

Inter-Registry Record Linkage without Releasing Patient Level Identifiers



Jill A. MacKinnon, PhD, CTR
Florida Cancer Data System

Co-Authors

- Paul Sterns - ACE
- Dr. Maria J. Schymura - NY
- Betsy Kohler - NAACCR



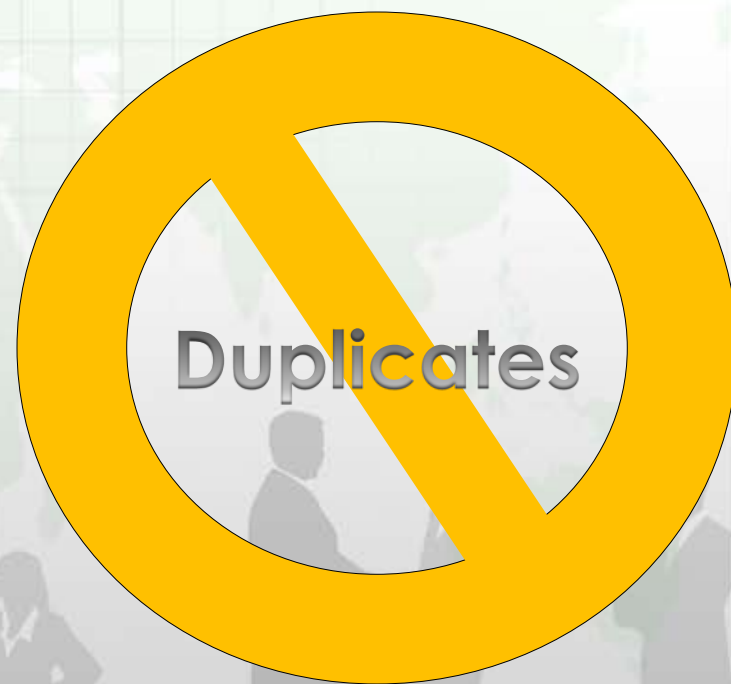
NAACCR Workgroup

- Betsy Kohler
- Glenn Copeland
- Jill MacKinnon
- Andy Lake
- Lilia O'Connor
- Maria Schymura
- Royal Anne Hinds
- David Roney



Background

- Patient/cancer duplicate reporting
 - Intra state duplication
 - CCR
 - Inter state
 - National duplication
- Incorrect rates



Magnitude

- What is the magnitude of over reporting
 - CCR
 - Claiming cases that aren't yours
 - National Level
 - Reporting same case from multiple CCR's



Multiple Challenges

- CCR authority to release patient identifiers
- Policies
 - How to do this
 - What to do once duplicates are found
 - Who's incident case is it



CCR Authority to Release Patient Identifiers



What is One-way Encryption

- A function that creates a digital summary of data (called 'hashes')
 - Most password files are stored this way
 - Confidential data fields will be comprised of hashes not data
- File/data can't be decrypted (unlike two-way encryption)

Encrypted Data

Smith

00159b415aa08a9590f5ee6503dbe1c004
cc7fd42894a57333284aa768670d36



Deterministic Linkage

First Test

- New York and Florida
 - Duplicates as identified by NDI match
 - 2,740 NY cases linked with 2,135 FL cases
 - Industry standard probabilistic match
- Permission by NY to use these as test cases (gold standard)
- Each file was run through same encryption routine

Variables in File

- Registry ID
- Record key
- First name (e)
- 1st letter of First name
- Last name (e)
- 1st letter of last name
- SSN (e)
- Date of birth (e)
- Year of Birth
- Sex
- Race

First Test Results

Central Clearing House - Detecting Duplicate Reports Between CCR's

Deterministic Matching Passes

Variables	1	2	3	4	5	6	7	8	Unmatched	Total
e-SS#	x	x	x		x	x	x	x		
e-last name	x	x		x			x			
e-First Name	x		x							
e-DOB	x	x	x	x	x					
Yr of Birth						x				
First Letter Last Name						x				
First Letter First Name						x				
N Cases Matched	1,703	91	66	104	7	100	50	4	10	2,135
% Matched by Pass	79.8	4.3	3.1	4.9	0.3	4.7	2.3	0.2	0.5	100.0

Second Test

- Larger datasets
 - How does the methodology perform
 - Tweaked passes
- FL patient file vs. FL 2007 mortality file
 - FCDS patient file n= 2,361,467
 - 2007 Mortality n = 173,657
- Found 59,050 of the 60,682
 - ~ 300 of 1,632 cases – DC procedural issues

Second Test Results

Central Clearing House - Detecting Duplicate Reports Between CCR's										
Deterministic Matching Passes - 1/12/10 - Results**										
Variables	1	2	3	4	5	6	7	8	9	
e-SS#	x	x	x	x	x	x	x	x	x	
e-last name	x	x		x				x	x	
e-First Name	x		x		x	x			x	
e-DOB	x	x	x		x		x		x	
Yr of Birth				x		x		x		
First Letter Last Name			x			x				
First Letter First Name				x						
Sex				x		x	x	x		
Race		x	x	x	x	x	x	x		
N Cases Matched	49,532	3,724	1,165	1,403	471	46	224	17	2,468	59,050
% Matched by Pass	83.9	6.3	2.0	2.4	0.8	0.1	0.4	0.0	4.2	
False Positives	0	0	0	0	0	0	0	0	0	
** Eliminated any unknown SS # from eligibility										
2007 Mortality File = 173,657										
FCDS Patient File = 2,361,467										
N Actual Matches 60,682 as of 11/16/09										
Known FCDS Matches but didn't match because of										
1) Unknown SS										
2) Procedural issue with FCDS										

As Epidemiologists We Must Have 2 x 2 Tables

	Truth per FCDS			
Encryption	Positive	Negative	Total	PV
Pos	59,050	0	59,050	100
Neg	1,632	112,975	114,607	98.5
	60,682	112,975	173,657	
Total	97.3 <i>Sensitivity</i>	100 <i>Specificity</i>		

Next Steps

- Expand encrypted linkage to two or more states to expand proof of concept
 - Challenges with approvals for raw data release



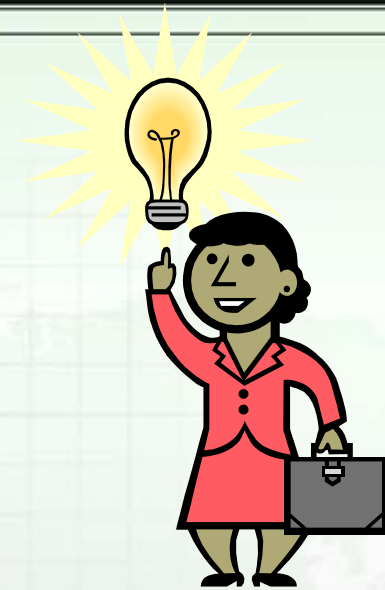
- NY and FL
 - Attempt to assign residence of the matched records

How to Determine the Incident Case?

- Rules/standards (To be established)
 - Electronic
 - Manual
- How much manual intervention between CCR



Prediction



Linking will be the easy part



Future???

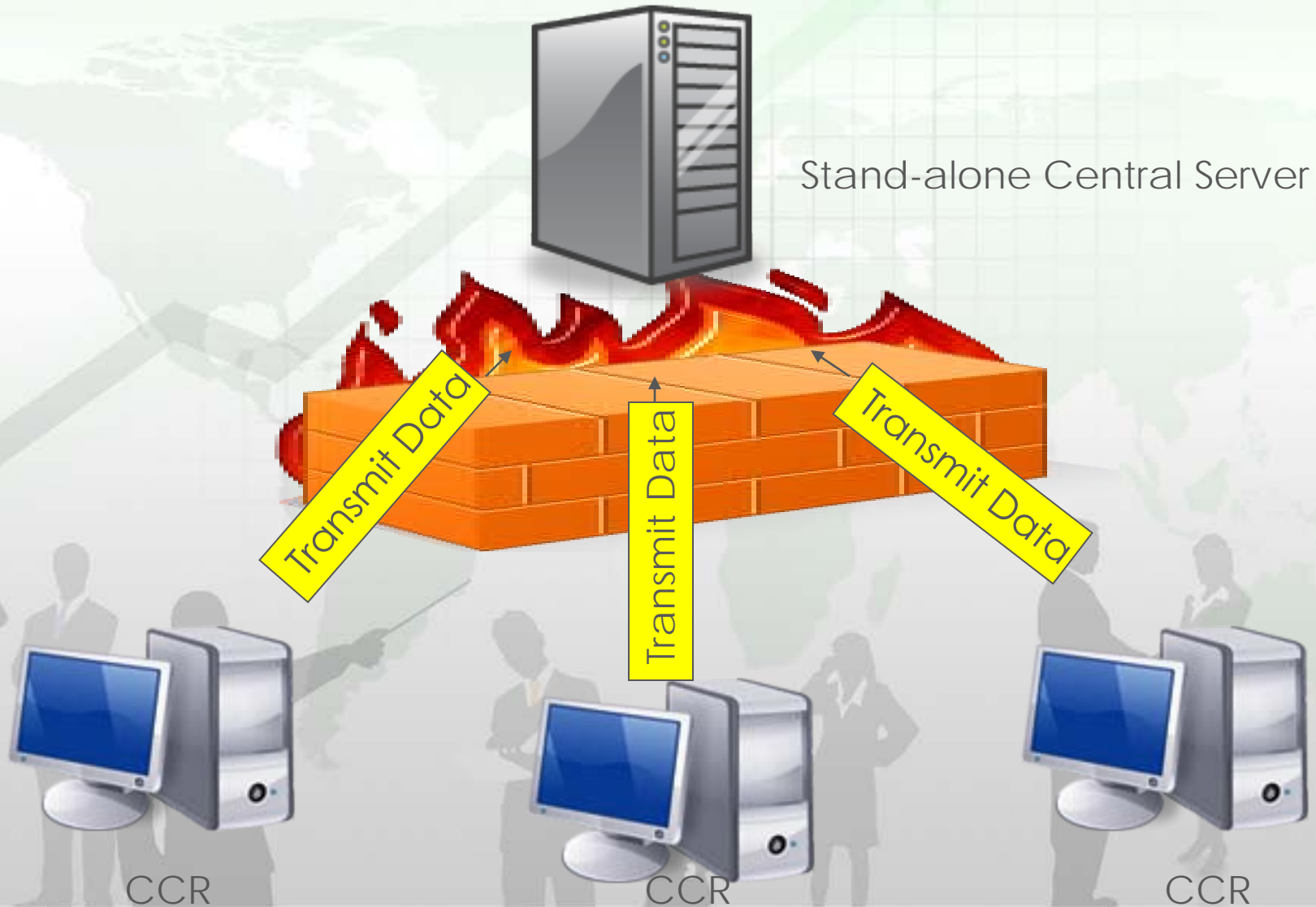
- Web-based
- Server behind firewall
- Data are one way encrypted
 - Encryption takes place on sender server
- Transmitted via SSL
- Authentication software

Data Request

- All CCR's encrypt data prior to transmitting to central server
 - Using same algorithm



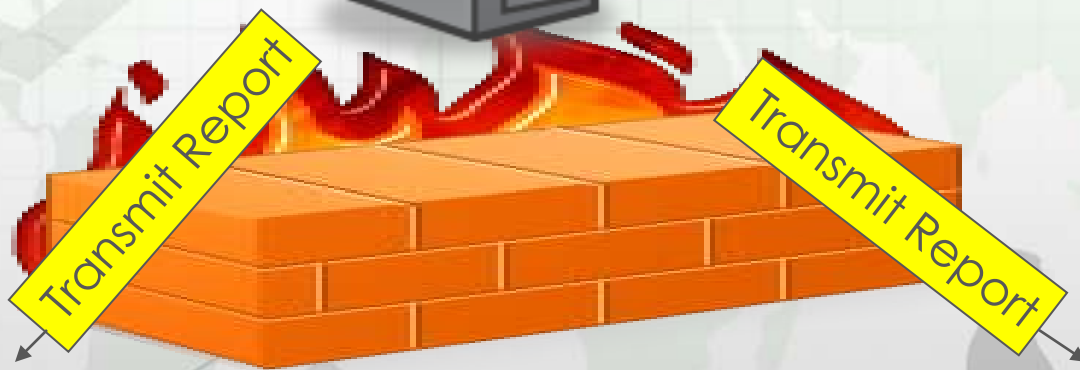
One Way Encryption



Results of Matching - Report



Stand-alone Central Server



Matched Registry



Matched Registry



Thank You