Linked Information Network of Colorado (LINC)
Project #19-01 Data Matching Report

Overview

This project involves matching personal identifiers from two different sets of administrative records. Approved data elements from each of the two systems were then added to the extract once matching was completed. Finally, all personal identifiers were stripped from the extract or anonymized to prevent reidentification of the individual. The data provided by the two systems were as follows:

Data source received	Organization	Data date ranges	Other data restrictions
Homelessness Management Information System (HMIS) Extract	Metro Denver Homeless Initiative (MDHI)	All MDHI enrollments where age > 18 as of 12/1/2018.	Only those records with a signed release of information (ROI) held by MDHI. This restriction reduced the study population by about 50%.
Trails Child Welfare Extract	Colorado Dept. of Human Services, Office of Children Youth and Families	All Trails clients from 2000 through 2019.	Used LINC Trails extract rather than a direct pull from Trails.

Identifier Field for Matching	Trails	Trails Notes	HMIS	HMIS Notes
First Name	Yes	0.1% had blank, 1 initial, or 2 letter/initial first names.	Yes	
Middle Name	Yes	42.5% were blank	Yes	40% were blank.
Last Name	Yes		Yes	
Generational Suffix	No		Yes	
Date of Birth	Yes		Yes	
SSN	Yes	32% were blank	Yes	Partial/no data for 15% of records
Gender	Yes	99.99% had an entry of male or female	Yes	98.2% had an entry of male or female

Data Deduplication and Matching Methodology

The identity resolution process began with removing all duplicates from the smaller of the datasets (HMIS). Deduplicating one of the datasets avoided a many-to-many set of relationships between the two systems. The HMIS dataset had only a few thousand records compared to Trails which had hundreds of thousands of records. Choosing to deduplicate the HMIS data first limited the time investment required to resolve each of the potential duplicates. Closely scrutinizing the HMIS data was also sensible because the loss or erroneous inclusion of a single HMIS record represented a far greater proportion of the HMIS data compared to a single record in the Trails data.

Deduplicating the 3,674 HMIS records was initiated by running the data through the Senzing commercial software package. Because multiple records for an individual might exist with slight variations (e.g., use of nickname instead of first name) or errors, the goal of the deduplication process was to recognize these slight differences so that the records could be consolidated into a set of unique individuals. Senzing uses a pre-trained analytical model that already understands how to identify these slight variations and how much weight to give to a similar first name, last name, date of birth, SSN, etc.

After generating a list of potential duplicates with Senzing, the two records in a pair of potential duplicates were compared. For example, if the name was unusual and matched exactly or nearly so, and the date of birth was identical, then it was decided that the pair was the same person. If, however, the name and date of birth were a complete mismatch but the SSN was an identical match, it was decided that one of the people had an incorrect SSN entry and that the two are actually unique individuals even though their SSNs are an identical match. This human quality check provided balance to the use of the pre-trained model.

If a pair was determined to be a duplicate pair, a final assessment was made to determine which of the two records were kept as the master record. To make that determination, each item of a pair was examined for whether it would then be the best candidate in the second round of matching when the HMIS records would be compared to the Trails dataset. For example, a more complete HMIS record would be a better matching candidate than an incomplete record. All substantive data from duplicate records were reassigned to the master record and thus all system involvement of the individual was preserved. Of the 3,674 HMIS records, 58 records (about 1.5%) were determined to be duplicates resulting in 3,616 HMIS records to be matched to Trails.

With a unique set of identifiers for the HMIS dataset, the next step was to run the identity resolution process with Trails with deduplication of the Trails dataset performed after the matching process was completed. Because of known data entry errors in Trails, it might be possible for one HMIS record to match only one of two duplicate Trails records. If deduplication was conducted first, one might not know which record to retain for purposes of matching to the other dataset and a non-match result might occur. By deduplicating the Trails data after matching, the match rate could be maximized.

Identity Resolution

For matching purposes, the Trails extract was restricted to those individuals who had system involvement as a minor and a date of birth within the HMIS date of birth range. This created a Trails identity resolution dataset of 832,929 individuals to be matched to the unique set of 3,616 HMIS individuals. The reasoning in using a large set of identifiers was based on the insight that linking a record to the other system might allow one to correct a data error that would otherwise exclude the record. For example, matching records where the date of birth did not agree between the two systems could be resolved to determine which date would be used. By applying the study criteria to a corrected date of birth, the record might be included in the study. While not a major factor, this approach resulted in the identification of several matches that would have otherwise been unmatched. By using a large set of identifiers, resolving match differences, and then applying the study criteria maximized the match rate.

Senzing creates three categories of matches with a diminishing likelihood of being a valid match: Duplicates, Possible Duplicates, and Possibly Related. After inspection and quality review, the following accuracies were calculated.

Category	Matches	Possible Matches	Possibly Related	Non-Match
Senzing Output	1,600 (44%)	330 (9%)	6 (.16%)	1,681 (46.4%)
Actual Matches	1,596 Accuracy: 99.75%	330 Accuracy: 100%	6 Accuracy: 100%	29 (Found manually)
Actual Non-Matches	4(0.25%)	0 (0%)	0 (0%)	1,652 Accuracy: 98%

This table shows that the Senzing software minimized false positives (i.e., avoiding a match if it might not be a true match). Senzing also avoided false negatives (i.e., missing a match), and only twenty-nine such matches (2%) were found. The method for identifying the matches missed by Senzing was to import the HMIS non-matches into Trails for sampling and examination as a quality check. Numerous SQL queries were run using partial data fields (e.g., only the first two letters of the person's first name, only the month/day of birth, only the last six digits of the SSN, etc.) By combining partial data fields in numerous ways, twenty-nine additional matches were identified. With only few additional matches discovered from the non-match group, there was a high level of confidence in the quality of the Senzing algorithmic matching process. These additional matches were included in the final match list to increase the final match rate.

After matching was completed, the Trails individuals who met the terms of the study or matched to HMIS individuals were exported for deduplication. The deduplication results for the Trails individuals (201,988) are displayed in the table below. Duplicate candidate were examined by hand, but flagged as duplicates only when strong evidence (e.g., a unique name) suggested a duplicate. Favoring uniqueness in the duplicate candidate pool was reasonable for several reasons:

1) the large Trails dataset resulted in numerous high-frequency names (e.g., John Brown) born on the same day due to random chance, making an accurate decision on these common names difficult;
2) the large Trails dataset with fewer than 1% duplicate candidates meant that an error of including a duplicate as a unique record would not affect the inferences drawn about the Trails data; and
3) the time required to scrutinize a large set of duplicate records would have been substantial.

The results of deduplicating and matching the two datasets is further delineated in the two tables and venn diagram below:

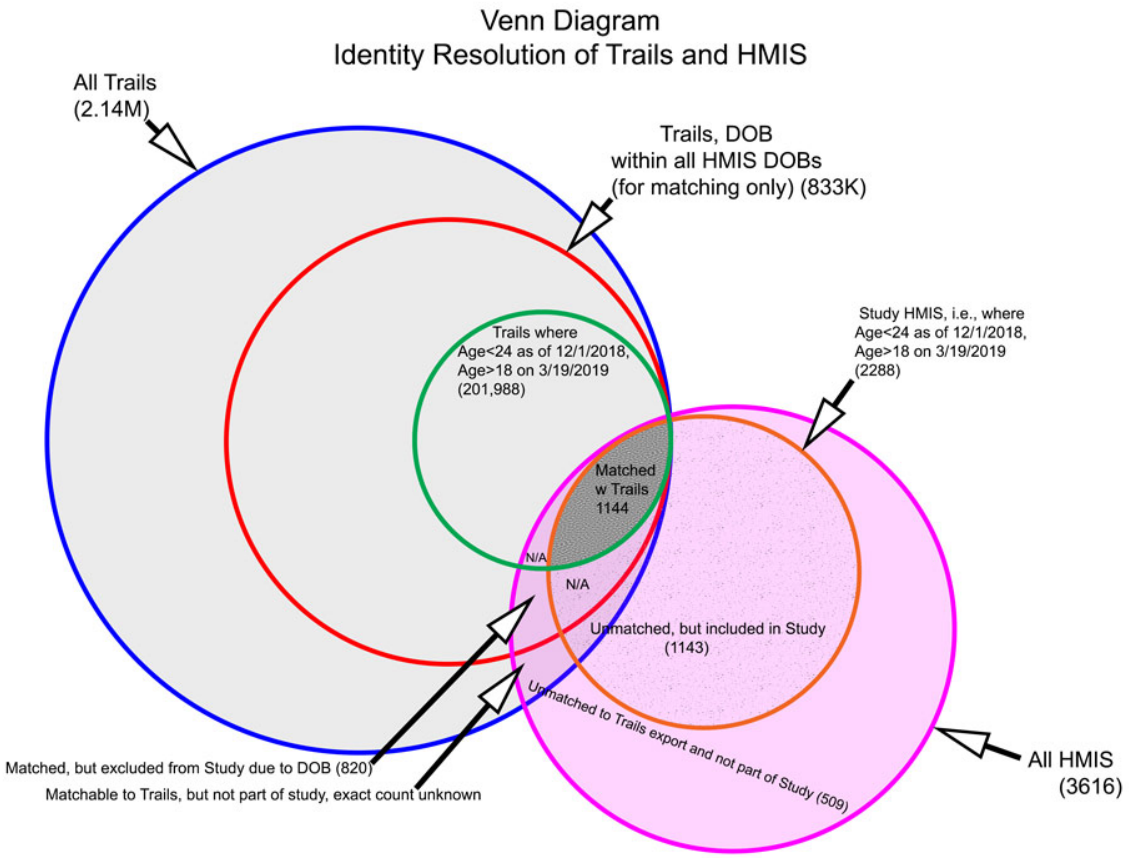
HMIS Data:

	Record count	Study inclusion/exclusion	Record count	
Full HMIS Data	3,674			
Duplicates	58 (1.6%)			
Unique Individuals	3,616(98.4%)			
Included in Study*		2,287 (63%)		
			1,128 (49%)	Matched by Senzing
			16 (1%)	Matched Manually
			1,143 (50%)	Not Matched
Excluded from Study*		1,329(37%)		
			808 (61%)	Matched by Senzing
			12 (1%)	Matched Manually
			509 (38%)	Not Matched

Trails Data:

	Record count	Study inclusion/exclusion	Record count	
Full Trails Data	201,988			
Duplicates	408 (.2%)			
Unique Individuals	201,580 (99.8%)			
Included in Study*		200,767 (99.6%)		
			1144 (.6%)	Matched
			199,623 (99.4%)	Not Matched
Excluded from Study*		813 (.4%)		

*Study criteria was Age>=18 as of 12/1/2018 and Age<=24 as of 3/19/2019 with additional criteria for Trails as having system involvement as a minor. This criteria resulted in a reduction of the Trails data from 2.14M to 201,988. The 813 "Excluded" Trails individuals are those individuals is the count of those Trails individuals matched to HMIS individuals who did not meet the date of birth criteria.



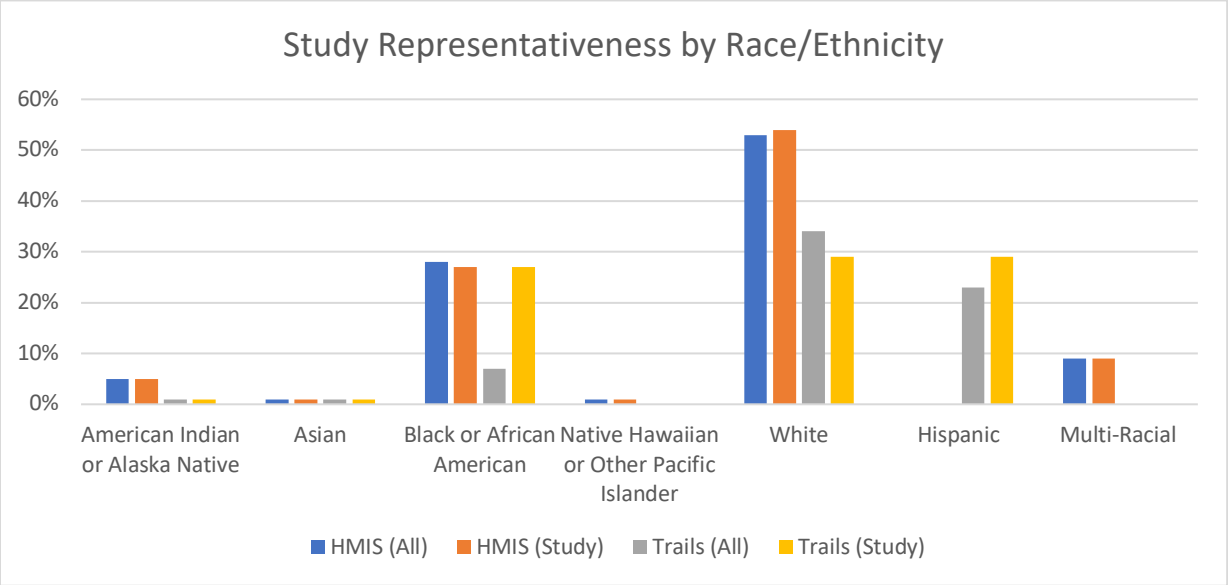
Quality Control

At every stage of the matching process, results were sampled and examined to avoid mismatches. The matches that were included in the study were subjected to a final set of scrutiny. Of the 1,148 matches, all but 107 could be approved via simple rules, with the remaining 107 of 1,148 pairs of records set aside for inspection on agreement of the match. Four pairs were deemed errors and removed, with 104

deemed correct with the match left intact. For example, a pair of records involving a person identifying as transgender with two separate records had initially been matched, but the issue was identified and resolved. A final quality check was run to detect possible bias in the matching. The results are summarized in the table below:

Race	HMIS (All)	%	HMIS (Study) Only	%	Trails (All)	%	Trails (Study) Only	%
American Indian or Alaska Native	164	5%	104	5%	1,146	1%	12	1%
Asian	25	1%	18	1%	2093	1%	8	1%
Black or African American	1,002	28%	624	27%	13,959	7%	314	27%
Native Hawaiian or Other Pacific Islander	35	1%	23	1%	416	0%	2	0%
White	1,918	53%	1,239	54%	68,087	34%	333	29%
Hispanic	-	-	-	-	45,484	23%	328	29%
Multi-Racial	330	9%	197	9%	-	-	-	0%
Client doesn't know	61	2%	26	1%	-	-	-	-
Client Refused	61	2%	39	2%	-	-	-	-
Data Not Collected	21	1%	18	1%	70,803	35%	148	13%
Grand Total	3,617	100%	2,288	100%	201,988	100%	1,145	100%

The below graph displays the information for those selecting a race/ethnicity category. The race/ethnicity percentages for those included in the study were very similar to the overall unique set of individuals from each data source, with only two discrepancies. Trails had 7% African Americans overall, but the matched population is 27%, which was similar to the HMIS demographic. The White category is much lower in Trails, but that is partially because Trails included Hispanic as a racial category. Overall, there is no reason to conclude that there is inherent bias in over-matching or under-matching any particular racial category.



Final Steps

With the identity resolution process completed and the population of the study established, the substantive data elements from the database could be extracted and the linkage identifiers be anonymized. All personal identifiers were deleted with the exception of date of birth which was anonymized by retaining the month and year of birth, but setting the day of birth to the first day of the month for all individuals.

Anonymized Data Set Dictionary

TRAILS CLIENTS	EXAMPLE DATA	NOTES
TrailsID	TRL000001	For Linking Records (masked)
LINCID	LINC0001	For Linking Records (masked)
DOB	6/1/1990	Anonymized
Gender	Female	
Hispanic	Yes	New Field
American_Indian	No	
Asian	No	
Hawaiian	No	
African_American	No	
Caucasian	Yes	
Primary_Ethnicity	Hispanic	New Field
Federal_Ethnicity	Hispanic	New Field
Mother_AgeBirth	40	
Mother_PrimaryEthnicity	Caucasian	New Field
Father_AgeBirth	45	
Father_PrimaryEthnicity	Caucasian	New Field
Has_CW_Involvement_as_Child	Yes	
Has_DYC_Involvement_as_Youth	Yes	New Field
Number_Referrals_Child	8	
Number_Assessments_Child	3	
Number_Founded_Assessments_Child	1	
Number_Cases_CW_Child	3	
Number_Cases_DYC_Youth	8	
Number_Placements_CW	5	
Number_Removals	4	
Date_First_Removal	5/8/2003	New Field
Age_First_Removal	13	
Days_OOH_CW	755	
Days_OOH_DYC	43	
Total_Days_OOH	798	
Has_DYC_Involvement	Yes	
Has_DYC_Detention	Yes	

Has_DYC_Commitment	No	
Adopted_from_Care	No	Replaces Ever_Adopted
Emancipated_from_Care_CW	No	
Date_Emancipated_from_Care_CW		New Field
Number_Children	5	Not originally requested by included because Has_Children was not provided in extract
TRAILS INVOLVEMENTS	EXAMPLE DATA	NOTES
TrailsID	TRL000004	For Linking Records (masked)
LINCID		For Linking Records (masked)
Involvement_ID	2546467	Note: Anonymized identifier. This table began as a one Involvement to many Involvement_Clients. It was linked and the data provided has a separate record for each Client.
Involvement_Type	Case - Child Welfare	
Involvement_Outcome	Parents - Return Home	
Involvement_Role	Child/Youth	Previously in Involement_Clients
Age_at_Open	1	Previously in Involement_Clients
Involvement_Program_Area	PA5 - Child Protection	Previously in Involement_Clients
Founded_Victim		New Field
Days_Placed_Involvement	315	New Field
Days_Placed_Involvement_CW	315	New Field
Number_Placements_Involvement	1	New Field
Number_Placements_Involvement_CW	1	New Field
Date_Child_Placed	1/14/1998	New Field
Open_Date	1/14/1998	
TRAILS INVOLVEMENT_CLIENTS	EXAMPLE DATA	NOTES
TrailsID	TRL048328	For Linking Records (masked)
LINCID	LINC000001	For Linking Records (masked)
Involvement_ID	247314	For Linking Records (masked)
Involvement_Type	Referral	
Open_Date	4/26/2005	
Close_Date	4/27/2005	Previously in Involvements
Open_County	Baca	Previously in Involvements
Closure_Outcome	Accepted	Previously in Involvements
Involvement_Type_Detail	PA5 - Child Protection	Previously in Involvements
Involvement_Findings	Inconclusive	Previously in Involvements
Case_ID		For Linking Records (masked)
Referral_ID		For Linking Records (masked)

Has_Risk_Assessment	N/A	Previously in Involvements
Risk_Assessment_Tool	N/A	New Field
Has_Safety_Assessment	N/A	Previously in Involvements
Issue_Neglect	Yes	New Field
Issue_Abuse	No	New Field
Issue_Physical_Abuse	No	Previously in Involvements
Issue_Substance_Abuse	Yes	Previously in Involvements
Issue_Homelessness	No	Previously in Involvements
Issue_Lack_Supervision	No	Previously in Involvements
Issue_Domestic_Violence	No	Previously in Involvements
Issue_Sexual_Abuse	No	Previously in Involvements
Issue_Educational_Neglect_Truant	No	New Field
Issue_Medical_Neglect	No	New Field
Issue_PA4_BCOP	No	New Field
Issue_PA4_Placement_Eval	No	New Field
Issue_Child_Disability		New Field
Issue_Substance_Exposed_Newborn	No	New Field
Risk_Level		Previously in Involvements
Risk_Abuse_Score		Previously in Involvements
Children_Placed_Open	No	Previously in Involvements
Children_Placed_During	N/A	Previously in Involvements
Date_Children_Placed		Previously in Involvements
TRAILS PLACEMENTS	EXAMPLE DATA	NOTES
TrailsID	TRL000026	For Linking Records (masked)
LINCID	LINC000001	For Linking Records (masked)
Case_ID	1530717	For Linking Records (masked)
RMVL_ID	1517908	For Linking Records (masked)
Start_Date	10/12/2002	
End_Date	5/29/2003	
Placement_Type	Residential	
Days_Placed	229	
Leave_Reason	Family Preservation Success	
County_Agency	Baca	
Paid_Placement	Yes	
Subsequent_Placements_Removal	1	
Subsequent_Placements_Total	1	
Discharge_Setting		New Field

TRAILS REMOVALS	EXAMPLE DATA	NOTES
TrailsID	TRL000026	For Linking Records (masked)
LINCID	LINC000001	For Linking Records (masked)
RMVL_ID	1517908	For Linking Records (masked)
Case_ID	1530717	For Linking Records (masked)
Removal_Begin_Date	10/11/2002	
Removal_End_Date	3/15/2004	
Removal_End_Reason	Reunification with Parents	
Days_Removed_Episode	521	
Age_at_Removal	8	
Number_Placements_During_Episode	2	
Removal_Family_Structure	Married Couple	Column renamed
Removal_Manner	EMERGENCY	
Removal_Substance_Parent	No	Column renamed
Removal_Neglect	Yes	Column renamed
Removal_Child_Behavior	Yes	Column renamed
Removal_Substance_Child	No	Column renamed
Removal_Housing	No	Column renamed
Removal_Cope	No	New Field
Removal_Physical_Abuse	No	Column renamed
Removal_Sexual_Abuse	No	Column renamed
Removal_Parent_Incarceration	No	New Field
Removal_Parent_Death	No	Column renamed
Removal_Child_Disability	No	Column renamed
Removal_Abandonment	No	New Field
Removal_Relinquish	No	New Field
Reentry_1year		
Caregiver1_Age		New field provided due as substitute for missing field in Involvements
Caregiver1_Relationship		New field provided due as substitute for missing field in Involvements
Caregiver2_Age		New field provided due as substitute for missing field in Involvements
Caregiver2_Relationship		New field provided due as substitute for missing field in Involvements

TRAILS SERVICES	EXAMPLE DATA	NOTES
TrailsID	TRL011547	For Linking Records (masked)
LINCID	LINC0809	For Linking Records (masked)
Involvement_ID	2769004	For Linking Records (masked)
Service_Type	Family Group Decision Making	
Service_Category	Core Services	
Days_Open	1197	
Core_Goal	REMAIN HOME	
Core_Outcome	Partially Successful	
OtherService	0	New Field
MISSING TRAILS FIELDS	TABLE	NOTES
Combined_Race	Clients	May use individual racial flags as filters
Has_Children	Clients	Not provided in the Trails extract, but Number_Children has been included and using Number_Children>0 is equivalent
Close_Date	Involvements	Moved to Involvement_Clients
Open_County	Involvements	Moved to Involvement_Clients
Involvement_Type_Detail	Involvements	Moved to Involvement_Clients
Involvement_Type_Findings	Involvements	Moved to Involvement_Clients
Has_Risk_Assessment	Involvements	Moved to Involvement_Clients
Has_Safety_Assessment	Involvements	Moved to Involvement_Clients
Physical_Abuse	Involvements	Moved to Involvement_Clients
Substance_Abuse	Involvements	Moved to Involvement_Clients
Mental Health	Involvements	Not provided in the Trails extract
Lack_Supervision	Involvements	Moved to Involvement_Clients
Domestic_Violence	Involvements	Moved to Involvement_Clients
Homelessness	Involvements	Moved to Involvement_Clients
Sexual_Abuse	Involvements	Moved to Involvement_Clients
Basic_Needs_Issues	Involvements	Not provided in the Trails extract
PCG_Age	Involvements	Not provided in the Involvements table of the Trails extract, but Caregiver1_Age provided in the Removals table and included among that data
PCG_Relation	Involvements	Not provided in the Involvements table of the Trails extract, but Caregiver1_Relationship provided in the Removals table and included among that data

Has_SCG	Involvements	Not provided in the Involvements table of the Trails extract, but Caregiver2_Age and Caregiver2_Relationship provided in the Removals table and included among that data
Reason	Involvements	Not provided in the Trails extract
Involvement_Role	Involvement_Clients	Moved to Involvements table
Age_at_Open	Involvement_Clients	Moved to Involvements table
Involvement_Program_Area	Involvement_Clients	Moved to Involvements table
Was_Child_Removed	Involvement_Clients	Not provided in the Trails extract - see Removals Fields
Date_Child_Removed	Involvement_Clients	Not provided in the Trails extract - see Removals Fields
Total_Days_Removed_Involvement	Involvement_Clients	Not provided in the Trails extract - see Removals Fields
Number_Placements_Involvement	Involvement_Clients	Not provided in the Trails extract - see Removals Fields
Placement_Level	Placements	Not provided in the Trails extract
Days_Removed_Total	Removals	Not provided in the Trails extract, but see Days_OOH_CW in Clients table
Age_at_First_Removal	Removals	Not provided in the Trails extract, but equivalent field in Clients table
Days_Foster_Episode	Removals	Not provided in the Trails extract, but see Days_Removed_Episode in Removals table
Days_congregate_episode	Removals	Not provided in the Trails extract, but see Days_Removed_Episode in Removals table
HMIS CLIENTS	EXAMPLE DATA	NOTES
LINCID	LINC0001	For Linking Records (masked)
MDHIID	MDHI0728	For Linking Records (masked)
UniqueMDHIID	MDHI0728	For Linking Records (masked)
ClientsDateofBirth	1/1/1993	
ClientsCurrentAge	27	
ClientsEthnicity	Hispanic/Latino	
ClientsRace	White	
MultiRacial-AmericanIndianorAlaskaNative	No	
MultiRacial-Asian	No	
MultiRacial-BlackorAfricanAmerican	No	

MultiRacial-NativeHawaiianorOtherPacificIslander	No	
MultiRacial-White	Yes	
ClientsGender	Male	
ClientProfileCreatedDate	9/20/2012	
ClientProfileLastUpdatedDate	6/21/2019	
TrailsMatch	0 or 1	
ServiceLookup	311	
Date	4/12/2018	
Notes		
HMIS ENROLLMENTS	EXAMPLE DATA	NOTES
MDHIID	MDHI0451	For Linking Records
LINCID	LINC0001	For Linking Records
Agency Name	St. Francis Center	
Project Name	SFC_Day Shelter_DS	
Project Type	Day Shelter	
Age at Project Start	22	
Enrollment ID	154101	For Linking Records (masked)
Household ID	154109	For Linking Records (masked)
Individual or Family Enrollment	Individual	
Project Start Date	6/14/2013	
Foster Care: Former Ward		
Foster Care: Number of Months		
Foster Care: Number of Years		
Housing Move-in Date		
Project Exit Date	6/14/2015	
Destination at Exit		
HMIS SERVICES	EXAMPLE DATA	NOTES
MDHIID	MDHI0277	For Linking Records (masked)
LINCID	LINC0001	For Linking Records (masked)
Enrollment ID	452953	For Linking Records (masked)
Household ID	452361	For Linking Records (masked)
Service Type	Long Term	
Service Date	8/13/2013	
Service Name	Outreach	
Service Item Name	Basic Needs	
Service Expense Amount		

Date identifiers will be deleted from LINC environment: October 15, 2020

Data anonymized file will be deleted from LINC environment: December 31, 2020

Issues encountered in data cleaning/matching (information for data providers):
Overall, the data in both systems were of high quality and suitable for matching. The owners of both systems might consider reconciling what appear to be duplicate records so that future extracts have fewer duplicates (deduplication results can be shared by LINC). The differences in names and dates of birth between the two systems suggested that one or both of the systems have some data validation and/or data entry issues. Errors of data validation and mis-entry may be difficult to detect; nevertheless, these kinds of data integrity issues are worth examination for process improvement.

One small change to the HMIS extract is recommended for future extracts. In the current extract, the UniqueClientIdentifier is generally alphanumeric, but sometimes is comprised of just digits, with a few cases of digits and a single letter. Importing this field into Excel is not usually a problem unless the sole letter is an "E" which Excel converts to exponential notation, thus ruining the identifier(e.g., 83456E30 gets converted by Excel to 8.3456E34). If this field were prepended with a letter(s) in addition to whatever has been created, Excel would always consider it to be a text entry and the conversion problem could be avoided.

Prepared by:
John Hokkanen, Senior Data Scientist
Email: John.hokkanen@state.co.us
Phone: 303-641-3031