Eurofound
European Foundation for the Improvement of Living and Working Conditions

The tripartite EU Agency providing knowledge to assist in the development of better social, employment and work-related policies

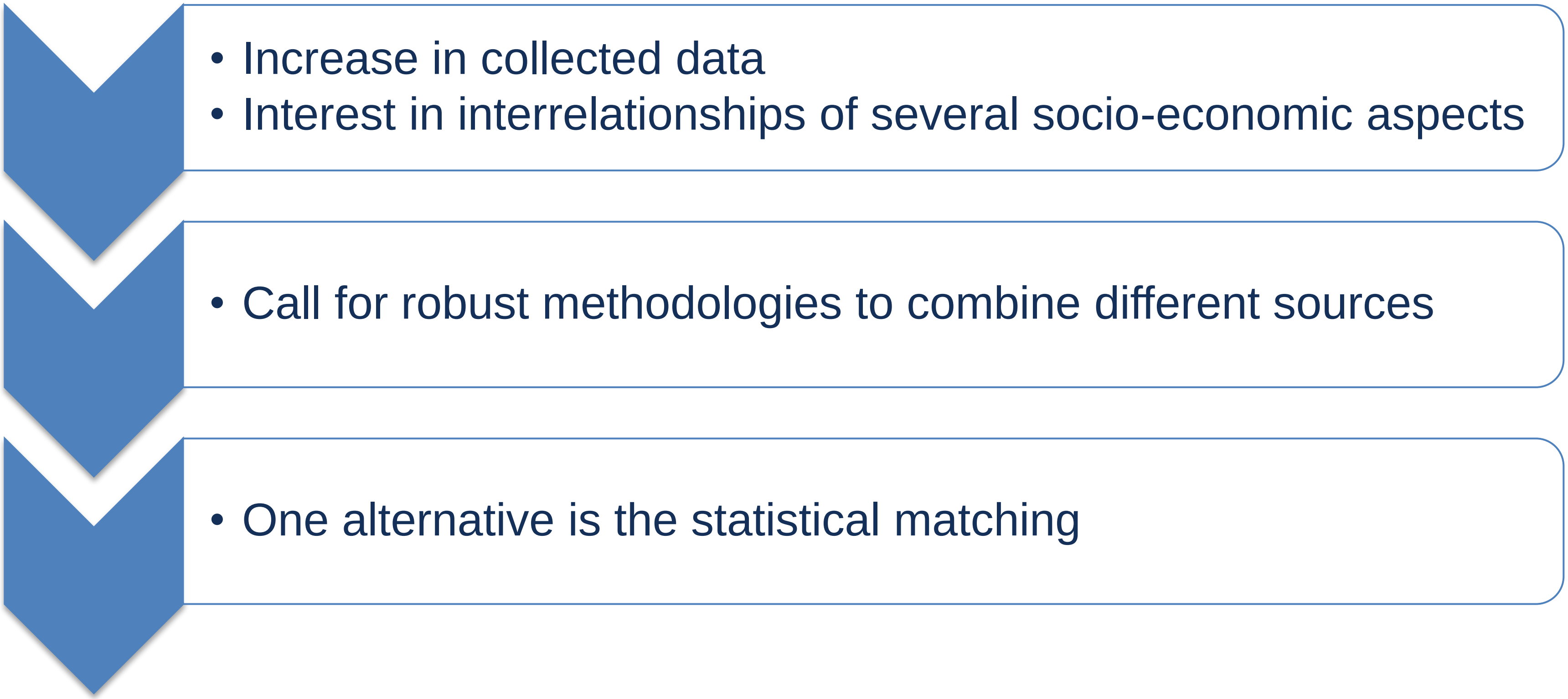
Statistical matching of EQLS and EU-SILC: A case study on public services

Daphne Ahrendt (Eurofound)
Tadas Leoncikas (Eurofound)
Irene Riobóo Lestón (Rey Juan Carlos University)

6th European User Conference for EU-Microdata
8 March 2019, Mannheim

Methodological features

Definition and relevance



- Increase in collected data
- Interest in interrelationships of several socio-economic aspects

- Call for robust methodologies to combine different sources

- One alternative is the statistical matching

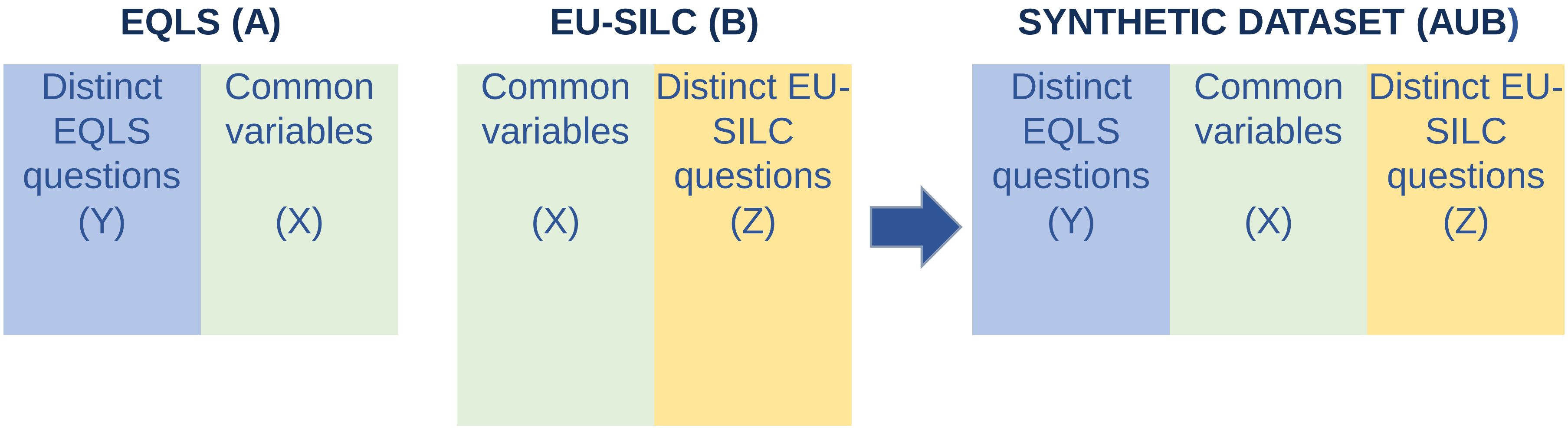
Statistical matching (SM, or data fusion or synthetic matching):

- **A series of statistical methods whose aims are:**
  - the integration of specific variables from two (or more) independent data sources referred to the same target population, using information shared between them as a link.
  - to study relationships among variables not jointly observed in a single data source.

Specific objectives of statistical matching:

The data sources share a subset of common variables (X) and, at the same time, each source observes distinctly other sub-sets of variables (Y and Z).

MICRO: Derive a synthetic dataset with X, Y and Z



EQLS (A)

Distinct EQLS questions (Y)

Common variables (X)

EU-SILC (B)

Common variables (X)

Distinct EU-SILC questions (Z)

SYNTHETIC DATASET (AUB)

Distinct EQLS questions (Y)

Common variables (X)

Distinct EU-SILC questions (Z)

MACRO: Estimation of parameters (correlation coefficient, regression coefficient,…).

Assumptions:

1. The records in both sources are drawn randomly and independently of each other from the same population.
2. The relationship between Y and Z is completely explained by X. This means that Y and Z are independent once conditioning on the X variables.

Conditional Independence Assumption (CIA)

A VERY STRONG ASSUMPTION!

Relevance for the quality of life research

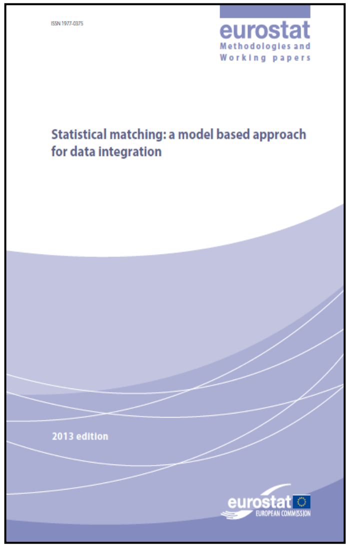
© Can Stock Photo - csp11663234

EQLS
EU-SILC
ESS
EUROBAROMETER

Benefits of the matching:

- possibility to expand the analysis incorporating different angles,
- without increasing response burden or costs.

EQLS and EU-SILC



ISSN 1977-0375

eurostat
Methodologies and
Working papers

Statistical matching: a model based approach for data integration

2013 edition

eurostat
EUROPEAN COMMISSION

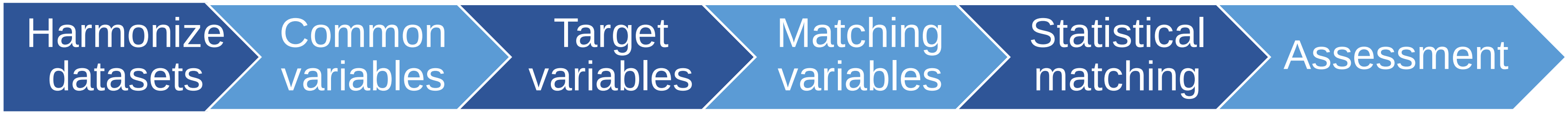
- Eurostat (2013) publication used the 2007 data to assess linking both sources

Target variables	
EQLS	**EU-SILC**
Overall life satisfaction	At-risk-of-poverty rate
Trust in institutions	Severe material deprivation rate
Recognition	Low work intensity rate
Social exclusion	

- The aim now is to analyse if the **new data** show improvements in the matching of both sources:
  - Harmonization of core variables
  - New modules on public services

Methodology

A six steps methodology



1. Harmonize data

Reference population:

- EU-SILC had to be adjusted to persons aged 18 and over.

2. Identify common variables

- Conceptual consistency
- Statistical consistency

Conceptual consistency

- **Direct,** when there is a direct equivalence between questions in the two sources.
- **Adjusted,** when questions need to be transformed in order to make them equivalent.
- **Incompatible,** when questions in both sources cover similar concepts but the correspondence is not possible.
- **Not available,** when a question in the EQLS covers a concept without a similar correspondence in the EU-SILC.

Statistical consistency

- **Similar distributions:**
  - Frequencies, response rates, missing values,...
  - Hellinger distance (EU and country level)
- **Similar associations:**
  - Spearman correlation coefficient
  - Chi-square test

Variable	Conceptual correspondence	Distribution (EU level)	Association	Conclusion
hhsize_c	Adjusted	Same	Similar	Common
Sex	Direct	Same	Similar	Common
Age_c	Adjusted	Same	Similar	Common
EmpStatus	Adjusted	Same	Similar	Common
TypeContract	Adjusted	Same	Similar	Common
Occupation2	Adjusted	Same	Similar	Common
HoursWorked_c	Direct	Same	Similar	Common
Hours2Job_c	Direct	Same	Similar	Common
HoursWorked_partner_c	Direct	Same	Similar	Common
Accommodation	Adjusted	Same	Similar	Common
RotDamp	Adjusted	Same	Similar	Common
MaritalStatus	Direct	Same	Similar	Common
Health	Direct	Same	Similar	Common
ChronicIll	Direct	Same	Similar	Common
Limitations	Adjusted	Same	Similar	Common
Edu_pst	Adjusted	Same	Similar	Common
Holiday	Adjusted	Same	Similar	Common
Meal	Adjusted	Same	Similar	Common
ArreasMort	Adjusted	Same	Similar	Common
ArreasBills	Adjusted	Same	Similar	Common
Childc	Direct	Same	Similar	Common

3. Determine target variables

SCENARIO 1

Target variables on quality of services and unmet needs

EQLS Questions	EU-SILC Questions
Q58 In general, how would you rate the quality of each of the following public services in your country? a. Health services b. Education system d. Childcare services e. Long term care services Q59 How do you rate the quality of the following two healthcare services in your country? a. GP, family doctor or health centre services b. Hospital or medical specialist services	PH040: Unmet need for medical examination or treatment PH050: Main reason for unmet need for medical examination or treatment PH060: Unmet need for dental examination or treatment PH070: Main reason for unmet need for dental examination or treatment HC240: Unmet needs for professional home care HC050: Unmet needs for formal childcare services PC110: Unmet needs for formal education

SCENARIO 2

Target variables on fairness and corruption and unmet needs

EQLS Questions	EU-SILC Questions
Q63 & Q66 To what extent do you agree or disagree with the following about GP, family doctor or health centre services (+hospital or medical specialist services) in your area? a. All people are treated equally in these services in my area b. Corruption is common in these services in my area Q75 To what extent do you agree or disagree with the following statements about long-term care services in your area? a. All people are treated equally in these services in my area b. Corruption is common in these services in my area Q83 To what extent do you agree or disagree with the following statements about childcare services in your area? a. All people are treated equally in these services in my area b. Corruption is common in these services in my area Q86 To what extent do you agree or disagree with the following statements about school services in your area? a. All people are treated equally in these services in my area b. Corruption is common in these services in my area	PH040: Unmet need for medical examination or treatment PH050: Main reason for unmet need for medical examination or treatment PH060: Unmet need for dental examination or treatment PH070: Main reason for unmet need for dental examination or treatment HC240: Unmet needs for professional home care HC050: Unmet needs for formal childcare services PC110: Unmet needs for formal education

SCENARIO 3

Target variables on subjective wellbeing and unmet needs

EQLS Questions	EU-SILC Questions
Q4 How satisfied would you say you are with your life these days? Please tell me on a scale of 1 to 10. Q5 Taking all things together on a scale of 1 to 10, how happy would you say you are? Q6 Could you please tell me on a scale of 1 to 10 how satisfied you are with each of the following items, where 1 means you are very dissatisfied and 10 means you are very satisfied? a. Your education c. Your present standard of living f. Your local area as a place to live Q32 On the whole, how satisfied are you with the present state of the economy in [country]? Please tell me on a scale of 1 to 10, where 1 means very dissatisfied and 10 means very satisfied.	PH040: Unmet need for medical examination or treatment PH050: Main reason for unmet need for medical examination or treatment PH060: Unmet need for dental examination or treatment PH070: Main reason for unmet need for dental examination or treatment HC240: Unmet needs for professional home care HC050: Unmet needs for formal childcare services PC110: Unmet needs for formal education

SCENARIO 4

Target variables on childcare services

EQLS Questions	EU-SILC Questions
Q58 In general, how would you rate the quality of each of the following public services in your country? d. Childcare services Q81 You mentioned that the main form of childcare received by the youngest child is [SERVICE]. How satisfied or dissatisfied you were with each of the following aspects a. Quality of the facilities (building, room, equipment) b. Expertise and professionalism of staff/carers c. Personal attention the child was given, including staff/carers’ attitude and time devoted d. Being informed or consulted about the child's care e. The curriculum and activities Q83 To what extent do you agree or disagree with the following statements about childcare services in your area? a. All people are treated equally in these services in my area b. Corruption is common in these services in my area	HC050: Unmet needs for formal childcare services HC060: Main reason for not making (more) use of formal childcare services RC010: Payment for the cost of formal childcare services RC020: Proportion of the cost of formal childcare services paid

4. Select matching variables

- Subset of common variables (X) that are at the same time connected with Y and Z.

- Subset to be used as a link between both sources for predicting the target variables in the synthetic dataset.

- Trade-off in choosing the number of matching variables:
  - The higher the number of matching variables, the more their power to explain Y and Z and therefore and the more plausible is the conditional independence assumption.
  - The higher the number of matching variables, the lower number of registers are suitable to be used in the matching.

Tools used:

- **Spearman correlation coefficients** for the common and target variables, that allow to identify the common variables that register the higher associations.

- **Random Forest** (regression and classification trees), which allows to study:
  - the predictive power that the set of common variables has for each target variable and
  - the importance that each common variable has in the prediction of the target variables, both individually and globally.

Sets of matching variables
Scenario 1 (quality of services and unmet needs**)**

Variable	Min (3)	Half (8)	Max (10)
Health	x	x	x
Holiday	x	x	x
HoursWorked_partner_c	x	x	x
Accommodation		x	x
ArreasBills		x	x
hhsize_c		x	x
HoursWorked_c		x	x
Occupation2		x	x
Childc			x
Edu_pst			x

5. Implement statistical matching

- Selecting one of the available dataset as **recipient** (the other is the **donor**). Usually the recipient is the smaller one (EQLS).
- Selecting the **donation classes** (homogeneous strata) according to the values of one or more categorical variables chosen among the available common variables ones: country and sex.
- The matching has been carried out through the **nearest neighbor distance hot deck**:
  - For each record in EQLS, it is selected the closest donor record in EU-SILC according to a distance computed on the matching variables.
  - Then the value of Z observed on the EU-SILC´s unit it is imputed in the EQLS.

Several matching models have been considered according to the three sets of matching variables.

Additionally, for each matching set two versions have been implemented based on the constrains on the use of donors:

- **Version constrained to one donor**, where a donor can be used just once.
- **Version unconstrained**, where a record in the donor file can be selected unlimitedly as a donor.

Hence, six has been the matching models implemented.

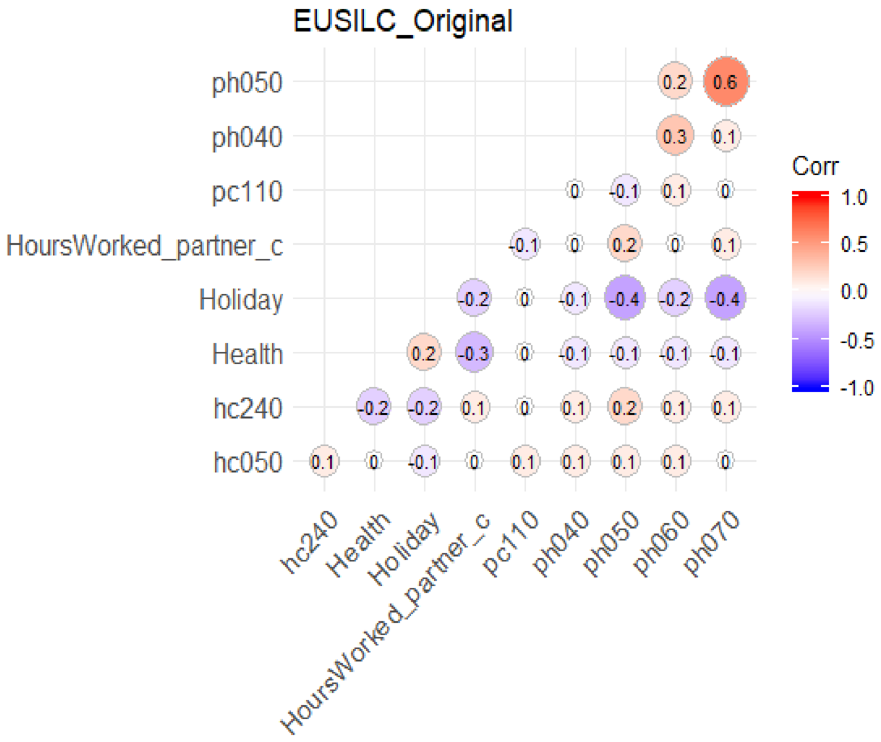
6. Evaluate results

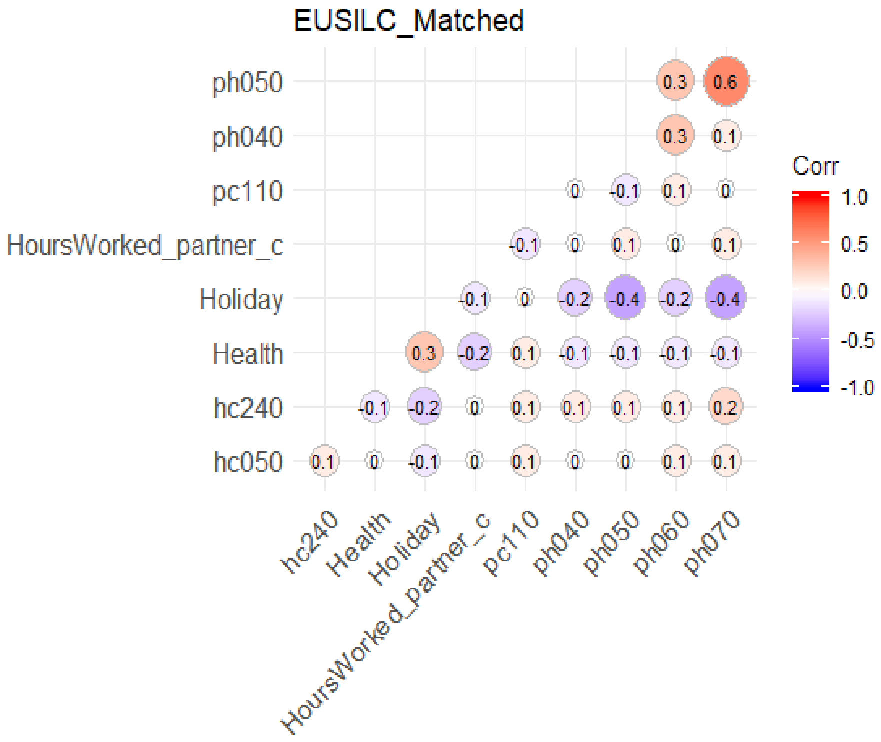
Quality assessment:

- Checking if the marginal distributions of the target variables observed in the original dataset is preserved in the synthetic dataset.
- Comparing the degree of similarity (HD) between matched records for each model.
- Checking the matching distances between each couple recipient-donor.
- Checking if the relation (strength and direction) between matching variables observed in the donor file is preserved in the synthetic dataset.

Spearman correlation coefficients for the original EU-SILC dataset

Spearman correlation coefficients for the matched EU-SILC dataset





Winner model in scenario 1: match.min1.NND.unc

Illustration of the substantive results (selected)

	Q58 a. Quality of health services		Q59 a. Quality of GP, family doctor or health centre services		Q59 b. Quality of hospital or medical specialist services	
PH050. Main reason for unmet need for medical examination or treatment	**Mean**	**Unweighted Count**	**Mean**	**Unweighted Count**	**Mean**	**Unweighted Count**
Could not take time because of work, care for children or for others	**6.9**	109	**7.3**	109	**6.6**	109
Other reasons	**6.5**	83	**6.8**	83	**6.8**	83
Wanted to wait and see if the problem got better on its own	**6.3**	189	**7.6**	189	**6.7**	189
Too far to travel/no means of transportation	**6.1**	52	**7.7**	52	**6.9**	52
Didn’t know any doctor or specialist	**5.8**	27	**6.0**	27	**6.3**	27
Waiting list	**5.8**	269	**6.5**	269	**6.1**	269
Fear of doctor/hospitals/ examination/treatment	**5.5**	41	**6.4**	41	**5.8**	41
Could not afford (too expensive)	**5.4**	618	**7.0**	618	**5.9**	618

Reference group: "Yes"	Life Satisfaction
PH040. Unmet need for medical examination	1.22** (0.51)
PH060. Unmet need for dental examination	-0.44 (0.52)
HC050. Unmet needs for formal childcare services	0.59* (0.32)
HC240. Unmet needs for professional home care	1.21** (0.55)
PC110. Unmet needs for formal education	-0.52 (0.68)
Constant	7.79*** (0.80)
Country fixed effects	Yes
Observations	128
R-squared	0.33

Notes: Clustered standard errors at the country level in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Concluding messages

- Statistical matching between EU-SILC and EQLS can work, even though there can be problems with particular survey waves.
- The new datasets open the door to investigate new research lines.

  This exercise helps to explore how specific reasons for unmet healthcare need relate to perceptions of its quality (lack of availability is critical in case of primary care; lack of affordability in case of hospital services and overall system ratings).
- The basic limitations:
  - Sample size of a smaller data set (and is particularly difficult for small subsamples of e.g. long-term care users in EQLS).
  - CI assumption could not be tested.
- Comments for finalisation of the working paper most welcome!
- Eurofound continues efforts on data matching (next exercise 2019).

EQLS 2016 Overview report:

http://bit.ly/EQLS-overview

EQLS source questionnaire:

http://bit.ly/EQLS-Q

More about the EQLS:

- http://bit.ly/EQLS-info
- EQLS 2016: Quality Assessment
- EQLS 2016: Technical and fieldwork report

Thank you

Contact: *tle@eurofound.europa.eu*