Supplementary Material 1

Description of other linkage tools

RecLink is a C++-based record linkage tool implementing probabilistic data linkage methods. Records are blocked based on soundex functions applied over name attributes, and further compared by three different similarity functions. The user can define cut-off points to customize matching decisions.

Febrl is a Python-based data linkage pipeline implementing data cleansing, deduplication and pairwise comparison. It has different standardizes for names, dates, and addresses. Febrl implements seven indexing methods which are customizable by the user. More than 20 different functions are supported for pairwise comparison, as well supervised and unsupervised classifiers are used to perform weight vector classification.

FRIL is a Java-based tool providing a set of highly customizable functions. Data integration (or reconciliation) is supported through different merging and splitting functions. The user can choose different strategies for pairwise comparison: nested loop joins, for small data sets, or sorted neighborhood or index search, for large data sets. Matching weights are defined based on the expectation-maximization (EM) method. FRIL runs transparently over multicore architectures.

AtyImo is a probabilistic data linkage tool, jointly developed by UFBA and CIDACS between 2013 and 2016, that runs distributed over Spark or in parallel over CUDA in hybrid (multicore+multi-GPU) architectures and is freely available on Github. It implements a pipelining comprising data preprocessing (cleansing, standardization, blocking and anonymization), pairwise comparison and accuracy assessment. Blocking is based on different predicates built with five linking attributes (name, mother's name, date of birth, sex and municipality).

In AtyImo, anonymization is based on a 128-bit Bloom filter, which guarantees privacy-preserving requisites related to sensitive (identifiable) data, allowing AtyImo to run within less protected environments if needed. AtyImo implements a two-round linkage step in which a mixture of deterministic and probabilistic methods can be used together to generate high accurate data marts (domain specific data). Finally, accuracy assessment can be performed manually, based on gold standards (when existent), to certify small data marts; or automatically, based on supervised machine methods, to certify big data marts.