CDC Health Information Innovation Consortium
August Forum

Brian Lee, MPH
Chief Public Health Informatics Officer
Office of Public Health Scientific Services

Tuesday, August 1, 2017
10:00-11:00am ET
Chamblee 106, Room 1A + Adobe Connect
Dial-in 1-800-593-0825; Participant code 7695550#
August Presenters

Matthew Hogben, PhD is the Branch Chief of the Social and Behavioral Research and Evaluation Branch. He has a PhD in Social Psychology from State University of New York. He has been a scientist in the Division of STD Prevention since the late 1990s. His research interests are sexual health, specifically promotion, care-seeking and provision; innovative models of partner services for persons infected with STD or HIV; and models and mechanisms for integrating program activity and program-focused research.
August Presenters

Karen Kroeger, PhD is a Research Anthropologist in the Social and Behavioral Research and Evaluation Branch of the Division of STD Prevention. She conducted ethnographic fieldwork in Indonesia among female sex workers and received her PhD in 2000 from Washington University in St. Louis. She began her career at CDC in 2001 as a postdoctoral fellow in STD prevention, and from 2004-2009, she was a Behavioral Scientist CDC’s Global AIDS program. She has conducted qualitative research, rapid ethnographic assessment, and evaluation among vulnerable populations of female sex workers, men who have sex with men, and persons who use drugs in the US and abroad.
Ellsworth (Ells) Campbell is a health scientist in the Laboratory Branch of the Division of HIV/AIDS Prevention. Ells holds bachelor’s and master’s degrees in biology from UC San Diego and is currently pursuing a PhD in biology at Penn State University. Ells began working at CDC as a PhD student and Oak Ridge Institute for Science Education (ORISE) fellow in 2013 and recently transitioned to a full-time Associate Service Fellowship.
INNOVATIVE SURVEILLANCE AND ASSESSMENT TECHNIQUES USING SOCIAL MEDIA TO INFORM STD AND HIV PREVENTION ACTION

Matthew Hogben, Branch Chief
Karen Kroeger, Team Lead
Social and Behavioral Research and Evaluation Branch
Division of STD Prevention
CDC Health Information Innovation Consortium Forum
August 1, 2017
Pre-exposure prophylaxis (PrEP) reduces HIV transmission, yet only 2-5% of Americans who may benefit from it are using it.

Men who have sex with men (MSM) using PrEP show high rates of sexually transmitted infections (STI).

Divergence of traditional STD prevention from HIV prevention raises new questions:
- Is STI an indicator for PrEP? (as an STD program intervention)
- What is the impact of PrEP on STI prevention?
PURPOSE OF THE PROJECT

- To evaluate social media site REDDIT as a novel data source for monitoring behaviors and attitudes relevant to PrEP
- To explore networks of attitudes and beliefs surrounding STD prevention in the context of PrEP
Within a subreddit, users see a list of submissions. These can be statements, questions, images or links to other websites.

In the PrEPared subreddit, we frequently find people asking about symptoms, experiences, sources and perceptions related to pre-exposure prophylaxis.
DISSECTING REDDIT

Within a submission, users can comment in a threaded, hierarchical branching format.

Note: We use the term ‘post’ to refer to submissions and comments.
Data collection and processing
- Refine research questions, identify concepts of interest
- Define sampling frame (subreddits - e.g. askgaybros, PrEPared)
- Develop search and processing taxonomies
  - 20 search terms (product names, STI names, concepts from STI prevention)
  - Up to 40 concepts (e.g. stigma, side effects, adherence)

Data analysis and output
- Process queries (one time .csv data file with meta-tagged data)
- Data explorer dashboard to conduct analysis/output

STEPS IN THE PROCESS
REFINED RESEARCH QUESTIONS,

- What were the main topics of discussion regarding PrEP in 2015?
- How much uptake/use of PrEP was there vs. information seeking?
- What are main barriers to PrEP uptake?
- How are MSM talking with sex partners about PrEP?
- When discussing PrEP how often are STDs other than HIV discussed, which STDs are mentioned, and in what context?
Temporal visualization of mentions of PrEP and STIs
Visualization of distribution of mentions across digital space
Visualization of relationships between topics
Provide a view of the “size or volume” of conversations about PrEP (but not the “shape or meaning” of the conversation)
ORGANIZING TERMS AND QUESTIONS

The matrix format helps us to organize our terms and questions into explicit sets and relations.

<table>
<thead>
<tr>
<th>Search and Classification</th>
<th>Recovery</th>
<th>Research Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Granular responses...</td>
<td></td>
<td>How can we...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>improve...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>outcomes?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>using...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>techniques?</td>
</tr>
</tbody>
</table>

| Granular responses...     |          | How can we...      |
|                          |          | improve...         |
|                          |          | outcomes?           |
|                          |          | using...            |
|                          |          | techniques?         |
### Search Terms and Concepts

#### Gathering Data

<table>
<thead>
<tr>
<th>Term</th>
<th>Search Term</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Tagging Concepts

<table>
<thead>
<tr>
<th>Term</th>
<th>Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partners</td>
<td>partner, partners, boy/girlfriend, husband, wife</td>
</tr>
<tr>
<td>Work</td>
<td>job, employment, occupation</td>
</tr>
<tr>
<td>Relationship</td>
<td>family, love, friendship</td>
</tr>
<tr>
<td>Communication</td>
<td>talk, conversation, dialogue</td>
</tr>
<tr>
<td>Medications</td>
<td>medicine, pill, drug</td>
</tr>
<tr>
<td>Side Effects</td>
<td>fatigue, dizziness, nausea</td>
</tr>
<tr>
<td>Focus and Benefits</td>
<td>motivation, goal, priority</td>
</tr>
</tbody>
</table>

- **Term**: Specific word or phrase associated with a concept.
- **Synonyms**: Alternative words or phrases that can be used interchangeably.
8363 Posts from 6 Subreddits
(2411 Mentions of “Prep”)
Jan 1 - Dec 31, 2015
https://demo.medwatcher.org/
DEMO EXAMPLE: MEDWATCHER
DEMO EXAMPLE: MEDWATCHER
NEXT STEPS

- Use completed dashboard to explore the volume and content of conversations about PrEP and STIs
  - Attitudes, perceptions
  - Barriers, facilitators
  - Context of PrEP use (e.g. behaviors, sex partners)
- Evaluate potential utility of subreddits and search terms for future surveillance use
- Evaluate potential for network analysis (topical content) and subsequent qualitative analysis
ACKNOWLEDGEMENTS

- DSTDP
  - Kyle Bernstein
  - Sevgi Aral
- Epidemic
  - Michael Gilbert
  - Chi Bahk
- CDC Health Information Innovation Consortium
Building Capacity for Automated Phylodynamic Analysis
A look at the 2015 HIV Outbreak in rural Indiana

Ells Campbell, MS
Computational Biologist
Division of HIV/AIDS Prevention
National Center for HIV/AIDS, Viral Hepatitis, STDs, and Tuberculosis Prevention

CDC Health Information Innovation Consortium
August 1st, 2017
Program Perspective – Data Silos

Surveillance  |  Laboratory  |  Epidemiology
Program Perspective – Data Lake

Surveillance

Laboratory

Epidemiology
Connecting the Dots

Incidence Surveillance

Molecular Surveillance
HIV-TRACE – Quantifying difference between sequences

- Independently align each sequence to a reference
  - Master Reference Sequence: HXB2
  - Sample Sequence

- Calculate distance between all possible pairs
  - Sequence A
  - Sequence B

- Threshold Filtration

<table>
<thead>
<tr>
<th>ID1</th>
<th>ID2</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>0.001</td>
</tr>
<tr>
<td>A</td>
<td>D</td>
<td>0.014</td>
</tr>
</tbody>
</table>
HIV-TRACE Results – Simple Example

<table>
<thead>
<tr>
<th>ID1</th>
<th>ID2</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>0.001</td>
</tr>
<tr>
<td>A</td>
<td>D</td>
<td>0.014</td>
</tr>
</tbody>
</table>

![Diagram showing distances between IDs A, B, and D]
Why build transmission networks?

Results directly comparable to traditional contact tracing data
• Can readily compare genetic distance and contact tracing

Fine-grain resolution of transmission dynamics
• Inform outbreak-specific intervention efforts
• Inform future prevention strategies

Identification of unreported direct or indirect transmission links
• Aid for partner services and outbreak investigations
Complementary Perspectives

Epidemiological

High-Risk
Sexual
Network
Complementary Perspectives

Epidemiological + Laboratory

High-Risk Sexual Network

HIV Genetic Distance Network
Complementary Perspectives

Epidemiological + Laboratory = Integrated

High-Risk Sexual Network + HIV Genetic Distance Network = Both
Analytics Enrichment

Integration + Informatics

Both Contact & Genetic

Minimum Spanning Tree (MST)
Decision Support

Integration

Informatics

Clinical

Both Contact & Genetic

Minimum Spanning Tree (MST)

Recency of HIV Infection

 Established

 Acute
Outbreak Response – HIV Outbreak among PWID in rural Indiana

Outbreak Details

- >220 HIV+ diagnoses
- >90% inject drugs
  - 4-15x/day
  - 1-6 partners/injection
- >90% HCV co-infection
- High prevalence of sex e:
Decision Support - Machine Learning
- HIV Outbreak in Rural Indiana, 2015

IDU ≤ 3

N=435
16.8% HIV-infected

IDU ≤ 3
N=357
9.8% HIV-infected

IDU ≤ 1
16/263
6.1% HIV-infected
1 partner

IDU > 1
19/94
20.2% HIV-infected
2+ partners

IDU ≥ 4
N=78
48.7% HIV-infected

IDU ≤ 1
11/47
23.4% HIV-infected
1 partner

IDU > 1
27/31
87.1% HIV-infected
2+ partners

IDU ≥ 4
109/117
93.2% HIV-infected

N = 552
32.8% HIV-infected

Color Scale
0% HIV Prevalence
100% HIV Prevalence
Outbreak Simulations
- HIV Outbreak in Rural Indiana, 2015
Outbreak Response – HIV Outbreak among PWID in rural Indiana

- Most HIV sequences are highly similar, representing rapid and recent transmission

Genetic Distance Network
polymerase region

1.5% distance threshold

Pairwise Genetic Distance Histogram
Outbreak Response – HIV Outbreak among PWID in rural Indiana

- Pruning links with a lower threshold reveals community structure at cost of historical resolution
- Subgroups may be evidence of biological, social, temporal, and/or geographic factors

0.1% distance threshold

![Genetic Distance Network](image)
Outbreak Response – HIV Outbreak among PWID in rural Indiana

- Minimum spanning trees (MSTs) select the most parsimonious links from the distribution of distances.

- There can be many equally true MSTs, so we combine many unique MSTs to illustrate uncertainty.

![Composite of 100 unique MSTs](image)
Complementary Perspectives
- HIV Outbreak in Rural Indiana, 2015

Close Genetic Links
- Mostly Uninformative Noise
- N > 10,000

High-Risk Contacts
- Probable Transmission Links
- N = 182

N > 1,500
Outbreak Response – HIV Outbreak among PWID in rural Indiana

Genetic Distance Network
1.5% distance threshold

Minimum Spanning Trees
Mean Distance <0.1%

Inferred Transmission Network
Integration is Complex and Resource Intensive

HIV Surveillance Data Integration Workflow
Data Prep
Trifacta Wrangler

- Summary statistics
  - Interactive visualization

- Data cleaning
  - Guided by machine learning
  - Live preview of changes

- Optimized for collaboration
  - Shareable scripts

- Personal & Enterprise versions available

https://www.trifacta.com/start-wrangling/
Exploratory Analysis
Alpine Chorus
Powerful collaborative and analytics tools via web forms

Intuitive workflow editor with drag and drop “operators”

Built-in analytics (decision trees, linear regression, PCA, k-means, and dozens more)

Custom code turned into drag & drop operators (R, Python, Java)

Open source and Enterprise

github.com/alpinedatalabs
Automated Solution

Enables reproducible machine learning analytics

...without the programmer!
Network Visualization
Centrifuge - Enterprise

Network visualization, exploration, and animation

Dynamic model generation
- Person \leftrightarrow Person
- County \leftrightarrow County

Variables mapped via dropdown
- Shape = Risk Factor
- Color = Infection Status
- Size = # of Reported Partners
- Link width = genetic distance

Pattern Recognition

Low-cost ‘lite’ version on the way for public health
MicrobeTRACE

Sequences + Epidemiologic Data

Interactive Exploration of Transmission Networks

github.com/cd cgov/microbetrace
Potential for Enhancement

**Business**
Partner with vendors for licensing innovation
  - Public health “ROI” doesn’t translate to increased budget

**Admin**
Network administration and security must catch up and keep pace

**Technology**
Dynamic redaction of PII
  - Submit necessary data for automated analysis
  - Rapidly integrate sensitive data locally
On the Shoulders of Giants

Indiana
- Indiana State Department of Health Scott County Health Department
- Clark County Health Department
- Foundations Family Medicine
- Indiana University School of Medicine, Division of Infectious Diseases
- University of Louisville School of Medicine, Division of Infectious Diseases
- NCHHSTP, Division of HIV/AIDS Prevention
- NCEZID, Division of Viral Hepatitis
- NCHHSTP, Division of Infectious Diseases
- Epidemic Intelligence Service

Surveillance
- HIV Incidence and Case Surveillance Branch (HICSB)
  - Molecular HIV Surveillance
- State, Territory, Local, and Tribal Public Health Departments

STOP Study
- NYC, NC, and SF public health
- NCHHSTP, Division of HIV/AIDS Prevention (DHAP)

Commercial Laboratories
- Quest
- LabCorp

Software and Support
- SciComp, ITSO and OCISO at CDC
- Leidos Inc.
  - Centrifuge Systems
  - Trifacta
  - Alpine Data Labs
- UCSD Viral Evolution Group (VEG)
- Stanford Visualization Group
- Seattle Interactive Data Lab
- Stanford HIV Drug Resistance DB Team
- Los Alamos National Laboratories
- Gephi Consortium
Items of interest

• Data Visualization Scientific Advisory Group [link]
• CDC GitHub organization - https://github.com/CDCgov/
• CodeForAtlanta – 8/13 – [link]
• CodeDay – Nov 11/12 – Registrations up [link]
• CDC Tracking Network Enviro Health App Challenge http://www.envirohealthchallenge.com/ - winners Sep 1
• Next Forum – November 7, 10am-11am ET- Share ideas you would like to have covered

• Please take 90 seconds to share your feedback, https://epiinfowebsurvey.cdc.gov/Home/aa515f48-55f7-4a14-9b82-1a5bd2f96124. We’d love to hear from you on improving the quarterly forums.
Thank You & Questions

For more information please contact Centers for Disease Control and Prevention

1600 Clifton Road NE, Atlanta, GA 30333
Telephone: 1-800-CDC-INFO (232-4636)/TTY: 1-888-232-6348
Visit: www.cdc.gov | Contact CDC at: 1-800-CDC-INFO or www.cdc.gov/info

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.