

Data integration in research system of refugees

Marek Cierpial-Wolan, Assoc. Prof., Statistics Poland, University of Rzeszow
Dominik Rozkrut, PhD, Statistics Poland, University of Szczecin

Agenda

1.

Background

2.

Integration of administrative registers

3.

Survey of refugees – WHO and Statistics Poland

4.

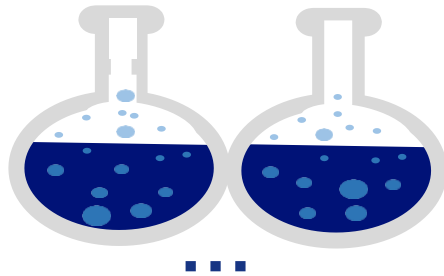
Use of big data sources

5.

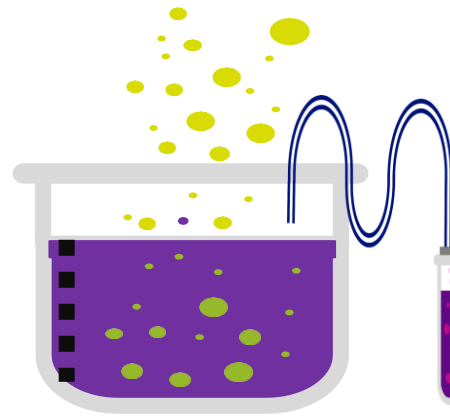
Conclusions

Specificity of movements of refugees

Data integration - methodological challenge to preserve high level of quality



Administrative Data



Census survey

Sample surveys

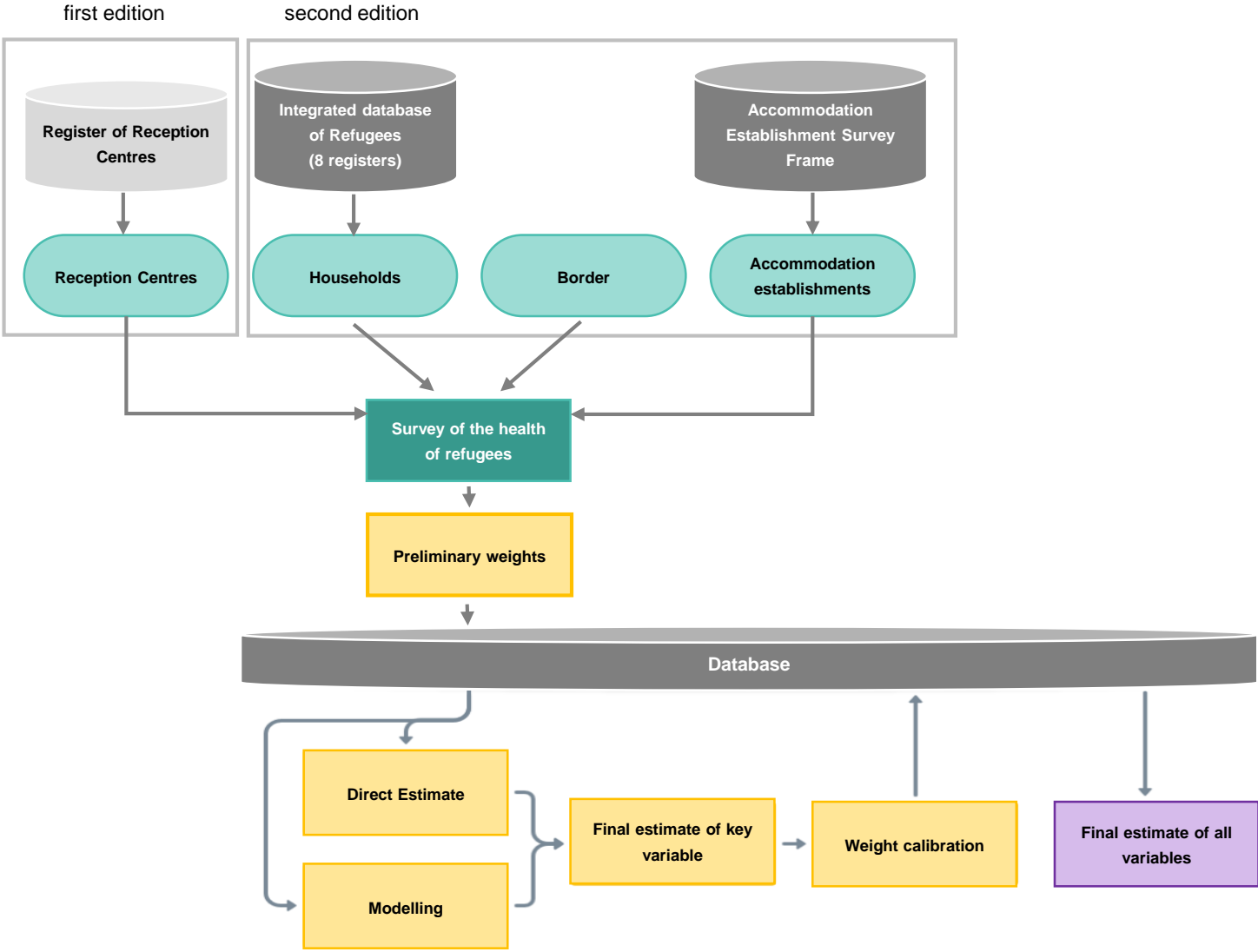


Big Data

How to integrate data from different sources?

Data integration model

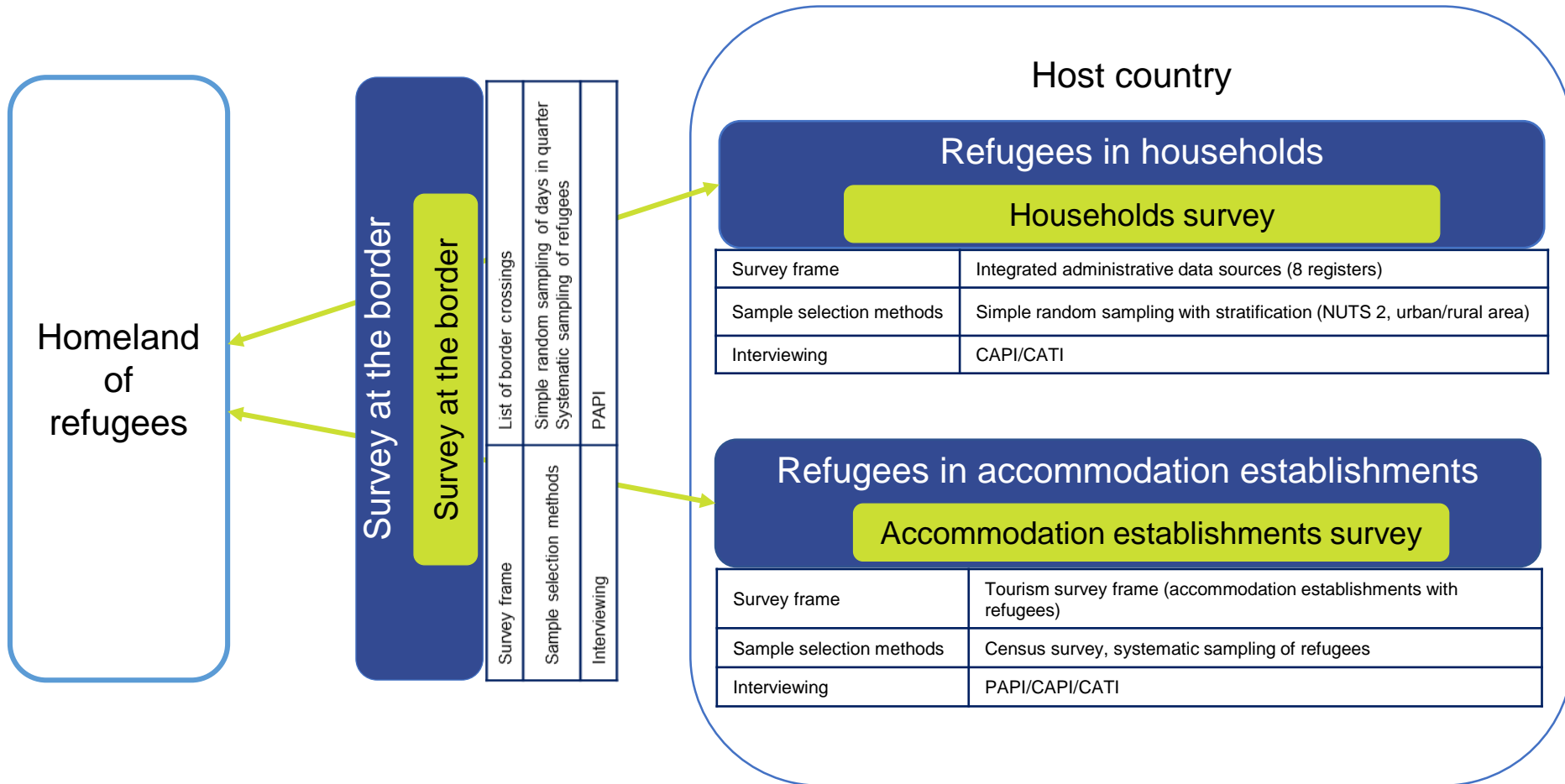
Health of refugees – comprehensive approach



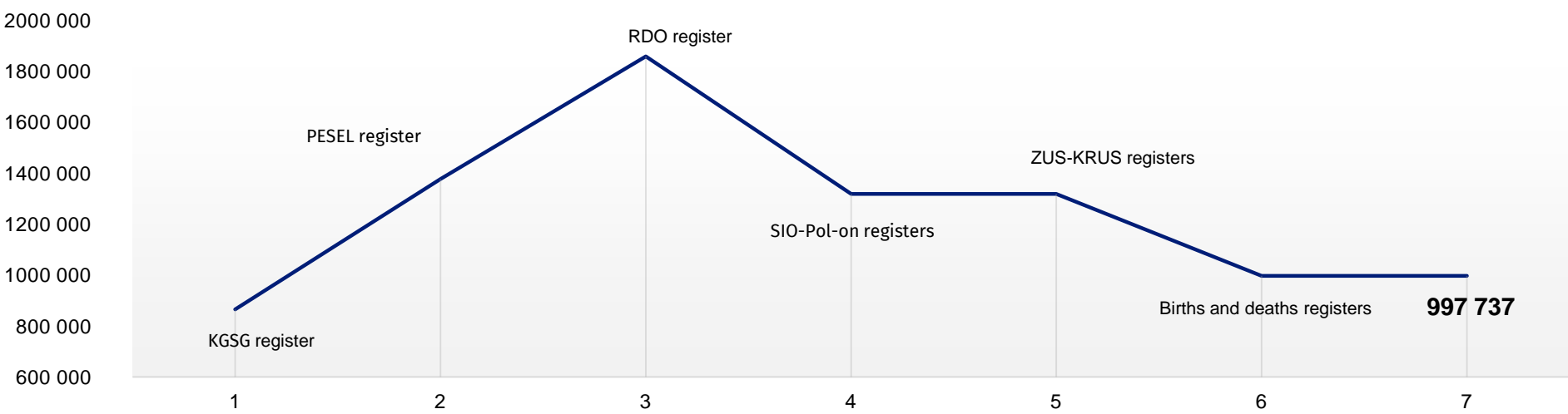
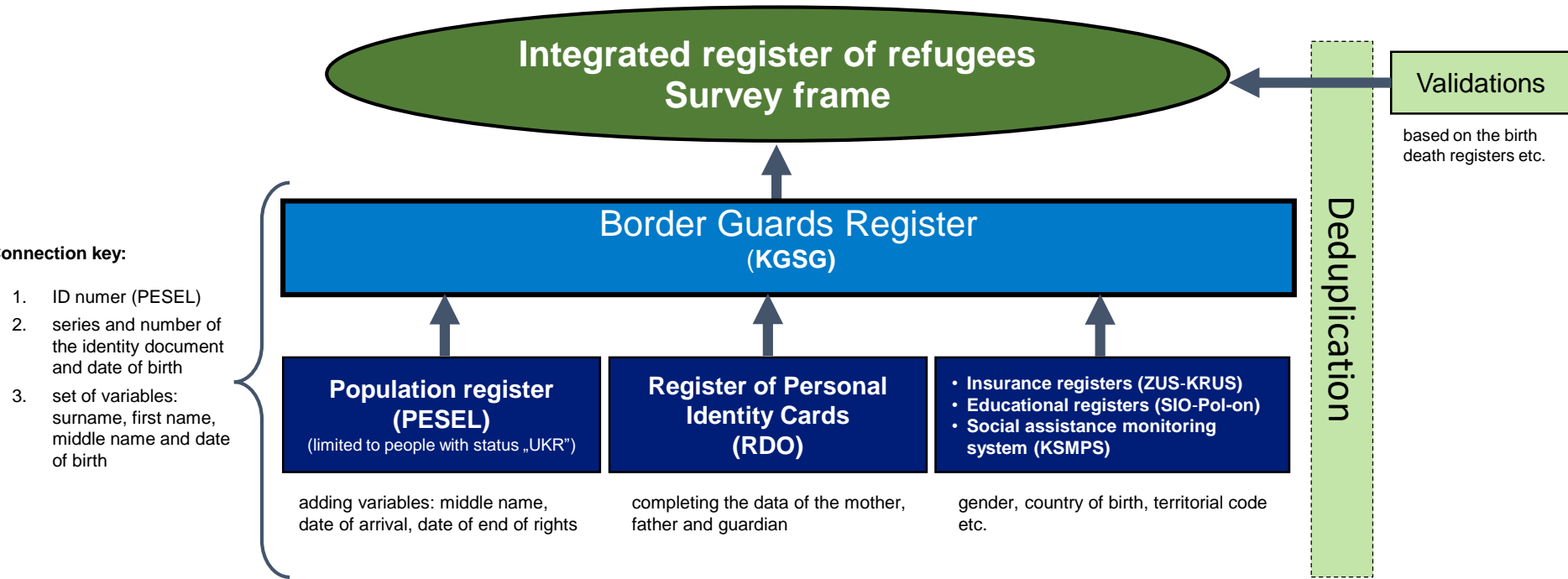
Sample surveys

WHO and Statistics Poland

Methodology – selected aspects (second edition)

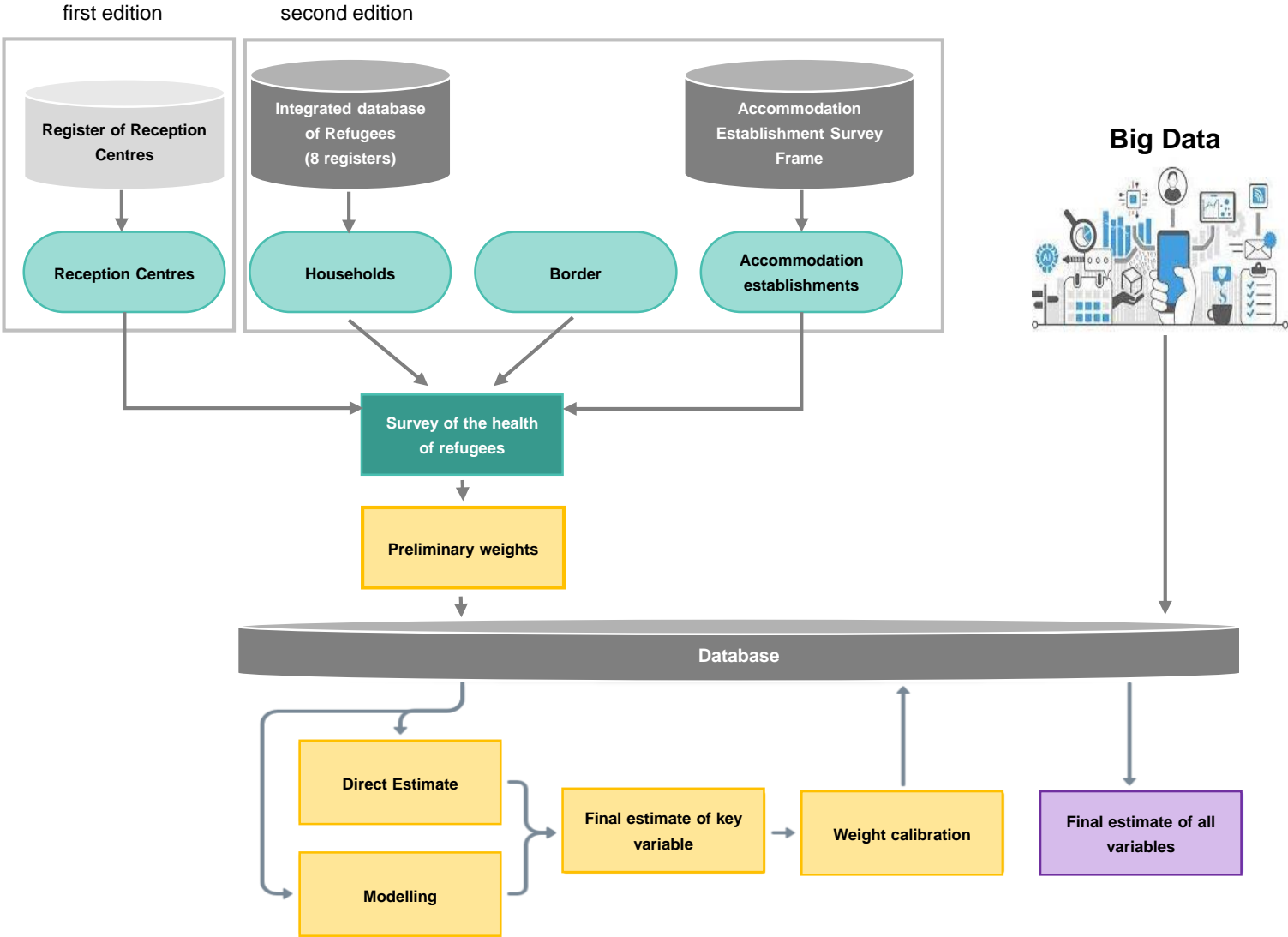


Developing integrated register of refugees from Ukraine



Data integration model

Health of refugees – comprehensive approach



Big data sources

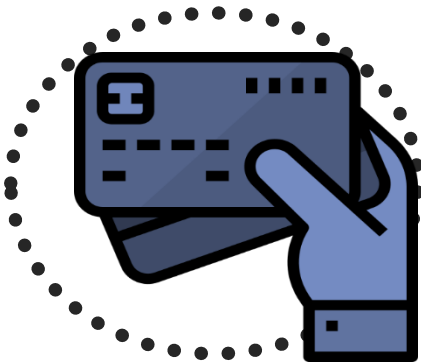
Mobile network operators

T-Mobile provides daily data



Payment/credit card operators


Samples of data



Big data – Mobile Network Operator (MNO)

Statistics Poland receives **daily** data on active SIM cards held by Ukrainian refugees

User ID	Date	Nationality	Starting Time	Ending Time	LON	LAT
123***	2023-01-01	Ukraine	00:14:10	09:57:00	52***	19***
123***	2023-01-01	Ukraine	09:47:20	10:45:20	52***	19***
123***	2023-01-01	Ukraine	11:35:50	14:29:10	52***	17***
...		Ukraine				
123***	2023-01-08	Ukraine	20:11:00	22:59:37	52***	19***



Date	District ID	Number of active SIM cards
2023-01-01	0201011	274
2023-01-01	0201022	34
2023-01-01	0201032	17
...		
2023-01-08	3263011	255

Mobility model

- MNO: SIM card must be active for at least **3 hours** in a given area - multiple counting



$$y_1 = x_1$$

y_i – active SIM cards with duplicates, x_i - unique active SIM cards, $p_{ij|k}^{(s)}$ - share of SIM card holders who moved in s -th step from i -th area to j -th area after visiting k -th area.

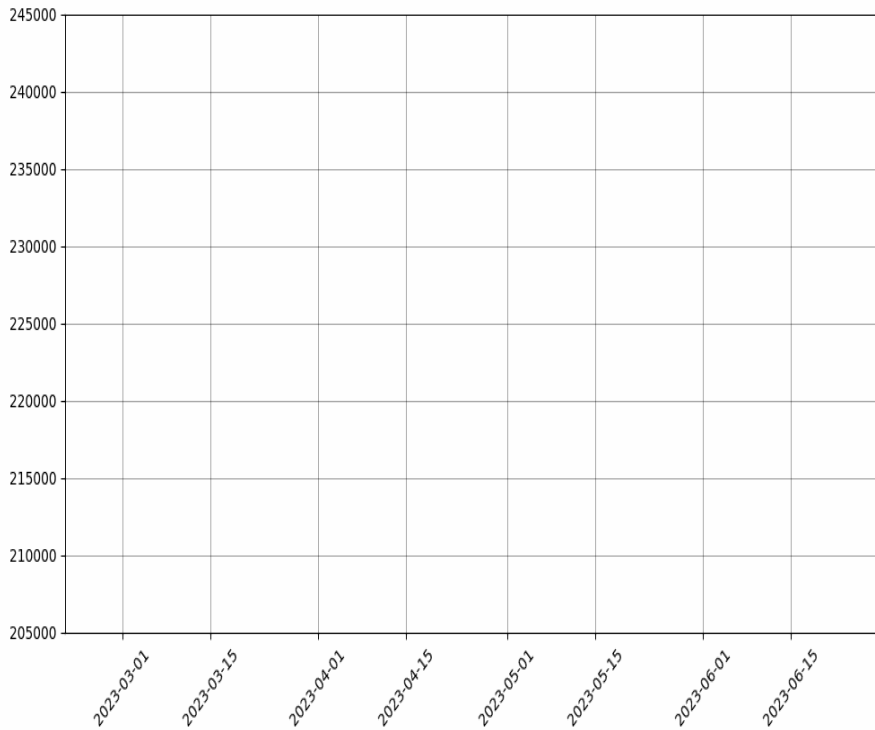
Two-stage procedure of estimation:

- Mobility model of SIM card users for deduplication and mobility assessment: based on the idea of the transition matrix of Markov process with parameters estimated with fixed point method;
- Estimator of total number of refugees: based on MNO's market share, digital literacy by age cohorts, average SIM cards per card user, age-sex structure of refugees from administrative data.

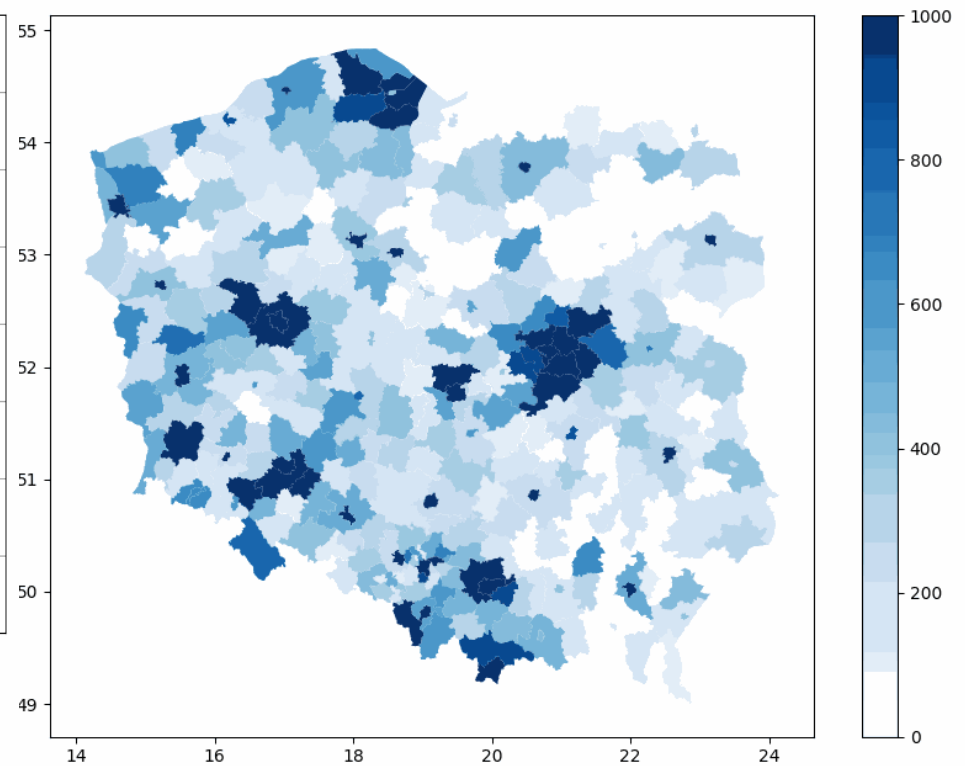
MNO data may „reveal” refugees not covered by administrative data sources.

The use of mobility model

Daily total number of active SIM cards in Poland

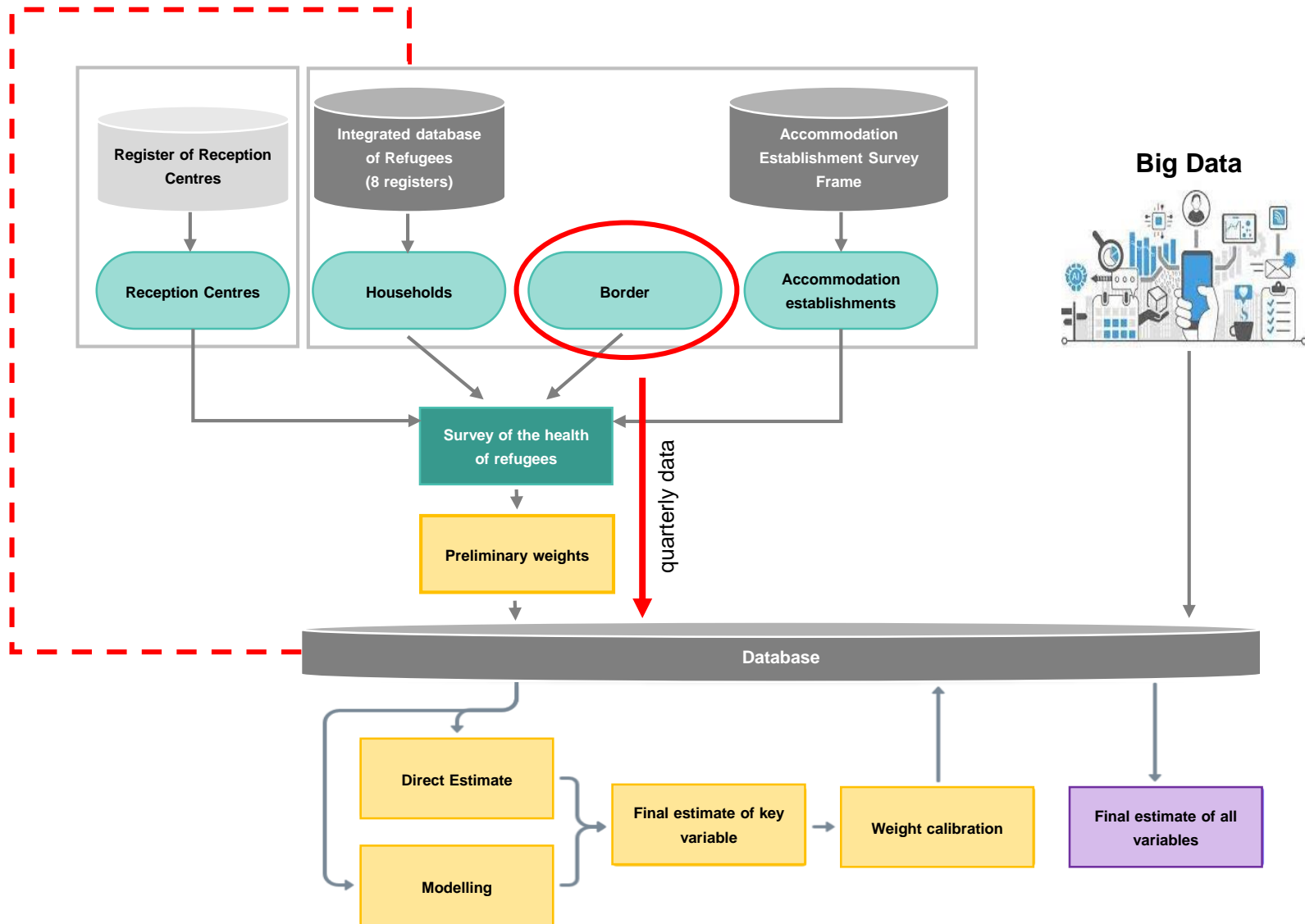


Movements of refugees
(February 2023 – June 2023)



Data integration model

Real-time picture on health of refugees

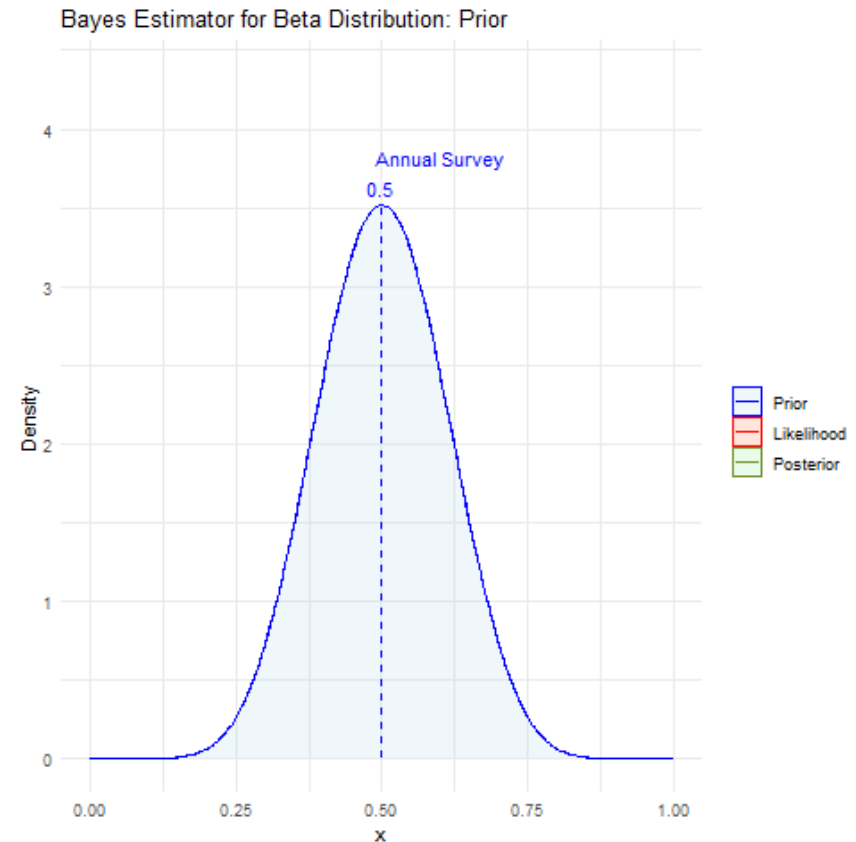


Updating survey results

Application of Bayes estimator

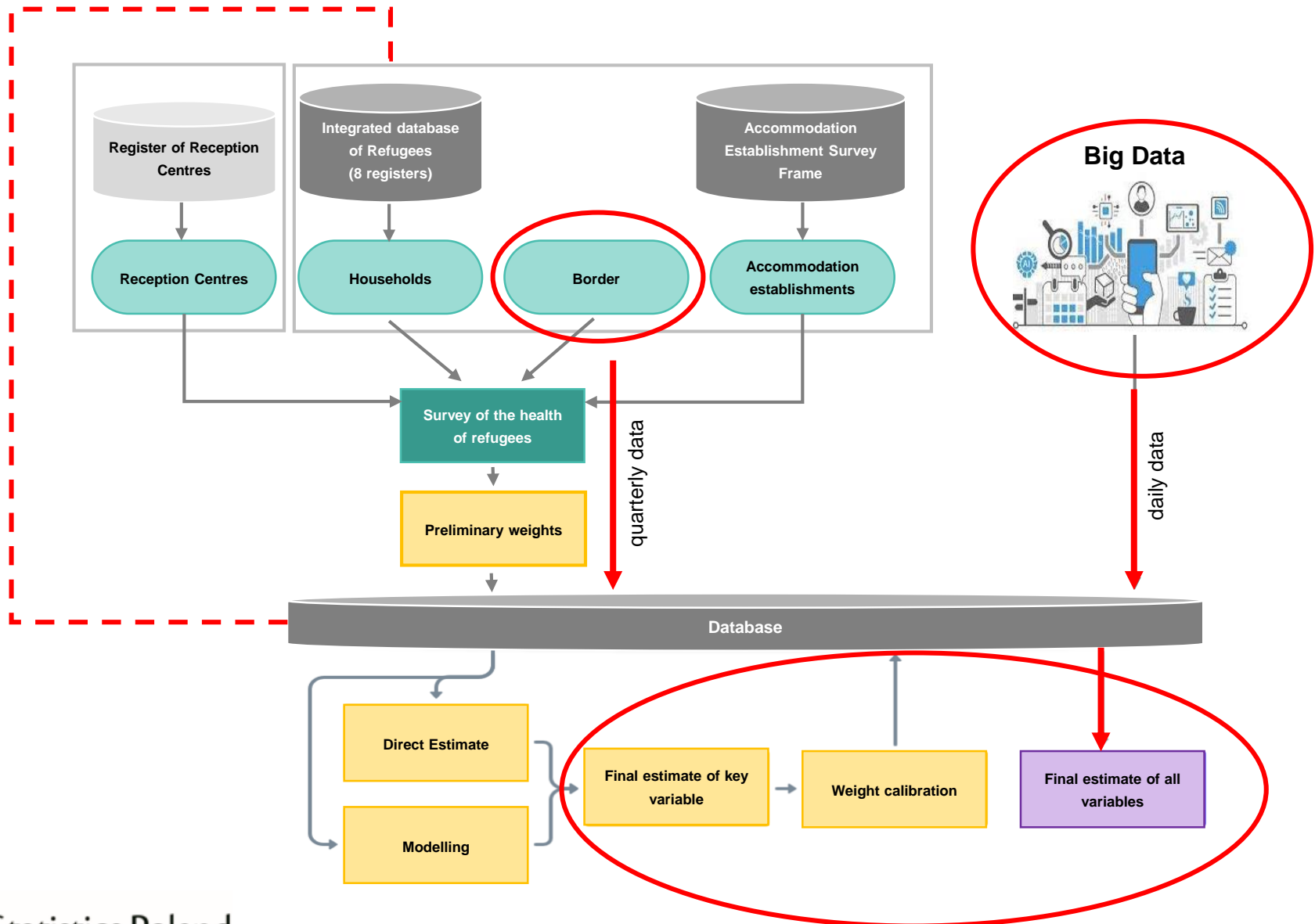
How to combine data from high-frequency survey with low-frequency survey

- The procedure of incorporating high-frequency sample surveys into the estimates goes as follows:
 - Derive statistics from high-frequency sample survey,
 - Obtain Bayesian estimates of key variables;
- In Bayesian approach researcher assumes the prior distribution of the parameter of interest θ describing the uncertainty about that parameter;
- The Bayes estimate $\hat{\theta}(x)$ of θ based on available data x is the mean of posterior distribution, which is a product of likelihood and prior.



Data integration model

Real-time picture on health of refugees



Monitoring health of refugees from Ukraine

Dashboard

Date of the last survey



1Q2024

Number of refugees



956995

Healthcare need

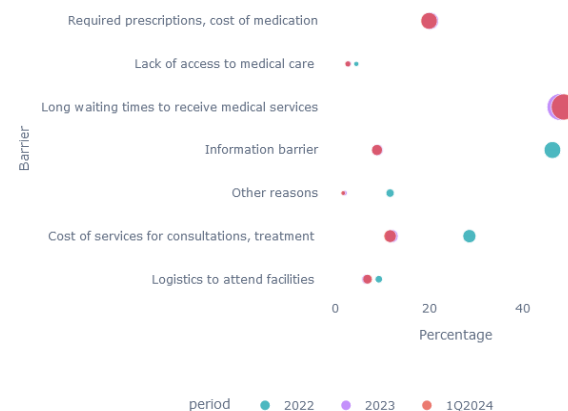
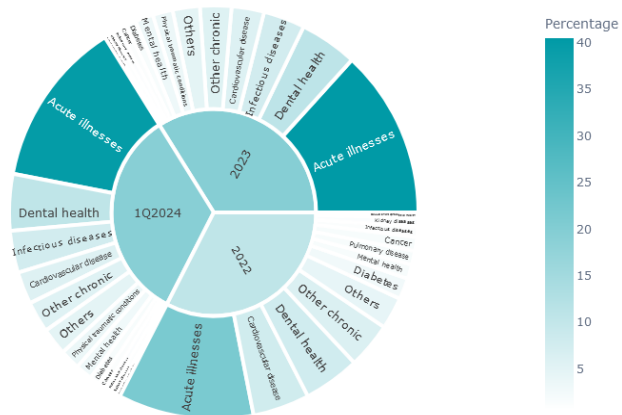
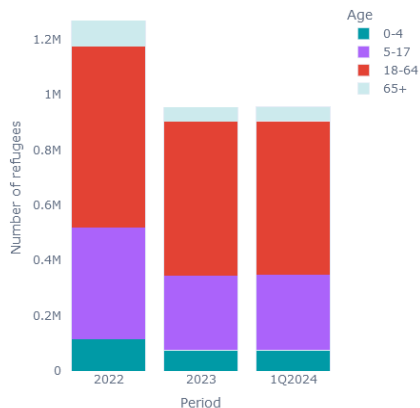


48.6%

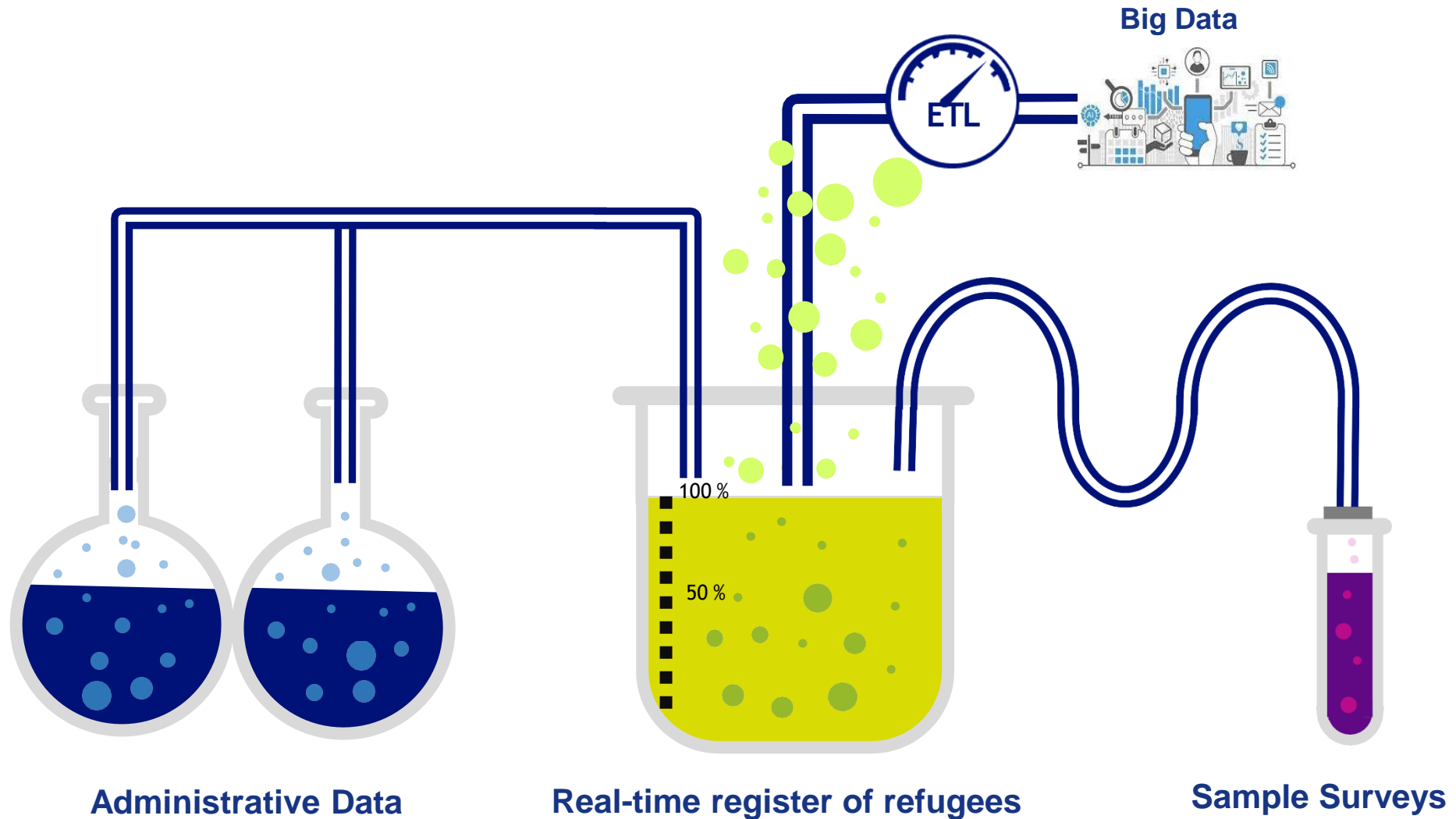
Mental health



10.0%



Developing integrated register of refugees from Ukraine



Conclusions

- Integration of data from various sources allows:
 - Creation of a refugee survey frame and update it systematically,
 - Analyses of the number of refugees in time and space, practically in real time,
 - Conducting new sample surveys, such as the health survey of Ukrainian refugees jointly with WHO, which allows updating the refugee frame,
 - Improving the quality of estimation in existing surveys.
- Implementation of more advanced methods of data integration;
- Using other sources, such as: Google Trends, smart city systems, drones; satellite images;
- Coherent research system for refugees - the need for research and analysis based on data from multiple sources, multi-method approach.

Thank you for your attention

Big data – payment cards (VISA)

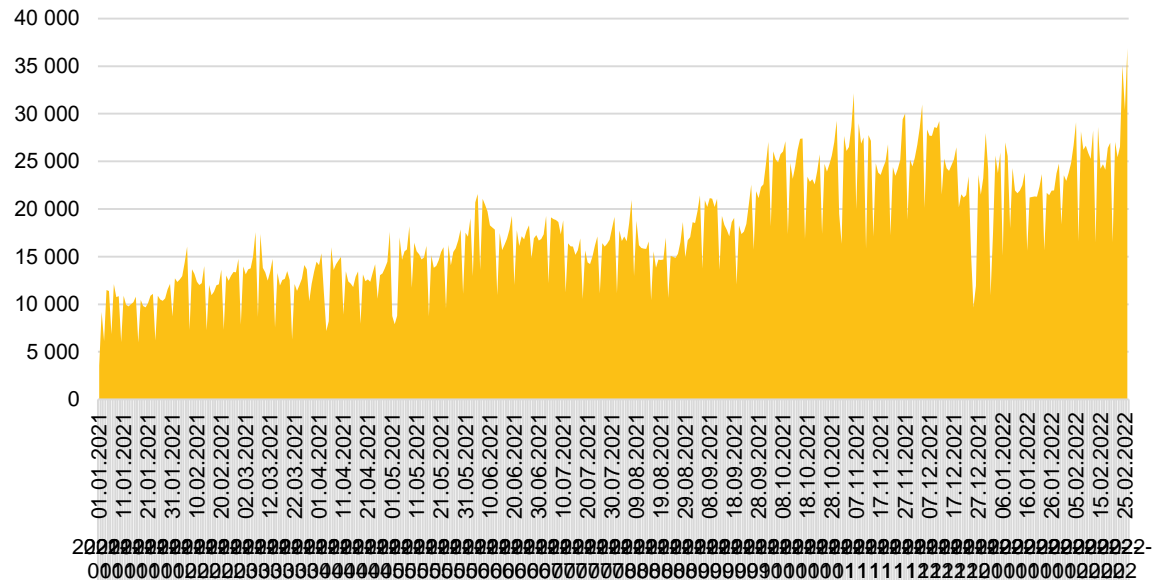
VISA monitors activity of payment cards issued in Ukraine but used in Poland

Sample of a daily data:

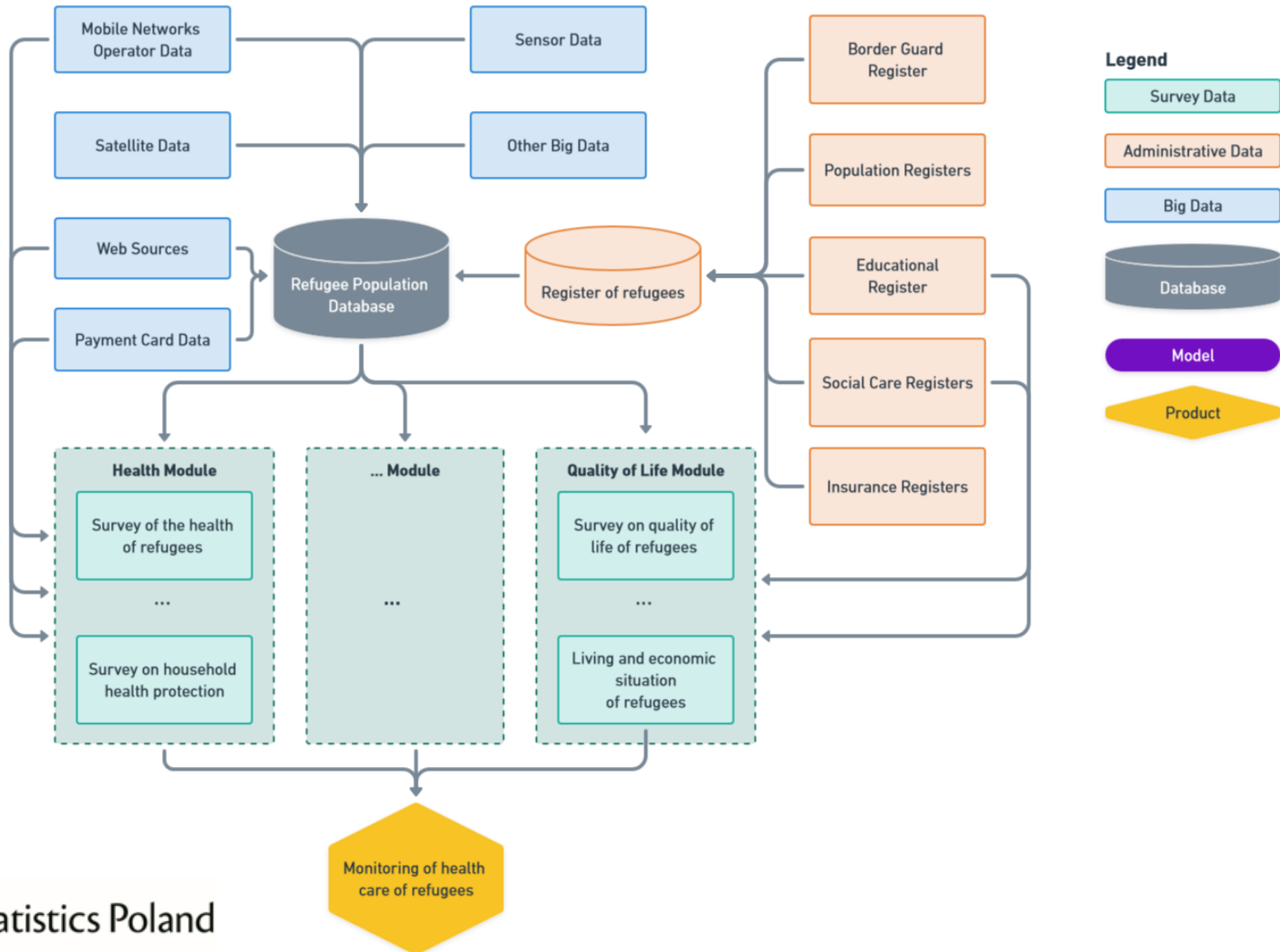
- spatial unit ID
- expense amount
- number of transactions
- MCC (Merchant Category Code):

8011 Doctors and Physicians
8021 Dentists and Orthodontists
8031 Osteopaths
8041 Chiropractors
8049 Podiatrists and Chiropodists
8050 Nursing, Home Healthcare
and Personal Care Facilities
8062 Hospitals
8099 Medical Services and Health Practitioners

Daily number of payment card transactions made by Ukrainians in Poland



Model of refugees information system with health module



Sample surveys

Survey of refugees – WHO and Statistics Poland

Methodology – selected aspects (the first edition)

Sample selection methods:

- Two-stage random sampling:
 - (1) simple random sampling (locations) with stratification (NUTS 2)
 - (2) systematic sampling (refugees)

Additional data sources:

- population register
- Border Guards register

Interviewing:

- CAPI/PAPI/CATI

Updating survey results with MNO data

- In the process of integrating the sample survey and the big data (MNO data), it must be kept in view that we deal with the problem of combining unbiased and possibly biased estimators.
- In such a case researchers propose among others shrinkage estimators (e.g. a James-Stein type estimator) which offer several advantages over traditional estimators, especially in scenarios involving high-dimensional data such as: improved estimation accuracy, bias reduction, robustness, etc.
- Using update MNO data reduced bias of estimates on NUTS 2 level by 0.44 percentage points on average (ranging from 0.07 to 1.47 percentage points).