FCDS Florida Cancer Data System
Florida Statewide Cancer Registry

Comparing the Linkage Performance of `fastLink`, `Splink`, and `Match*Pro` at the Florida Cancer Registry Using Simulated `pseudopeople` Data (2024 Monograph)

Anders Alexandersson

06/30/2024

This 2024 monograph describes a performance study at the Florida Cancer Data System (FCDS) of the software `fastLink`, `Match*Pro`, and `Splink` for probabilistic record linkage (PRL). The software are tested on 670,214 simulated records from the `pseudopeople` package. 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 to use `Splink` for PRL. The main limitation is that `Splink` requires beginner-level skills in Python and initial Python setup which is labor intensive.

Introduction

This 2024 monograph describes a performance study at the Florida Cancer Data System (FCDS) of the software fastLink (Enamorado, Fifield, and Imai 2019; Enamorado 2019, 2021), Match*Pro, and Splink (Linacre et al. 2022) for probabilistic record linkage (PRL). The monograph consists of two components:

- The main text, this document: a high-level non-technical overview (13 pages)

- A supplement: a low-level technical showcase using 670,214 artificial records (62 pages). See part 1 (S1) for the `pseudopeople` data, part 2 (S2) for the `fastLink` results, part 3 (S3) for the `Splink` results, and part 4 (S4) for the `Match*Pro` results.

The FCDS usually uses the R package `fastLink` for PRL. The FCDS in 2021 and 2022 developed two `fastLink` templates which are R Markdown document templates on how to use `fastLink` within the RStudio IDE. The FCDS 2023 monograph suggested that it is feasible to use `Splink` for PRL, because `Splink` was 50 times faster and more accurate than `fastLink`. The North American Association of Central Cancer Registries (NAACCR), through the Virtual Pooled Registry Cancer Linkage System (VPR-CLS), recommends `Match*Pro` for PRL.

PRL is usually described as one step in a data cleaning pipeline with these four steps (e.g., Binette and Steorts 2022): 1) attribute alignment, 2) blocking, 3) PRL, and 4) canonicalization. PRL can have two sets of errors: predicted match and actual non-match (false positive, FP), and predicted non-match and actual match (false negative, FN). The sections below will compare the performance of the three software in the four steps, and it will conclude with recommendations.

The Comparative Performance Study Design

The objective is to compare the performance of `fastLink`, `Splink` and `Match*Pro` for large PRL at the FCDS. `fastLink`, `Match*Pro`, and `Splink` currently are the most relevant PRL software to the FCDS. The test data are created from simulated "Rhode Island" pseudopeople 0.8.3 data (Haddock et al. 2024). On the test data, there are 670,214 linkable records, with 660,227 actual matches and 9,987 actual non-matches. For easier analyses, all actual non-matches come from the smaller second dataframe, `df2`. It is critical to use large PRL because as a dataset grows, the risk for FP and FN grows quadratically. The main advantage of the `pseudopeople` data is that it is designed for large and realistic reproducible PRL, which includes labelled data with the actual match status.

A requirement was no expected FP after a limited clerical review. For performance metrics, we used the confusion table (confusion matrix), and especially correct matches (true positives, TP) since FP is expected to be 0 after the clerical review. A comparison of derived performance metrics, such as Matthews correlation coefficient (MCC or phi), was beyond the scope of the study.

At a minimum one of the following combinations are required to link records with the FCDS:

1) First Name, Last Name, Sex, Date of Birth, Zip Code and Street Address
2) First Name, Last Name, Sex, Date of Birth, and Social Security Number (SSN)

Therefore, we select those seven variables in the `pseudopeople` test data.

Because of the importance of SSN, we test the performance with and without SSN. The supplement reports many results. In this main text, we report only the overall best test results for the three software, so six results in total. See below, section `Linkage Performance` Table 1 and Table 2 for the linkage performance.

We used the latest versions of the three software: `fastLink` 0.6.1, `Splink` 3.9.14, and `Match*Pro` 2.4.4. We used `fastLink` on R 4.3.0 for Windows. We used `pseudopeople` and `Splink` on Python 3.11.0 for Windows 10. `Match*Pro` is a standalone software written in Java for Windows. `Match*Pro` is often easier to use than `fastLink` and `Splink` by having a graphical user interface (GUI). However, `fastLink` and `Splink` are easier for automated reporting using Quarto Markdown.

For access to the `pseudopeople` data, please make a data request on the project Github page. Reproducibility can be defined (e.g., Brodeur 2024, 14) as the examination of whether the results and conclusions of original studies can be duplicated using the original studies' data. There are three types of reproducibility: computational reproducibility, recreate reproducibility, and robustness reproducibility. Computational reproducibility (transparency) requires both data and code. `fastLink` and `Splink` as free and open source software (FOSS) are computationally reproducible (transparent) whereas `Match*Pro` as proprietary software is not.

Recreate reproducibility do not require data and/or code. Recreate reproducibility is possible by comparing match probability because the evidence for a match is commonly represented as a probability. `fastLink` and `Splink` use an iterative optimization algorithm for PRL known as Expectation Maximization (EM). `Match*Pro` by default instead uses a frequency-based `EpiLink`-like (Contiero et al. 2005) algorithm or "fuzzy matching" for calculating "total score", not match probability. `Match*Pro` has an EM algorithm but it is optional, and it does not provide a variable for match probability. Robustness reproducibility (sensitivity analysis) uses alternative plausible analytical decisions. We support robustness reproducibility by analyzing the effect of SSN, and by sensitivity analyses in the supplement.

The four linkage steps can be complex and require user decisions, each of which has an impact on the end result.

Step 1: Attribute alignment

Step 1, attribute alignment, is the pre-processing before step 2, blocking. An attribute is an elementary feature of an entity such as address, date of birth, gender or name. All three software require aligned attributes.

`fastLink` has the least features for PRL, which makes `fastLink` more dependent on R packages for attribute alignment. The FCDS `fastLink` templates use the R package PGRdup for Double Metaphone of first and last name. The FCDS `fastLink` templates also create two tables for exploratory analyses, with the first table showing summary completeness and the second table showing detailed missingness.

Splink has extensive documentation on statistical best practices, which for attribute alignment includes a tutorial and a topic guide. Splink has two built-in interactive charts for exploratory analysis: a completeness chart, and profile columns. Typically, the completeness chart is the most useful. Figure 1 belows shows the non-interactive completeness graph for Splink on the test data when SSN is available:

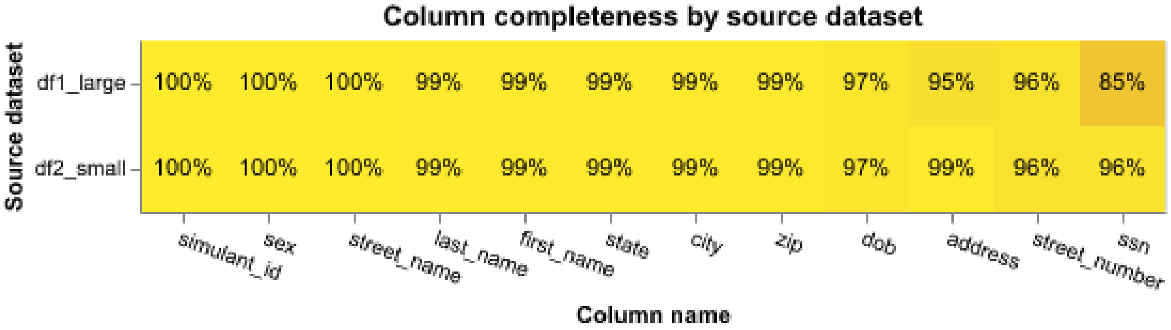


Figure 1: Completeness Chart

Figure 1 succinctly indicates that a major issue is the completeness of SSN in df1, with 85% completeness and, equivalently, 15% missingness.

Match*Pro has relatively many linkage comparators. For example, it has Double Metaphone. This makes attribute alignment easier. The typical way to configure Match*Pro is by creating and then re-use a Match*Pro linkage configuration file (".mplc"). Match*Pro can rename input variables but Match*Pro has no built-in graph or table for exploratory analyses.

Step 1, attribute alignment, is the most difficult to automate of the four steps. This step would benefit from more standardized linkage variables, especially for names, SSN, and address (preferably an implementation of the Project US@ Technical Specification). For error checking of step 1, Splink is the most user-friendly software thanks to its built-in charts.

Step 2: Blocking

Step 2, blocking, is optional in both Splink and fastLink but required in Match*Pro. Blocking can lead to FN but it enables faster, more accurate and larger record linkages. A promising blocking method is probabilistic blocking. Probabilistic blocking means using record linkage in the blocking step (Enamorado and Steorts 2020), which is rather difficult in all three software. The fastLink developer Ted Enamorado is developing a new blocking algorithm based on locality-sensitive hashing (LSH)(source: comment on issue #83).

The fastLink blocking was done with the blockData() function using exact matching on sex (2 values) and k-means clustering on first name (3 clusters), with a total of 6 blocks. The output of the fastLink file for blocking is a blocking object.

Splink requires one of three link type settings: dedupe_only, link_only, and link_and_dedupe. The setting names are self-explanatory except that Splink requires extra code to remove duplicates. Splink has two types of blocking rules, namely, for prediction and for model training. The aim of prediction blocking rules are to capture as many true matches as possible, and to reduce the total number of comparisons being generated. Splink has a blocking chart which counts the number of comparisons generated by blocking for prediction. The blocking for prediction works cumulatively, as shown in Figure 2 below, which has three SQL blocking rules:

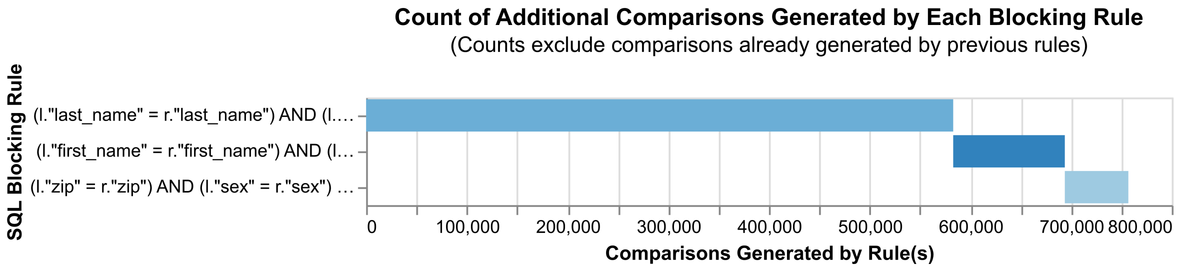


Figure 2: Blocking Chart

The Splink blocking for model training is described in step 3, record linkage. For step 2, blocking, Splink and Match*Pro are better than fastLink by offering OR (disjunctive) blocking which reduces the risk of FN. Match*Pro blocking is difficult to troubleshoot because it does not provide separate blocking output like fastLink or Splink.

Step 3: Record linkage

For fastLink, extracting matches from record linkage when using blocking is an issue with a workaround. In the fastLink templates, Table 3 is frequencies of linkage pattern and Table 4 is the confusion table. A similar frequency table is less useful for Splink because Splink allows more comparison values. A confusion table is usually very difficult to create for Splink because it requires available labeled data. With fastLink, a limitation is that only *one* string distance method (of four) is allowed. For example, you may select using either Jaro-Winkler similarity on first name or Damerau-Levenshtein distance on SSN but not both.

In Splink, the record linkage step consists of two sub-steps: 1) estimate model parameters, and 2) predict results. Having two sub-steps enables more complicated models which often are more accurate. Estimate model parameters means to estimate:

- λ (`lambda` or prior): The probability that two random records (with no blocking) match.
- `u`: The probability that wrong matches match (FP).
- `m`: The probability that true matches match (TP).

Two ways to get faster and more accurate results in `Splink` than using the default settings are to set `lambda` manually, and to estimate `lambda` and `u` directly.

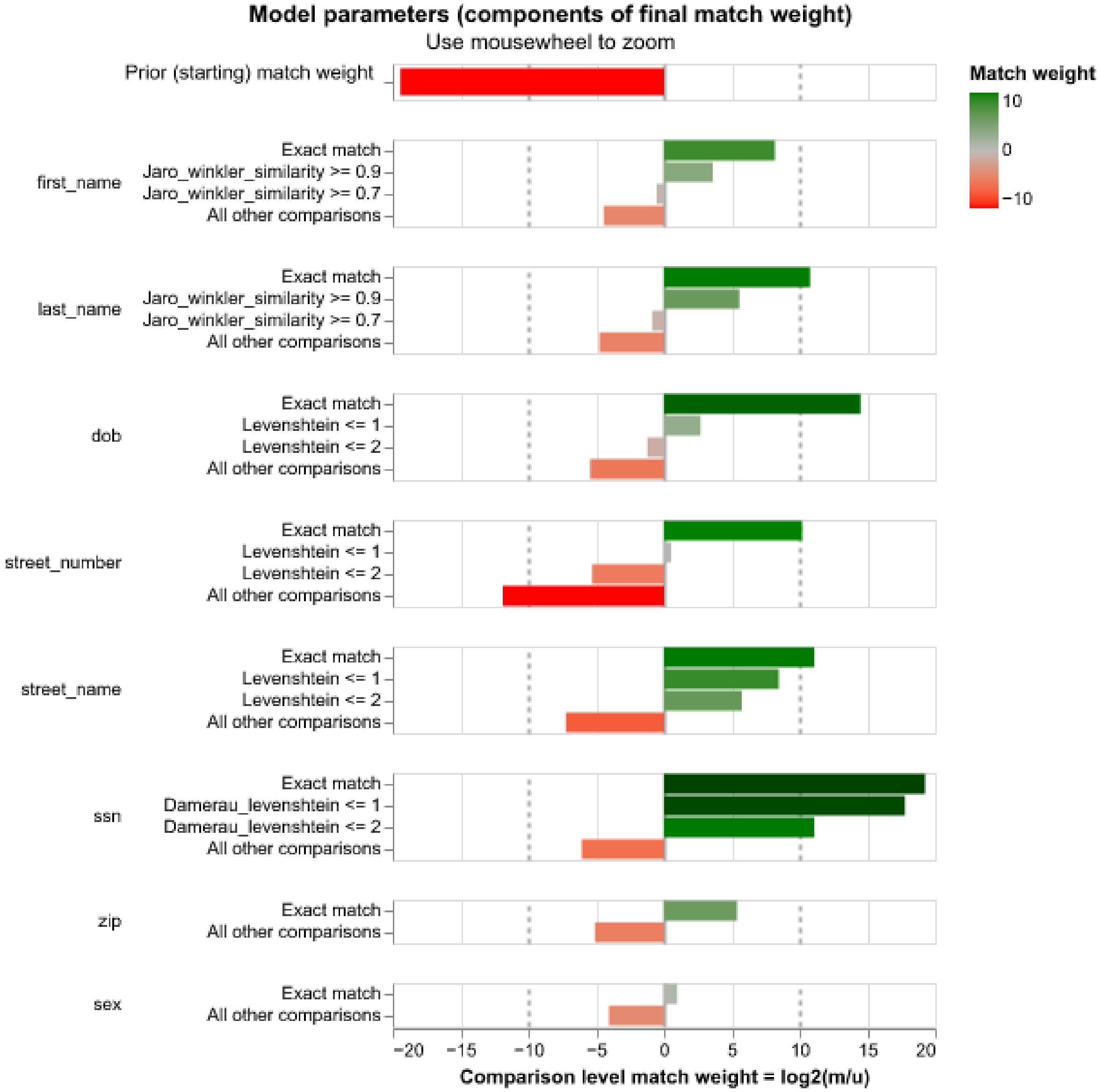
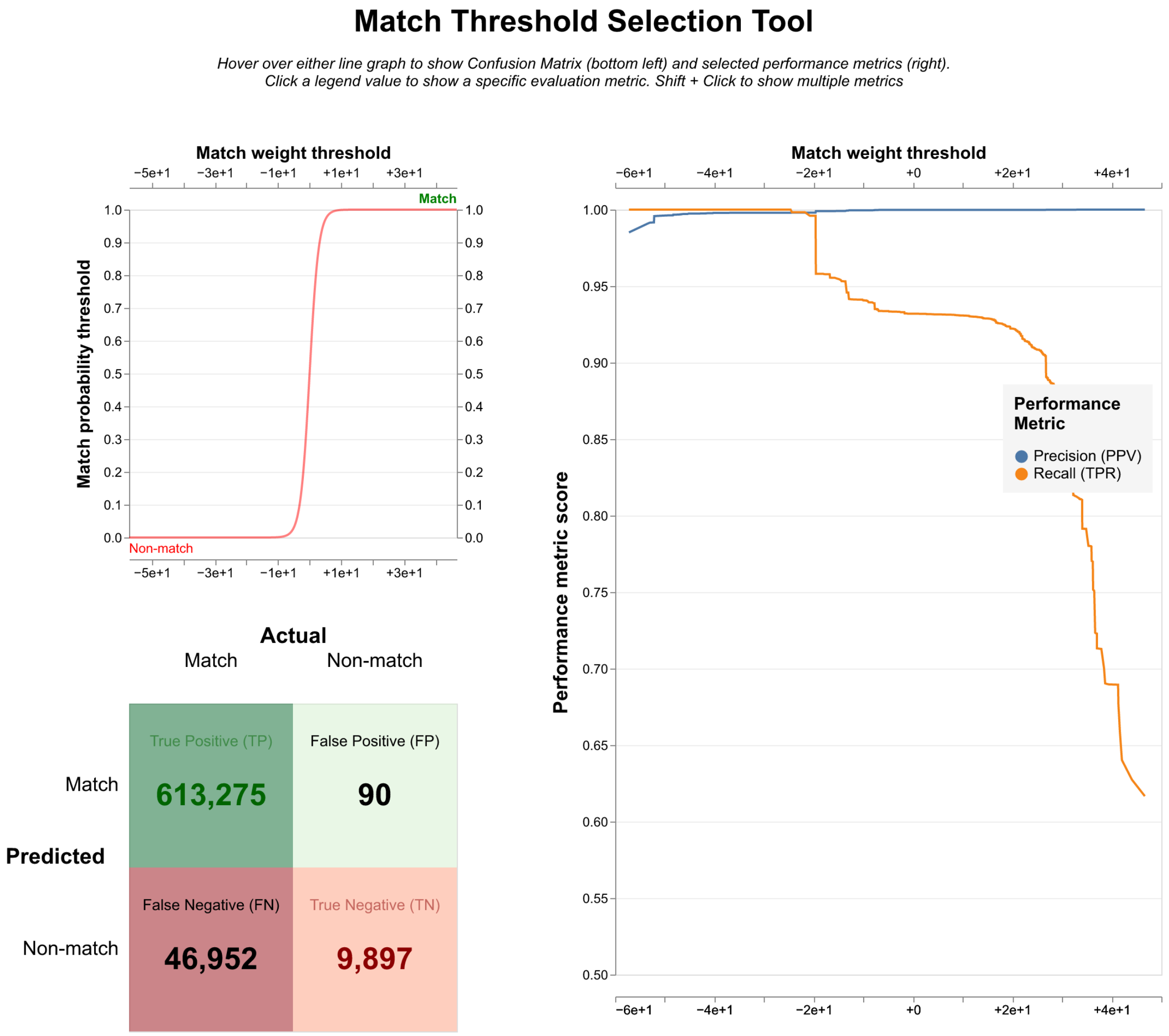


Figure 3: Match Weight Chart

Predict results mostly means to calculate the overall match probability. The `Splink` match weight chart (as PNG) in Figure 3 shows the results of a trained `Splink` model. The exact match comparison takes priority. Records that do not fall within a comparison level are allocated to the rest of the comparison levels.



Match Threshold Selection Tool

Hover over either line graph to show Confusion Matrix (bottom left) and selected performance metrics (right).
Click a legend value to show a specific evaluation metric. Shift + Click to show multiple metrics

		Actual	
		Match	Non-match
Predicted	Match	True Positive (TP) 613,275	False Positive (FP) 90
	Non-match	False Negative (FN) 46,952	True Negative (TN) 9,897

Figure 4: Match Threshold Selection Tool

Threshold selection is a key decision point within a linkage pipeline. One of the major benefits of PRL versus a deterministic (i.e. rules-based) approach is the ability to choose the amount of evidence required for two records to be considered a match (i.e. a threshold). As more predicted matches become non-matches at the higher threshold, TP become FN, but FP become TN. This demonstrates the trade-off between FP (Type 1) and FN (Type 2) errors when selecting a match threshold, or precision vs recall. Only Splink has a match threshold selection tool. As Figure 4 shows, a reasonable threshold here is high, at least 0.99.

This match threshold selection tool is especially useful interactively, which requires HTML output (chart), not PDF output (graph). The developers are working on some issues with the tool (see issue 2070 and pulls 2120, 2150, and 2187).

The largest weakness with the Match*Pro default is that it fails to predict the data quality of the overall dataset: it does not quantify the relative importance of non-matches. This is why PRL is more accurate than fuzzy matching, and why Match*Pro needs to provide match probability to assess accuracy. The optional Match*Pro EM algorithm also does not provide a variable for match probability.

Step 4: Canonicalization

Step 4, canonicalization, is the post-processing after the PRL. By default, fastLink deduplicates the matches into representative or "canonical" records. Unfortunately, the fastLink deduplication seems to only occur on the first dataset, see issue 78. The most time consuming part of step 4 is the clerical review. The showcase could not fully test the clerical review since the test data are simulated (artificial). The FCDS uses the RStudio Data Viewer for the clerical review of uncertain matches (of adults) against the records in the online tool Accurint from LexisNexis.

Choice of threshold can have a significant impact on the final linked data produced (i.e. clusters). For example, if we increase the threshold from 0.95 to 0.99, then linked records with threshold 0.96 are discarded. It breaks the records into two clusters. See Figure 5. For how to interpret the graph further, see the topic guide linked data as graphs.

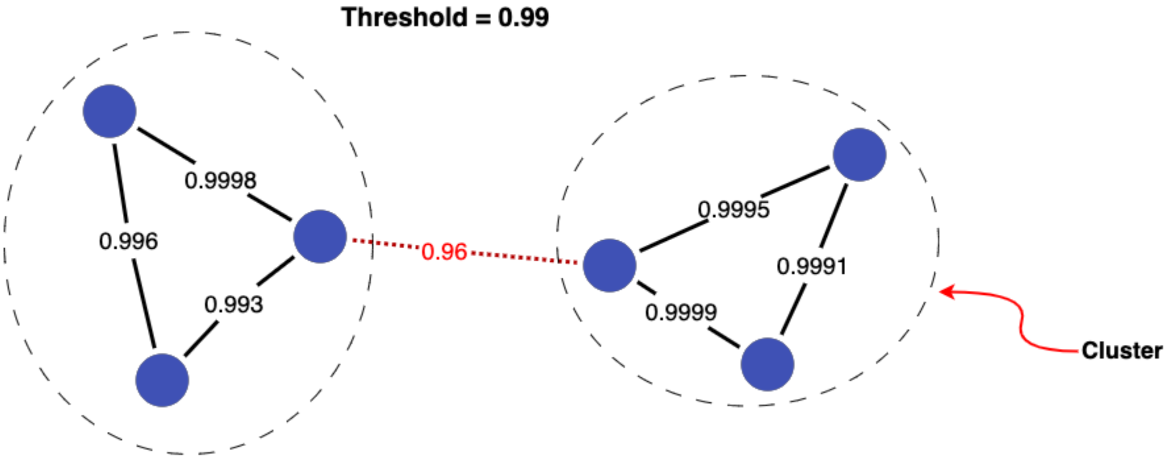
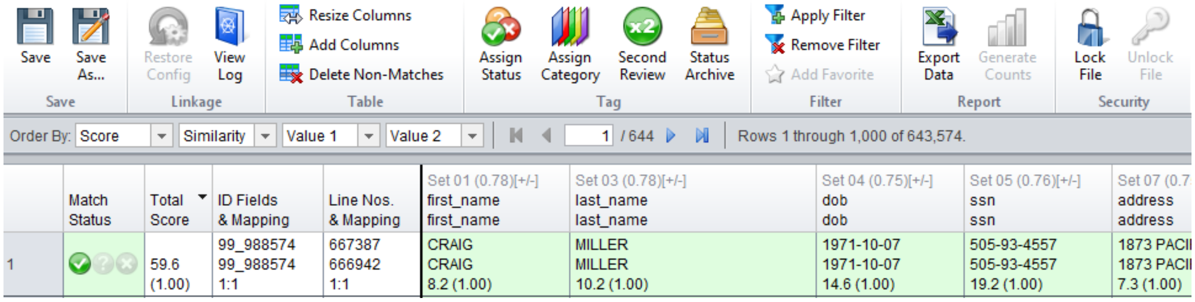


Figure 5: Linked Data as Graph

Similar to step 1, attribute alignment, Splink has no feature for deduplication in step 4. Unlike fastLink, Splink requires extra code for removing possible duplicates from the record linkage. How to best do the clerical review in Splink is uncertain. The waterfall chart which shows the breakdown of the match probability for pairs of records, is useful for spot checking but not for reviewing many records. A clerical labeling tool is in testing (beta) but progress has stalled due to the current focus on Splink 4. For now, the best practice seems to be to use the RStudio Data Viewer, similar to fastLink.

Match*Pro does not document how it does deduplication. The Match*Pro Classification tab is a set of currently 98 rules that adjusts the "total score" into "Match", "Non-Match" or "Uncertain". It is not clear, how is the Match*Pro "total score" related to match weight and match probability? Match*Pro does have a useful clerical review tool, as shown in Figure 6 below:



Order By: Score | Similarity | Value 1 | Value 2 | 1 / 644 | Rows 1 through 1,000 of 643,574.

	Match Status	Total Score	ID Fields & Mapping	Line Nos. & Mapping	Set 01 (0.78)[+/-] first_name first_name	Set 03 (0.78)[+/-] last_name last_name	Set 04 (0.75)[+/-] dob dob	Set 05 (0.76)[+/-] ssn ssn	Set 07 (0.7 address address
1	✓	59.6 (1.00)	99_988574 99_988574 1:1	667387 666942 1:1	CRAIG CRAIG 8.2 (1.00)	MILLER MILLER 10.2 (1.00)	1971-10-07 1971-10-07 14.6 (1.00)	505-93-4557 505-93-4557 19.2 (1.00)	1873 PACI 1873 PACI 7.3 (1.00)

Figure 6: Clerical Review Tool

Linkage Performance

The PRL performance compared in this report is the use case of a "large PRL" using simulated pseudopeople data. Other possible use cases are not compared here such as a small-to-medium PRL, deduplication, real-time PRL, and privacy-preserving record linkage (PPRL). Because of the importance of SSN, we report the linkage performance of the large PRL both with and without SSN as a linkage variable (ssn).

Splink has the fastest EM algorithm by using the backend database DuckDB. Typical run times are: fastLink 4 hours, Splink 20 minutes (12 times faster), Match*Pro default algorithm 60 minutes (4 times faster). All computations were done on a Dell Precision 3431 desktop with an Intel Core i9-9900 CPU, which was high end in 2019. Similar run times should be expected from a modern (2024), fast, and repairable laptop such as a Framework Laptop 13.

Regardless if SSN is available, the FCDS expects the clerical review to be <= 500 records for fastLink and Splink and <= 1,000 records for Match*Pro. The expected clerical review for Match*Pro is larger for two competing reasons: Match*Pro is easier to configure which frees

Table 1: Confusion Table If Using ssn

(a) fastLink

	True Matches	True Non-Matches	Total
Links	571,568 TP	19 FP	571,587
Non-Links	88,659 FN	9,968 TN	98,627
Total	660,227	9,987	670,214

(b) Splink

	True Matches	True Non-Matches	Total
Links	627,315 TP	4 FP	627,319
Non-Links	32,912 FN	9,983 TN	42,895
Total	660,227	9,987	670,214

(c) Match*Pro

	True Matches	True Non-Matches	Total
Links	436,494 TP	4 FP	436,498
Non-Links	223,733 FN	9,983 TN	233,716
Total	660,227	9,987	670,214

up time for the clerical review (good), and Match*Pro records are difficult to rank without match probability which requires more clerical review (bad).

When SSN is available, the FCDS recommends that fastLink and Splink require a match on first_name and sex. Match*Pro does not require a match on first_name and sex thanks to the rules in the Classification tab. Table 1 shows the resulting confusion tables when SSN is available.

Table 1 shows that Splink (Table 1b) performed best when SSN is available, followed by fastLink (Table 1a) and Match*Pro (Table 1c). Compared with the fastLink TP (571,568), Splink predicted 9.8% more TP (627,315), and Match*Pro predicted 23.6% fewer TP (436,494). The threshold match probability was 0.98 for fastLink, 0.999999999 (9 decimals) for Splink, and 0.95 with "Total Score > 40" for Match*Pro regardless if SSN was available. The higher threshold of "Total Score > 40" was required to compensate for the uncertainty of not having match probability in Match*Pro, which worsened the performance.

Table 2 shows the same ranking when SSN is not available. When SSN is not available, the FCDS recommends that fastLink, but not Splink, require a match not only on first_name and sex but also on dob. Again, Splink (Table 2b) performed best, followed by fastLink (Table 2a) and Match*Pro (Table 2c). Compared with the fastLink TP (532,650), Splink predicted 8.6% more TP (578,726), and Match*Pro predicted 17.0% fewer TP (442,338).

Table 2: Confusion Table If *Not* Using `ssn`

(a) `fastLink`

	True Matches	**True Non-Matches**	**Total**
Links	532,650 TP	109 FP	532,759
Non-Links	127,577 FN	9,878 TN	137,455
Total	660,227	9,987	670,214

(b) `Splink`

	True Matches	**True Non-Matches**	**Total**
Links	578,726 TP	3 FP	578,729
Non-Links	81,301 FN	9,984 TN	91,485
Total	660,227	9,987	670,214

(c) `Match*Pro`

	True Matches	**True Non-Matches**	**Total**
Links	442,338 TP	3 FP	442,341
Non-Links	217,889 FN	9,984 TN	227,873
Total	660,227	9,987	670,214

Overall, when compared with `fastLink`, `Splink` predicted 8.6%-9.8% more TP and `Match*Pro` predicted 17.0%-23.6% fewer TP. In summary, `Splink` predicted about 9% more TP than `fastLink` and `Match*Pro` predicted about 20% fewer TP. `Splink` predicted the least FP.

Recommendation

`Splink`: Recommended as the new FCDS default - Best in test

Based on the linkage performance on the `pseudopeople` test data, the recommendation is to change the FCDS default software for PRL from `fastLink` to `Splink`. The main limitation is that `Splink` requires beginner-level skills in Python (e.g., Harrison 2023). More advanced users would benefit from knowledge of Pandas (e.g., Harrison 2024), SQL, DuckDB, Altair, and Quarto. A comprehensive tutorial book on `Splink` 3.9.5 is “Hands-On Entity Resolution” (Shearer 2024).

`fastLink`: Recommended as a backup alternative - Best in practice

The FCDS has successfully used `fastLink` for linkage data requests since 2019. Unfortunately, `fastLink` performed about 9% worse than `Splink` on the test data. The other main limitation is that `fastLink` requires beginner-level skills in R. The recommendation is to use `fastLink` as a backup alternative until the FCDS has created a template for using `Splink`.

`Match*Pro`: Not recommended - Worst in test, and not transparent

The FCDS does not recommend `Match*Pro` for linkage data requests. `Match*Pro` performed about 20% worse than `fastLink` and about 30% worse than `Splink`. The `Match*Pro` test results will be more directly comparable and they might improve if `Match*Pro` provides a variable for match probability. A related issue is that `Match*Pro` is not transparent because users do not have the freedom to study how the program works. Access to the code is a precondition for this user freedom.

What's Next

A proposed topic of the next PRL project at the FCDS (for example, the 2025 monograph) is titled "The FCDS Record Linkage Template using `Splink`". The basic idea is to create a template of best practices for using `Splink` in FCDS linkage data requests. In order of priority, these are the specific recommendations:

- Update and test Splink 4. `Splink` 4 is expected in fall 2024. It will require less code, allow pre-processing without a linker, have simpler function imports, autocomplete configuration, and less backend-specific code. All new features will be in `Splink` 4.
- Improve best practices for using `Splink`. Consider changing the integrated development environment (IDE) from RStudio to Visual Studio Code (VS Code) or to Positron (in beta) because VS Code is the standard IDE for Python. Try collaboration using Live Share in VS Code. Improve step 4, canonicalization (post-processing), in `Splink`. An example is the clerical labeling tool in beta. How to best handle duplicates is debatable.
- Begin to "standardize" variables in the FCDS database for PRL. An example is to create variable `std_ssn` with only "valid" SSN. Standardized variables should be useful to the FCDS also outside of PRL.

The FCDS will remain vigilant in evaluating the relative performance of PRL software. Ideally, the FCDS should also try disruptive innovation. Two examples are Project Bluefin (Bluefin) and FastLink.jl. Bluefin is cloud-native computing on the Linux desktop. Development, such as the FCDS template, is done in Dev Containers. `FastLink.jl` is `fastLink` in the Julia language for faster PRL. Bluefin is generally available whereas `FastLink.jl` is not yet released.

To test `fastLink` again, the FCDS requires an update with the promised improved blocking, Active Learning or equivalent. To test `Match*Pro` again, thanks to the institutional support of NAACCR, the FCDS only requires a variable for match probability and the associated code. The FCDS hopes to create better simulated `pseudopeople` data, especially Florida-specific data, for any such additional testing. In the meanwhile, it is prudent for the FCDS to focus on the prioritized recommendations for "The FCDS Record Linkage Template using `Splink`".

References

Binette, Olivier, and Rebecca C. Steorts. 2022. "(Almost) All of Entity Resolution." *Science Advances* 8 (12): 1–14. https://doi.org/10.1126/sciadv.abi8021.

Brodeur, Abel et al. 2024. "Mass Reproducibility and Replicability: A New Hope." *Working Paper, I4R Discussion Paper Series* 107: 23. https://hdl.handle.net/10419/289437.

Contiero, Paolo, Andrea Tittarelli, G. Tagliabue, A. Maghini, Sabrina Fabiano, P. Crosignani, and R. Tessandori. 2005. "The Epilink Record Linkage Software: Presentation and Results of Linkage Test on Cancer Registry Files." *Methods of Information in Medicine* 44 (1): 66–71. https://doi.org/10.1055/s-0038-1633924.

Enamorado, Ted. 2019. "Active Learning for Probabilistic Record Linkage." *SSRN e-Print*, 1–37. http://dx.doi.org/10.2139/ssrn.3257638.

———. 2021. "A Primer on Probabilistic Record Linkage." In *Handbook of Computational Social Science, Volume 2*, edited by Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu, and Lars Lyberg, 95–107. Routledge. https://doi.org/10.4324/9781003025245-8.

Enamorado, Ted, Bejamin Fifield, and Kosuke Imai. 2019. "Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records." *American Political Science Review* 113 (2): 353–71. https://doi.org/10.1017/S0003055418000783.

Enamorado, Ted, and Rebecca C. Steorts. 2020. "Probabilistic Blocking and Distributed Bayesian Entity Resolution." In *Privacy in Statistical Databases*, edited by Domingo-Ferrer J. and Muralidhar K., 224–39. Cham, Switzerland: Springer Lecture Notes in Computer Science. Volume 12276. https://doi.org/10.1007/978-3-030-57521-2_16.

Haddock, Beatrix, Alix Pletcher, Nathaniel Blair-Stahn, Os Keyes, Matt Kappel, Steve Bachmeier, Syl Lutze, et al. 2024. "Simulated Data for Census-Scale Entity Resolution Research Without Privacy Restrictions: A Large-Scale Dataset Generated by Individual-Based Modeling [Version 1; Peer Review: Awaiting Peer Review]." *Gates Open Research* 8-36: 1–10. https://doi.org/10.12688/gatesopenres.15418.1.

Harrison, Matt. 2023. *Learning Python for Data: Fundamental Python Skills for Starting with Data*. Independently published. https://store.metasnake.com/python-for-data-digital-book.

———. 2024. *Effective Pandas 2: Opiniated Patterns for Data Manipulation*. Independently published. https://store.metasnake.com/effective-pandas-book.

Linacre, Robin, Sam Lindsay, Theodore Manassis, Zoe Slade, and Tom Hepworth. 2022. "Splink: Free Software for Probabilistic Record Linkage at Scale." *International Journal of Population Data Science* 7 (3): 23. https://doi.org/10.23889/ijpds.v7i3.1794.

Shearer, Michael. 2024. *Hands-On Entity Resolution: A Practical Guide to Data Matching with Python*. Sebastopol, CA: O'Reilly. https://www.oreilly.com/library/view/hands-on-entity-resolution/9781098148478/.