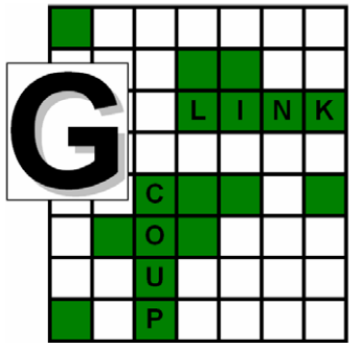
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Statistical Office FSO

Swiss Statistics

Swiss data linkage project using G-Link



Sylvie Berrut

2016.04.04 Euro-Peristat Meeting in Paris

Contents

Data linkage in Switzerland

Swiss databases used for Euro-Peristat 2010 report

Our data linkage project

Matching variables

Two-steps matching of babies' data:

- Deterministic matching
- Probabilistic matching with G-Link

Conclusion

Data linkage in Switzerland

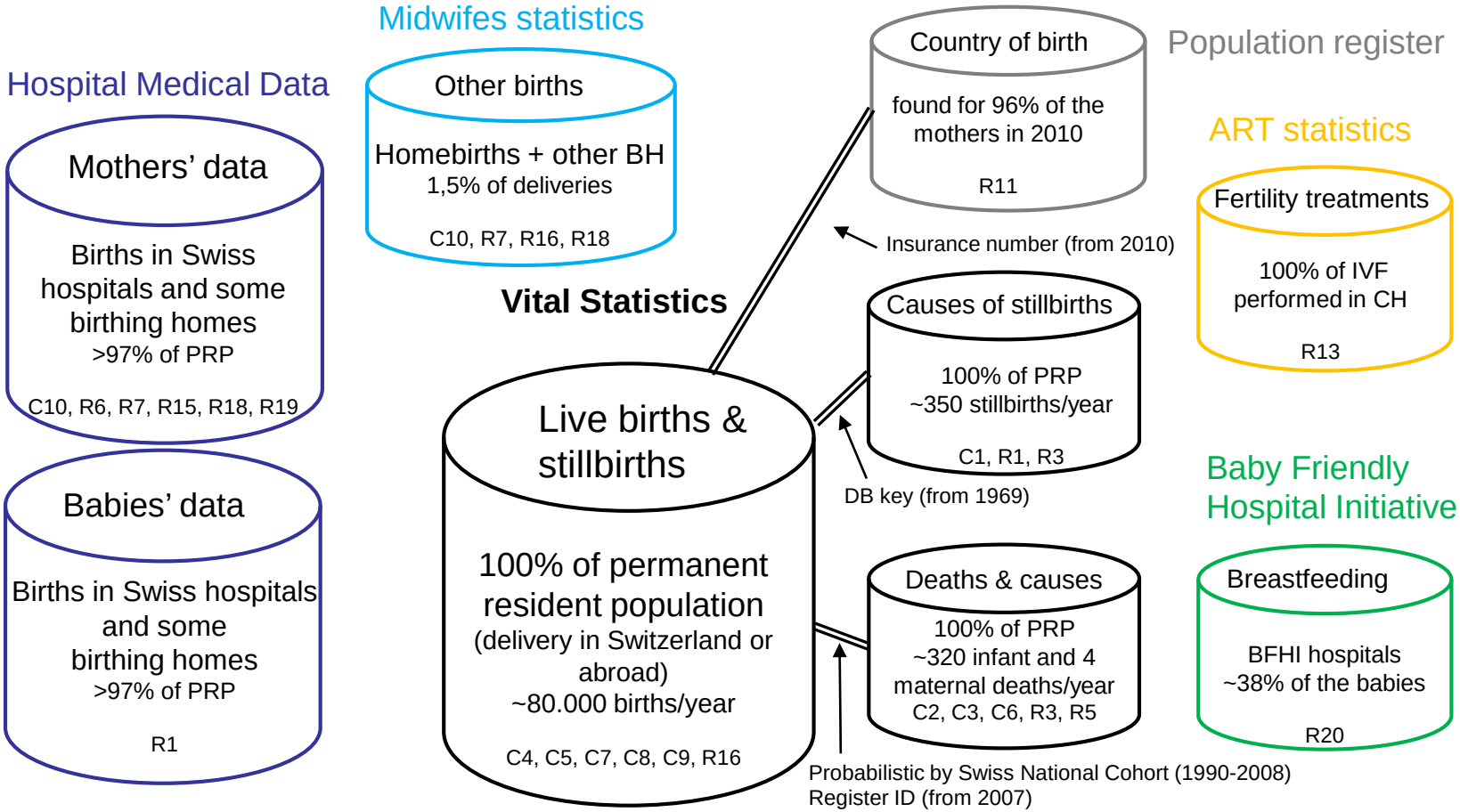
Data linkage is a very actual topic in Switzerland and for the Federal Statistical Office (FSO).

Legal: Modification of the law on federal statistics (2006) and of the ordinance on the execution of federal statistical surveys (2014) with rules on data linkage. In 2014 the Federal Department of Home Affairs had also published a new ordinance on linkage of statistical data.

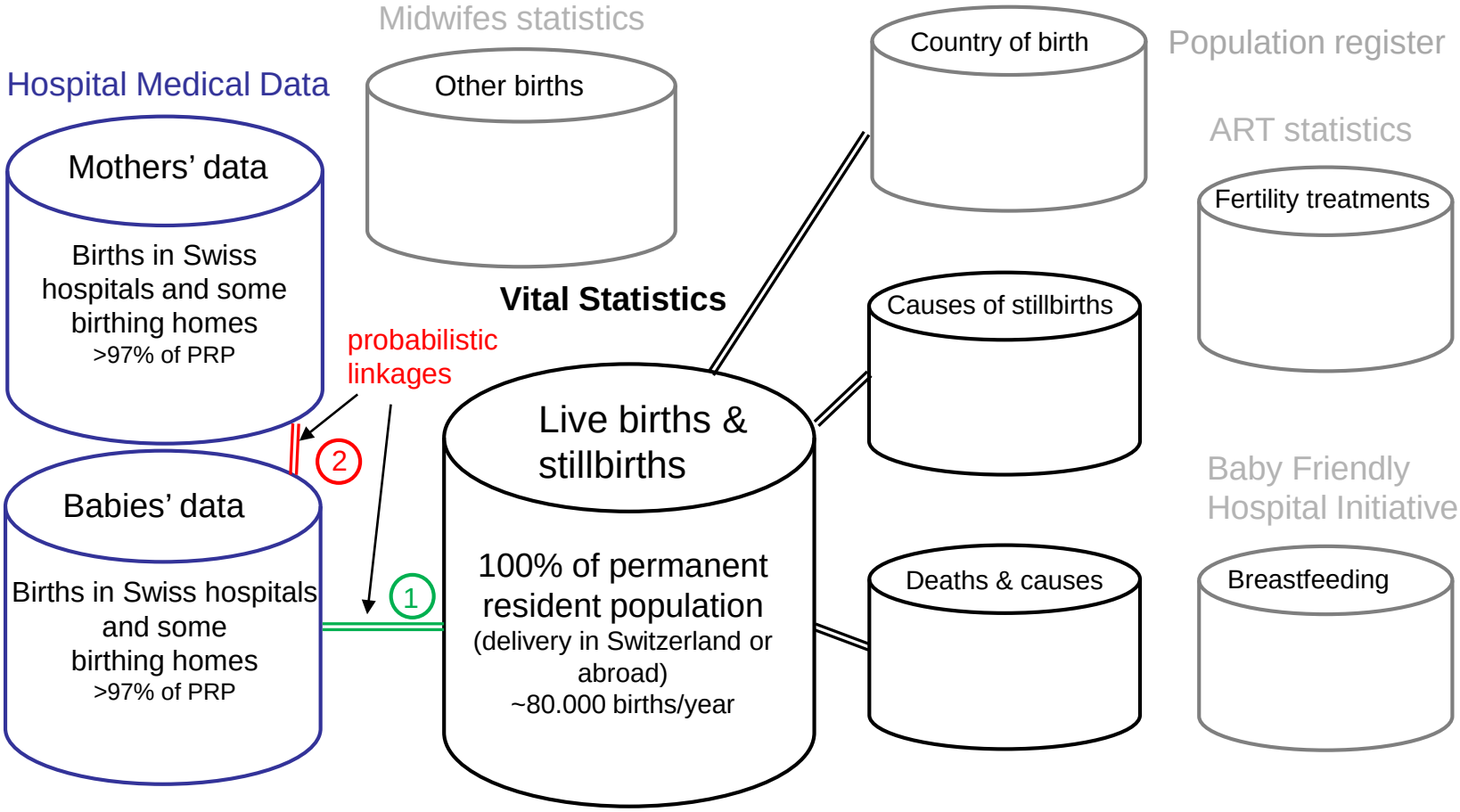
Organizational: New processing policy by the FSO (2014/2015)

Practical: Current implementation of the processing policy and growing use of data linkage by the FSO, but until now very little use of probabilistic data linkage.

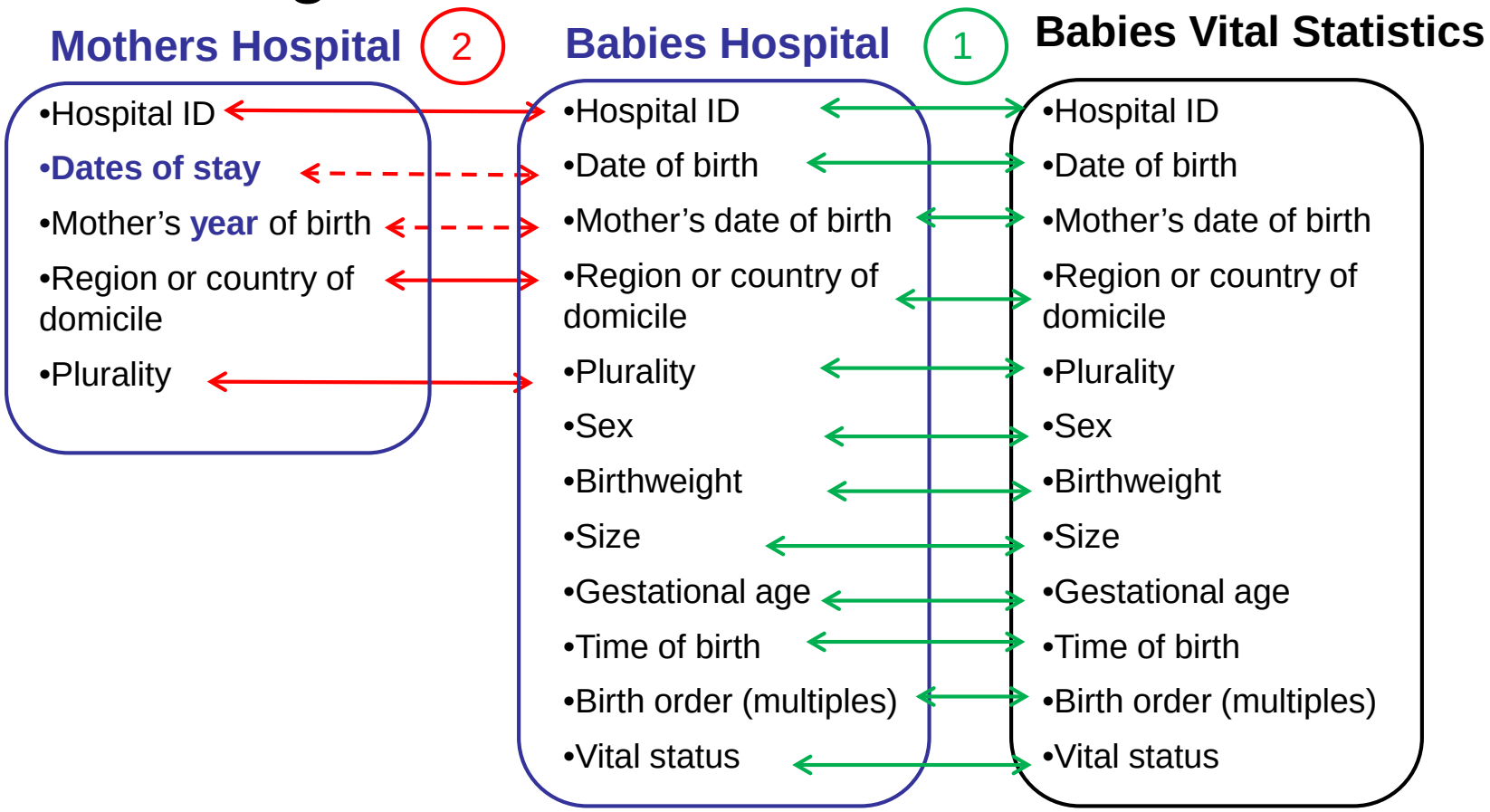
Swiss databases used for the 2010 report



Our data linkage project



Matching variables



Two-step matching of babies' data

First step: deterministic matching

If sex, region or country of domicile, hospital ID, date of birth, mother's date of birth, birth order, gestational age, birthweight and size not missing

→ new variable by concatenation of the variables (e.g. MCHE 847129227320110621198407061403240 49)

If key is unique, used to link the two data files.

→ 85% of the 2010-2014 babies linked at this step.

→ non-linked cases exported for probabilistic linkage

Second step: probabilistic linkage with G-Link

G-Link: software developed by Statistics Canada
to do probabilistic data linkage on large databases.
(statistical software SAS needed to use it)

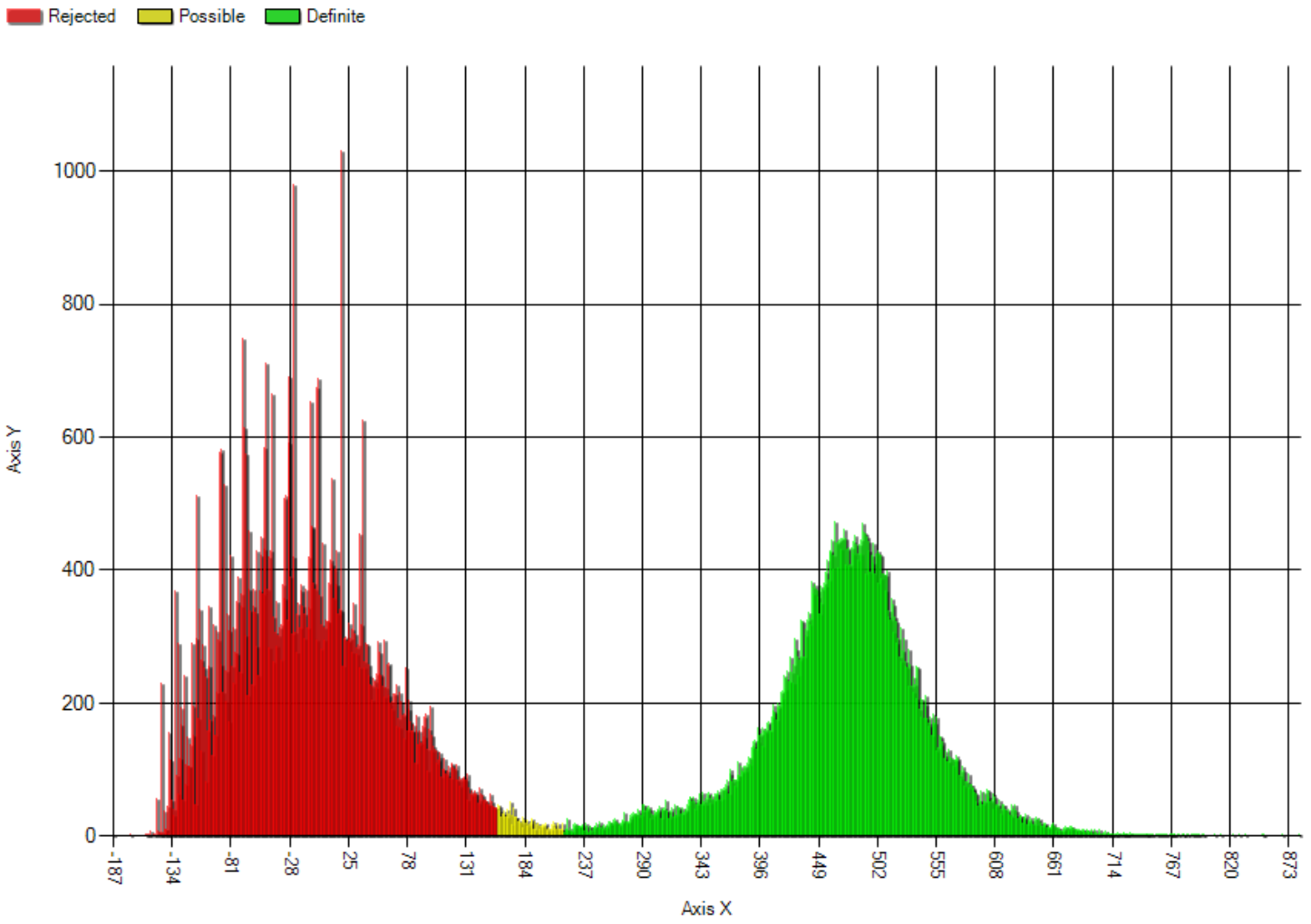
Various steps in G-Link:

- Minimal criterion for possible pairs (=blocking variables)
- Rules applied on matching variables. Various rules for each data-type (number, text or date) are already implemented.
- Outcome weights depending on the quality of the data
- Frequency weights

For each possible pair, the software computes a total weight.

Iterative process to identifies optimal rules, weights and thresholds.

Distribution of weights for babies' linkage



Conclusion

Probabilistic data linkage is often the only way to link large anonymous data files without ID.

A good knowledge of the data is needed for good quality matching. Data pre-processing can be time consuming.

A dedicated software such as G-Link is very useful, but a lot of time is needed to understand and master each matching step and to do the «fine tuning» for a given dataset.

If you are interested to use G-Link:

Statistics Canada Statistique Canada

Benoît Fontaine benoit.fontaine@canada.ca