ANALYSIS OF CHARM AND DOHMPI MATCHING

19TH ANNUAL EARLY HEARING DETECTION & INTERVENTION MEETING

KANSAS CITY, MO
MARCH 9, 2020

INTRODUCTION

- In public health care, federated or integrated data systems can provide
  - Authorized users with access to data from multiple sources
  - Timely alerts about unusual situations
  - Opportunities to validate data or to fill-in missing pieces of information
- The success of a federated or integrated data system involving person-centric data depends on accurately matching person records across its data sources
- In Utah, EHDI is a data source for two different federated data systems:
  - Child-health Advanced Record Management (CHARM)
  - Department of Health Master Patient Index (DOPMPI)

THE MATCHING QUALITY PROBLEM

- As with most federated or integrated data systems with person-centric data, CHARM and DOHMPI both include software components for matching person records
- Problem: How can the CHARM or DOHMPI teams measure the quality of their matchers
- Three common approaches
  - Match rates
  - Manual review of a sampling of matching decisions
  - Run matchers against known testbeds

THE MATCHING QUALITY PROBLEM

- Approach #1 – Match rates
  - Measure the number of records from one data source that match records from another
  - Track the measurements over time, looking at
    - Whether they are close to the expected overlap in the data source's population
    - Whether there are any sudden increases or decreases
  - Pros:
    - Easy to automate
  - Cons
    - Does not measure matching quality directly
    - Significant changes are only indication of potential changes in matching quality that warrant further investigation

BACKGROUND

Matching Rates

Given two data sources: A and B

Metric	Equation
Matching Rate of A /wrt to B	$\frac{\lvert A \text{ records matching } B \text{ records} \rvert}{\lvert A \rvert}$
Matching Rate of B /wrt to A	$\frac{\lvert B \text{ records matching } A \text{ records} \rvert}{\lvert B \rvert}$

- Example:
  - Let A contain 10000 records
  - Let B contain 15000 records
  - Let 9900 distinct records in A match 9990 distinct records in B
  - Matching Rate of A /wrt to B = 99.9%
  - Matching Rate of B /wrt to A = 66.6%

BACKGROUND

Classification Analysis of Matching Decisions

Classification of pairs of records between A and B:

Matcher's Classification	**True Classification**	
	Match (P)	Non-Match (N)
Match	True Positive (TP)	False Positive (FP)
Non-Match	False Negative (FN)	True Negative (TN)

Metric	**Equation**
Sensitivity (recall, hit rate, or true positive rate)	TP / P
Specificity (selectivity or true negative rate)	TN / N
Precision	TP / (TP + FP)
Accuracy	(TP + TN) / (P + N)

THE MATCHING QUALITY PROBLEM

- Approach #2 – Manual review of sample matching decisions
  - Select a set of sample matching decisions
  - For each selected matching decision, manually determine if is a true match (TP), false match (FP), true non-match (TN), or missed match (FN)
  - Use TP, FP, TN, TN to compute
    - Sensitivity – How well the matcher is finding TP's and avoiding FN's
    - Specificity – How well the matcher is finding TN's and avoiding FP's
    - Precision – How well the matcher is finding TP's while avoiding FP's
    - Accuracy – How well the matcher is doing in make correct decisions, i.e., finding TP's and TN's, while avoiding FP's or FN's
  - Pros
    - Can provide excellent indicators of matching quality
  - Cons
    - Can be timing consuming and costly, even if some the review process can be automated
    - Sensitive the size and distribution of selected sample set

THE MATCHING QUALITY PROBLEM

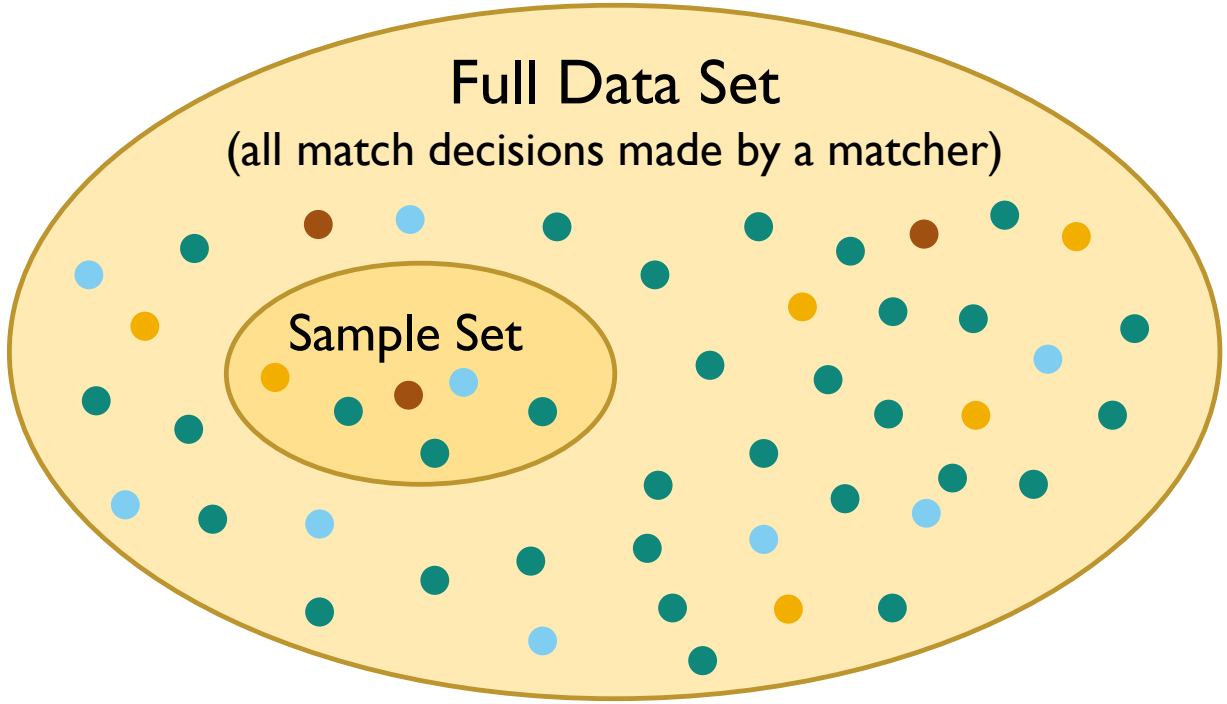
- Approach #3 – Run against a known testbed
  - Feed test records to the matcher
  - Compare matching decisions against expected (known) matching decisions for the testbed
  - Track true matches (TP), false matches (FP), true non-matches (TN), and missed matches (FN)
  - Compute sensitivity, specificity, precision, and accuracy
  - Pros
    - Can provides a good indicator of matching quality, if the testbed is representative the real records in the data sources
  - Cons
    - Very costly to setup a testbed that is truly representative of the records in the real data sources
    - Can be hard to wire the testbed into the matcher

AN ALTERNATIVE: APPROACH #4

- Compare the decisions made by two independent matchers and then sample conflicting match decisions
  - Pros
    - Can provide excellent indicators of matcher quality
    - Doesn't require extensive sampling, like Approach #2
    - Doesn't require creating a testbed that is representative of records from each data source and doing manually matching for all those records, like #3
  - Cons
    - Requires two independent matchers
- CHARM and DOHMPI both match records from the EHDI database and from a birth registry, referred to here as Vital Statistics (VS)

BACKGROUND

Classification Analysis via Sampling



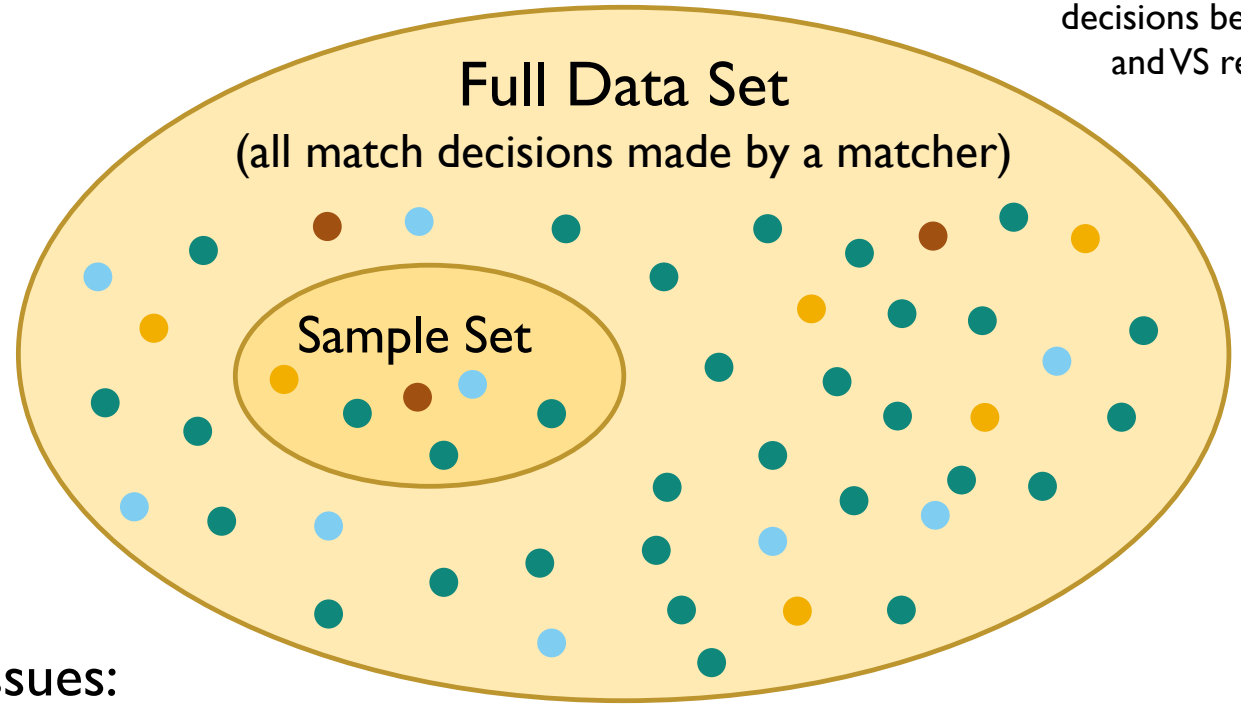
General Process

1. Select a sample set
2. Evaluate each match decision, i.e., determine whether it is a TP, FP, FN, or TN
3. Generalize to entire data set

BACKGROUND

Classification Analysis via Sampling

Note: in Utah, there ~100 billion match decisions between EHDI and VS records in the study period (5.5 years)

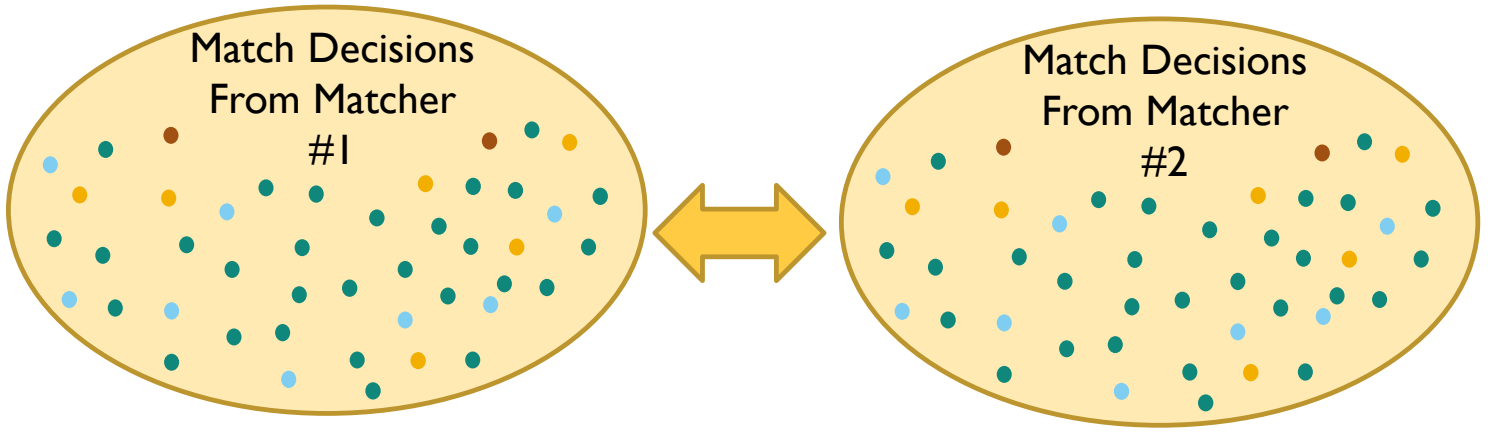


Issues:

- The Sample Set needs to be BIG
  - The classification of the matches is not a normal distribution
  - FP + FN is expected to be very small compare to the number TP + TN
- The bigger the sample set, the more costly it is to do the manual analysis of the match determination

APPROACH #4: THE METHOD

Use two independent matchers to classify "most" of the match decisions automatically



General Process

1. For each record from each data source, compare the match decision(s) made by each matcher (#1 and #2) for that record
   - If the match decisions are the same, then classify them as TP and TN
   - If the match decisions are different, mark match decisions as "potential problems"

APPROACH #4

Use two independent matchers to classify "most" of the match decisions automatically

General Process (continued)

2. Categorize the potential problems as
   - Matched by #1 but not #2
   - Matched by #2 but not #1
   - #1's matches are a subset of #2's
   - #2's matches are a subset of #1's
   - #1's and 2's matches intersect
   - #1's and 2's matches are completely different
3. Randomly select samples from each potential-problem category
4. Review the match decisions in each sample set and classify them as TP, FP, FN, and TN; also, record observations about the data
5. Generalize results to all matching decisions for both matchers

APPROACH #4

Specific steps from a technical perspective

1. Extract sample records from each data source
2. Extract match decisions
3. Categorize match decisions
4. Select sample sets for each category that represents potential problems
5. Extract detailed data for each record in each sample set
6. Manual review the potential problems and classify each associated match decision as TP, FP, FN, or TN
7. Analyze and generalize the results
8. Formulate recommendations for improvement

APPROACH #4

Automated steps

1. Extract sample records from each data source
2. Extract match decisions
3. Categorize match decisions
4. Select sample sets for each category that represents potential problems
5. Extract detailed data for each record in each sample set
6. Manual review the potential problems and classify each associated match decision as TP, FP, FN, or TN
7. Analyze and generalize the results
8. Formulate recommendations for improvement

STEP 1

Extract sample records from each data source

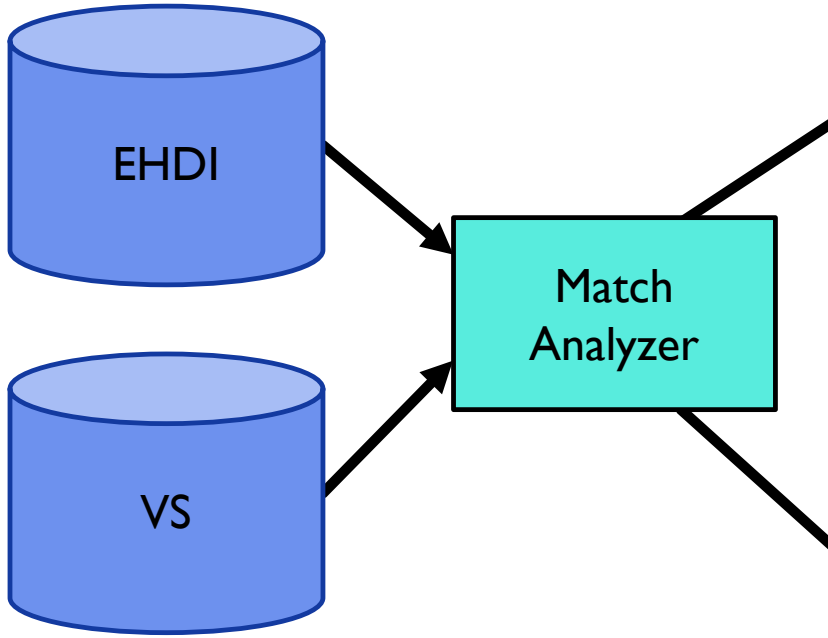
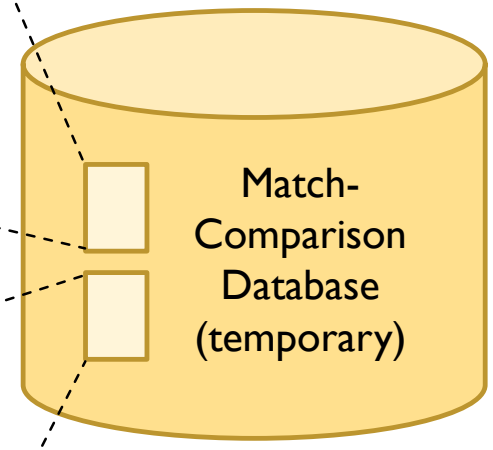


Table A - EHDI Records

EHDI Id	Birth Year
142	2016
325	2013
459	2015
637	2016
716	2016
...	

Table B - VS Records

VS Id	Birth Year
343	2014
452	2015
632	2016
673	2016
743	2017
...	



STEP 2
Extract Match Decisions

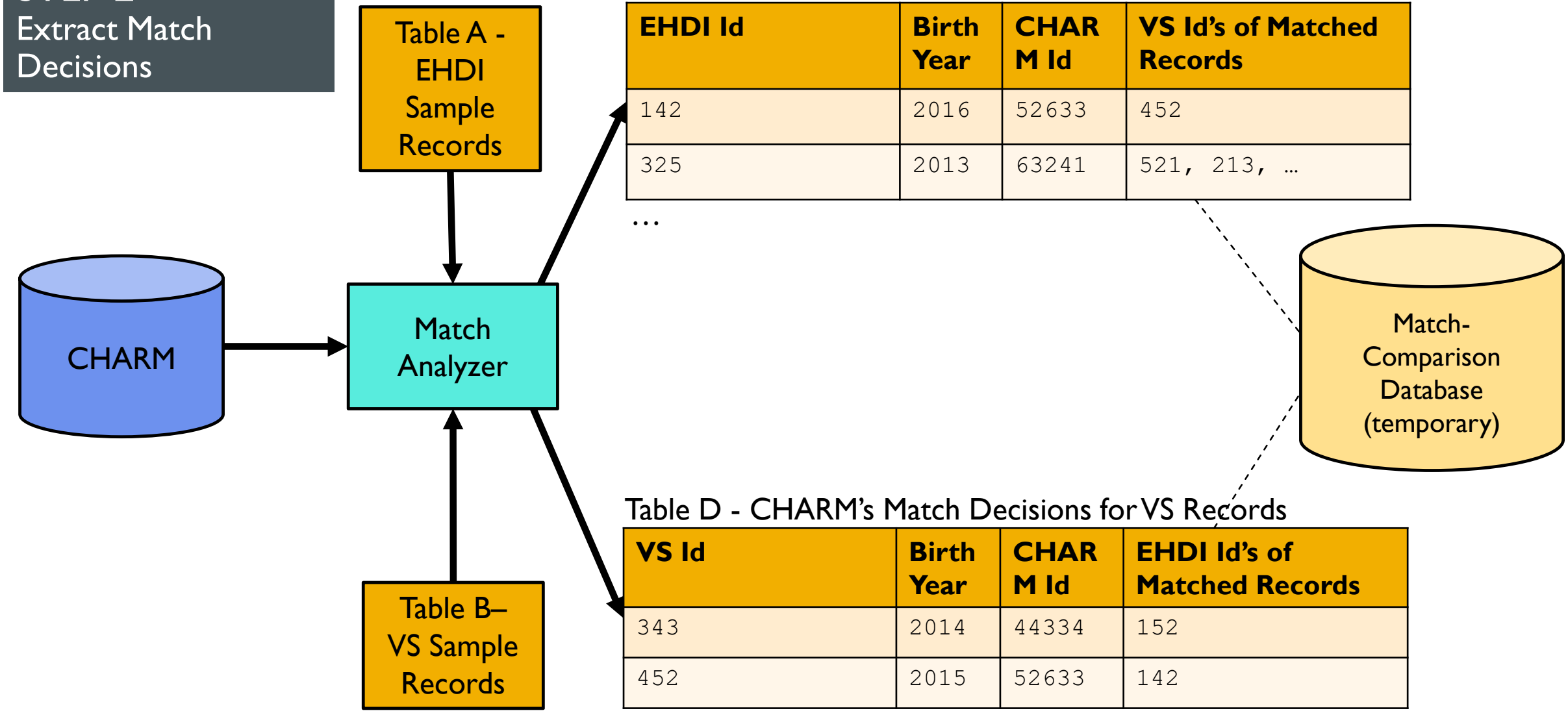


Table C - CHARM's Match Decisions for EHDI Records

EHDI Id	Birth Year	CHARM Id	VS Id's of Matched Records
142	2016	52633	452
325	2013	63241	521, 213, …

…

Table D - CHARM's Match Decisions for VS Records

VS Id	Birth Year	CHARM Id	EHDI Id's of Matched Records
343	2014	44334	152
452	2015	52633	142

…

STEP 2
Extract Match Decisions

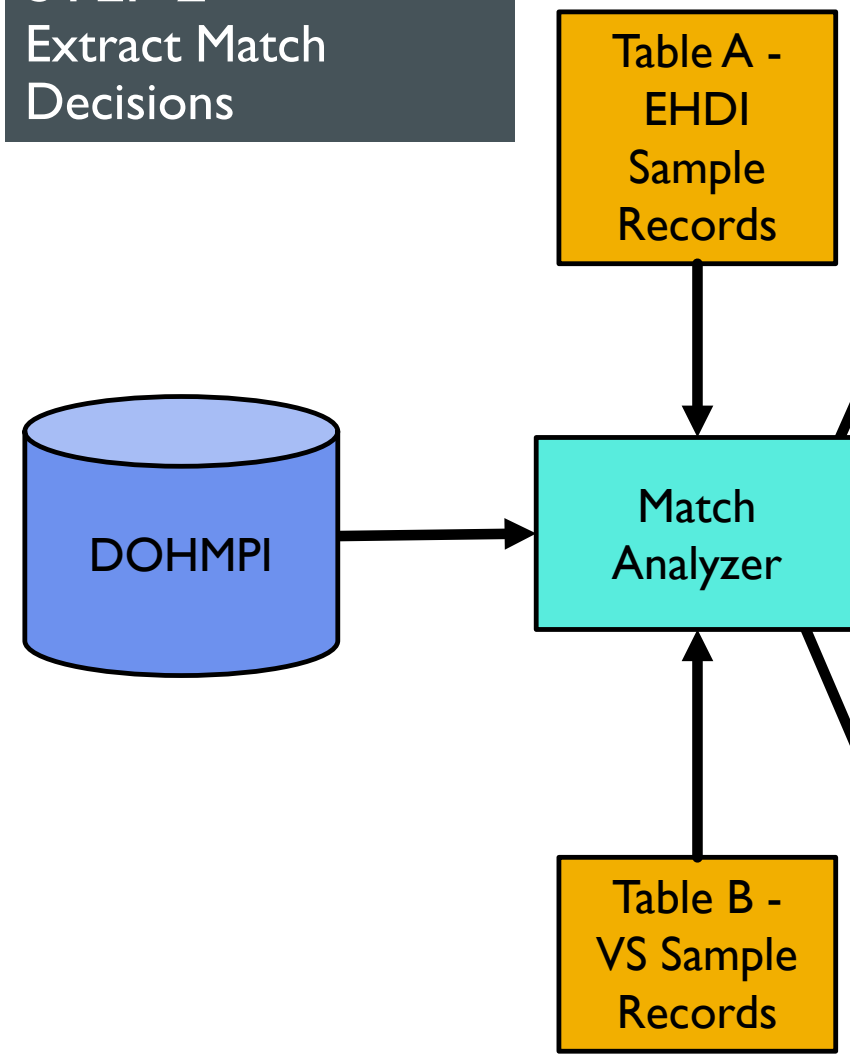


Table E - DOHMPI's Match Decisions for EHDI Records

EHDI Id	Birth Year	DOHMPI Id	VS Id's of Matched Records
142	2016	46799	452
325	2013	10032	

...

Match-Comparison Database (temporary)

Table F - DOHMPI's Match Decisions for VS Records

VS Id	Birth Year	DOHPI id	EHDI Id's of Matched Records
343	2014	74123	421
452	2015	46799	142

...

STEP 3

Categorize match decisions

- The Match Analyzer compares the match decisions in Table C with Table E and those in Table D with Table F – record by record
  - Because a record from one data source can match zero or more record from the other, these are set comparisons
- The Match Analyzer places each record in one of the following categories:
  - MCC_1 - Same Matches in Both Systems
  - MCC_2 - Neither System found a Match
  - MCC_3 - Matched by CHARM but Not DOHMPI
  - MCC_4 - Matched by DOHMPI but Not CHARM
  - MCC_5 - CHARM Matches Subset of DOHMPI Matches
  - MCC_6 - DOHMPI Matches Subset of CHARM Matches
  - MCC_7 - Matches Intersect
  - MCC_8 - Matches Differ
- For MCC_1 and MCC_2, the Match Analyzer automatically classifies associated match decisions as TP's and TN's

STEP 3

Examples

Table C - CHARM's Match Decisions for EHDI Records

EHDI Id	Birth Year	CHARM Id	VS Id's of Matched Records
142	2016	52633	452
325	2013	63241	521, 213, …

Table E - DOHMPI's Match Decisions for EHDI Records

EHDI Id	Birth Year	DOHMPId	VS Id's of Matched Records
142	2016	46799	452
325	2013	10032	

- Since both matchers linked EHDI record 142 to VS record 452, EDHI record 142 is categorized as MCC_1 and
  - The (EHDI 142,VS 152) pairing are classified as TP match decisions for both matchers.
  - Also, all other (EHDI 142,VS x) pairings, where x is one of the 316,847 other records in VS for the study period, are automatically considered TN match decisions for both matchers.

STEP 3

More Examples

Table C - CHARM's Match Decisions for EHDI Records

EHDI Id	Birth Year	CHARM Id	VS Id's of Matched Records
142	2016	52633	452
325	2013	63241	521, 213

Table E - DOHMPI's Match Decisions for EHDI Records

EHDI Id	Birth Year	DOHMPI Id	VS Id's of Matched Records
142	2016	46799	452
325	2013	10032	

- Since the CHARM matcher linked EHDI record 325 to VS records 521 and 213 and DOHMPI did not link 325 at all, this EDHI record is categorized as MCC_3 and a subsequent manual review will determine
  - if the (EHDI 325, VS 521) and (EHDI 325, 213) pairings are TP for CHARM and FN for DOHMPI, or
  - if they are FP for CHARM and TN for DOHMPI

STEP 3

Results of matching categorization

Population Counts by Year

Year	# of EHDI Records	# of VS Records
2013	52840	51907
2014	53576	52211
2015	52577	51769
2016	52088	51550
2017	50295	49667
2018	48642	48221
2019*	11644	11523
Totals	321662	316848

*2019 was only for the months of Jan - May

STEP 3

Results of matching categorization

Categorizations matchings for EHDI Records

Year	MCC_1	MCC_2	MCC_3	MCC_4	MCC_5	MCC_6	MCC_7	MCC_8
2013	39790	3200	3242	6216	357	4	0	31
2014	35970	4012	4121	9174	257	2	0	40
2015	38266	787	73	13362	74	4	0	11
2016	39855	482	76	11598	56	4	0	17
2017	39460	469	69	10213	66	1	0	17
2018	39831	249	81	8423	44	0	0	14
2019	9472	130	101	1916	20	1	0	4
Totals	242644	9329	7763	60902	874	16	0	134

MCC_1 - Same Matches in Both Systems
MCC_2 - Neither System found a Match
MCC_3 - Matched by CHARM but Not DOHMPI
MCC_4 - Matched by DOHMPI but Not CHARM
MCC_5 - CHARM Matches Subset of DOHMPI Matches
MCC_6 - DOHMPI Matches Subset of CHARM Matches
MCC_7 - Matches Intersect
MCC_8 - Matches Differ

STEP 4

Select sample sets for each category that represents potential problems

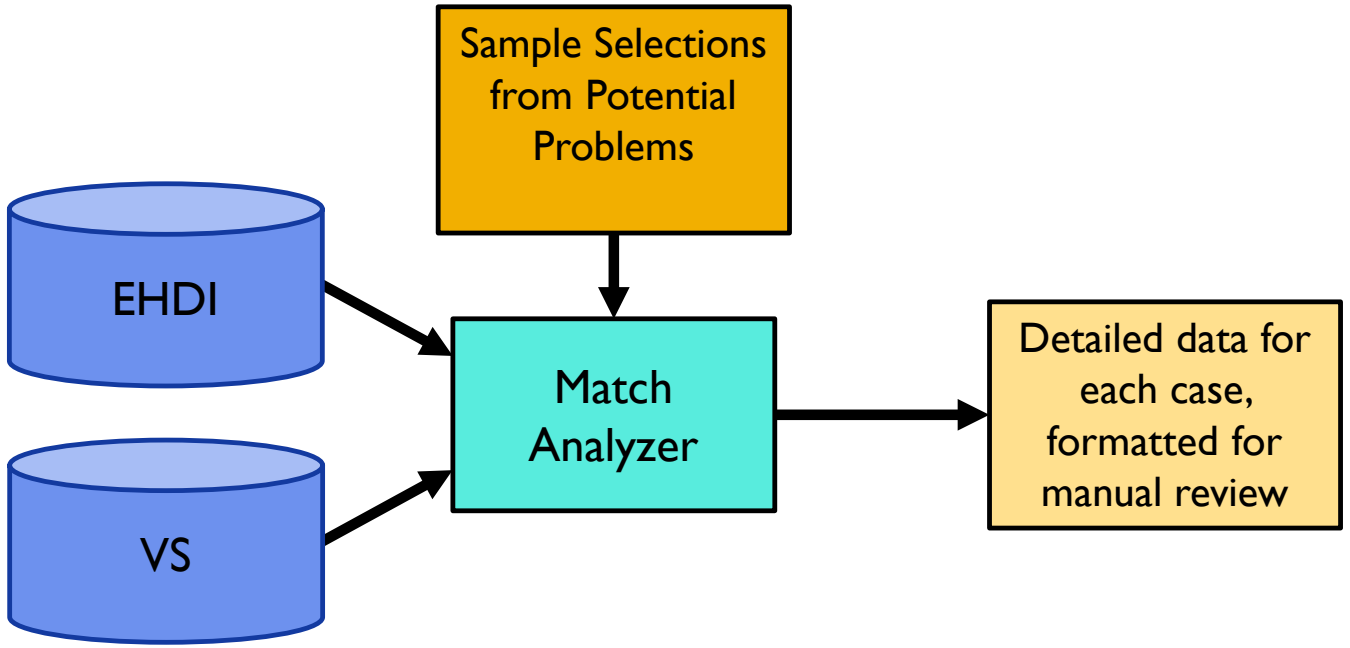
- The Match Analyzer randomly selects a percentage of the match comparisons from each potential problem category {MCC_3, …, MCC_8}

Comparison with EHDI as Basis						
Year	**MCC_3**	**MCC_4**	**MCC_5**	**MCC_6**	**MCC_7**	**MCC_8**
2013	3242	6216	357	4	0	31
2014	4121	9174	257	2	0	40
2015	73	13362	74	4	0	11
2016	76	11598	56	4	0	17
2017	69	10213	66	1	0	17
2018	81	8423	44	0	0	14
2019	101	1916	20	1	0	4
Totals	7763	60902	874	16	0	134

STEP 5

Extract detailed data for each potential problem in the sample sets

- For each selected match comparison, the Match Analyzer extracts detailed data for all involved child records
- These detailed data are (were) stored that in a temporary file on a secure server within the UDOH intranet



STEP 6

Manual review those potential problems and classify each associated match decisions as TP, FP, FN, or TN

- In this step, we systematically studied all the randomly selected match comparisons from each potential-problem MCC and identify the following for each match or missed match, for each matcher:
  - Whether a match made by each matcher is a TP or FP
  - Whether a match made by the CHARM matcher but not the DOHMPI matcher is a TN or FN for DOHMPI
  - Whether a match made by the DOHMPI matcher but not the CHARM matcher is a TN or FN negative for DOHMPI
  - The characteristics of the data may have led to a bad decision, i.e., a FP or FN

RESULTS

Matches Found by CHARM but Not DOHMPI (MCC_3)

From EHDI's Perspective

Year	# Case for manual review	CHARM: Manual Matching of Sample Pairings				DOHMPI: Manual Matching of Sample Pairing			
		TP	FP	TN	FN	TP	FP	TN	FN
2013	34	34	0	0	0	0	0	0	34
2014	61	61	0	0	0	0	0	0	61
2015	2	2	0	0	0	0	0	0	2
2016	0	0	0	0	0	0	0	0	0
2017	2	2	0	0	0	0	0	0	2
2018	0	0	0	0	0	0	0	0	0
2019	1	1	0	0	0	0	0	0	1
Total	**100**	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**100**

From VS's Perspective

Year	# Case for manual review	CHARM: Manual Matching of Sample Pairings				DOHMPI: Manual Matching of Sample Pairing			
		TP	FP	TN	FN	TP	FP	TN	FN
2013	50	50	0	0	0	0	0	0	50
2014	59	61	0	0	0	0	0	0	61
2015	0	0	0	0	0	0	0	0	0
2016	1	1	0	0	0	0	0	0	1
2017	0	0	0	0	0	0	0	0	0
2018	1	1	0	0	0	0	0	0	1
2019	2	2	0	0	0	0	0	0	2
Total	**113**	**115**	**0**	**0**	**0**	**0**	**0**	**0**	**115**

RESULTS

Matches Found by DOHMPI but Not CHARM (MCC_4)

From EHDI's Perspective

Year	# Case for manual review	CHARM: Manual Matching of Sample Pairings				DOHMPI: Manual Matching of Sample Pairing			
		TP	FP	TN	FN	TP	FP	TN	FN
2013	8	2	0	2	6	6	4	0	0
2014	17	0	0	0	17	17	0	0	0
2015	24	0	0	2	22	22	2	0	0
2016	18	0	0	0	18	18	0	0	0
2017	16	0	0	0	16	16	0	0	0
2018	14	0	0	1	13	13	1	0	0
2019	3	0	0	0	3	3	0	0	0
Total	**100**	**2**	**0**	**5**	**95**	**95**	**7**	**0**	**0**

From VS's Perspective

Year	# Case for manual review	CHARM: Manual Matching of Sample Pairings				DOHMPI: Manual Matching of Sample Pairing			
		TP	FP	TN	FN	TP	FP	TN	FN
2013	12	0	0	1	11	11	1	0	0
2014	8	0	0	4	5	5	4	0	0
2015	30	0	0	1	29	29	1	0	0
2016	24	0	0	2	22	22	2	0	0
2017	15	0	0	1	14	14	1	0	0
2018	16	0	0	0	16	16	0	0	0
2019	3	0	0	0	3	3	0	0	0
Total	**108**	**0**	**0**	**9**	**100**	**100**	**9**	**0**	**0**

RESULTS

CHARM's Matches are Subset DOHMPI's Matches (MCC_5)

From EHDI's Perspective

Year	# Case for manual review	CHARM: Manual Matching of Sample Pairings				DOHMPI: Manual Matching of Sample Pairing			
		TP	FP	TN	FN	TP	FP	TN	FN
2013	16	16	0	22	0	16	22	0	0
2014	14	14	0	25	0	14	25	0	0
2015	3	3	0	5	0	3	5	0	0
2016	2	2	0	3	0	2	3	0	0
2017	3	3	0	3	0	3	3	0	0
2018	2	2	0	2	0	2	2	0	0
2019	2	2	0	2	0	2	2	0	0
Total	**42**	**42**	**0**	**62**	**0**	**42**	**62**	**0**	**0**

From VS's Perspective

Year	# Case for manual review	CHARM: Manual Matching of Sample Pairings				DOHMPI: Manual Matching of Sample Pairing			
		TP	FP	TN	FN	TP	FP	TN	FN
2013	21	21	0	21	3	23	22	0	0
2014	20	20	0	20	2	22	20	0	0
2015	9	14	0	8	4	18	8	0	0
2016	5	5	0	3	2	7	3	0	0
2017	7	7	0	5	3	11	4	0	0
2018	5	5	0	5	2	7	5	0	0
2019	1	1	0	2	0	1	2	0	0
Total	**68**	**73**	**0**	**64**	**16**	**89**	**64**	**0**	**0**

RESULTS

DOHMPI's Matches are Subset CHARM's Matches (MCC_6)

From EHDI's Perspective

Year	# Case for manual review	CHARM: Manual Matching of Sample Pairings				DOHMPI: Manual Matching of Sample Pairing			
		TP	FP	TN	FN	TP	FP	TN	FN
2013	1	2	0	0	0	1	0	0	1
2014	1	1	1	0	0	1	0	1	0
2015	0	0	0	0	0	0	0	0	0
2016	0	0	0	0	0	0	0	0	0
2017	0	0	0	0	0	0	0	0	0
2018	0	0	0	0	0	0	0	0	0
2019	0	0	0	0	0	0	0	0	0
Total	**2**	**3**	**1**	**0**	**0**	**2**	**0**	**1**	**1**

From VS's Perspective

Year	# Case for manual review	CHARM: Manual Matching of Sample Pairings				DOHMPI: Manual Matching of Sample Pairing			
		TP	FP	TN	FN	TP	FP	TN	FN
2013	4	8	0	0	0	4	0	0	4
2014	7	17	0	0	0	8	0	0	9
2015	9	19	4	0	0	9	0	4	10
2016	7	13	4	0	0	7	0	4	6
2017	9	17	1	0	0	9	0	1	8
2018	1	2	0	0	0	1	0	0	1
2019	1	1	1	0	0	1	0	1	0
Total	**38**	**77**	**10**	**0**	**0**	**39**	**0**	**10**	**38**

RESULTS

CHARM's Matches intersect DOHMPI's Matches (MCC_7)

From VS's Perspective

Year	# Case for manual review
2013	4
2014	7
2015	9
2016	7
2017	9
2018	1
2019	1
Total	**38**

CHARM: Manual Matching of Sample Pairings

TP	FP	TN	FN
8	0	0	0
17	0	0	0
19	4	0	0
13	4	0	0
17	1	0	0
2	0	0	0
1	1	0	0
77	**10**	**0**	**0**

DOHMPI: Manual Matching of Sample Pairing

TP	FP	TN	FN
4	0	0	4
8	0	0	9
9	0	4	10
7	0	4	6
9	0	1	8
1	0	0	1
1	0	1	0
39	**0**	**10**	**38**

RESULTS

CHARM's Matches are completely different from DOHMPI's Matches (MCC_8)

From EHDI's Perspective

Year	# Case for manual review	CHARM: Manual Matching of Sample Pairings				DOHMPI: Manual Matching of Sample Pairing			
		TP	FP	TN	FN	TP	FP	TN	FN
2013	4	2	2	0	0	0	4	0	0
2014	9	8	1	0	0	1	8	0	0
2015	2	1	1	0	0	1	1	0	0
2016	3	0	3	0	0	2	1	0	0
2017	0	0	0	0	0	0	0	0	0
2018	3	3	0	1	0	0	4	0	0
2019	1	1	0	1	0	0	2	0	0
Total	**22**	**15**	**7**	**2**	**0**	**4**	**20**	**0**	**0**

From VS's Perspective

Year	# Case for manual review	CHARM: Manual Matching of Sample Pairings				DOHMPI: Manual Matching of Sample Pairing			
		TP	FP	TN	FN	TP	FP	TN	FN
2013	1	1	0	0	1	1	0	0	1
2014	4	4	0	0	4	4	0	0	4
2015	5	4	1	0	3	3	2	1	2
2016	2	2	0	0	1	1	1	0	1
2017	5	2	3	0	2	3	2	1	1
2018	1	1	0	0	1	1	0	0	1
2019	0	0	0	0	0	0	0	0	0
Total	**18**	**14**	**4**	**0**	**12**	**13**	**5**	**2**	**10**

STEP 7

Analyze and generalize the results

- Analyze the data from multiple perspectives and looking for patterns.
- Estimate the sensitivity, specificity, precision, and accuracy for each matcher
- Two approaches:
  A. Generalize the results of the manual review to predicate the classification of all match decisions in the potential problem categorizes
  B. Generalize the sampling of potential problems to create corresponding samplings of the MCC_1 and MCC_2 categories, for each the classifications are automatically computed

GENERALIZATION

Method A - Generalize of classifications to all match decisions

Base Numbers for Generalization Method A

	Total Number of Potential Problem Records	Number of Match Pair Among Potential Problems	Total Number of Records	Total Number of Possible Match Pairs
From EHDI's Perspective	69,689	87,504	321,662	103,466,120,582
From VS' Perspective	68,949	102,147	316,848	100,392,338,256

GENERALIZATION

Method A - Generalize of classifications to all match decisions

Match-Decision Classifications for Generalization Method A

		Overall Estimated Match-Decision Classifications			
		TP	**FP**	**TN**	**FN**
From EHDI's Perspective	**CHARM**	285,086	2,096	103,465,808,511	24,889
	DOHMPI	278,385	23,317	103,465,792,419	26,461
From VS's Perspective	**CHARM**	300,003	2,750	100,392,007,805	27,698
	DOHMPI	292,191	15,322	100,391,997,545	33,198
Average	**CHARM**	292,545	2,423	101,928,908,158	26,294
	DOHMPI	285,288	19,320	101,928,894,982	29,830

GENERALIZATION

Method A - Generalize of classifications to all match decisions

Estimated Matching Quality Using Generalization Method A

		Overall Estimated Match-Decision Classifications			
		Sensitivity	Specificity	Precision	Accuracy
From EHDI's Perspective	CHARM	91.970643%	99.999998%	99.270149%	99.999974%
	DOHMPI	91.319880%	99.999977%	92.271513%	99.999952%
From VS's Perspective	CHARM	91.547783%	99.999997%	99.091669%	99.999970%
	DOHMPI	89.797442%	99.999985%	95.017446%	99.999952%
Average	CHARM	91.753336%	99.999998%	99.178554%	99.999972%
	DOHMPI	90.533848%	99.999981%	93.657576%	99.999952%

GENERALIZATION

Method B - Generalize of potential-problem sampling to the good categories

Base Numbers for Generalization Method B

	Total Records	Total Potential Problems	Samples from the Potential Problems	% of Potential Problems Samples
From EHDI's Perspective	321662	69689	266	0.38170%
From VS' Perspective	316848	68949	351	0.50907%

GENERALIZATION

Method B - Generalize of potential-problem sampling to the good categories

Match-Decision Classifications for Generalization Method B

		Overall Estimated Match-Decision Classifications			
		TP	FP	TN	FN
From EHDI's Perspective	**CHARM**	1,088	8	105	95
	DOHMPI	1,069	89	37	101
From VS's Perspective	**CHARM**	1,518	14	113	141
	DOHMPI	1,487	78	52	169
Average	**CHARM**	1,303	11	109	118
	DOHMPI	1,278	84	45	135

GENERALIZATIONS

Method B - Generalize of potential-problem sampling to the good categories

Estimated Matching Quality Using Generalization Method B

		Overall Estimated Match-Decision Classifications			
		Sensitivity	Specificity	Precision	Accuracy
From EHDI's Perspective	CHARM	91.969569%	92.920354%	99.270073%	92.052469%
	DOHMPI	91.367521%	29.365079%	92.314335%	85.339506%
From VS's Perspective	CHARM	91.500904%	88.976378%	99.086162%	91.321389%
	DOHMPI	89.794686%	40.000000%	95.015974%	86.170213%
Average	CHARM	91.695989%	90.833333%	99.162861%	91.628812%
	DOHMPI	90.445860%	34.765625%	93.867058%	85.820896%

STEP 8

Formulate recommendations for improvement

- This step involves identifying patterns in the results and making some meaning recommendations for each matcher and for improving the quality of the source data

RECOMMENDATIONS

For CHARM

- Data Cleaning Improvements
  - Ensure that bad Newborn Screening Numbers (ones that don't match the expected formats) are converted to null
  - Reformat all Newborn Screening Numbers to a standard format for easy comparison
  - Convert a birth order to null, if it is 0 or if the multiple-birth flag is false
  - Convert a bad birth weight to null
- Tweak Matching Rules
  - Ensure that there are no positive and less negative match influences for missing birth orders and birth weights
  - Have more negative match influence for mismatched birth orders
  - Have more positive match influence for matched birth weights
- Enhance CHARM so it can do re-matching

RECOMMENDATIONS

For DOHMPI

- Data Cleaning Improvements
  - Standardize the format of newborn screening numbers
- Tweak Matching Rules
  - Raise the "join" threshold for all data sources (i.e. the match threshold for a record to link to a different data source record).
  - Require a minimum of 3 non-gender data elements to link with another identity record.
  - Add more restrictive matching rules related to different last names where other fields match
- Re-match updated records
- Implement an explicitly unlink and re-match feature

RECOMMENDATIONS

For EHDI Data

- EHDI's data appears to have improved over time
- Continue efforts to improve the quality of
  - First and middle names
  - Birth weight
  - Birth orders for multiple births
  - Newborn screening numbers for multiple births
- Consider changing the sequencing of newborn screening kits to make it less likely for newborn screening numbers to be mixed up for multiple births

SUMMARY

- Accurate matching of person records is critical to the success of any federated or integrated data system
- Traditional approaches to measuring matching quality are either very time consuming, expensive, or limited
- If there are two independent matchers, then there is a cost-effective alternative, where
  - A significant portion of the work can be automated
  - The amount manual reviewing of matching decisions need is relatively small
  - There are two viable approaches to generalizing the results of the manual reviews to estimated overall match quality

QUESTIONS?