Quantifying the churn effect in the DC metropolitan region using a novel privacy and data sharing technology

or The Black Box

Georgetown University
District of Columbia Dept of Health
Maryland Dept of Health
Virginia Dept of Health

Anne Rhodes, PhD (VDH)
Outline

• Background
• Planning process
• Pilot
• Results
• Next Steps/Considerations
Surveillance is the conscience of the epidemic
- Dr. James Curran
National HIV Care Continuum

In the US, 1.2 million people are living with HIV. Of those:

- **Diagnosed**: 86%
- **Engaged in Care**: 40%
- **Prescribed ART**: 37%
- **Virally Suppressed**: 30%

**DATA and PROGRAM**

- People move among jurisdictions
- Labs/care markers not reported consistently
- Not all deaths reported

(Source: CDC National HIV Surveillance System and Medical Monitoring Project, 2011)
Why is “connecting-the-dots” so hard?

- **Plumbing**: Massive logistics problem to integrate thousands of government/non-government data systems at scale
  
  *Different standards, models, security, infrastructure, procedures, policies, networks, access, compartments, applications, tools, protocols, etc.* ... all at immense scale!

- **Protection**: Large-scale integration of data resources increases cyber security risks
  
  *Prevention of adversary exploitation of strategic national assets.*

- **Patterns**: Lack of analytic algorithm techniques to automatically detect data patterns and alert
  
  *Transition from “analytic dumpster diving” to early-warning indication and real-time notification*

- **Privacy**: Significant tension between security and liberty
  
  *Who trusts the “watchers”?
  Who watches the watchers?*
DC Metropolitan Area

Chart 1. Geographic designations for the 22 counties in the Washington metropolitan area

Component
- Core
- Suburban
- Exurban
Virginia Data to Care Results
Pre-Black Box

N = 83

- In Care: 43%
- Out of State: 23%
- Deceased: 5%
- Not In Care: 6%
- Not Found: 23%

2/3 of clients did not need intervention (data issue)
NIH pilot study
Examination of HIV care “churn effect” across DC metropolitan region
Pilot study security methods

Physical Security
- Tier 3 Data Center monitored 24/7
- Two Layers of Secured Doors
- Motion Activated Cameras throughout the facility
- Controlled Entry and recording of access to facility

Network Security
- Firewall and VPN for all equipment connected at data center
- Enforcing Policies and Guidelines on all equipment in data center applying Principles of Least Privilege
- Comprehensive Event Correlation
Matching Categories and Variable Definitions

- **Exact** = last_name, first_name, dob, ssn, sex, race
- **Very high** = last_name, first_name, dob, sex or m.ssn
- **High** = last_name, first_name and m.dob and (m.sex or m.race) then m.score := high;
- **Medium high** = last_name, first_soundex, dob and sex
- **Medium (1st definition)** = last_name, dob, sex, race
- **Medium (2nd definition)** = last_soundex, first_soundex, dob, (sex or race)
- **Medium low** = last_soundex, first_soundex, partial_dob, partial_ssn, (sex or race)
- **Low** = last_soundex and (partial_dob and partial_ssn) and (sex or race)
- **Very low** = last_soundex and (partial_dob or partial_ssn)
Example input data, logs, and results

| DC-HARS000002 | MD-HARS000002 | VERY_HIGH | 1 2  Last First DOB Sex Race |
| DC-HARS000003 | MD-HARS000003 | VERY_LOW  | 9 1  Soundex(Last) Partial(DOB) |
| DC-HARS000004 | MD-HARS000004 | MEDIUM    | 2 1  Last DOB Sex Race       |
| DC-HARS000005 | MD-HARS000005 | EXACT     | 1 1  Last First DOB Sex      |
| DC-HARS000028 | MD-HARS000028 | VERY_HIGH | 9 2  Last First Partial(SSN) Sex Race |
| DC-HARS000040 | MD-HARS000040 | MEDIUM_LOW| 1 2  Last First Partial(SSN) Sex Race |
| DC-HARS001299 | MD-HARS001299 | VERY_HIGH | 1 1  Last First DOB Sex Race  |
| DC-HARS001401 | MD-HARS001401 | VERY_HIGH | 9 2  Last First Partial(SSN) DOB Sex Race |
| DC-HARS001554 | MD-HARS001554 | MEDIUM_HIGH| 1 2  Last Soundex(First) DOB Sex |
| DC-HARS001840 | MD-HARS001840 | MEDIUM_LOW| 1 2  Last Soundex(First) Partial(SSN) Parti DOB Sex Race |
| DC-HARS004523 | MD-HARS004523 | MEDIUM_LOW| 1 2  Last First DOB Sex Race  |
| DC-HARS003257 | MD-HARS003257 | HIGH      | 2 2  Last First DOB Race      |
| DC-HARS006382 | MD-HARS006382 | MEDIUM    | 2 1  Last DOB Sex Race       |
| DC-HARS004805 | MD-HARS004805 | VERY_LOW  | 9 9  Soundex(Last) Soundex(First) Partial(SSN) Sex Race |
| DC-HARS007579 | MD-HARS007579 | LOW       | 2 1  Soundex(Last) Partial(SSN) Parti DOB Race |

2015/03/25 11:11:06 VA data loaded
2015/03/25 11:11:06 Matching DC & MD
2015/03/25 11:11:08 Matches found: 15
2015/03/25 11:11:08 Matching DC & VA
2015/03/25 11:11:09 Matches found: 16
2015/03/25 11:11:09 Matching MD & VA
2015/03/25 11:11:11 Total Matches: 31
2015/03/25 11:11:11 Report created: /Data/DC/private/DC2015032511111111.txt
Pilot system “churn effect” summary

Output of person-matching across DC, MD, and VA eHARS databases:

<table>
<thead>
<tr>
<th>Person matches across jurisdictions:</th>
<th>Exact</th>
<th>Very High</th>
<th>High</th>
<th>Medium High</th>
<th>Medium</th>
<th>Very Low</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC-MD*</td>
<td>4013</td>
<td>5907</td>
<td>53</td>
<td>268</td>
<td>645</td>
<td>482</td>
<td>11 368</td>
</tr>
<tr>
<td>MD-VA*</td>
<td>856</td>
<td>2343</td>
<td>11</td>
<td>117</td>
<td>377</td>
<td>865</td>
<td>4569</td>
</tr>
<tr>
<td>VA-DC*</td>
<td>1064</td>
<td>3340</td>
<td>15</td>
<td>149</td>
<td>438</td>
<td>529</td>
<td>5535</td>
</tr>
<tr>
<td>Total</td>
<td>5933</td>
<td>11 590</td>
<td>79</td>
<td>534</td>
<td>1460</td>
<td>1876</td>
<td>21 472</td>
</tr>
</tbody>
</table>

*Bidirectional reporting results (i.e., DC-reported MD matches were equal to MD-reported DC-matches; etc.)
VALIDATION OF RESULTS

Each jurisdiction validated the results using their own internal matching processes, including LinkKing, LinkPlus, and SAS programs for matching.

Found that over 90% of matches in exact, very high and high categories were matches according to jurisdictional methods.

About half of the matches found by the Black Box had not been previously determined as matches through RIDR or other means.
Utilization of Data

• Updating eHARS in each jurisdiction to include:
  • Address
  • Demographics
  • Labs
  • Vital Status
Next Steps

• Expansion to other data systems that house care markers for HIV including medical site EMRs

• Expansion to other states

• Journal article in Public Health and Surveillance
Thoughts/Wrap Up
Acknowledgements

• **Georgetown Team:** Jeff Collman, Joanne Michelle Ocampo, Jay Smart, Raghu Pemmaraju

• **George Washington University:** Amanda Castel

• **CDC Staff:** Benjamin Laffoon, Dr. Irene Hall

• **DC Department of Health:** Michael Kharfen, Garret Lum, Auntrre Hamp

• **Maryland Department of Health:** Colin Flynn, Reshma Bhattacharjee

• **Virginia Department of Health:** Sahithi Boggavarapu, Sharon Carter, Diana Jordan, Jeff Stover, Jean Cadet