Supplementary Material 2

Creation of gold standard data set

Figure 1 depicts the approach used to create the gold standard:

A. Only records related to 2015 were used in the experiment: 3,013,228 live births and 1,293,219 deaths;

B. Records without DN number were removed from SIM, as they would not be linked to SINASC. Names with generic or invalid information (e.g. unknown or stillbirth) or with only one name (e.g. Maria) were also removed. We also removed records with void and invalid names from SINASC, as the majority of babies did not have a name yet;

C. Both SIM and SINASC records of the same child were linked via SINASC registration number (DN) by exact match, adding up to 3,028 records. Thus, stillbirths who had the same DN number on both systems, SIM and SINASC, were linked;

D. Each linkage tool was configured to return a data set at the size of the smaller one and to evaluate the performance of each tool the SINASC registration number was used during the experiment.

E. A number of non-linked records from SIM and SINASC were added to the final linked data in order to simulate a number of matches usually found in CIDACS data sets. We added 3,429 records to SIM data set and 10,017 to SINASC data set expected to be non-linkable to evaluate linkage performance tools for identifying false matches;

F. Each of the five linkage systems produces a resulting file with the matched pairs and their respective similarity scores. Except from FRIL and RecLink, all other tools produce a resulting data mart of size equals to the smallest data set linked. FRIL and RecLink require a threshold to be set prior to execution and they include any pair with a score higher than the threshold in the result, while AtyImo, CIDACS-RL and Febrl keeps only the pair with the highest score found amongst all candidate pairs.

G. When a record from the smallest data set is matched with two different records on the larger data set, only the pair with the highest similarity score was considered for metrics calculation;

H. After selecting cut-off points for each record linkage tool, sensitivity, specificity and PPV metrics were calculated.

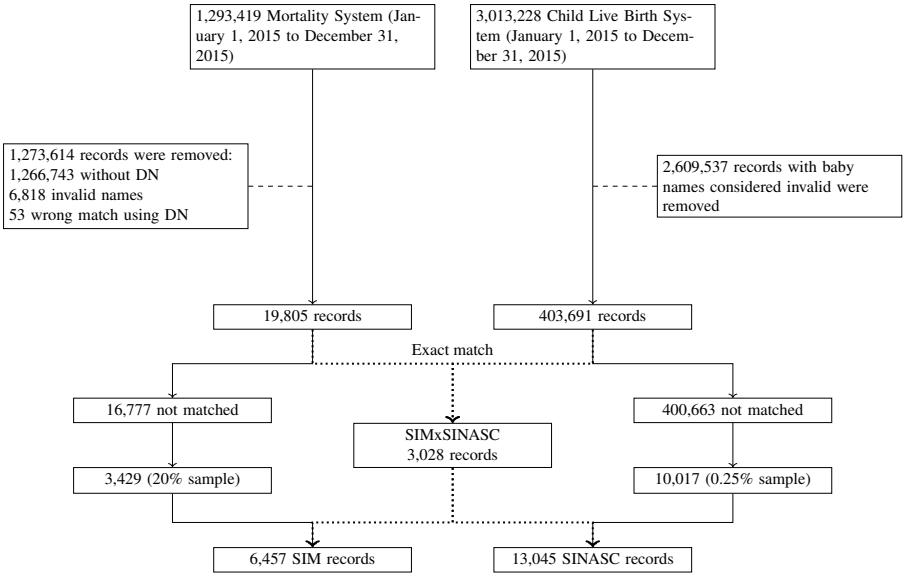

Figure 1: Construction of the gold standard data set