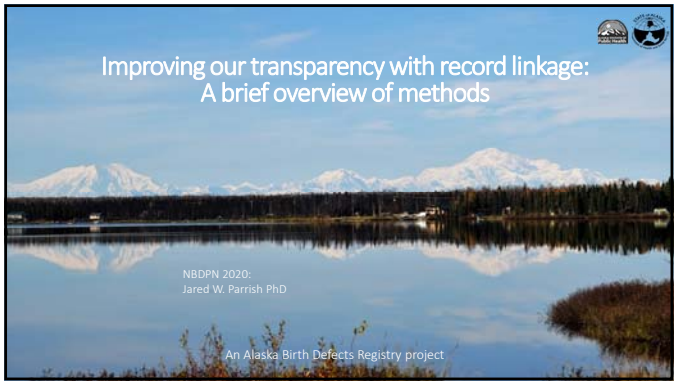
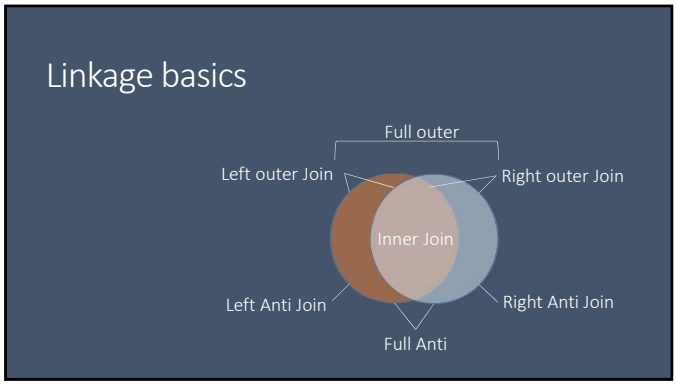
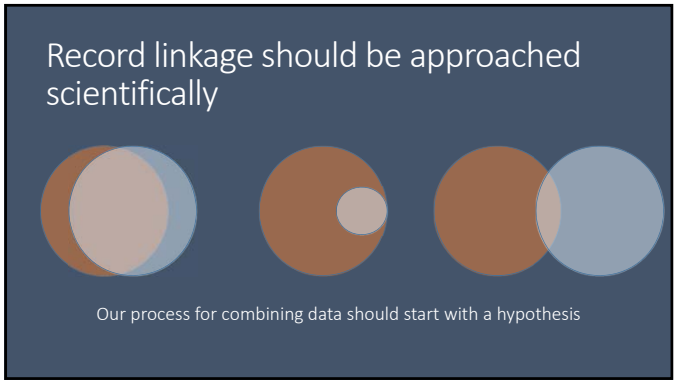

Improving our transparency with record linkage:
A brief overview of methods
NBDPN 2020:
Jared W. Parrish PhD
An Alaska Birth Defects Registry project

1


Linkage basics
Full outer
Left outer Join
Right outer Join
Inner Join
Left Anti Join
Right Anti Join
Full Anti

2


Record linkage should be approached scientifically
Our process for combining data should start with a hypothesis

3

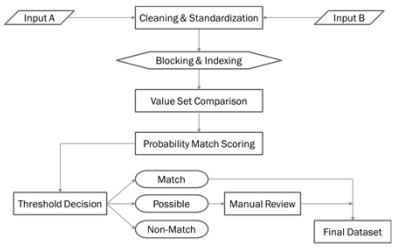
Image source: unbiasedresearch.blogspot.com

Statistics are based on probabilities.

Statistical conclusions are subject to error!

		True Match Status		
		Matches	Non-matches	
Link status (assigned by computer)	Links	True Links (True Positives)	False Links (False Positives / Type I error)	Total links
	Non-Links	Missed Links (False Negatives / Type II error)	True Non-Links (True Negatives)	Total Non-links
		Total Matches	Total Non-matches	Total record pairs

4



Both Systematic and Random error must be considered

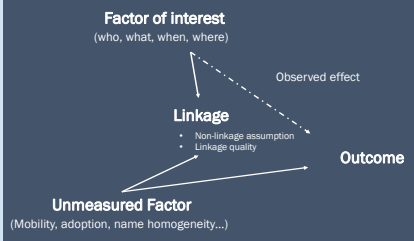
In General

- Systematic error occurs at decision points
- Random error occurs during scoring and matching

5

Poor linkages can result in bias

Bias can occur when linkage quality is differential by any unmeasured factor "U".



6

Quick example of distance algorithms

Modified Levenshtein algorithm
Q-grams algorithm
Jaro-Winkler Algorithm

Compare “Chlid” and “Child”
Compare “Cihld” and “Child”
Compare “John-Adams” and “Adams-John”

7

Modified Levenshtein Algorithm

Test how many operations needed for string “A” to = string “B”

$$ED_{w\,(s1,s2)} = \begin{cases} if\ e(s1,s2) > d * \max(L(s1),L(s2))\ then\ 0 \\ if\ e(s1,s2) < a * \max(L(s1),L(s2))\ then\ 1 \\ \left(\dfrac{[d * \max(L(s1),L(s2))] - e(s1,s2)}{(d-a) * \max(L(s1),L(s2))}\right) \end{cases}, otherwise$$

$e(s1,s2)$ = edit distance score a = approve level
$(s1,s2)$ = Strings 1 and 2 L = String length
d = Disapprove level

$$\begin{bmatrix} 0 & C & h & l & i & d \\ C & 0 & & & & \\ h & & 0 & & & \\ i & & & 1 & & \\ l & & & & 1 & \\ d & & & & & 1 \end{bmatrix} \begin{bmatrix} 0 & C & i & l & h & d \\ C & 0 & & & & \\ h & & 1 & & & \\ i & & & 1 & & \\ l & & & & 2 & \\ d & & & & & 2 \end{bmatrix}$$

$$ED_{w\,(Child,Child)} = \begin{cases} if\ 1 > 0.4 * 5\ then\ 0, False \\ if\ 1 < 0.1 * 5\ then\ 1, False \\ \left(\dfrac{[0.4 * 5] - 1}{(0.4 - 0.1) * 5}\right) \end{cases} = \frac{1}{1.5} = 0.67.$$

$$ED_{w\,(Cihld,Child)} = \begin{cases} if\ 2 > 0.4 * 5\ then\ 0, True \\ if\ 2 < 0.1 * 5\ then\ 1, False \\ \left(\dfrac{[0.4 * 5] - 2}{(0.4 - 0.1) * 5}\right) \end{cases} = \frac{0}{1.5} = 0$$

8

Q-Grams Algorithm

Substrings string “A” and “B” into sets of length “q”

$$Q_{Bw\,(S1,S2)} = \begin{cases} if\ D[Gs(s1),Gs(s2)] > R[a * (|Gs(s1)| + |Gs(s2)|)]\ then\ 0 \\ if\ D[Gs(s1),Gs(s2)] < R[a * (|Gs(s1)| + |Gs(s2)|)]\ then\ 1 \\ \left(1 - \dfrac{D[Gs(s1),Gs(s2)] - R[a * (|Gs(s1)| + |Gs(s2)|)]}{R[d * (|Gs(s1)| + |Gs(s2)|)] - R[a * (|Gs(s1)| + |Gs(s2)|)]}\right) \end{cases}, otherwise$$

D = Disjoint grams a = approve level
Gs = grams of length ‘q’ d = Disapprove level
$(s1,s2)$ = Strings 1 and 2 R = Round

$$Q_{Bw\,(S1,S2)} = \begin{cases} if\ 6 > R[0.4 * (|10| + |10|)]\ then\ 0 \\ if\ 6 < R[0.1 * (|G10| + |10|)]\ then\ 1 \\ \left(1 - \left(\dfrac{6 - R[0.1 * (|10| + |10|)]}{R[0.4 * (|10| + |10|)] - R[0.1 * (|10| + |10|)]}\right)\right) \end{cases}, otherwise$$

$$= \begin{cases} if\ 6 > 2\ then\ 0, False \\ if\ 6 < 8\ then\ 1, False \\ \left(1 - \left(\dfrac{6-2}{8-2}\right)\right) = 1 - \dfrac{4}{6} = 0.33 \end{cases}$$

For example, setting q=2 the strings “John-Adams” and “Adams-John”:
- have the following 2-gram vectors {Jo, oh, hn, n-, -A, Ad, da, am, ms, s} and {Ad, da, am, ms, s-, -J, Jo, oh, hn, n}.
- The two vectors share the 2-grams of {Jo, oh, hn, Ad, da, am, ms}, with the disjointed 2-grams of {n-, -A, s, s-, -J, n}.

9

Jaro-Winkler Algorithm

Scores the number of elements in common

$$d_{jw} = \left(\frac{1}{3}\left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m}\right)\right) + l * f(1 - \left(\frac{1}{3}\left(\frac{m}{|s2|} + \frac{m}{|s2|} + \frac{m-t}{m}\right)\right))$$

$m = matchin\ characters$ $l = length\ of\ common\ characters\ (n < 5)$
$(s1, s2) = Strings\ 1\ and\ 2$ $f = scaling\ factor\ for\ common\ characters$
$t = \#of\ transpositions/2$

Using the same example Edit Distance example above we can calculate the *JW* distance for the two string comparison of "CHILD" and "CHLID" which experiences a single letter transposition between strings.

$S_i = \left\lfloor \frac{MAX(|S1|,|S2|)}{2} \right\rfloor - 1 = 1.5,$

m = 5, and $t = \frac{2}{2} = 1$ based on the characters I/L and L/I which are less than 1.5 characters apart

Thus,

$$d_{jw} = \left(\frac{1}{3}\left(\frac{5}{|5|} + \frac{5}{|5|} + \frac{5-1}{5}\right)\right) + 2 * 0.1\left(1 - \left(\frac{1}{3}\left(\frac{5}{|5|} + \frac{5}{|5|} + \frac{5-4}{5}\right)\right)\right) = 0.95$$

10

Comparing these algorithms

	Levenshtein	Q-grams (g=2)	Jaro-Wikler
Child Chlid	0.67	0.00	0.95
Child Clhd	0.00	0.00	0.88
John-Adams Adams-John	0.00	0.33	0.30

11

Setting manual review thresholds

Setting a threshold for manual review depends on the value of missing or making an erroneous match

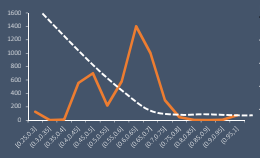
- Trying to replicate a longitudinal prospective cohort missing a linkage (loss to follow-up) may be paramount
- Conducting population level surveillance in a large population balancing erroneous linkages and missed linkages while mitigating differential matching between subpopulations may be paramount

- Some helpful guidelines:
  - *Use machine learning tools on a subset of data with manually reviewed known parameters to develop a starting point*
  - *Training datasets from your input sources can help establish sensitivity of specificity of decision rules and comparison algorithms*
  - *The generalized Petro Distribution can be leveraged to establish minimum thresholds*

12

Threshold example

Two Fictitious data sets with 100 records each-
-Strings: DOB, Last Name, First Name,
-Block: Sex
-Method Jaro-Winkler



Threshold		Classification			Identified	Truth
Upper	Lower	N	P	L		
0.99	0.8	4944	33	55	70	71
0.95	0.8	4944	14	74	75	71
0.9	0.8	4944	7	81	81	71
0.99	0.75	4898	79	55	71	71
0.99	0.7	4598	379	55	71	71

13

Linkage projects in process

Child Protective Services

- Are children reported for birth defects over-represented in Child Welfare
- Are children reported for NAS over-represented in Child Welfare

Permanent Fund Dividend – In process

- Map birth defects to identify clusters and overlay environmental exposures
- Map location of pregnancy and population movement

14

Child Protection linkage example

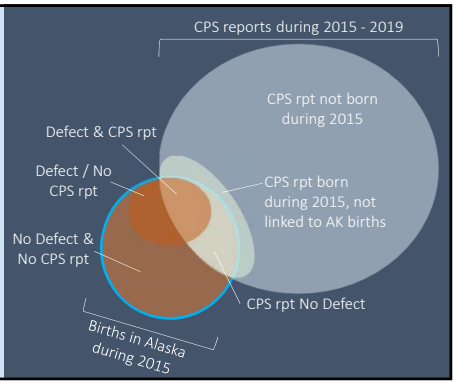
Linkage Approach

1. Base population all births in Alaska during 2015
2. Left outer join (All birth records and matching CPS records)
3. Elements: First & Last name, Date of birth, Sex (manual review on middle name, residence, maternal name)
4. First subset to multiple births, Second non multiple births
   a. Linkage: Deterministic (reconcile duplicates)
   b. Linkage: Probabilistic (Jaro-Winkler, auto at 1.0, manual 0.89 – 0.99, block on year of birth)
5. Manual review: middle name match priority with DOB, minor typos, First name priority (name probability <=1%)
6. Duplication reconciliation

15

Linkage to 2015 birth cohort

- 11, 356 Live births in Alaska
- 58,455 Reports to AK CPS between 2015 – 2019
- 2,942 Reports to AK CPS born during 2015
  - 2,780 AK births linked to AK CPS reports born during 2015
  - 194 CPS reports not linked to AK births



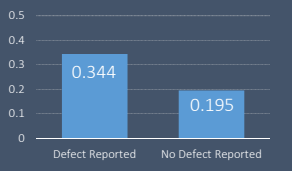
16

Proportion reported to CPS by 2019

Finding:

Relative to children born without a birth defect in Alaska during 2015, children born with a defect, are 76% more likely to be reported to CPS by 2019.

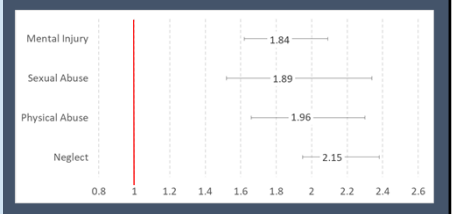
	CPS rpt	No CPS rpt	
Defect Reported	1301	2480	3781
No Defect Reported	1479	6096	7575
	2780	8576	11356



Prevalence Ratio: 1.76 (1.65, 1.87)

17

Relative proportion reported to CPS by maltreatment type



18

Wrap up

In general in a manuscript describe:

Methods:
- The base population
- Type of join
- Linkage methodology and assumptions
- Identifiers used for linkages
- General data cleaning rules
- Manual review methods and thresholds

Results:
- Base population, proportion linked, non-linked, final study population
- If doing something fancy, provide validation results

Discussion:
- Limitations and potential / quantified bias
- Strengths of approach

Appendix/online only:
- Expand upon methods
- Algorithms used and how you determined
- Any ad-hoc analysis demonstrating linkage validity
- Full linkage specification and results

19

Questions?

Jared W. Parrish PhD
Alaska Division of Public Health
jared.parrish@alaska.gov
(907) 269-8068

ABDR website:
http://www.dhss.alaska.gov/dph/wcfh/Pages/mchepi/abdr/default.aspx

20