FCDS Florida Cancer Data System

Florida Statewide Cancer Registry

Executive Summary

This 2024 FCDS monograph is titled *Comparing the Linkage Performance of `fastLink`, `Splink`, and `Match*Pro` at the Florida Cancer Registry Using Simulated `pseudopeople` Data.* fastLink is an R package, Match*Pro is a Java-based Windows application, and Splink is a Python package. The software are tested on 670,214 simulated records from the Python package pseudopeople. A requirement is no expected false positives (FP) after a clerical review. Compared with fastLink, Splink predicts about 9% more true positives (TP) and Match*Pro predicts about 20% fewer TP. Therefore, the FCDS recommends using Splink for probabilistic record linkage (PRL). The main limitation is that Splink requires beginner-level skills in Python. The monograph consists of the main text (13 pages), and of a technical showcase (62 pages).

PRL has four main steps, which do not include software installation and configuration. Match*Pro is much easier to install and configure than R/RStudio/Quarto/LaTeX/fastLink, which is easier than Python/RStudio/Quarto/LaTeX/Splink. These are the comparative results for the four main steps:

1) pre-processing (known as "attribute alignment"): The performance is similar. Splink has a useful completeness chart feature. All three software would benefit from more standardized linkage variables, especially for name, Social Security Number (SSN), and address.

2) blocking: Match*Pro and Splink are better thanks to "OR (disjunctive)" blocking.

3) PRL: Splink is more accurate by enabling more complicated models. Splink was typically 12 times faster than fastLink (20 minutes vs 4 hours).

4) post-processing (known as "canonicalization"): Splink is more difficult to use. A more user-friendly Splink 4 is due to be released in fall 2024.

The following are three FCDS recommendations, in order of suggested priority:

- Gradually replace fastLink with Splink 4 for linkage data requests. For more comparative testing at the FCDS, fastLink would need a new major release, and Match*Pro would need a variable for "match probability".

- Create an FCDS template for using Splink. Test changing the Integrated Development Environment (IDE) from RStudio to Visual Studio Code (VS Code) or to Positron (in beta). Try collaboration using Live Share in VS Code. Try development using Dev Containers in Project Bluefin.

- Begin to standardize linkage variables in the FCDS database.