

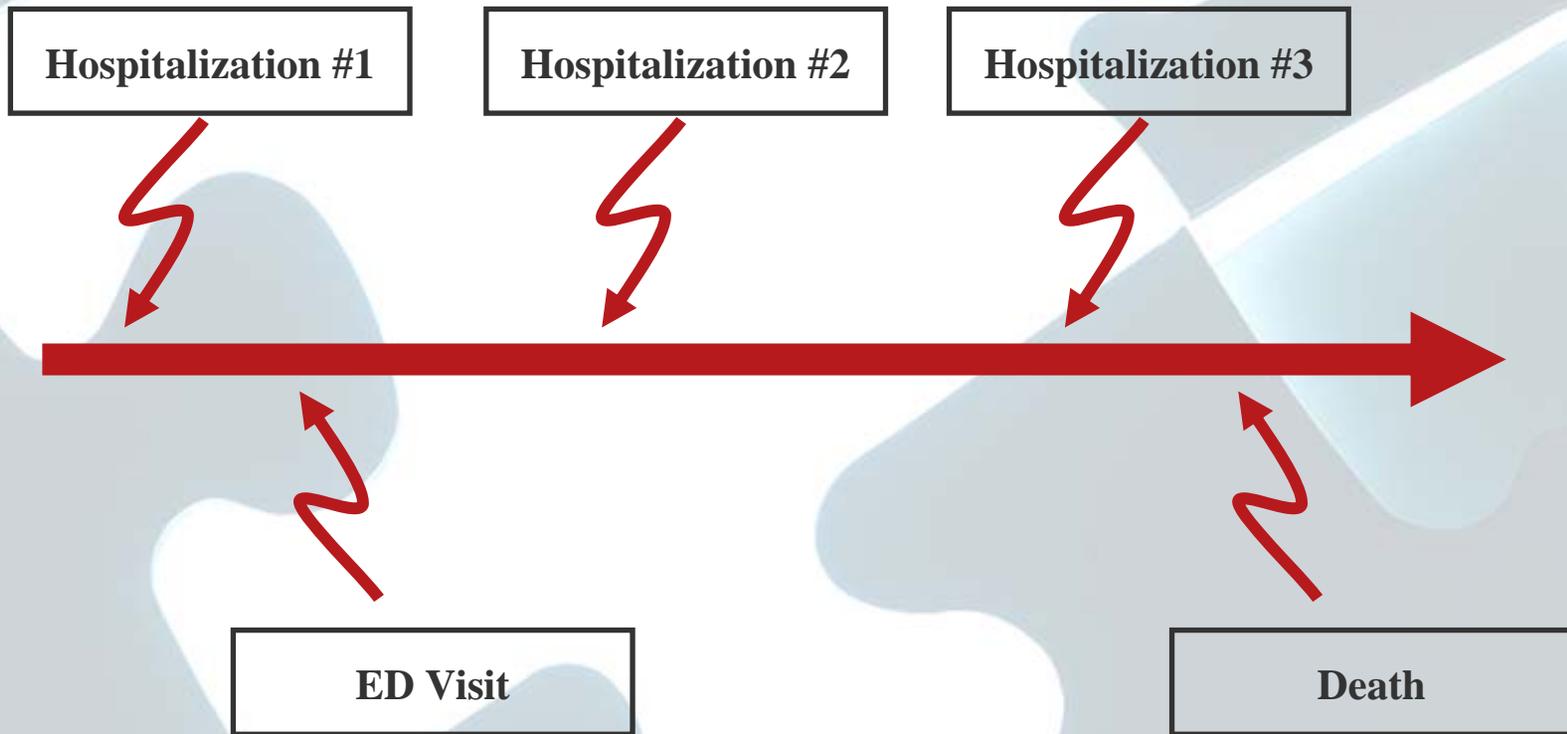
# Putting the Pieces Together

**The Linked PDD-Death Product**  
*More than you want to know*

**David Zingmond, MD, PhD**  
**Division of General Internal and Health  
Services Research**  
**UCLA School of Medicine**



# The Holy Grail: Creating a Longitudinal Description of Care





# Aims

Create a linked PDD-Death data product that would be

- validated
- available to end users
- manageable in size
- easily updated as new data became available



# Patient Discharge Database

- All hospital discharges from acute care hospitals in California
  - Includes all levels of care
  - Excludes Federal facilities and prison hospitals
- Results are compiled and released annually by OSHPD



# Death Statistical Master File

- All deaths occurring within California borders
  - Includes deaths of California residents that occur within other states
  - Excludes deaths of California residents that occur outside of the U.S.
  - Excludes non-California residents who die outside of California
- Results are maintained and released by the Department of Public Health, Office of Vital Statistics



# The PDD-Death Data Product

- **Probabilistic linkage** between the PDD and Death Statistical Master File using available personal identifying and ancillary information in each data set
- Each death reported in the DSMF is linked to exactly **one** discharge abstract from the PDD (the last identifiable hospitalization)
- Each file contains all hospitalizations in a given year where there was a subsequent death



# Deterministic and Probabilistic Linkage

- **Deterministic linkages** are based on exact matching of all merge variables
- **Probabilistic linkages** are based on exact matching of some merge variables (**blocking**) with scoring partial matches on the other merge variables



# The Challenge to Data Linkage

- Lack of a unique universal patient I.D.
  - Each health provider / payer assigns their own unique identifiers for internal use
  - Use secondary patient identifying information to assess care and outcomes within and across sites
  - Potential for errors in identifying information interferes with longitudinal assessment



# Patient Identifying Information

- Patient Discharge Database
  - SSN, gender, birth date, race/ethnicity, **Zip code**, expected source of payment, date of admission, **date of discharge, hospital death**, principle diagnosis
  - Hospital code, Hospital Zip code
- Death Statistical Master File
  - Name, SSN, gender, birth date, race/ethnicity, **Zip code, date of death**, location of death, cause of death.

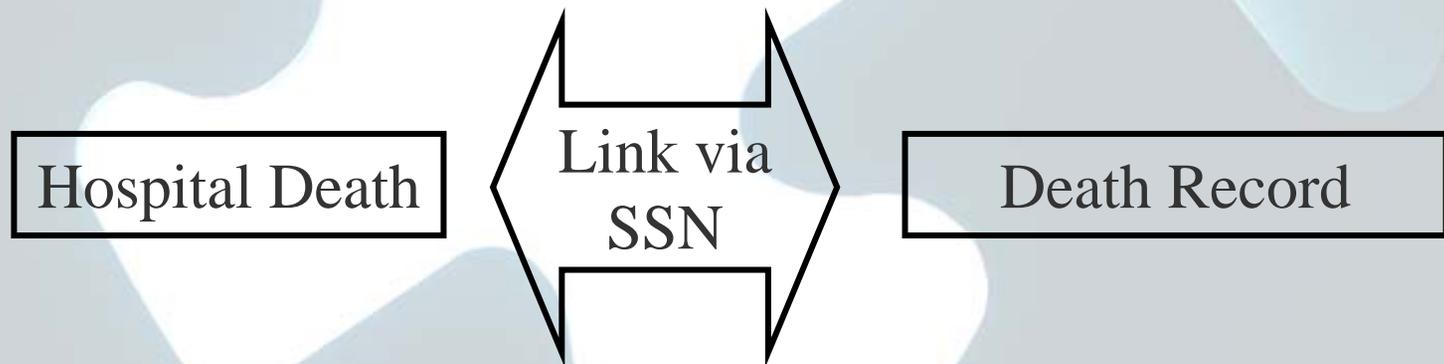


Why not just use a deterministic data linkage?



# Examine Hospital Deaths with Deterministic Linkage

- Linkage of PDD to DSMF death records deterministically using SSN + gender
- Evaluate linkage results for two years of date: 1995 and 1996





# Evaluation of Deterministic Linkage to Hospital Death

<b>Category</b>	<b># Records</b>	<b>% of Total</b>	<b>% with SSN</b>
Total Records	159,247	100.0%	
No SSN	7,221	4.5%	
With SSN	152,026	95.5%	100.0%
With SSN - DSMF Match	139,349	87.5%	91.7%
With SSN - No DSMF Match	12,677	8.0%	8.3%



# Deterministic Linkage Failure

1. SSN error in either DSMF or PDD
  - DSMF - death certificates transcribed
  - PDD - patient SSN incorrectly transcribed at first admission at a particular hospital
  - Use of SSN by spouse or other family member
2. SSN missing in either DSMF or PDD
  - unreported or invalid SSN



# Potential Sources of Bias

- Unmatched in-hospital deaths were more likely to be from:
  - minorities (11.2% AfrA vs 8.0 % white)
  - women (10.5% vs 6.0 % for men)
  - younger (18% for young adults)
  - Certain DRGs
    - trauma, stroke, severe mental illness
  - public hospitals



# Probabilistic Linkage - Methods



# Probabilistic Linkage - Methods

1. **Assume** - Unique identifiers have errors



# Unique identifiers have errors

- **Substitution Error**

*Correct Sequence*

\_\_\_ A \_\_\_

*Current Sequence*

\_\_\_ B \_\_\_



**Error**

- **Transposition (Switching) Error**

*Correct Sequence*

\_\_\_ A B \_\_\_

*Current Sequence*

\_\_\_ B A \_\_\_



**Switch Error**

- **Translocation (Insertion) Error**

*Correct Sequence*

A B C D E F G H I

*Current Sequence*

X A B C D E F G H



**Insertion Error, Other Digits Shift**





# Probabilistic Linkage - Methods

1. **Assume** - Unique identifiers have errors
2. **Find** - Other identifying information



# Find Other identifying information

- If unique, create alternative selection criteria
- If not unique, create additional evidence in support of matches by SSN that are close
- Other identifying information
  - Gender, Birth date, Race, Ethnicity, Zip code, Date of hospital admission



# Probabilistic Linkage - Methods

1. **Assume** - Unique identifiers have errors
2. **Find** - Other identifying information
3. **Create** - Potential matches via blocking



# Blocking

- In blocking, a subset of the identifying information (e.g. last four of SSN plus birth date) is exactly matched to create a limited set of records to compared.
- The subset of variables is the **Blocking Variables**
- The remaining identifying information (variables) are compared to determine a match.



# Potential Matches without Blocking

Database A
Record 1 (SSN <sub>1</sub> , S <sub>1</sub> , BD)
Record 2 (SSN <sub>2</sub> , S <sub>2</sub> , BD <sub>2</sub> )
Record 3 (SSN <sub>3</sub> , S <sub>3</sub> , BD <sub>3</sub> )

**X**

Database B
Record 1 (SSN <sub>4</sub> , S <sub>4</sub> , BD)
Record 2 (SSN <sub>5</sub> , S <sub>5</sub> , BD <sub>2</sub> )
Record 3 (SSN <sub>6</sub> , S <sub>6</sub> , BD)



SET OF ALL POTENTIAL MATCHES
Record <sub>A1</sub> - Record <sub>B1</sub>
Record <sub>A1</sub> - Record <sub>B2</sub>
...
Record <sub>A3</sub> - Record <sub>B3</sub>

**Number  
Potential  
Matches**

**A · B = 9**

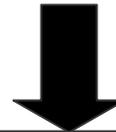


# Blocking\*

Database A
Record 1 (SSN <sub>1</sub> , S <sub>1</sub> , <b>BD</b> <sub>1</sub> )
Record 2 (SSN <sub>2</sub> , S <sub>2</sub> , <b>BD</b> <sub>2</sub> )
Record 3 (SSN <sub>3</sub> , S <sub>3</sub> , <b>BD</b> <sub>3</sub> )

X

Database B
Record 1 (SSN <sub>4</sub> , S <sub>4</sub> , <b>BD</b> <sub>1</sub> )
Record 2 (SSN <sub>5</sub> , S <sub>5</sub> , <b>BD</b> <sub>2</sub> )
Record 3 (SSN <sub>6</sub> , S <sub>6</sub> , <b>BD</b> <sub>1</sub> )



SET OF POTENTIAL MATCHES BY BD
Record <sub>A1</sub> - Record <sub>B1</sub>
Record <sub>A1</sub> - Record <sub>B3</sub>
Record <sub>A2</sub> - Record <sub>B2</sub>

**Number  
Potential  
Matches**

$$\sum_i A_i \cdot B_i = 3$$



# Blocking

- Choice of Blocking is important
  - If non-specific, it will create too many potential matches  $\Leftrightarrow$  computationally intensive
- Multiple steps of Blocking must be used
  - If a true match record (e.g. DSMF or PDD) has an error in the blocking variable(s), then the true match will not appear in the set of potential matches



# Examples of Blocking Arrangements

- SSN alone
- SSN plus Sex
- Last Four of SSN + Birth Year
- Gender + Birth Date + Race + Ethnicity



# Probabilistic Linkage - Methods

1. **Assume** - Unique identifiers have errors
2. **Find** - Other identifying information
3. **Create** - Potential matches via blocking
4. **Grade** - Matches via scoring algorithm



# Grading Potential Matches

- After potential matches are generated, each record pair must be graded as to the goodness of the match.
- Goodness of match is rated by scoring the agreement between variables not used in blocking (the **Matching Variables**).



# Probabilistic Linkage - Methods

1. **Assume** - Unique identifiers have errors
2. **Find** - Other identifying information
3. **Create** - Potential matches via blocking
4. **Grade** - Matches via scoring algorithm
5. **Select** - Matches via selection score



## Decide if Potential Match is a Probable Match

- Keep highest scoring record pair for each PDD and each death record
- Accept linked record pairs whose match score is greater than a pre-determined threshold score equivalent to a match on SSN, Birth date, and Gender



# Probabilistic Linkage - Methods

1. **Assume** - Unique identifiers have errors
2. **Find** - Other identifying information
3. **Create** - Potential matches via blocking
4. **Grade** - Matches via scoring algorithm
5. **Select** - Matches via selection score
6. **Correct** - **Remove inconsistent matches**



# Implementation



# Overview of Linkage Algorithm

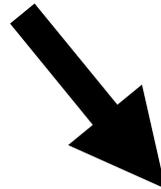
**By DSMF Year**

**By PDD Year**

**Match DSMF to PDD**



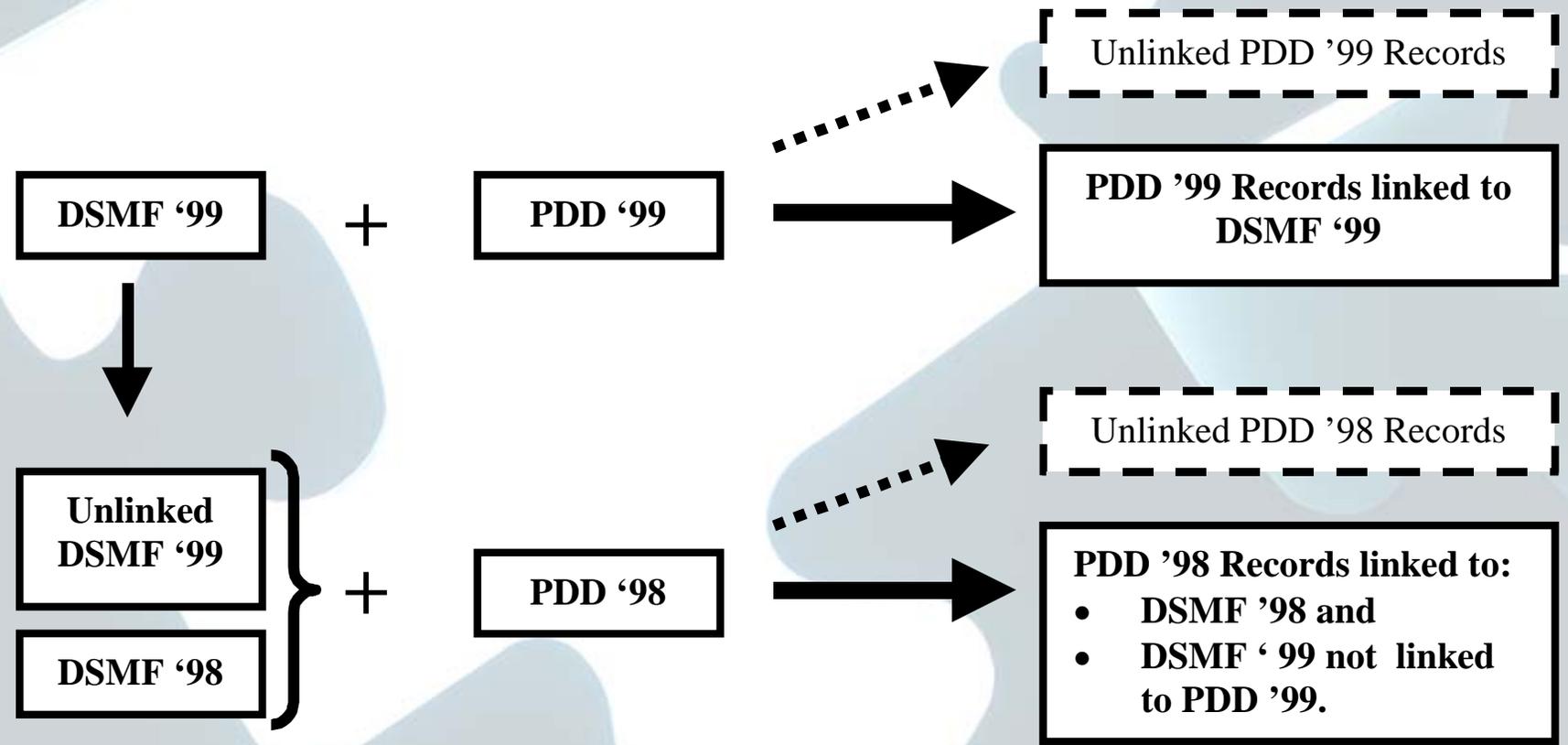
**Remove  
Matched  
Records**



**Unmatched DSMF records  
used for matching to  
earlier years  
("Remainder")**

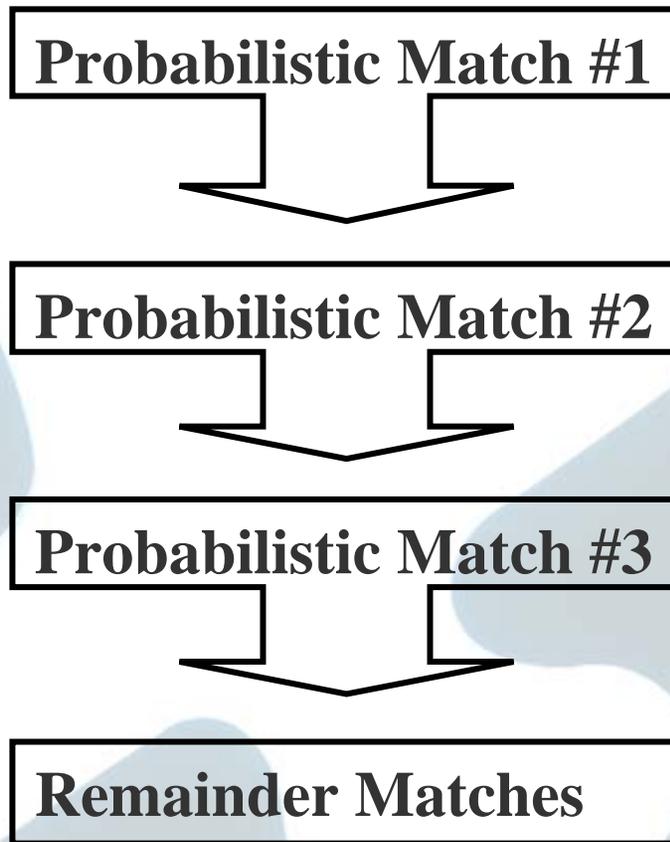


# Example of Linkage Algorithm





# Matching Within Year



**Sequential matching using different blocking variables**

**Attempt to match the unmatched hospital deaths.**



# Implementation - Matching

	<b>Blocking Variables</b>	<b>Matching Variables</b>
1.	9 Digit SSN, Gender	Birth Date, Race/Ethnicity, Zip code
2.	Birth Date, 1 <sup>st</sup> Three of Zip	1 <sup>st</sup> Five of SSN, Last Four of SSN, Race/Ethnicity
3.	1st Five of SSN, Year of Birth	Last Four of SSN, Gender, Race/Ethnicity, Zip code
4a.	Birth Date, Death Date, Gender, Hospital Death	SSN, Race/Ethnicity, Zip code
4b.	Birth Date, Death Date, Gender, Hospital Death	Race/Ethnicity, Zip code



# Linkage Results



# Linkage Performance

Blocking	Hospital Death		Discharged Alive		Total Records
	SSN	No SSN	SSN	No SSN	
#1	63,738	-	64,926	-	128,664
#2	1,513	-	1,738	-	3,251
#3	355	-	334	-	689
#4a	804	-	-	-	804
#4b	-	1,415	-	-	1,415
<b>Overall</b>	66,410	1,415	66,998	-	134,823
<b>Unmatched</b>	3,347	4,265	1,721,703	-	1,729,315
<b>Total</b>	69,757	5,680	1,788,701	-	1,864,138

Comparison of linkage of death records to known alive and dead (hospital death) patients for 1 year - 1997 hospitalization to 1997 & 1998 death. **Hospital Death** = Death reported during hospitalization in the PDD; **Discharged Alive** = Patient reported as discharged alive.



# Accuracy of the Linkage Algorithm

- Compare linkage results to a gold standard of “known deaths” and “known survivors”
  - Known deaths = hospital deaths
  - Known survivors = individuals with readmissions
- Estimate sensitivity, specificity, and accuracy of the linked records



# Gold Standard Evaluation

Blocking	True Positive	False Positive	False Negative	True Negative	Sens	Spec	Accu
#1	63,738	14	5,993	372,605	0.91406	0.99996	0.98642
#2	1,513	29	4,480	372,576	0.25246	0.99992	0.98809
#3	355	42	4,125	372,534	0.07924	0.99989	0.98895
#4a	804	0	3,321	372,534	0.19491	1.00000	0.99118
<b>Overall</b>	66,410	85	3,321	372,534	0.95237	0.99977	0.9923

#1: SSN + Gender; #2: Birth Date + 1<sup>st</sup> 3 Digits of Zip; #3: Last 4 of SSN + Year of Birth; #4a: Hospital Death + Birth Date + Death Date + Gender

**True Positive (TP)** – Linked Hospital Death; **False Positive (FP)** – Linked Known Alive; **False Negative (FN)** – Unlinked Hospital Death; **True Negative (TN)** – Unlinked Known Alive; **Sens** – Sensitivity (TP/TP+FN); **Spec** – Specificity (TN/TN+FP); **Accu** – Accuracy (TP+TN / TP+TN+FP+FN)

**Gold Standard:** Known Deaths (Hospital Deaths) + Known Alive (Readmitted in 1998)



# Demographics of Unlinked Hospital Deaths

	Number of Hospital Deaths					% Hospital Deaths Unmatched to DSMF				
	White	Black	Hisp	Asian	Other	White	Black	Hisp	Asian	Other
<b>Men</b>										
< 1 Year	12	6	34	5	4	20	0	13	0	33
1 to 34 Years	566	175	488	95	57	5	10	9	6	16
35 to 64 Years	5,336	1,129	1,817	657	267	3	6	6	4	5
65 to 84 Years	14,364	1,112	2,057	1,374	424	2	4	4	3	2
85 Years & Older	4,256	261	546	387	124	2	5	3	2	3
<b>Women</b>										
< 1 Year	14	7	34	9	5	13	13	21	10	29
1 to 34 Years	376	126	259	78	36	6	7	8	5	8
35 to 64 Years	4,223	991	1,179	493	180	3	6	5	3	5
65 to 84 Years	12,969	1,263	1,925	1,182	398	6	8	6	4	7
85 Years & Older	6,664	505	770	359	159	8	9	8	6	9



# Data Product



# Data Product Format

- Files are arranged by year (y) of PDD.
- PDD records are linked to every subsequent year (x) of DSMF data that are available.
  - PDD(y) linked to DSMF (x:  $x \geq y$ )
  - e.g. PDD 1999 would be linked to DSMF 1999 to DSMF 2007
- Annual updates replace prior versions due to greater information (knowledge of subsequent readmissions and deaths).



# Final Merged Data Product Format

- PDD - Full PDD Record of last known hospitalization
- DSMF - Full DSMF Record
- Derived Variables
  - Time to Death (Continuous, Categorical)
  - Blocking Variable (1 to 4B)
  - Probability Matching Score



# Conclusions

- For records with SSN, data linkage captures ~95% of known deaths.
  - Unmatched records occur more often in the young, in the very old, in minorities, and in women.
  - Inconsistencies in linkages may be more apparent on specific projects
- Data linkage is unreliable for records without SSN
  - Improved SSN requirements and reporting may help
  - Additional identifying information (e.g. names) would improve matches



os**h**pd

Office of Statewide Health Planning & Development



# Grading Potential Matches

- The score for agreement on each matching variable is calculated as:

$\log (M/U)$  if there is a match

$\log ((1-M)/(1-U))$  if there is not a match

M = Mismatch Rate = 1 - Error Rate

U = Probability of a Match



# Grading Potential Matches

- Total Score for a Potential Match
  - Add up the scores for each matched and unmatched digit in matching variables
  - Overall Score =  $\text{Score}_{\text{var}\#1} + \text{Score}_{\text{var}\#2} + \dots$

## SET OF POTENTIAL MATCHES BY BD

$\text{BD} = \text{BD}_1$        $\text{Record}_{\text{A}1} - \text{Record}_{\text{B}1} : \text{Score}_{11}$

$\text{Record}_{\text{A}1} - \text{Record}_{\text{B}3} : \text{Score}_{13}$

$\text{BD} = \text{BD}_2$        $\text{Record}_{\text{A}2} - \text{Record}_{\text{B}2} : \text{Score}_{22}$



## Select Potential Match with Highest Score

### **SET OF POTENTIAL MATCHES BY BD**

- BD = BD<sub>1</sub>**
1. Record<sub>A1</sub> - Record<sub>B1</sub> : **Score<sub>11</sub>**
  2. Record<sub>A1</sub> - Record<sub>B3</sub> : **Score<sub>13</sub>**

**Score<sub>11</sub> > Score<sub>13</sub> ...**

**Match #2 is the Best “Potential Match” for records matching on BD<sub>1</sub>**



# Unique Matches

- Key requirement: DSMF-PDD matches must be unique (one-to-one match)
  - Death record corresponds to a single discharge record
- First, by death record, retain highest scoring potential DSMF-PDD match
- Then, by PDD record, retain highest scoring potential DSMF-PDD match