PA-NEDSS: Advances in Deduplication of Patients

Kirsten Waller, MD, MPH
Steven Kowalewski
Kevin Jastrzemski
Paul Heter

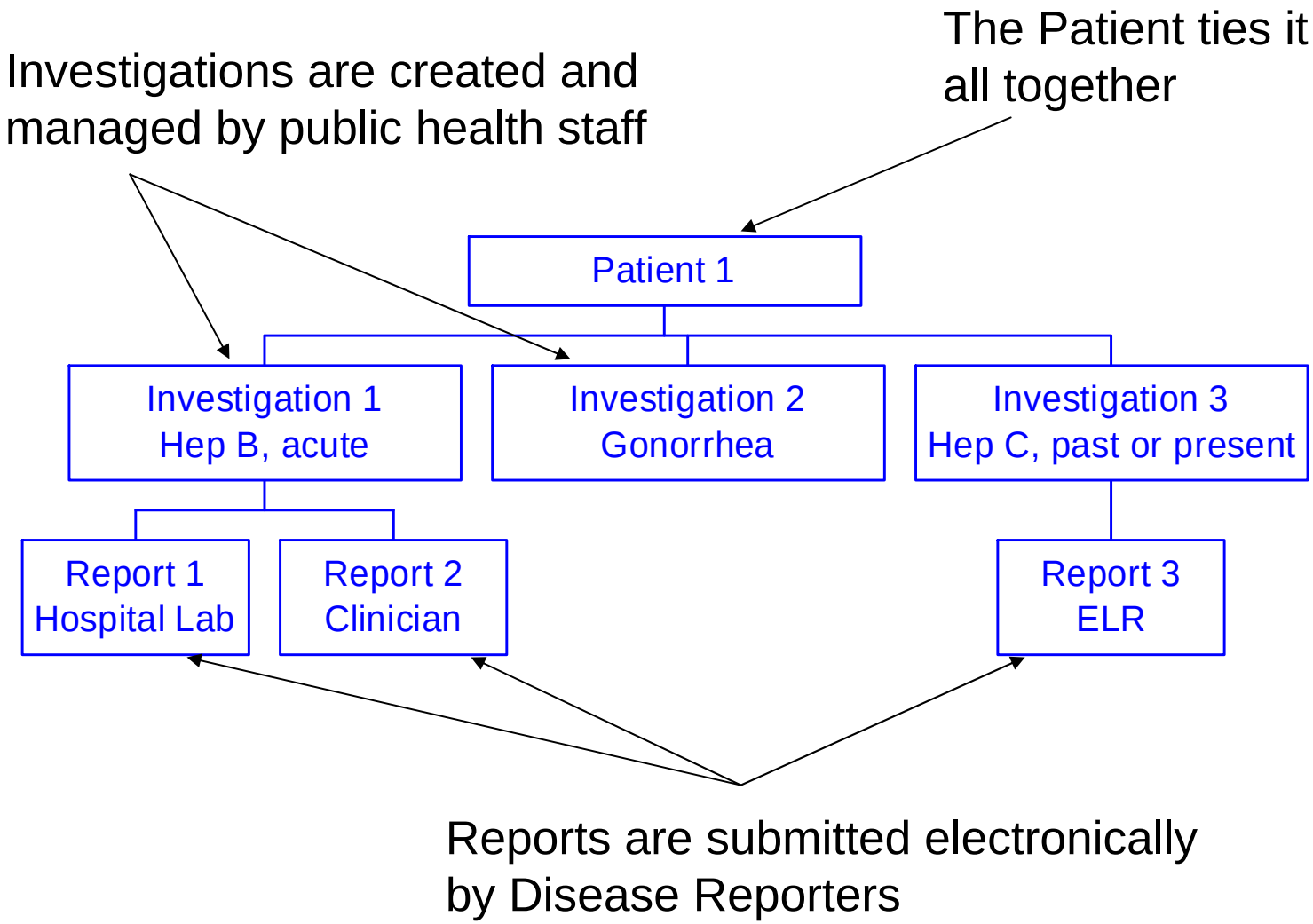
PA-NEDSS Features

- On-line reporting by laboratories, hospitals, and clinicians
- Integrated Electronic Laboratory Reporting
- Accepts all PA reportable infectious diseases except HIV/AIDS, also includes lead
- Used “for real” since Jan 2003

Lots of Information

- 350+ hospitals
- 200+ labs
- 40 ELR labs
- Approx 20,000 new Reports received per month

Functional Organization of PA-NEDSS



PA-NEDSS: Online Disease Reporting - Microsoft Internet Explorer provided by PA Dept of Health

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.nedsstraining.state.pa.us/nedss/asp/diseasereports/patient_hra_management.asp Go Links

Welcome | Administration | Disease Reporting | Alerts | Analysis & Reports | Log Off

Search | Inboxes | Patient ID: 5404

User: MarkInv **Patient:** Fred Flintstone **DOB:** 09/01/1955 **Session Expires:** 29 minutes Renew

Location: Disease Reporting > Patient Reports/Investigations Management

Patient Reports/Investigations Management

Patient Name: Fred Flintstone **Patient ID:** 5404 **Gender:** Male **Birth Date:** 09/01/1955

Primary Residence: 100 Main St, Pittsburgh, PA 15215 **Phone Number:** None Available

Unlinked Reports

Report ID	Condition	Date of Onset	Reported Date	Initial Reporter	Responsible Jurisdiction	Link To
No unlinked reports						

Investigations

Investigation ID	Condition	Investigation Status	Case Status	Assigned To	Responsible Jurisdiction
3100 [View] [Edit]	Botulism, foodborne	Open	Under investigation	Jennifer Hollingsworth	Allegheny
3092 [View] [Edit]	Hepatitis A	Closed	Confirmed case	Mark Investigator2	Dauphin

To log a new Investigation for this patient, please select the Condition and press the "New Investigation" button.

- Condition - New Investigation

Linked Reports

Report ID	Condition	Date of Onset	Reported Date	Initial Reporter	Responsible Jurisdiction	Linked Investigation ID
3091 [View] [Edit]	Hepatitis A	01/09/2004	01/30/2004	Dr Jack Brown	Dauphin	3092 Unlink

Internet

Start | Inbo... | N:\S... | Ded... | PA-E... | Welc... | A&R ... | PHIN... | SAS ... | PA-N... | ELRe... | 3:35 PM

Problem

- How does PA-NEDSS decide if a newly-entered patient is the same as an existing patient?
- If “sameness” is not recognized, patient’s history will not be accessible to investigator
- Result:
  - Duplicate Investigations started
  - Cases double-counted
  - Wasted staff time

4 Points Where Deduplication Occurs

1. At patient entry (automatic)
2. At Inbox review (investigator-managed)
3. Batch review (central office-managed)
4. Ad hoc Custom/expert deduplication tool

Automatic Deduplication

- New patients are merged with existing patient if the following fields match exactly:
  - First and last name, DOB, Gender

  * ELR: First and last name, DOB, Gender match to deduped patient

  * ELR: First and last name, gender, accession number, lab (within 6 mos)

  # First/last name reversals, DOB, gender

  # SSN, DOB

Limitations of Automatic Deduplication

- Many users inputting data = typos, variant spellings, missing data
- Have resisted probabilistic matching of patients

Investigator-Managed Deduplication

- Entry in Reports Inbox is flagged as a *Possible Duplicate*
- Investigator needs to decide if patients are same or different before proceeding

PA-NEDSS: Online Disease Reporting - Microsoft Internet Explorer provided by PA Dept of Health

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.nedsstraining.state.pa.us/nedss/asp/diseasereports/report_list.asp?Filter=ALL&SubFilter=ALL_UNLINKED Go Links

Welcome | Administration | Disease Reporting | Alerts | Analysis & Reports | Log Off

Search | Inboxes

User: MarkInv **Session Expires:** 29 minutes Renew

Location: Disease Reporting > Inbox: All Reports > All Unlinked

Inbox: All Reports > All Unlinked

View Active Inbox Filters

Patient Name	Condition	Report Source	Reported Date	Report ID	Assigned To	Linked Investigation ID	Inv Status	Case Status	Juri
KimOne Catrall	Tuberculosis	SYSTEM ACCOUNT	05/09/2004	3420 [View] [Edit]	Unassigned Assign	Unlinked Start Investigation	N/A	N/A	Chester
Kim Catrall	Tuberculosis	SYSTEM ACCOUNT	05/09/2004	3418 [View] [Edit]	Unassigned Assign	Unlinked Start Investigation *Possible Duplicates*	N/A	N/A	Chester
Cole Schaeffer	Syphilis	SYSTEM ACCOUNT	05/09/2004	3406 [View] [Edit]	Unassigned Assign	Unlinked Start Investigation *Possible Duplicates*	N/A	N/A	Chester
JohnOneOneOne Jones	Typhoid fever	SYSTEM ACCOUNT	05/09/2004	3404 [View] [Edit]	Unassigned Assign	Unlinked Start Investigation *Possible Duplicates*	N/A	N/A	Lancaster
ColeOneOne Schaeffer	Syphilis	SYSTEM ACCOUNT	05/09/2004	3398 [View] [Edit]	Unassigned Assign	Unlinked Start Investigation	N/A	N/A	Chester
JohnOneOne Jones	Typhoid fever	SYSTEM ACCOUNT	05/09/2004	3399 [View] [Edit]	Unassigned Assign	Unlinked Start Investigation	N/A	N/A	Lancaster
ColeOne Schaeffer	Syphilis	SYSTEM ACCOUNT	05/09/2004	3393 [View] [Edit]	Unassigned Assign	Unlinked Start Investigation	N/A	N/A	Chester
JohnOne Jones	Typhoid fever	SYSTEM ACCOUNT	05/09/2004	3394 [View] [Edit]	Unassigned Assign	Unlinked Start Investigation	N/A	N/A	Lancaster
Mike Ditka	Syphilis	Renee Laboratory	04/27/2004	3336 [View] [Edit]	Arlene Muszynski	Unlinked Start Investigation	N/A	N/A	Chester

Done Internet

Start Inbox... N:\Su... Dedu... PA-EL... PA-N... A&R ... PHIN ... SAS -... PA-N... 4:18 PM

PA-NEDSS: Online Disease Reporting - Microsoft Internet Explorer provided by PA Dept of Health

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.nedsstraining.state.pa.us/nedss/asp/diseasereports/dupl_patients_for_report.asp Go Links

Welcome | Administration | Disease Reporting | Alerts | Analysis & Reports | Log Off

Search | Inboxes | Patient ID: 5687 | Report ID: 3404

User: MarkInv **Patient:** JohnOneOneOne Jones **DOB:** 01/01/1979 **Session Expires:** 29 minutes Renew

Location: Disease Reporting

You may click on the patient name to see full details.

Patient(s) already in PA-NEDSS

Name: Jones, JohnOneOne **Gender:** Male Combine with this Patient

Birth Date: 1/1/1979 **Race:** Black or African American **Ethnicity:** Not Hispanic or Latino

Location Information

Address Type	Address	Verified	Phone/Fax	Start Date	End Date
House/Home Primary Residence	251 S QUEEN ST LANCASTER PA 17603	Yes	[p] 717-874-9955		

Previous Names and Alias Information

Previous Names	Alias Names
No Previous Names	No Aliases

Patient on the New Report:

Name: Jones, JohnOneOneOne **Gender:** Male Set As Different Patient

Birth Date: 1/1/1979 **Race:** Black or African American **Ethnicity:** Not Hispanic or Latino

Location Information

Address Type	Address	Verified	Phone/Fax	Start Date	End Date
House/Home Primary Residence	251 S QUEEN ST LANCASTER PA 17603	Yes	[p] 717-874-9955		

Previous Names and Alias Information

Previous Names	Alias Names
No Previous Names	No Aliases

Internet

Start | Inbox... | N:\Su... | Dedu... | PA-EL... | PA-N... | A&R ... | PHIN ... | SAS -... | PA-N... | 4:24 PM

Investigator-Managed Deduplication

- Patients flagged as possible duplicate if:

Soundex of first and last name match and DOB algorithm is met:

New pt	Compared pt	Criterion
Exact	Exact	2/3 DOB components
Exact	Est.	+/- 3 years
Est.	Exact	+/- 3 years
Est.	Est.	+/- 3 years

Soundex

- First initial, followed by 3 numbers representing consonants
  1 B P F V
  2 C S K G J Q X Z
  3 D T
  4 L
  5 M N
  6 R
- Kirsten = K623 Waller = W460

Disadvantages of Soundex

- Most errors we receive are typos, not sound-alike errors
- Pairs:
  Mark / Mowers
  Samuels / Smalich
  Spengler / Spangenberg
- Misses:
  Truong / Trvong
  Kamlick / Kramlick
  Lander / Zander

Soundex Alternatives

- Edit distance, probabilistic matching, etc.
- Computationally intensive
- Not perfect
- COTS vs. home-grown

Batch Deduplication

- Searches with different criteria run nightly
- All pairwise comparisons =

  $$\sum_{i=1}^{n} (i-1) \qquad \text{(i.e. big number)}$$

- Opening up this functionality to more users
- Works best if reviewed daily

PA-NEDSS: Application Administration - Microsoft Internet Explorer provided by PA Dept of Health

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.nedsstraining.state.pa.us/nedss/asp/administration/deduplication.asp Go Links

HEALTH

Electronic Disease Reporting

Welcome | Administration | Disease Reporting | Alerts | Analysis & Reports | Log Off

User: MarkInv **Session Expires:** 26 minutes Renew

Location: Administration > Patient De-Duplication Search

Patient De-Duplication Search

Batch De-Duplication

Please choose from one of the search types below. After choosing a search type you will be able to further limit the search results that you want to see by First Name, Last Name, and when the patient record was created.

Last Name	First Name	Date of Birth	Choose
Exact	Exact	Exact	○
Exact	Exact	+/- 3 Years	◉
Exact	Soundex	+/- 3 Years	○
Exact	Exact	*Not applicable*	○
Soundex	Soundex	+/- 3 Years	○
Soundex	Soundex	*Not applicable*	○
Soundex	First 3 Character Match	*Not applicable*	○

Limit Result Groups By

☐ Last Name

☐ First Name

☑ Date Created From - Month - - Day - YYYY To May 19 2004

Search Reset

Done Internet

Start In... C:... W... we... Re... He... PH... SA... Re... PA... PH... 10:40 PM

PA-NEDSS: Application Administration - Microsoft Internet Explorer provided by PA Dept of Health

File Edit View Favorites Tools Help

Back Search Favorites Media

Address https://www.nedss.state.pa.us/nedss/asp/administration/deduplication.asp Go

Commonwealth of Pennsylvania
Electronic Disease Reporting

Welcome | Administration | Disease Reporting | Alerts | Analysis & Reports | Log Off

User: kwaller **Session Expires:** 28 minutes Renew

Location: Administration > Patient De-Duplication Search

Patient De-Duplication Search

Batch De-Duplication

Please choose from one of the search types below. After choosing a search type you will be able to further limit the search results that you want to see by First Name, Last Name, and when the patient record was created.

Last Name	First Name	Date of Birth	SSN	Choose
Exact	Exact	Exact		○
-	-	-	Exact	○
Exact	Exact	Close	-	○
Soundex	Soundex	Close	-	○
First 5 char match	First 1 char match	Close	-	○
Exact	Exact	+/- 3 Years	-	○

Limit Result Groups By

- ☐ Last Name
- ☐ First Name
- ☐ Date Created
- ☐ County

Done Trusted sites

Start | Sheetz - M... | PA-NEDS... | Salm 2005... | C:\Outbre... | PA-NEDSS... | FW: Sheet... | SAS - [VIE... | 10:21 PM

Possible Patient Duplicates

Patient Group 1 (Select for Auto Merge ☐) Merge

Patient ID	Name	Sex	Birth Date	Race	Ethnicity	Address	Exclude
1900	Brown, John	Male	2/10/1960	American Indian or Alaskan Native		PA 17101	
5593	Brown, John	Male	8/12/1962			PA 17101	

Patient Group 2 (Select for Auto Merge ☐) Merge

Patient ID	Name	Sex	Birth Date	Race	Ethnicity	Address	Exclude
5597	Catrall, Kim	Female	5/2/1961 [estimated]	Asian		114 Laurel Dr Coatesville PA 19320	
5689	Catrall, Kim	Female	5/2/1961	Asian		114 LAUREL DR COATESVILLE PA 19320	

Patient Group 3 (Select for Auto Merge ☐) Merge

Patient ID	Name	Sex	Birth Date	Race	Ethnicity	Address	Exclude
5613	ditka, mike	Male	5/12/1945 [estimated]	White		112 Laurel Dr Coatesville PA 19320	
5615	Ditka, Mike	Male	5/12/1945 [estimated]	White		112 Laurel Dr Coatesville PA 19320	

Patient Group 4 (Select for Auto Merge ☐) Merge

Patient ID	Name	Sex	Birth Date	Race	Ethnicity	Address	Exclude
3632	Dyer, Jack	Male	4/1/1997	Black or African American	Hispanic or Latino	PA 17104	
5416	Dyer, Jack	Male	1/7/1996	Native Hawaiian or Other Pacific Islander		PA 17101	

Patient Group 5 (Select for Auto Merge ☐) Merge

PA-NEDSS: Application Administration - Microsoft Internet Explorer provided by PA Dept of Health

File Edit View Favorites Tools Help

Back Address Go

Auto Merge De-duplication Rules - Microsoft Internet Explorer provided by PA Dept of Health

Auto Merge De-duplication Rules

The rules for the Automated Merge are implemented for each particular data item type. These rules are as follows.

Information Item Type Fields	Merge Rules
Patient Prefix	The patient prefix is irrelevant and will be set to null
First, Last Name, Birth Date, Gender, Race, Ethnicity, Middle Name, Highest Level of Education, Suffix	These fields will be transferred to the resulting de-duplicated patient from one of the patients being auto-de-duplicated. A scoring algorithm will be used to determine from which patient these fields should be taken. This algorithm is described in detail below this table.
Primary Language, Marital Status, US Citizen, Country of Citizenship, US Entry Date, US Entry Class	Only one value will be picked for each of these data items. Information with a non-null entry will take precedence over all null entries for all the data items. If more than one non-null entry is present, the entry from the patient from whom the demographic information was kept will be kept.
Name of School Attended	Information with a non-null entry will take precedence over all null entries. If more than one non-null entry is present, the entry from the patient from whom the demographic information was kept will be kept.
Name of School District	Information with a non-null entry will take precedence over all null entries. If more than one non-null entry is present, the entry from the patient from whom the demographic information was kept will be kept.
Daycare Attended	Information with a non-null entry will take precedence over all null entries. If more than one non-null entry is present, the entry from the patient from whom the demographic information was kept will be kept.
Employment History (all the information that is currently collected for each employment information item)	All the employment items for all the selected patients will be kept on the merged patient.

Home | Top | Print Page Contact Us | DOH Security Officer Help

manual_dedupe.asp :: 7.2.2 :: WC2 :: 1292

Trusted sites

Start | Sheet... | PA-NE... | Salm 2... | C:\NE... | PA-NE... | FW: S... | SAS - ... | Dedup... | Auto ... | 10:28 PM

PA-NEDSS: Application Administration - Microsoft Internet Explorer provided by PA Dept of Health

File Edit View Favorites Tools Help

Back Search Favorites Media

Address https://www.nedss.state.pa.us/nedss/asp/administration/deduplication.asp Go

Commonwealth of Pennsylvania
Electronic Disease Reporting

Welcome | Adr | Log Off

User: kwaller **Session Expires:** 30 minutes Renew

Location: Administration > Patient De-Duplication S

Patient De-Duplication Search

Batch De-Duplication

Please choose from one of the search types b limit the search results that you want to see by First Name, Last Name, and when the

Last Name	First	SSN	Choose
Exact	Exact	-	
-	-	Exact	
Exact	Exact	-	
Soundex	Soun	-	
First 5 char match	First	-	
Exact	Exact	-	

Limit Result Groups By

- Last Name
- First Name
- Date Created
- County

PA-NEDSS De-duplication: Auto Merge - Microsoft Intern...

Auto Merge

Your Auto Merge request has been initiated.
An email will be sent to you and all investigators currently assigned to concerned patients.

This window will be closed automatically after 10 seconds.

[Close]

Done Trusted sites

Start Shee... PA-N... Salm... C:\N... PA-N... FW: ... SAS ... Ded... PA-N... 10:31 PM

PA-NEDSS: Application Administration - Microsoft Internet Explorer provided by PA Dept of Health

File Edit View Favorites Tools Help

Duplicate Patient Details - Microsoft Internet Explorer provided by PA Dept of Health

Duplicate Patient Details

General Demographics

#	Name	Gender	Birth Date	Races(s)	Ethnicity	Highest Education
[1]	Quest, Johnny	Male	10/10/1960	Black or African American		
[2]	Quest, Johnny	Male	10/10/1960	White		
[3]	Quest, Johnny	Male	10/10/1957	White		

Merge Details

From the choices below, please select all of the data items you would like to keep for the merged patient from Patient Group 10.

General Information

Prefix
[2] ◉ Mr

First Name
[1] ○ Johnny
[2] ○ Johnny
[3] ◉ Johnny

Middle Name
None

Last Name
[1] ○ Quest
[2] ○ Quest
[3] ◉ Quest

Suffix
None

Gender
[1] ○ Male
[2] ○ Male
[3] ◉ Male

Birth Date or Estimated Date of Birth
[1] ○ 10/10/1960
[2] ○ 10/10/1960
[3] ○ 10/10/1957

Race
[1] ○ Black or African American
[2] ○ White
[3] ○ White

Ethnicity
None

Highest Education
None

School Attended
None

School District
None

Daycare Attended
None

Birth Country
[2] ◉ Unknown

Country of Citizenship
[2] ◉ Unknown

US Entry Date

Start | In... | C:... | W... | we... | Re... | He... | PH... | SA... | Re... | PA... | PH... | Mi... | Du... | 10:46 PM

PA-NEDSS: Application Administration - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address https://www.nedss.state.pa.us/nedss/asp/administration/deduplication.asp Go Links

Custom De-Duplication

Please choose any of the fields below to perform the de-duplication search. For the First Name and Last Name fields, chose the type of de-duplication search that you would like to perform. If you choose an N-Character search, you must select the number of characters to match. If you wish to match on Birth Date, you must also chose to match on at least one of the name fields. Please use the Limit Date criteria to retrieve patients that were added on or after the given date.

Custom De-Duplication Filter Criteria

- [ ] First Name
- [ ] Last Name
- [ ] Combined First, Middle and Last Name Soundex
- [ ] Birth Date
- [ ] First 3 digits of Zip Code
- [ ] City
- [ ] County
- [ ] Limit by Date Created

Search Reset

Expert De-Duplication

Please enter a list of Patient IDs to de-duplicate separated by commas in the box below.

Suspected Duplicate Patients

List of Patients [separated by a comma]

Done Trusted sites

Start C:\NEDSS Microsoft PowerPoint - [D... PA-NEDSS: Application ... 11:57 PM

Conclusions

- Duplicates are unavoidable
- Resolution requires significant staff time
- Need to continue to make algorithms more sensitive, specific