the time of data collection and helps limit biases introduced by undercoverage of target population.
The margins we introduce are the usual margins use to calibrate most of household surveys on general population implemented by Official Statistics offices in France :
- population by sex and groups of age : 15-24 years old, 25-44 years old, 45-64 years old, 65-74 year sold, 75 years old or more ;
- population by maximum level of education in four groups : more two years after baccalaureate, baccalaureate and baccalaureate plus two years, diplomas of a level inferior to baccalaureate, no diploma ;
- population by aggregated social category, in seven categories : clerks and workers ; former clerks and workers ; executive and technicians ; former executives and technicians ; independent workers, managers, former independent workers and managers ; farmers and former farmers ; persons having never worked ;
- population inhabiting a priority neighbourhood for urban policy ,
- population by nationality, in two modalities : french and not french ;
- population by groups of nuts2, in 8 groups : Île de france (FR10) ; Bourgogne (FRC1), Centre (FRB0), Champagne-Ardennes (FRF2), Basse Normandie (FRD1), Haute Normandie (FRD2), Picardie (FRE2) ; Nord Pas de Calais (FRE1) ; Alsace (FRF1), Franche Comté (FRC2), Lorraine (FRF3) ; Bretagne (FRH0), Pays de la Loire (FRG0), Poitou-Charentes (FRI3) ; Aquitaine (FRI1), Limousin (FRI2), Midi-Pyrénées (FRJ2) ; Auvergne (FRK1), Rhône Alpes (FRK2) ; La,guedoc-Roussillon (FRJ1), Provence Alpes Côte d'Azur (FRL0), Corse (FRM0).
- population by size of the urban unit : rural town, urban unit of less than 20 000 inhabitants, urban unit of less than 100 000 inhabitants, urban unit of less than 2 000 000 inhabitants, Paris urban unit.

Calibration was applied with the icarus R package (https://cran.r-project.org/web/packages/icarus/index.html), using the logit method imposing that the weights ratios are between 0.6 and 1.7.

3.5.2.4. Specification of other weight adjustments

No other weighting adjustments were used than those described in the previous sections (initial estimations weights deriving from the sampling design, total non response adjustments and calibration). Especially, non treatment of influential units is applied to data.

3.6. Adjustment

[not requested]

4. Quality management

Top

4.1. Quality assurance

The fieldwork company is certified ISO 20252

4.2. Quality management - assessment

The EHIS 2019 in France is recognised by the LABEL committee of the National Statistical Information Centre (CNIS) as of general interest and of statistical quality, and received the visa No.2019X070SA from the ministry of Economy and Finance.

5. Relevance

Top

5.1. Relevance - User Needs

Main users are :

The French ministry of Health.

For the ministry, EHIS is the main source giving information in the general population on the determinants of health, especially the body mass index, on health unmet needs and on certain health consumptions that are not reimbursed by the social security (for instance consultations with psychologists and psychotherapists). EHIS Data on health consumptions are otherwise not used to describe consumptions in the population, as they are less precise than social security data, but are used in analyses with other data collected in the sample. EHIS data will be matched with social security data on health consumption to study the differences between declared and reimbursed health consumptions.

In 2019, data were also collected in french overseas departments (who are out of scope of EHIS according to the European regulation) with a questionnaire close to the one used for EHIS in mainland France. It enabled us to pubish results on perceived health, unmet needs and health determinants in French ovserseas departments compared to mainland France.

The French ministry of urban policy.

The ministry of urban policy will use the survey data to compare health status, consumptions, unmet needs and determinants in the general population and in the prioriy neighbourhoods of urban policy.

The observatory of indoor air quality.

The OQAI (observatoire de la qualité de l'air intérieur, observatory of indoor air quality) proposed to EHIS's respondents an non compulsory ancillary survey on air quality. EHIS voluntary respondents could benefit from a measure of the indoor air quality of their dwelling that can be linked to the data collected on their health in EHIS.

Researchers on subjects link with health

Some research subjects are directly linked to questions that were introduced into the french versio of the questionnaire but that were not concerned by the European regulation :

- a subset of questions on litteracy in health from the internationally validated questionnaire HLQ. These questions will be analysed by researchers from Inserm (Institut national de la santé et de la recherche médicale, National institute of health and medical research) and Drees.
- a set of questions on supplementary health insurance, that will be analysed by the Irdes (Institut de Recherche et de Documentation en Economie de la Santé, Insitute of research and documentation on health economics) with Drees,

EHIS microdata are available to all researchers interested via the secur data access center (https://www.casd.eu/en/). Access to the data, due to their subject, is however only possible if autorised by the Cnil (commission nationale de l'informatique et des libertés, national commission for computer science and liberty, https://www.cnil.fr/en/home).

5.2. Relevance - User Satisfaction

No step has yet been taken to measure the users' satisfaction.

5.3. Completeness

All of the variables required for transmission have been included in the microdata

5.3.1. Data completeness - rate

[not requested]

6. Accuracy and reliability

Top

6.1. Accuracy - overall

Sources of errors in the disseminated data are numerous, as in all survey data :

- sampling errors : data are only collected on a small part of the population, which implies that estimates based on the sample differ from the real values of the parameters in the population. Due to the fact the sampling design used to select the sample is random, sampling errors are random errors generating estimation variance, and not systematic errors generating bias. Sampling errors are estimated through standard techniques
- non sampling errors : non sampling errors come to various sources :

Undercoverage of the sampling frame : the sampling frame we use, based on fiscal files, has a really high coverage rate of the population, so that ist undercoverage error can be neglected. However, the differences between the population at the time of data collection, which the survey's target population, and the population at the time the sampling frame is fixed, entails problems of overcoverage (units in the sampling frame belonging to the population of interest in the sampling frame but not in the target population) and undercoverage (units of the target population that did not belong to the population of the sampling frame). Overcoverage is detected through data collection, while undercoverage is treated through calibration on margins describing the target population.

Processing errors : Data are collected through cati and capi interviews, whose implementations have been tested prior to the real data collection.Numerous checks on data have been applied.

Measurement errors : Data are always collected through interviews, either by phone of by face to face, that can check that the interviewee understand the question and the associated concepts. It normally enables limiting satisficing but could introduce a social-desirability bias. The formation of interviewers insisted on the need of neutrality to avoid that bias.

Nonresponse errors : The data collection was managed to limit nonresponse and nonresponse bias. The data collection effort was for instance spread over all units of the sample thanks to a list of rules the data collection process by phone had to follow. After data collection, nonresponse was treated through reweighting involving two steps, modelling of nonresponse with auxiliary variables available for respondents and nonrespondents, followed by calibration on margins describing the target population. Hot-deck imputation was also used to treat item nonresponse.

6.2. Sampling error

Note: *Sampling errors* are the part of the difference between a population value and an estimate thereof, derived from a random sample, which is due to the fact that only a subset of the population is enumerated.

See **ANNEX 4** Table of sampling errors for selected variables.

Annexes:

Annex 4

6.2.1. Sampling error - indicators

[not requested]

6.3. Non-sampling error

Note: *Non-sampling errors* are errors in survey estimates which cannot be attributed to sampling fluctuations. Such errors can either be coverage errors, measurement errors, non-response errors, processing errors or model assumption errors.

6.3.1. Coverage error

Note: *Coverage errors* are errors that express the quantitative divergence between the sampling frame population and the target population due to, for example, remoteness, age, multiple entries; coverage of different sub-populations.

6.3.1.1. Over-coverage - rate

[not requested]

6.3.1.2. Common units - proportion

[not requested]

6.3.2. Measurement error

Note: *Measurement errors* are errors that occur during data collection and cause recorded values of variables to be different from the true ones.
See **ANNEX 5** Table of measurement errors from proxy interviews, survey questionnaire, interviewer, and quality control during fieldwork.

Annexes:
Annexe 5 - Measurement errors

6.3.3. Non response error

Note: *Non response errors* are errors that occur when the survey fails to get a response to one or possibly all of the questions.

6.3.3.1. Unit non-response - rate

See **ANNEX 6** Unit non-response and item non-response.

Annexes:
Annex 6 - Unit nonresponse and item nonresponse

6.3.3.2. Item non-response - rate

See above the **ANNEX 6** Unit non-response and item non-response in the concept 6.3.3.1.

6.3.4. Processing error

Note: *Processing error* is the error in final data collection process results arising from the faulty implementation of correctly planned implementation methods.

6.3.4.1. Imputation - rate

Imputation rate per variable are available in the attached file. For each variable, the imputation rate is equal to the number of observations imputed divided by the nomber of observations concerned by the question. Observations for which the question is not relevant are not taken into account in the number of observations for which the original values are kept.

Annexes:
Imputation rates

6.3.5. Model assumption error

Note: *Model assumption errors* are errors due to domain specific models needed to define the target of estimation.

6.4. Seasonal adjustment

[not requested]

6.5. Data revision - policy

[not requested]

6.6. Data revision - practice

[not requested]

6.6.1. Data revision - average size

[not requested]

7. Timeliness and punctuality

Top

7.1. Timeliness

Note: *Timeliness* is a measure for the length of time between data availability and the event or phenomenon the data describe.

7.1.1. Time lag - first result

11 months and fifteen days (the first microdata were sent the 15th december 2020)

7.1.2. Time lag - final result

15 months and 2 days (final results were sent on 2d April 2021)

7.2. Punctuality

Note: *Punctuality* measures the time lag between the actual delivery of the data to Eurostat and the target date when it should have been delivered.
There is a 6 months and 2 days lag between the date microdata files were due and the date the final results were sent. The French ministry of health, which produced EHIS 2019 in France, has been strongly mobilized in the management and statistical reporting of the Covid-19 crisis. That explains the delay with which microdata have been sent to Eurostat.

7.2.1. Punctuality - delivery and publication

6 months and 2 days

8. Coherence and comparability

Top

Note:
Coherence means the adequacy of statistics to be reliably combined in different ways and for various uses.
Comparability means the measurement of the impact of differences in applied statistical concepts, measurement tools and procedures where statistics are compared between geographical areas or over time.

8.1. Comparability - geographical

The results are fully comparable at the intra-national level. The questionnaire is the same in all metropolitan regions. However the sample if not big enough to disseminate results on every region (at NUTS1 or NUTS2 level) in France.

8.1.1. Asymmetry for mirror flow statistics - coefficient

[not requested]

8.2. Comparability - over time

Some major changes were introduced with EHIS 2019 in comparison to previous waves of the survey.
Starting with wave 3, EHIS waves will be produced by the statistical service of the ministry of health, Drees, with the help of Irdes.
Until EHIS 2014, EHIS sample was selected directly into files of persons affiliated to social security services for the reimbursement of their medical consumption. Sampled persons were first contacted by phone or by face to face. All the household members in the scope of EHIS were interrogated by a PAPI survey, which enabled covering a larger part of the population. The questionnaire was either sent by mail or left at home by the interviewer. Due to the fact that all members of a family respond to the survey, there was a cluster effect in the results of the survey.
Weights were calculated using weight share method and calibration techniques, with, after the application of the generalized weight share method due to the use of indirect sampling, a computation of final weights in one step by direct calibration, according to the terminology introduced by Haziza and Lesage (Haziza, D., Lesage, E., A discussion on weighting procedures for unit nonresponse, Journal of Official Statistics, vol.32, 2016).
Starting with wave 3, the sampling design was completely changed to match the methods used in other surveys of the Official Statistical System in France. The sampling frame was changed to the fiscal data files, that are now used on the main sampling frame for all household surveys on general population conducted in France by the National Statistical Institute. To enable data collection through face to face interviews, the sampling design is now a multistage design, with the direct selection of individuals inside primary unit of neighbouring dwellings. We directly sample individual, with at most one individual per household, to lighten response burden for each household and also to prevent cluster

effects in our estimates. Data collection modes also changed, as we moved from PAPI to CAPI and CATI surveys. Nonresponse treatment now involves two steps, a direct nonresponse treatment through reweighting based on auxiliary variables available for respondents and nonrespondents, folllowed by a calibration on reference margins. Data are collected through phone and face to face interviews.
Is is possible that all these changes introduce differences in the results of EHIS wave 2 and EHIS wave 3 that undermine the direct comparability of the results of these two surveys.

8.2.1. Length of comparable time series

[not requested]

8.3. Coherence - cross domain

Note: *Coherence* – cross domain is the extent to which statistics are reconcilable with those obtained from other data sources or statistical domains.

Annexes:
EHIS 2019 - Coherence across domains

8.4. Coherence - sub annual and annual statistics

[not requested]

8.5. Coherence - National Accounts

[not requested]

8.6. Coherence - internal

[not requested]

9. Accessibility and clarity

Top

Note: *Accessibility* and *clarity* mean the simplicity and ease, the conditions and modalities by which users can access, use and interpret statistics, with the appropriate supporting information and assistance.

9.1. Dissemination format - News release

The statistical service of the ministry of health published a news release to announce the dissemination of the first results of EHIS 2019 the 9/4/2021 : https://drees.solidarites-sante.gouv.fr/communique-de-presse/enquete-de-sante-europeenne-une-sante-percue-plus-degradee-dans-les

9.2. Dissemination format - Publications

The first results of EHIS 2019 were published by the statistical service of the ministry of health, Drees, the 9/4/2021 (in French) :
Leduc, A. et al.. : Premiers résultats de l'enquête santé européenne (EHIS 2019), Dossiers de la Drees n°78, avril 2021 (https://drees.solidarites-sante.gouv.fr/publications/les-dossiers-de-la-drees/premiers-resultats-de-lenquete-sante-europeenne-ehis-2019-metrople-guadeloupe-martinique-guyane-la-r%C3%A9union-mayotte)

9.3. Dissemination format - online database

Large aggregate results of EHIS 2019 are accessible online on the opendata site of the statistical service of the ministry of health : https://data.drees.solidarites-sante.gouv.fr/explore/dataset/indicateurs-ehis/table/

9.3.1. Data tables - consultations

[not requested]

9.4. Dissemination format - microdata access

EHIS 2019 microdata are accessible via the secure access data center (https://www.casd.eu/en/). Access is however limited to researchers whose project has been validated by the Cnil (https://www.cnil.fr/en/home, national commission on information and liberty) and the secret comity of Official Statistics (https://www.comite-du-secret.fr/).

9.5. Dissemination format - other

EHIS 2019 results were also used as reference data for survey data on covid 19, especially for mental health subjects. The publications dedicated to these subjects and that used EHIS 2019 data are the following (in French) :
Hazo, J.B. et al.. : Confinement du printemps 2020 : une hausse des symptômes dépressifs, surtout chez les 15-24 ans, Etude et Résultats n°1185, mars 2021 (https://drees.solidarites-sante.gouv.fr/publications/etudes-et-resultats/confinement-du-printemps-2020-une-hausse-des-syndromes-depressifs)
Hazo, J.B. et al.. : Une dégradation de la santé mentale chez les jeunes en 2020, Etudes et Résultats n°1210, octobre 2021 (https://drees.solidarites-sante.gouv.fr/publications/etudes-et-resultats/une-degradation-de-la-sante-mentale-chez-les-jeunes-en-2020)
EHIS 2019 are also used in other projects, whose results have not been published yet.

9.6. Documentation on methodology

Documentation on methodology is still being written and is not yet available to the public

9.7. Quality management - documentation

Documentation on quality is still being written and has not been published yet.

9.7.1. Metadata completeness - rate

[not requested]

9.7.2. Metadata - consultations

[not requested]

10. Cost and Burden

Top

Note: *Cost* associated with the collection and production of the statistical product and burden on respondents.

10.1. Cost of the survey

3 455 800 €

10.2. Time for answering the survey; if possible by data collection mode

10.2.1. Average interview duration for the EHIS questions (in minutes)

44

10.2.2. Minimum interview duration for the EHIS questions (in minutes)

13

10.2.3. Maximum interview duration for the EHIS questions (in minutes)

388

10.3. Measures taken to reduce the cost and burden of the survey

The content of the survey was limited to the questions necessary to produce european variables and some questions of special interest for France, such as the questions on complementary health insurances, which are very important to assess what share of health costs remains borne by the patients.

We also decided to match the EHIS respondents' file with the administrative social security data, which contain informations on all health consumptions reimbursed by the social security. This will help compare the answers given by the patients to the real health consumptions registered in the files. It will help us identify variables for which the two informations are close enough to replace the answers to the question by the administrative information in the new editions of EHIS.

11. Confidentiality

Top

Note: Steps carried out to prevent access to EHIS national microdata from unauthorised persons at each step of the production chain.

11.1. Confidentiality - policy

EHIS data collection had to abide by the principles on confidentiality defined in the following laws :

- the General Data Protection Regulation (GDPR) ;
- the law n°78-17 said "Informatique et Libertés", in which are defined the national provisions mentioned in the GDPR ;
- the law n°51-711 on obligation, coordination and secret in the matter of statistics. This law defines the legal conditions in which data collection is realised by the national statistical administrations in France.

The conformity of EHIS data collection process to the obligations defined in these laws has been checked by the CNIS (national council on statistical information) and the CNIL (national commission on computer science and liberty). EHIS was autorised by the CNIL the 29 april 2019 (decision n°DR-2019-116). It received the label of statistical quality and general interest by the label comity of the French Official Statistics (decision n°2019-7034).

11.2. Confidentiality - data treatment

The whole data process was reviewed and validated by the CNIL, whose goal is to check the citizen rights as far as their personal data are guaranteed, and especially that these data are well protected. The CNIL demands concerning data on health are very high.

Especially the data collection process was organised so that :

- the ministry of health, which will receive the answers to the survey, has no access to information directly identifying the respondents ;
- all transmissions of information are made according to protocols respecting the state of the art for security ;
- the servers of the firm in charge of the data collection are protected by protocols respecting the state of the art in termes of security and the operations of the persons working on the data on these servers are registered ;
- the servers of the ministry of health are secured and protected by state of the art security protocols. Especially, the list of persons having access to EHIS microdata files is restricted to authorized persons.

12. Comment

Top

Related metadata

Top

Annexes

Top

annex 1
annex 2
annex 3
annex 5
annex 4