CDDA Record Matching Overview

CDDA Overview

The purpose of the Centralized Demographics Dataset Administrator (CDDA) is to facilitate and enhance data exchange across the agencies of the Illinois Longitudinal Data System (ILDS). These agencies include the following:

- Illinois Department of Commerce and Economic Opportunity (DCEO)
- Illinois Board of Higher Education (IBHE)
- Illinois Community College Board (ICCB)
- Illinois Department of Employment Security (IDES)
- Illinois Department of Human Services (IDHS)
- Illinois Student Assistance Commission (ISAC)
- Illinois State Board Education (ISBE)

Additionally, for some agencies multiple sources of data are provided to the CDDA. Each data source is treated as a separate entity for record matching and matching analysis purposes. Current examples of separate data sources that submit data to the CDDA are as follows:

- Illinois Network of Child Care Resources and Referral Agencies (INCCRRA)
- Parents Too Soon (PTS)
- Early Intervention (EI)
- Healthy Families Illinois (HFI)
- Child Care Assistance Program (CCAP)
- Maternal, Infant, Early Childhood Home Visiting (MIECHV)

In order to facilitate and enhance data exchange across agencies, and/or between different data sources, the CDDA receives records with personal identifiable information (PII) from the agencies per data source, links records across the various files, and produces the Master Client Index (MCI). The MCI contains the PII for each record received from each agency, along with the additional field that is created as part of the matching process that links records together. This record linkage field contains the CDDA-ID. These CDDA-IDs, or IDs for short, are provided back to the agencies for each record in the MCI and allows agencies to match records across agency systems and between agencies.

DATA STANDARDIZATION

Prior to record linkage, all records received from the agencies are standardized. Through data standardization the majority of differences in data fields are addressed. For example, one agency may allow an apostrophe to be used in a Last Name, such as O'Hare, while for another agency an apostrophe is not allowed, but a space is, and for a third agency neither special character is allowable. This can result in the same Last Name appearing as O'Hare, O Hare, and OHare. Data standardization also accounts for instances where a name may appear in different formats, as the previous example, within a single agency's records. As part of data standardization special characters are removed from the data fields, non-English letters/formats are removed (such as accent marks), names are all made upper case, gender is recoded as M (male) and F (female), race/ethnicity is standardized, DOB is formatted the same across all records, and SSN is also formatted the same across data files/records.

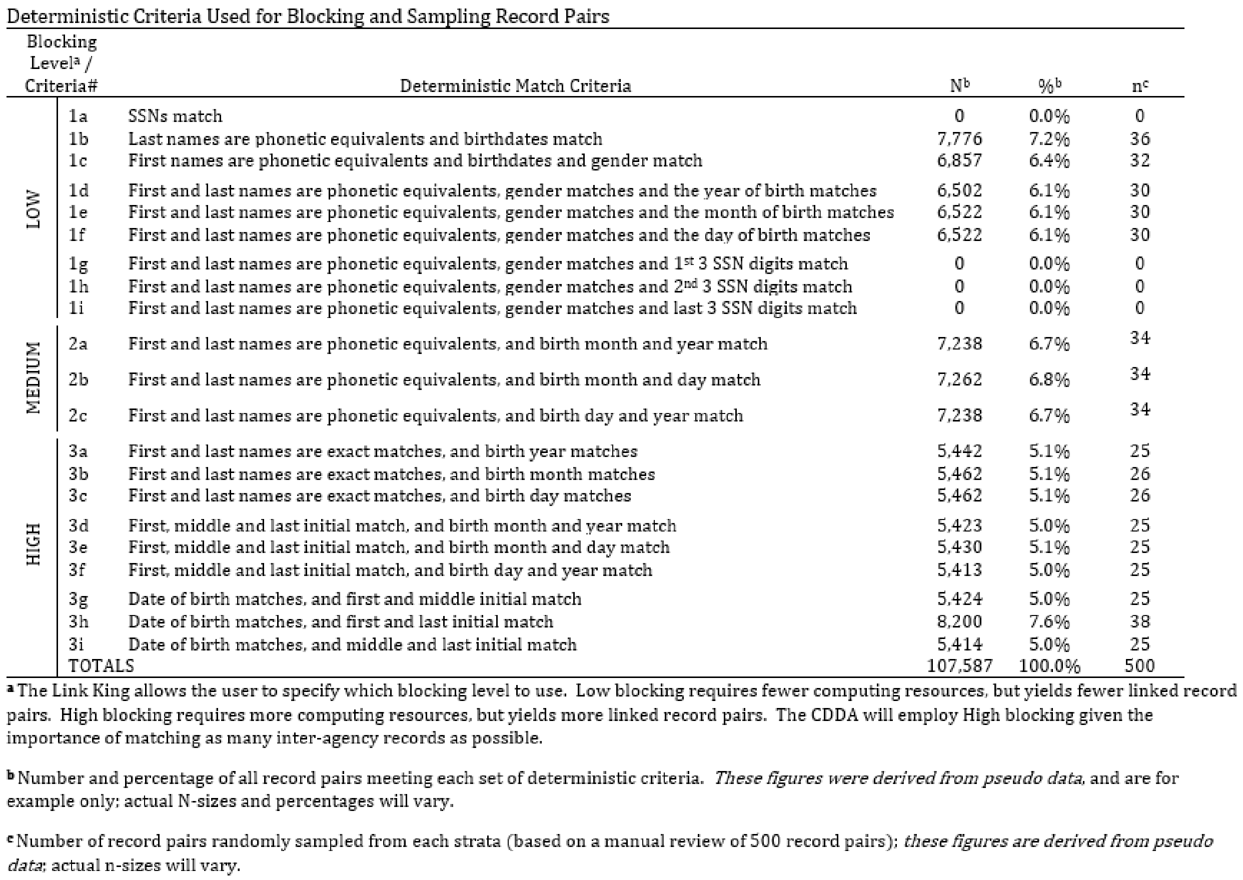
RECORD LINKAGE

The record linkage process, or record matching, uses two basic algorithm types. These algorithm types are Deterministic and Probabilistic. Deterministic is defined by the specific criteria for matching/record linkage and allows for fuzzy matches. For example, matching based on phonetics. Probabilistic is defined by the use of statistical analysis to compare data similarities (pairing of records based on these similarities), along with the use of a derived formula and scores to create cut-points to identify matches (where a match is definite, possible, or a non-match/unmatch).

The Link King

Matching is done with the use of the Link King, a freely available software, and SAS. The Link King uses both Deterministic and Probabilistic algorithms. First name, middle name, last name, date of birth (DOB), Social Security Number (SSN), gender, and race/ethnicity are all data elements used as part of the Link King algorithms. This software also allows for use of maiden name and a flex variable, such as a student ID or unique ID. The Link King is used as a first step in the record matching process.

The Link King allows for weights to be added to each of the variables that are used in the algorithms to link records. Additionally, the Link King allows for thresholds to be set on labeling record linkages as definite matches, possible matches, and non-match/unmatch. The thresholds that are set are high, meaning that through the Link King, generated record linkages are accepted only where there are definite matches or definite non-matches. These matches are based on Deterministic Match Criteria of a High Blocking Level. Probabilistic Weight is also set high for appropriate variables. Additional documentation for this is available via the Link King, but the Blocking Levels are provided below for reference



Deterministic Criteria Used for Blocking and Sampling Record Pairs

Blocking Level[a] / Criteria#		Deterministic Match Criteria	N[b]	%[b]	n[c]
LOW	1a	SSNs match	0	0.0%	0
	1b	Last names are phonetic equivalents and birthdates match	7,776	7.2%	36
	1c	First names are phonetic equivalents and birthdates and gender match	6,857	6.4%	32
	1d	First and last names are phonetic equivalents, gender matches and the year of birth matches	6,502	6.1%	30
	1e	First and last names are phonetic equivalents, gender matches and the month of birth matches	6,522	6.1%	30
	1f	First and last names are phonetic equivalents, gender matches and the day of birth matches	6,522	6.1%	30
	1g	First and last names are phonetic equivalents, gender matches and 1st 3 SSN digits match	0	0.0%	0
	1h	First and last names are phonetic equivalents, gender matches and 2nd 3 SSN digits match	0	0.0%	0
	1i	First and last names are phonetic equivalents, gender matches and last 3 SSN digits match	0	0.0%	0
MEDIUM	2a	First and last names are phonetic equivalents, and birth month and year match	7,238	6.7%	34
	2b	First and last names are phonetic equivalents, and birth month and day match	7,262	6.8%	34
	2c	First and last names are phonetic equivalents, and birth day and year match	7,238	6.7%	34
HIGH	3a	First and last names are exact matches, and birth year matches	5,442	5.1%	25
	3b	First and last names are exact matches, and birth month matches	5,462	5.1%	26
	3c	First and last names are exact matches, and birth day matches	5,462	5.1%	26
	3d	First, middle and last initial match, and birth month and year match	5,423	5.0%	25
	3e	First, middle and last initial match, and birth month and day match	5,430	5.1%	25
	3f	First, middle and last initial match, and birth day and year match	5,413	5.0%	25
	3g	Date of birth matches, and first and middle initial match	5,424	5.0%	25
	3h	Date of birth matches, and first and last initial match	8,200	7.6%	38
	3i	Date of birth matches, and middle and last initial match	5,414	5.0%	25
	TOTALS		107,587	100.0%	500

[a] The Link King allows the user to specify which blocking level to use. Low blocking requires fewer computing resources, but yields fewer linked record pairs. High blocking requires more computing resources, but yields more linked record pairs. The CDDA will employ High blocking given the importance of matching as many inter-agency records as possible.

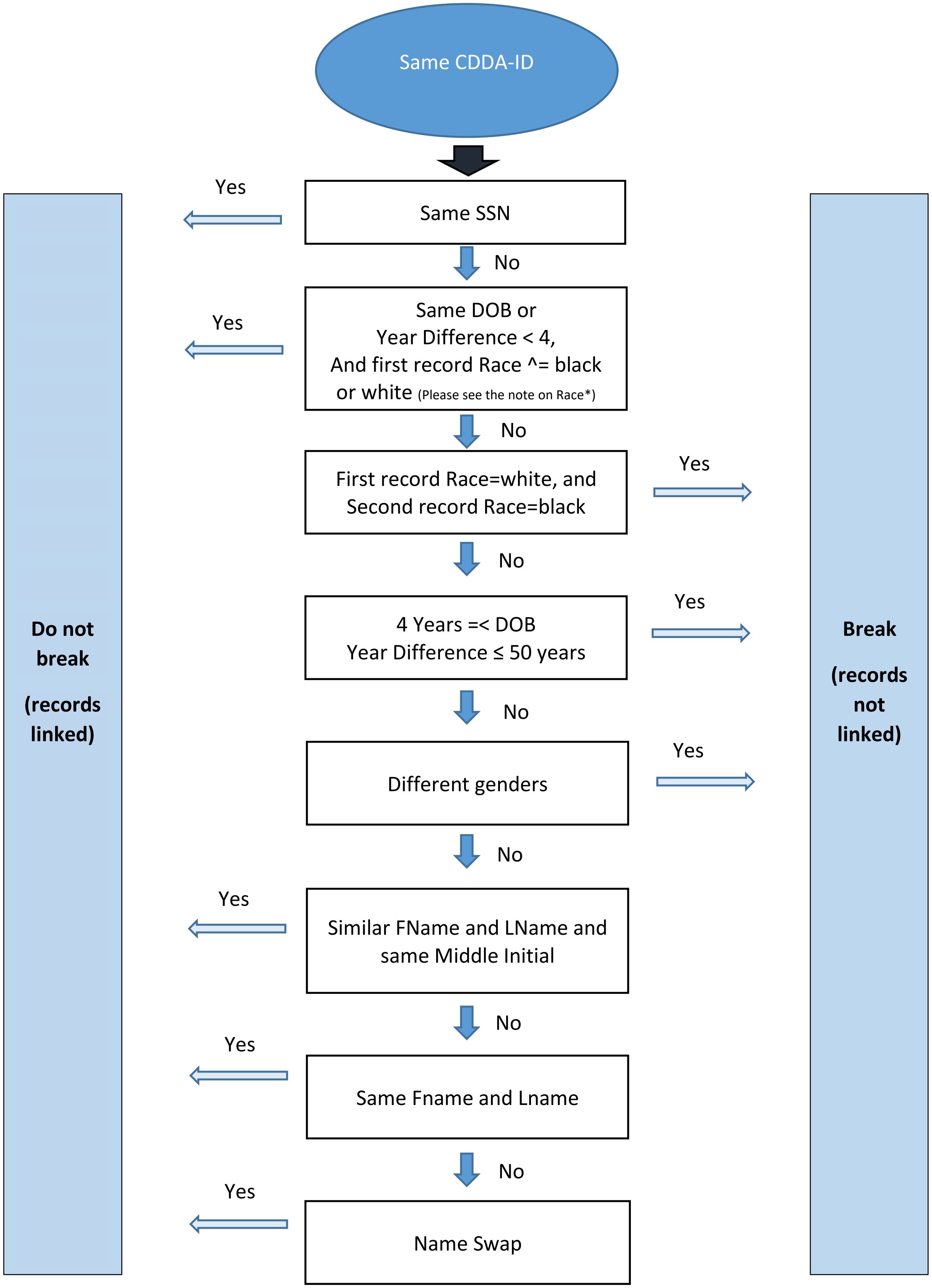
[b] Number and percentage of all record pairs meeting each set of deterministic criteria. *These figures were derived from pseudo data*, and are for example only; actual N-sizes and percentages will vary.

[c] Number of record pairs randomly sampled from each strata (based on a manual review of 500 record pairs); *these figures are derived from pseudo data*; actual n-sizes will vary.

SAS

The remaining record linkages, those that are neither definite matches or definite non-matches, go through a second set of algorithms that are coded through SAS by the CDDA. In this step of the record matching process each agency/data source may have variations in the algorithm used in determining record linkages. The reason for the algorithm variations is to accommodate for the differences in the data provided to the CDDA, but also based on whether records are considered probable matches or non-matches based on the Link King results. For example, ISBE does not use a SSN, but instead provides a student ID. ICCB provides SSN. The algorithms for ISBE and ICCB cannot be the same since one has a SSN variable and one does not. The basic algorithm used is provided on the following pages.

For the below chart, in cases where SSN is considered different, the first four and last four digits are different. With regards to similar FName and similar LName, if the first three letters of the FName or LName match, the names are considered the same. While the below example indicates a starting point of Same CDDA-ID, which is to cut down on confusion, but the process is similar when the CDDA-IDs are different for a pair of records. ***Please be aware of the following regarding race.** In cases where there is a different race for each record, such difference is only used as a deciding factor when one record indicates African American/Black and one indicates White. Differences in race, other than the aforementioned, are not used in the decision process.


Same CDDA-ID
Same SSN
Yes
No
Same DOB or
Year Difference < 4,
And first record Race ^= black
or white (Please see the note on Race*)
Yes
No
First record Race=white, and
Second record Race=black
Yes
No
4 Years =< DOB
Year Difference ≤ 50 years
Yes
No
Different genders
Yes
No
Similar FName and LName and
same Middle Initial
Yes
No
Same Fname and Lname
Yes
No
Name Swap
Yes
Do not
break
(records
linked)
Break
(records
not
linked)