HBM4EU

science and policy
for a healthy future

The theory behind record linkage

Obtaining health information for HBM studies from different sources

Workshop

4th and 5th October 2021

Hanna Tolonen (THL)
Helle Margrete Meltzer (NIPH)

Starting point, HBM studies: cross-sectional data

Analyses of nutrients, environmental chemicals and toxic substances.

- Biological material:
  - Urine
  - Blood
  - Hair
  - Breastmilk
- Complementary data:
  - Questionnaires or interviews
  - Anthropometric measures

Figure 2. Example of Food Frequency Questionnaire

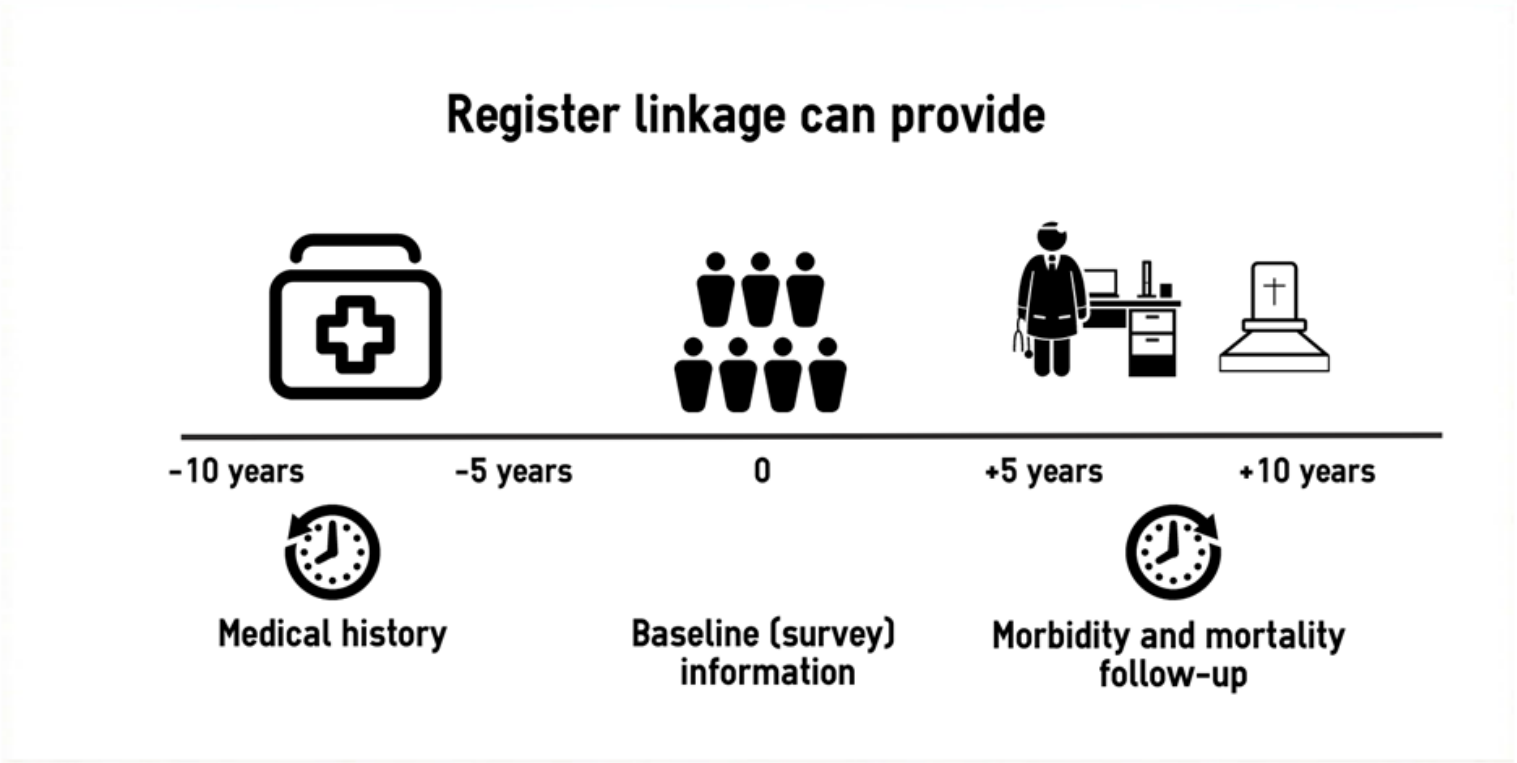
	Never	Once per week	2-4 per week	5-6 per week	Daily	Once per month	Once per 3 months	Once per year
Milk, yogurt, regular fat (1 cup)	○	○	○	○	○	○	○	○
Milk, yogurt, lowfat (1 cup)	○	○	○	○	○	○	○	○
Spinach, kale, other green leafy vegetables (1/2 cup)	○	○	○	○	○	○	○	○
Carrots (1 medium)	○	○	○	○	○	○	○	○
Beef (3 oz)	○	○	○	○	○	○	○	○
Rice, white (1 cup)	○	○	○	○	○	○	○	○
Rice, brown (1 cup)	○	○	○	○	○	○	○	○
Cookies (2 -2" diameter)	○	○	○	○	○	○	○	○
Ice cream, regular fat (1/2 cup)	○	○	○	○	○	○	○	○

Learn about past situation, e.g. diseases or medication use

Follow-up prospectively

Obtain more information of present situation

Theoretical example



What is a record linkage?

(From Eurostat, definition)

- *Record linkage* is the task of finding records in a data set which refer to the same entity across different *Data sources*.
- *Record linkage* is necessary when joining data sets based on entities that may or may not share a common *Identifier*, which may be due to differences in record shape, storage location, or curator style or preference.
- A data set that has undergone *Record linkage*-oriented reconciliation may be referred to as being 'cross-linked'.
- *Record linkage* is called *Data linkage* in many jurisdictions, but is the same process.
- *Record linkage* of administrative and survey data is increasingly used to generate evidence to inform policy, services and research

Why to conduct record linkage?

- Linkage offers a relatively quick and low cost means of capturing information from large administrative data-sets for service planning, delivery and evaluation, surveys and censuses, and research.
- It enriches, updates or improves the information stored in different sources.
- Allows to study the relationship among variables reported in different sources.
- Reduces participant burden during the data collection.

- Data minimization principle of the GDPR, Article 5.

The Norwegian Mother, Father and Child Cohort Study






- A nationwide population based pregnancy cohort
  - Biological samples
  - Questionnaires at regular intervals
  - Linkage to national health registries - long time follow-up
- Inclusion period: 1999-2008. In total 114 500 children, 95 200 mothers (participating with 106 900 pregnancies), and 75 200 fathers. Participation rate: 40.6%

Data collection – Questionnaires, Biological materials, linkage to Health Registries

The birth record from the Medical Birth Registry of Norway, which includes maternal health during pregnancy as well as procedures around birth and pregnancy outcomes, is integrated in the MoBa database.

National health registries that have been linked to the MoBa study

- Medical Birth Registry
- National Patient Registry
- Cause of Death Registry
- Prescription Database
- Vaccination Registry
- Cancer Registry

Added value, MoBa, of linkages

- Saves space in questionnaires, reduces participant burden
- Allows to generate new data by combining existing data sources
- Can in principle follow the participants until they die

- By using the unique identification number given to all residents in Norway, all participants can be linked to a number of health registries to allow a more complete follow-up for many diseases.

Data processing

- Record linkage is highly sensitive to the quality of the data being linked,
- All data sets under consideration (particularly their key identifier fields) should ideally undergo a data quality assessment prior to record linkage.
- Many key identifiers for the same entity can be presented quite differently between (and even within) data sets, which can greatly complicate record linkage unless understood ahead of time.

Data set	Name	Date of birth	City of residence
Data set 1	William J. Smith	1/2/73	Berkeley, California
Data set 2	Smith, W. J.	1973.1.2	Berkeley, CA
Data set 3	Bill Smith	Jan 2, 1973	Berkeley, Calif.

Record linkage methods

Deterministic linkage

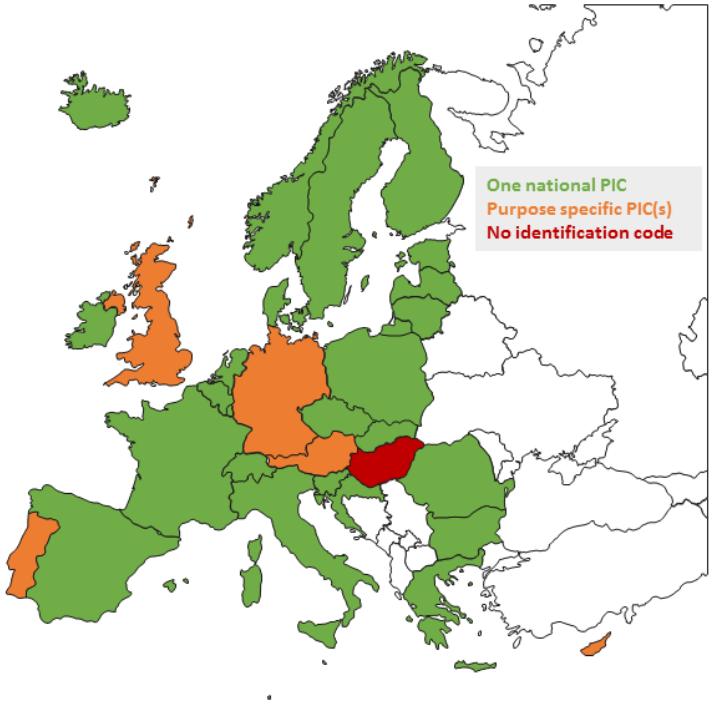
- Based on exact matching of data sets using unique identifiers or a combination of data fields that uniquely identify individuals
- Assumes that there is no missing information on identifiers used for linkage

Probabilistic linkage

- Uses number of identifiers, in combination, to identify and evaluate links
- Requires several steps to be completed
- Based on record linkage algorithms such as Fellegi-Sunder Method, Machine Learning and Bayesian Record Linkage techniques

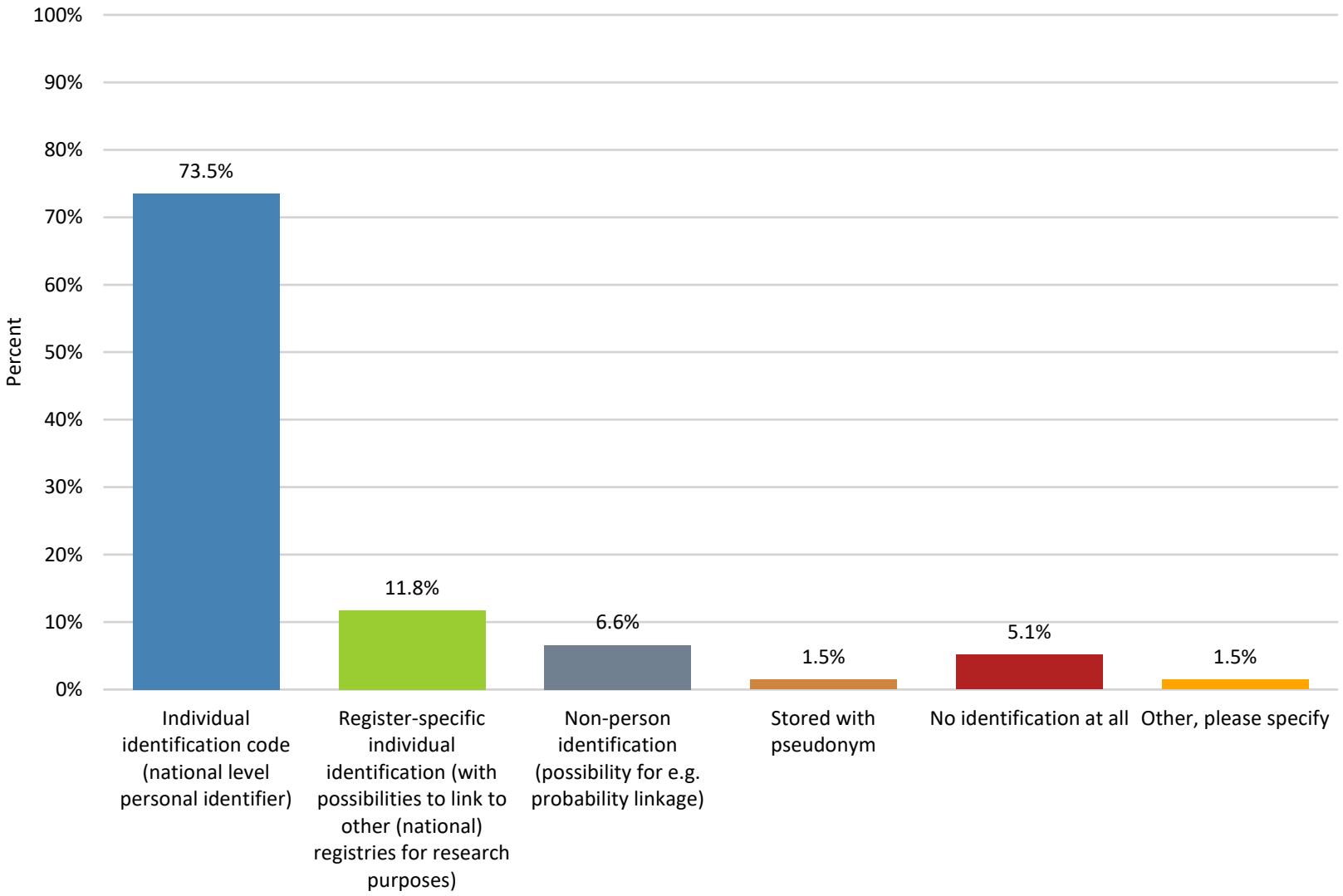
National Identification Number

- Most of the EU MSs + EEA have a unique national identification number
- Not in all countries this PIC is used systematically in different administrative data sources
- National legislation may prevent use of PIC for record linkage



https://en.wikipedia.org/wiki/National_identification_number#Europe

Type of unique identifier



Requirements for record linkage

- In case of survey data, based on informed consent, consent has to cover also record linkage
- Required permissions vary be country and sometimes also between register owners within country
- Implications of GDPR
  - Detailed definitions of uses of data need to be provided
  - Details of data sources to be linked to survey data needs to be provided
  - Lawfulness of processing has to be defined

Benefits of combining HBM and health studies

(from workshop in Brussels June 2018)

- Use on existing survey infrastructure on
  - recruitment of participants,
  - collection of biological samples and conducting health examinations, and
  - collection of data through questionnaires on wide range of topics from health, health behaviours, socio-economic position and exposure pathways
- Synergies in public relations activities during the fieldwork
- Reputation and awareness of the studies in the public
- Chance to combine data on exposure, health behaviours, health and socioeconomic position to large, comprehensive data sets
- Reduced costs of public resources

Challenges for combining HBM and health study (from workshop in Brussels June 2018)

- Partners involved in HBM and health studies may have different priorities
- Finding a balanced compromise between them is often needed
- Preparation phase may take longer time than for individual HBM and health study due to the need for more negotiations, meetings and agreements (rights, duties, sharing of costs, data protection and sharing, etc.)
- Target population, time lines, extent of the full survey etc. may be defined by HES study. Accommodating a HBM part to this may be challenging
- Volume of collected blood samples is limited. How is the use of the collected samples prioritised in relation to HBM and health parts' requirements?
- The large data sets with information on exposures and health that are generated by the combined studies are often underutilized

Linkage error

Perspectives

GUILD: GUidance for Information about Linking Data sets†

Ruth Gilbert[1], Rosemary Lafferty[1], Gareth Hagger-Johnson[1], Katie Harron[2], Li-Chun Zhang[3], Peter Smith[3], Chris Dibben[4], Harvey Goldstein[1]

Journal of Public Health 2016

Analysts have a range of methods for dealing with data quality issues, including linkage error, provided they are made aware of the problem.

Thanks to all Task 11.5 members!

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733032.

Partners

Austria: Hanns Moshammer & Eva Schernhammer, Medizinische Universität Wien (MUW)

Czech Republic: Ondrej Majek, Institute of Health Information and Statistics (IHIS)

Denmark: Tina Kold Jensen & Louise Dalsager, University of Southern Denmark (SDU)

Finland: Hanna Tolonen & Elsi Haverinen & Hanna Elonheimo, Finnish Institute for Health and Welfare (THL)

Norway: Helle Margrete Meltzer, Norwegian Institute of PublicHealth (NIPH)

Sweden: Marika Berglund, Agneta Åkesson, Karolinska Insitutet (KI), Maria Wennberg UMU