

International Journal of Population Data Science

Journal Website: www.ijpds.org



A Fully Automated End-to-End Framework for Scalable Record Linkage

Barret A. Monchka¹, Cole Chuchmach², Elizabeth Stoughton², Jingwen Li¹, and Lisa M. Lix¹

¹University of Manitoba, Winnipeg, Canada; George & Fay Yee Centre for Healthcare Innovation, Winnipeg, Canada

²University of Manitoba, Winnipeg, Canada

Background

Automated record linkage is often defined as applying a single threshold to classify candidate record pairs as matches or non-matches, in contrast to the Fellegi–Sunter two-threshold approach that relies on manual review for uncertain pairs. However, eliminating manual review alone does not constitute fully automated linkage, which also requires automation of data preprocessing, parameter selection, linkage quality evaluation, and report generation—steps that remain largely manual in practice and contribute substantially to the time and cost of data linkage projects.

transparency—facilitating more efficient, equitable, and sustainable linkages for routinely updated datasets.

Objective

To develop an open-source, fully automated, record linkage framework that supports scalable, reproducible, and transparent workflows.

Implementation

We developed three R packages, publicly available at <https://github.com/CHIMB>, that together form a fully automated linkage pipeline: *datastan*, *autolink*, and *linkrep*. The *datastan* package automates data standardization and preprocessing by transforming each data source into a consistent format prior to linkage, including variable naming, date formatting, variable categorization, standardization of name spelling variations, and handling of missing data. *Autolink* enables users to define and execute complex linkage algorithms, supporting deterministic, probabilistic, fuzzy, and phonetic matching. *Linkrep* automatically generates comprehensive linkage quality reports. Both the data standardization rules and linkage algorithms are easily editable through a graphical user interface and can be defined once, saved, and set to execute automatically for subsequent linkages (e.g., when data updates become available), without further user intervention.

Implications

By automating the full linkage lifecycle, this framework reduces turnaround times while improving reproducibility and

