MATCH*PRO STATUS UPDATE

Year in Review

SEER*DMS/SEER Data Quality Meeting

NCI Shady Grove Campus
Rockville, MD

September 26th 2018

Information Management Services, Inc.

WHAT IS MATCH*PRO ?

Match*Pro is a Java-based application developed by Information Management Services, Inc. (IMS) to conduct probabilistic record linkages based on the Fellegi and Sunter model.

It provides a user-friendly tabbed interface for the configuration of linkages and the manual review of uncertain matches. These linkages can be configured to find matches between two separate files or to find duplicates within a single file.

Match*Pro can run on any fixed-width or delimited data file.

WHAT IS MATCH*PRO ?

The linkage configuration engine is flexible and allows users to filter incoming data, specify blocking and matching methods, define unknown values, adjust weights, identify cutoff points, and more.

The manual review screen makes use of traffic lighting to help users discern the match status of linked pairs.

Users have access to several tools during the manual review process including a Query Editor that can be used to find specific linked pairs amongst the larger set of pairs that qualified for the review, and a Report Generator which can be used to create frequency distributions from the linkage data.

CONFIGURING A LINKAGE

The linkage configuration screen in Match*Pro provides users with an interface to define parameters for a probabilistic record linkage or deduplication.

On the linkage configuration screen you can define:

- the file types and locations of the input files
- any filters that you would like to place on the incoming data
- the blocking criteria and matching parameters
- any filters that you would like to place on the outgoing data (linked pairs) in order to classify pairs as a match, a non-match, or uncertain.
- options that tweak various aspects of how the linkage is run

MANUAL REVIEW

Match*Pro's manual review screen provides users with a mechanism for viewing and adjudicating the linked pairs in a results set.

Users can assign the **match status** of each linked pair and/or place each linked pair into 1 of 8, potentially overlapping, color-based **categories**.

The columns in the table can be sorted and re-arranged. Additional fields (columns) can be added to the table.

Users can place a **filter** on the table in order to restrict the view to only those linked pairs that meet the filter criteria. Filters/Views can be added to a list of favorites and revisited later.

Data from the manual review screen can be used to generate **frequency reports**. Users can create 1-way, 2-way, or 3-way frequency tables that can reference fields from the linkage (e.g. match status) or fields from each of the input files.

BETA

Took place from August 2017 until May 2018

Over 30 registries participated

Feedback collected from users

User suggestions integrated into subsequent releases

New releases every 2-4 weeks

VERSION 1.0

Released June 2018

Website created to track users and distribute software

NAACCR 2018 Workshop

Linkage quality evaluated by NCI using synthetic data

F-measures ranged from 0.96 ~1.0 depending on the number of errors introduced into the data and the settings used.

Outperformed LinkPlus in all tests where both applications used comparable settings.

VERSION 1.1

Released July 2018

Minor update

Addressed a couple of bugs in version 1.0

VERSION 1.2

Released September 2018 - First major update

G-Zip support

Street Address, Postal Code, Telephone, Sorensen-Dice, and Overlap comparators added

Name swapping feature

Color/format of text and cells on the manual review screen can be changed by users

Query Builder updates (faster loading, new logic builder, copy/paste)

Quality of life improvements (larger fonts, UI adjustments)

Increases on the number of matching parameters, categories, and acceptable input patterns for dates.

VERSION 1.3 AND BEYOND

Version 1.3 slated for Q4 2018

New double review feature

New tools for tracking/maintaining lists of non-duplicates

NAACCR XML input

Version 1.4 slated for Q1 2019

Data validation/edits features (custom configuration, reporting, etc.)

GETTING MATCH*PRO

Match*Pro can be downloaded for free.

Go to **https://surveillance.cancer.gov/matchpro/download**

Complete the registration form.

Check your email for a link to download the software.