cidacs
Centro de Integração de Dados e Conhecimentos para Saúde

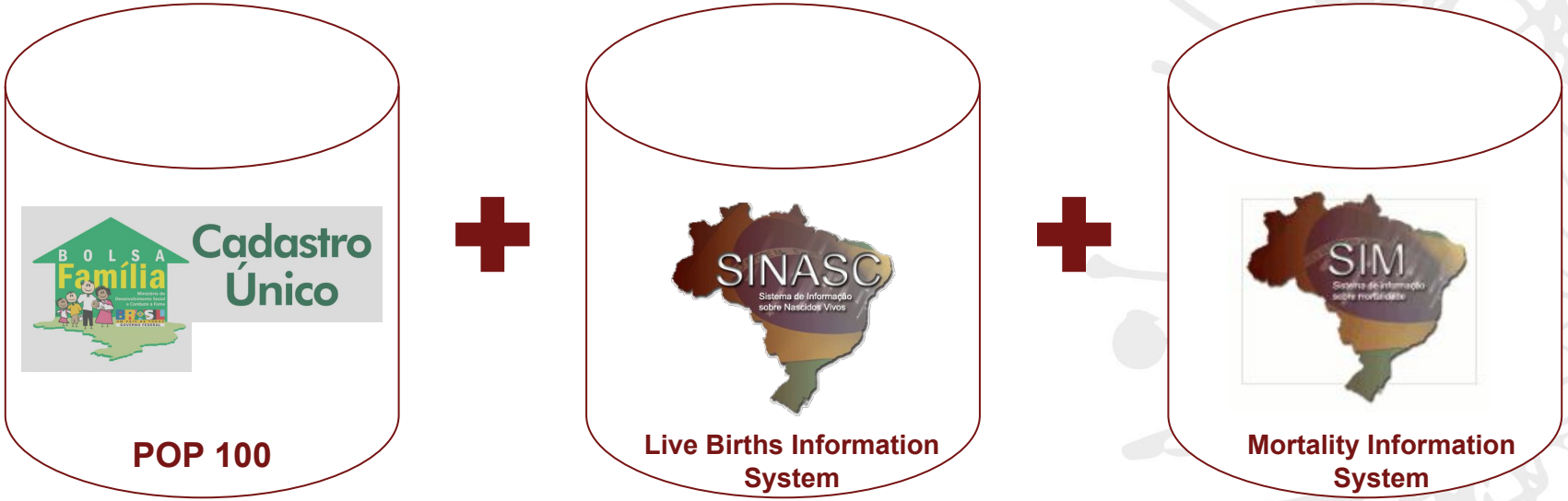
Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz
Instituto Gonçalo Moniz

Construction of Cohort 100M SINASC-SIM

Summary

1. Overview of cohort 100M SINASC-SIM
2. Description of datasets
3. Pre-processing datasets
4. Harmonization
5. Record linkage

Overview of cohort 100M SINASC-SIM



Description of datasets

	POP 100		SIM	SINASC
	CADU	BOLSA FAMILIA		
Nº registers	114.008.179	27.376.582	17.829.111	44.485.274
Nº variables	259	21	138	86
Coverage	2001-2015	2004-2015	2000-2015	2001-2015

Pre-processing datasets

Standardization: Integer variables

- Removal of special characters like:
  - (double) blank spaces;
  - *, -, ‘, “
- The value becomes invalid when there is letters between the numbers

Standardization: Integer variables

Input	Output
00123-000	123000
654*009	654009
01 0607	10607
01**tgp**09	null

Standardization: Date variables

- Removal of separators like:
  - -, /, space
- Conversion to date format

- The value becomes invalid when it cannot be converted to a known date format (yyyyMMdd, ddMMyyyy, etc…)

Standardization: Date variables

Input	Output
01/01/2001	01-01-2001
11 12 2016	11-12-2006
2016-08-13	13-08-2016
40503799	null

Standardization: Categorical variables

- Checks if each category agrees with the dictionary.

Standardization: Categorical variables (sex)

Dictionary	
Gender: M – Male F – Female I – Ignored	Gender: 0 – Null 1 – Male 2 – Female 88 – Ignored 99 – Inconsistent

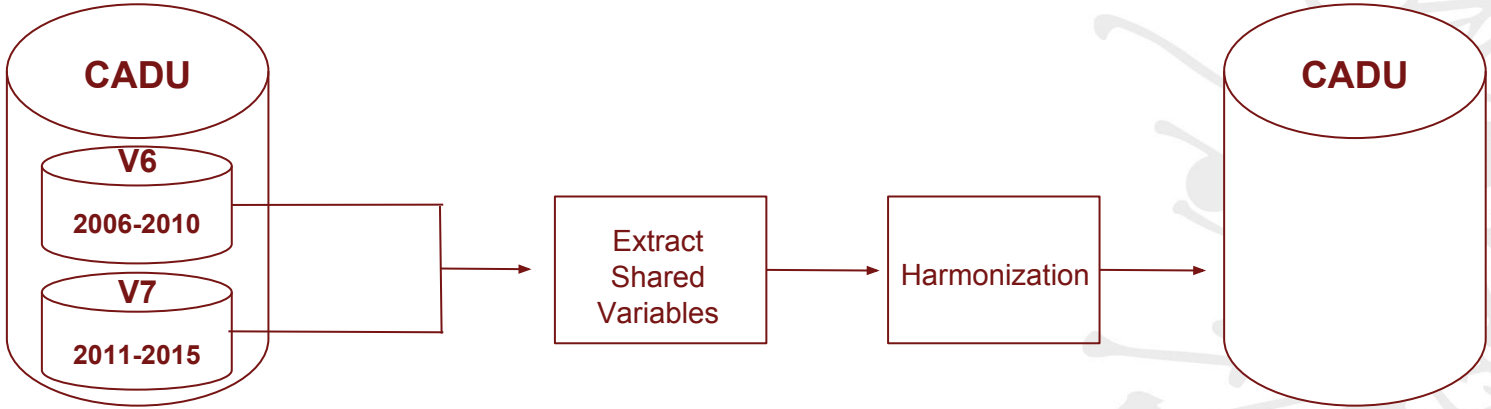
Input	**Output**
M | 51000 F | 54000 I | 5800 O | 2 null | 700	0 | 700 1 | 51000 2 | 54000 88 | 5800 99 | 2

Pre-Processing: String type variables

- To improve the Linkage process
- To improve the analysis
- Depending on the dataset

NAME	NAME_CORR
IGNORADO	NULL
JOAO SA*NTOS	JOAO SANTOS
MARIA2 DA SI8LVA	MARIA DA SILVA

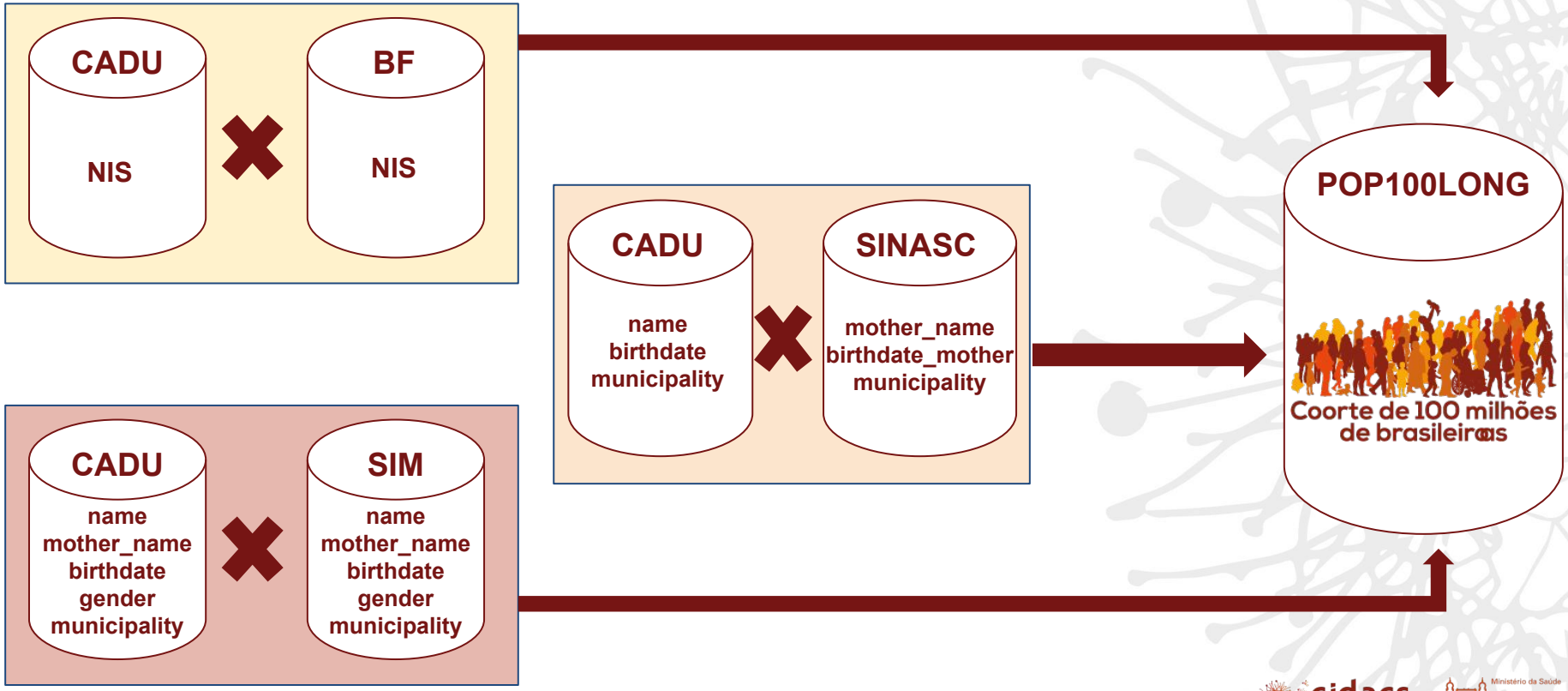
Harmonization



GEN_V6	GEN_V7	GEN_H
M	1	1
F	2	2

Record Linkage

Record Linkage: POP100



cidacs
Centro de Integração de Dados e Conhecimentos para Saúde

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz
Instituto Gonçalo Moniz

Thanks !

COLABORADORES CIENTÍFICOS

APOIADORES

Parque Tecnológico da Bahia
Rua Mundo, 121, Trobogy
Salvador - Ba, CEP 41745-715

www.cidacs.bahia.fiocruz.br
cidacs@bahia.fiocruz.br

fiocruzbahia.cidacs
cidacs_fiocruz
Cidacs Fiocruz

Applications

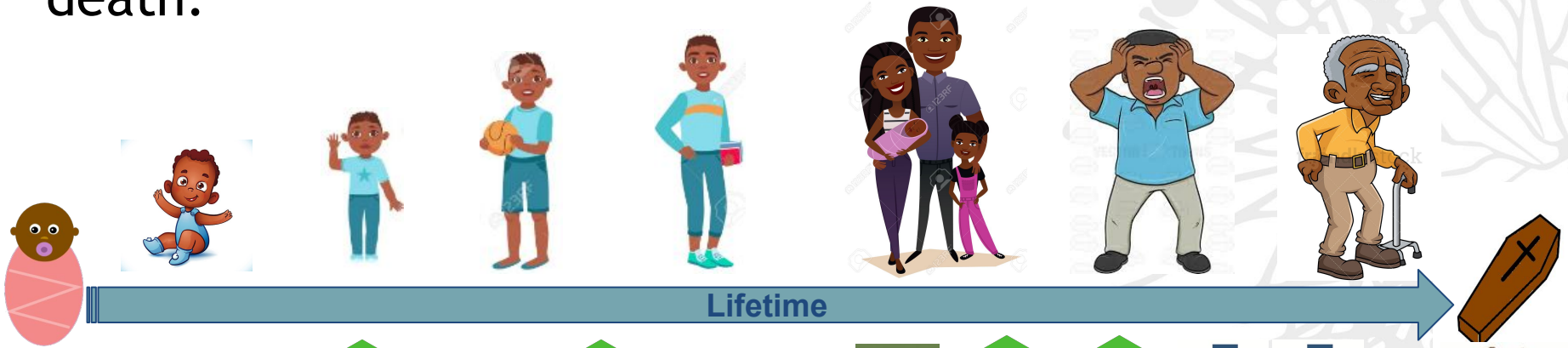
Applications

From this dataset it is possible to know:

- Profile of people registered at CADU since birth to death;
- Factors (sociodemographic, maternal characteristics and childbirth) related to infant death;
- Living conditions of families whose children were born during the study period..

Applications

To follow the impact of conditional cash transfer programme (Bolsa Familia) of some individual from the birth to the death.



Standardization: Categorical variables

Dictionary	
Method used. Values: 1 – Physical exam 2 – Another method 9 – Ignored	Method used. Values: 0 – Null 1 – Physical exam 2 – Another method 88 – Ignored 99 – Inconsistent

Input	Output
1 | 79000 2 | 1200 3 | 97 9 | 600 null | 100	0 | 100 1 | 79000 2 | 1200 88 | 600 99 | 97

Standardization: Categorical variables which starts with 0

Dictionary	
Schooling 2010. Valores: 0 – Without Schooling; 1 – Fundamental I (1ª a 4ª year); 2 – Fundamental II (5ª a 8ª year); 3 – Secondary (Old second degree); 4 – Incomplete undergraduated; 5 – Ungraduated; 9 – Ignored.	Schooling 2010. Valores: 0 – Null 1 – Without Schooling 2 – Fundamental I (1ª a 4ª série) 3 – Fundamental II (5ª a 8ª série) 4 – Secondary (Old second degree) 5 – Incomplete undergraduated; 6 – Ungraduated 88 – Ignored 99 – Inconsistent.

Standardization: Integer variables

Dictionary	
Birth hour	Birth hour (HHMM)

Input	Output
08:59	0859
09/36	0936
085:1	null

Standardization: Binary Categorical variables

Dicionário	
Status of new DO variable: 1 – Yes 0 – No	Status of new DO variable: 0 – Null 1 – Yes 2 – No 99 – Inconsistent

Entrada	Saída
1 | 45000 0 | 6700 3 | 22 null | 700	0 | 700 1 | 45000 2 | 6700 99 | 22

Correction on string type variables

Resultados	
Recovered names of children	0
Unrecovered names of children	419.557
Recovered names of mothers	5.228
Unrecovered names of mothers	1.932