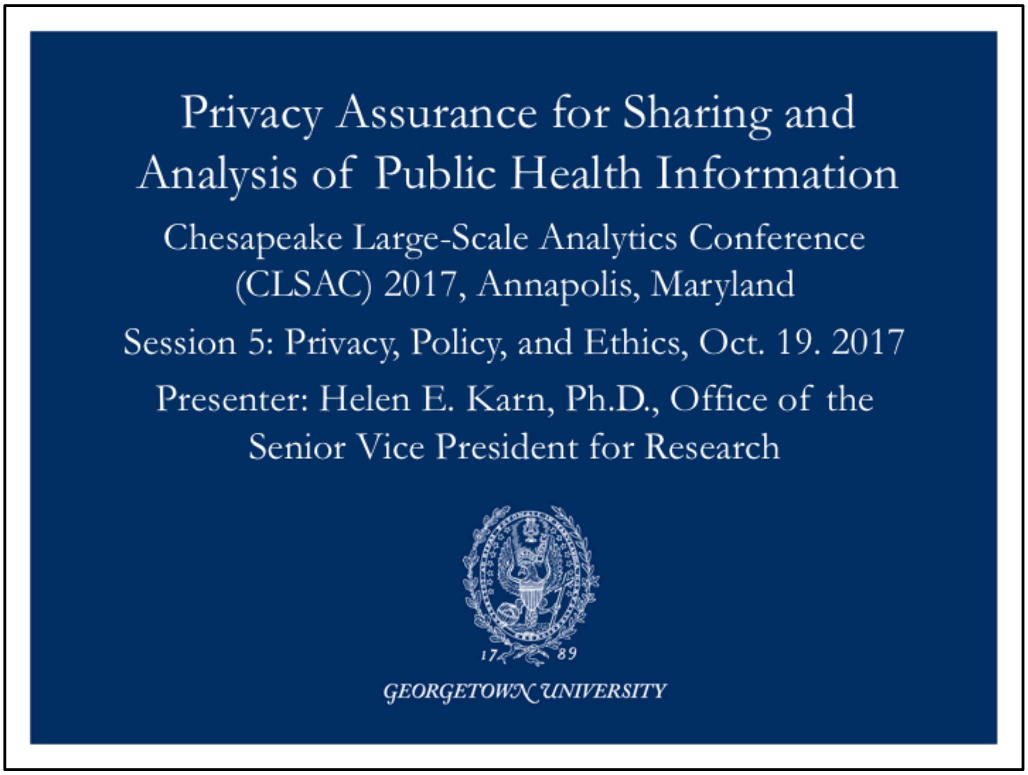


Privacy Assurance for Sharing and Analysis of Public Health Information

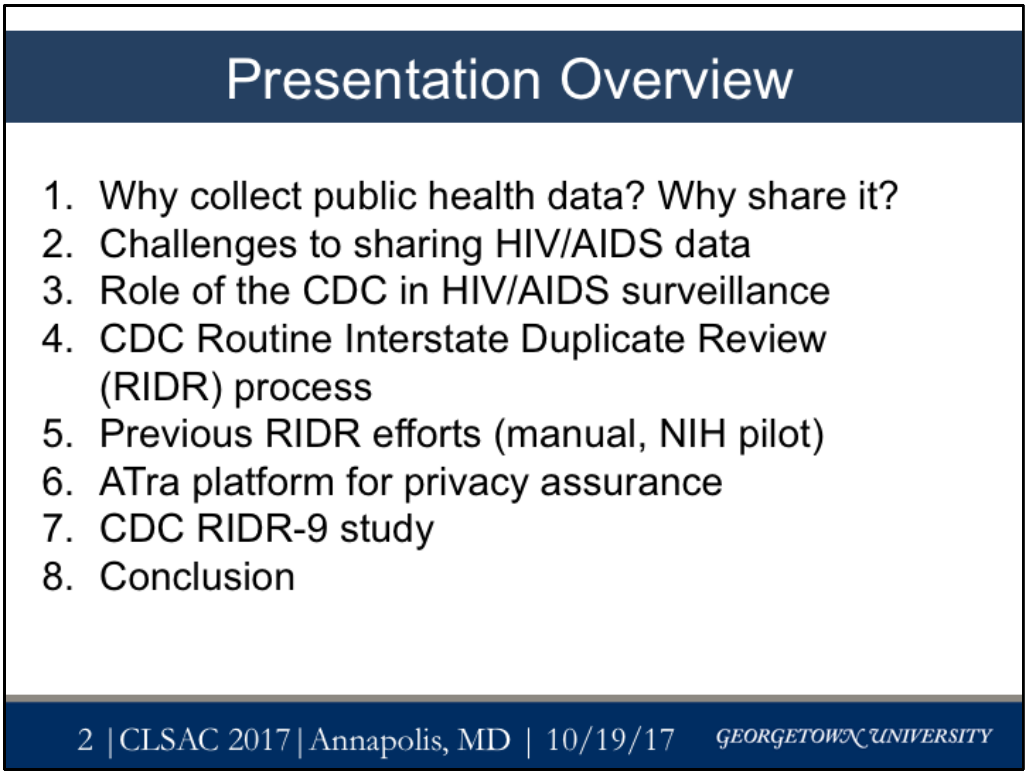
Chesapeake Large-Scale Analytics Conference (CLSAC) 2017, Annapolis, Maryland

Session 5: Privacy, Policy, and Ethics, Oct. 19. 2017

Presenter: Helen E. Karn, Ph.D., Office of the Senior Vice President for Research

1789

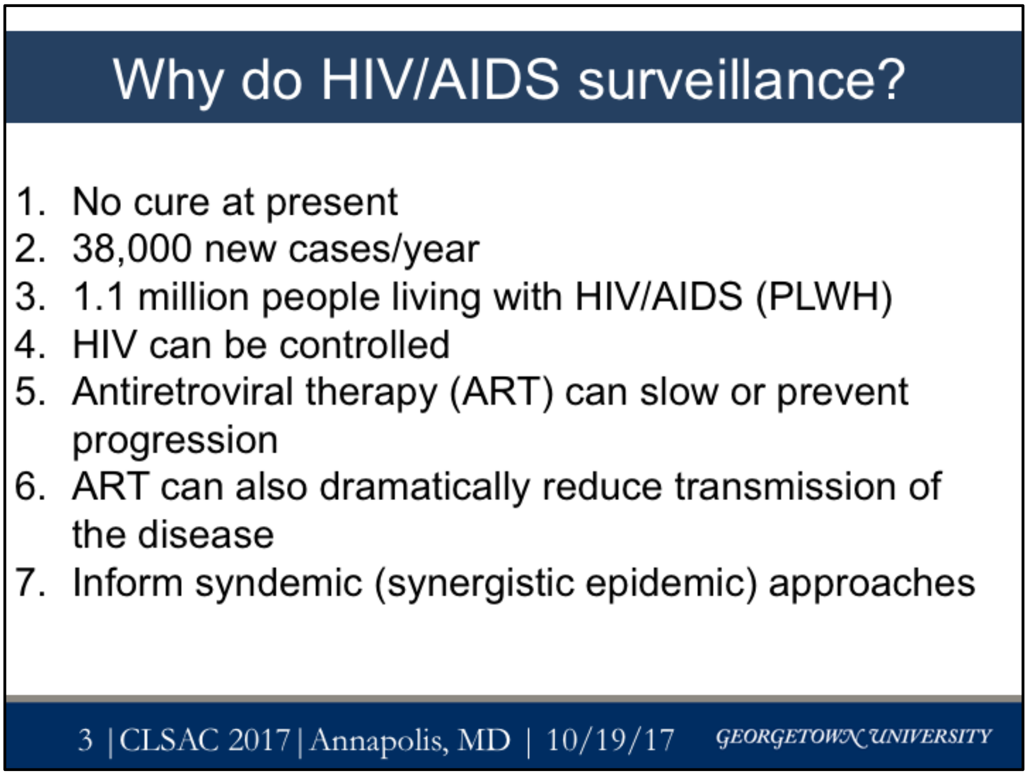
GEORGETOWN UNIVERSITY



Presentation Overview

1. Why collect public health data? Why share it?
2. Challenges to sharing HIV/AIDS data
3. Role of the CDC in HIV/AIDS surveillance
4. CDC Routine Interstate Duplicate Review (RIDR) process
5. Previous RIDR efforts (manual, NIH pilot)
6. ATra platform for privacy assurance
7. CDC RIDR-9 study
8. Conclusion

CDC = Centers for Disease Control and Prevention (CDC)



Why do HIV/AIDS surveillance?

1. No cure at present
2. 38,000 new cases/year
3. 1.1 million people living with HIV/AIDS (PLWH)
4. HIV can be controlled
5. Antiretroviral therapy (ART) can slow or prevent progression
6. ART can also dramatically reduce transmission of the disease
7. Inform syndemic (synergistic epidemic) approaches

3 | CLSAC 2017 | Annapolis, MD | 10/19/17 GEORGETOWN UNIVERSITY

1-3 – That's the bad news.

The good news is that a great amount of progress has been made over the past 35 years in the treatment of HIV/AIDS.

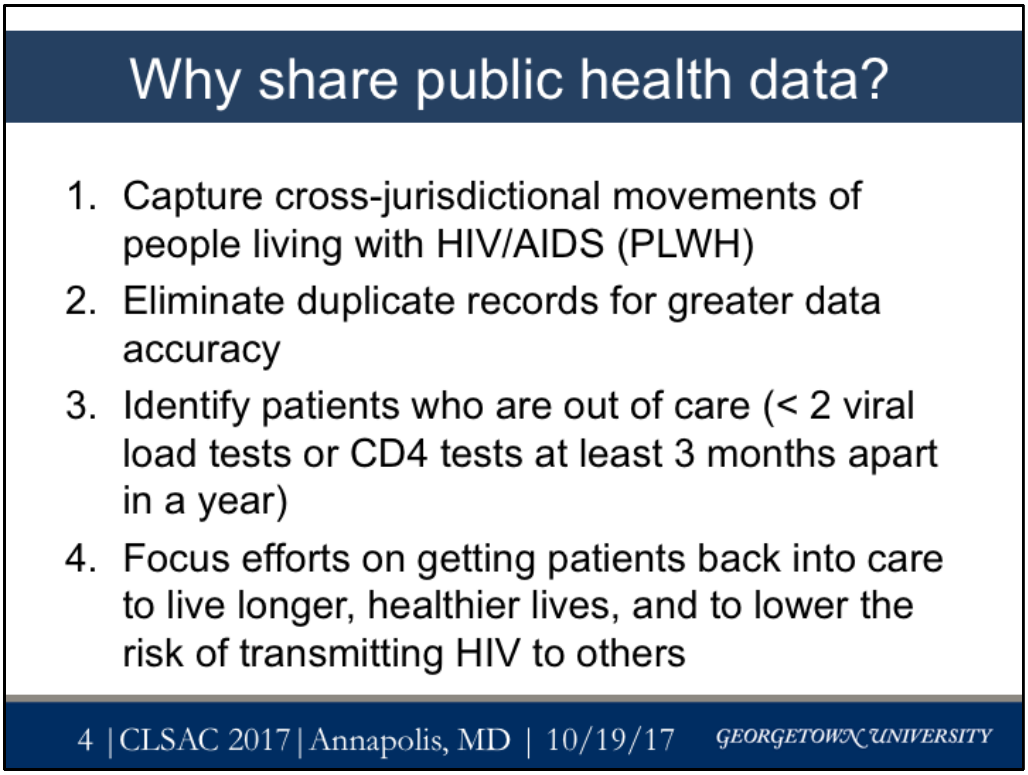
With proper medical care, the Centers for Disease Control and Prevention (CDC) reports that patients living with HIV can live nearly as long as those without HIV if it is diagnosed early and he or she receives proper medical care, and their quality of life can be nearly as healthy as those without HIV.

The primary treatment for HIV/AIDS consists of antiretroviral therapy, which can slow or prevent the disease from progressing. Importantly, ART dramatically prevents the virus from spreading.

HIV/AIDS can co-occur with other infectious diseases, such as other sexually transmitted diseases and hepatitis.

In addition to biomedical approaches to preventing and treating diseases, there are syndemic approaches as well that investigate demographic factors.

So data from HIV surveillance can inform synergistic epidemic approaches as well.



Why share public health data?

1. Capture cross-jurisdictional movements of people living with HIV/AIDS (PLWH)
2. Eliminate duplicate records for greater data accuracy
3. Identify patients who are out of care (< 2 viral load tests or CD4 tests at least 3 months apart in a year)
4. Focus efforts on getting patients back into care to live longer, healthier lives, and to lower the risk of transmitting HIV to others

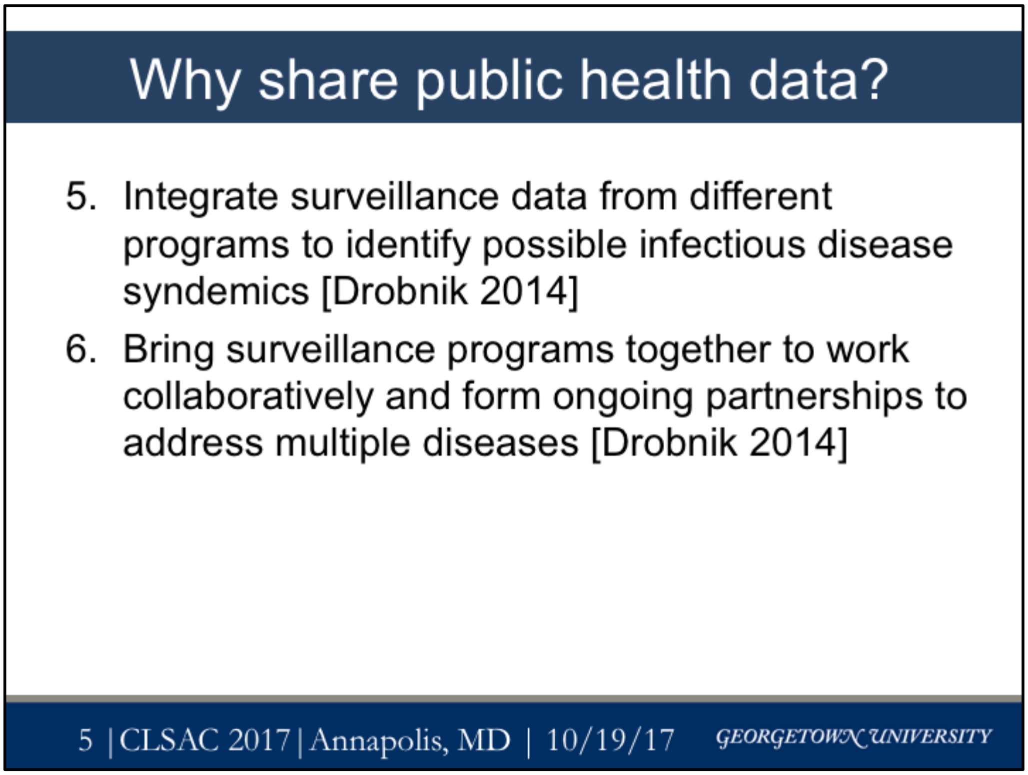
Why do we want to share public health data?

First, sharing data lets us capture cross-jurisdictional movements of people living with HIV/AIDS. This is particularly important in large metropolitan regions that have large numbers of people living with HIV/AIDS. A very good example is Washington, DC, Maryland, and Virginia, which has a large population of people, a relatively high percentage of people living with HIV/AIDS, and very close borders. For example, it's not unusual for someone to live in Northern Virginia, work in Washington, DC, and receive medical care at Johns Hopkins in Maryland.

Second, sharing data across jurisdictions helps eliminate duplicate patient records. This results in more accurate data.

Third, sharing data allows jurisdictions to identify patients who are out of care. "Out of care" is defined by the Centers for Disease Control and Prevention as someone who has not had at least 2 tests of their CD4 cells – the cells that are destroyed by HIV -- (or) 2 viral load tests at least 3 months apart in 1 year. People who have two or more viral load tests or CD4 cell tests at least 3 months apart in a year are considered to be "retained in care".

Fourth, sharing data allows local public health agencies and healthcare providers to focus their efforts on the people who are truly "out of care" as opposed to living in another part of the country.

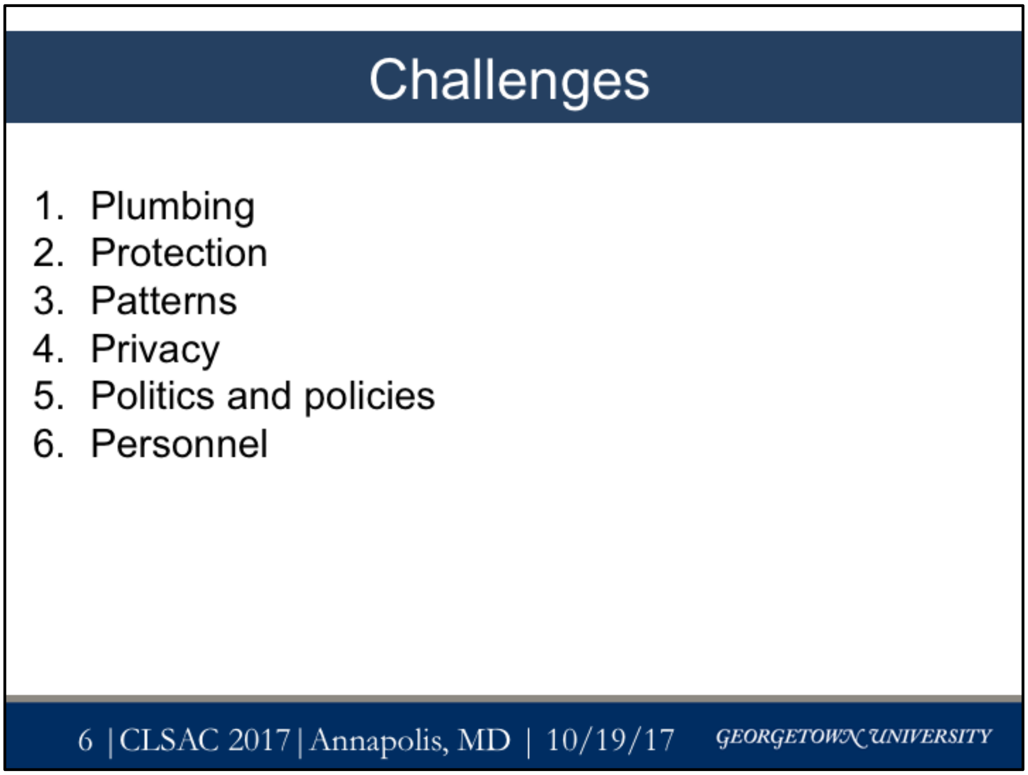


Why share public health data?

5. Integrate surveillance data from different programs to identify possible infectious disease syndemics [Drobnik 2014]
6. Bring surveillance programs together to work collaboratively and form ongoing partnerships to address multiple diseases [Drobnik 2014]

"The exchange of information between health officials in different jurisdictions has been demonstrated to be essential to managing outbreaks of well-understood diseases, such as measles, and identifying and responding to new and emerging threats, such as severe acute respiratory syndrome and pandemic influenza A (H1N1). However, the very developments in information technology and health care policy that make it increasingly possible to rapidly transmit health information, such as electronic medical records and health information exchanges, continue to raise questions about the possible inappropriate use and lack of protection of personally identifiable health information. [O'Connor and Matthews 2011]

"In highly charged situations, such as the recent influenza A (H1N1) pandemic, variations in state laws and incomplete understanding among jurisdictions can easily lead to inconsistent public health disclosure practices, resulting in media questions about the integrity of information access."



Challenges

1. Plumbing
2. Protection
3. Patterns
4. Privacy
5. Politics and policies
6. Personnel

6 |CLSAC 2017|Annapolis, MD | 10/19/17 GEORGETOWN UNIVERSITY

The "5 Ps" [Smart 2017] plus Policies and Personnel:

1. Plumbing: There are different standards, models, security, infrastructure, procedures, policies, networks, access, compartments, applications, tools, protocols, etc.

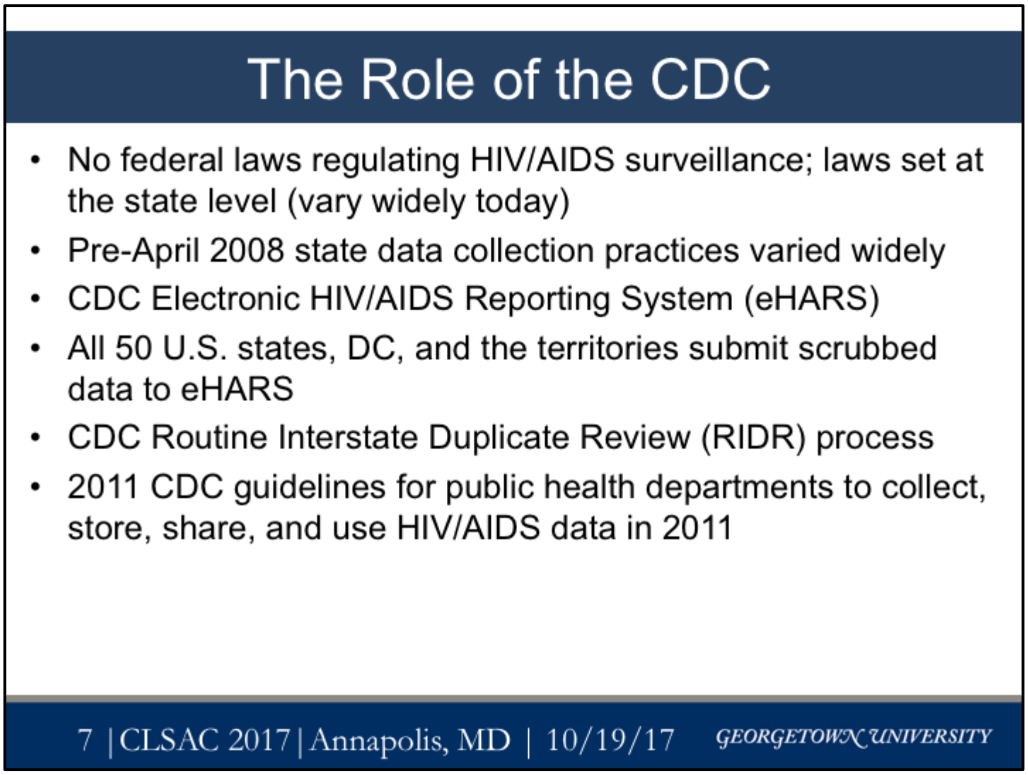
2. Protection: Integrating data resources on a large scale increases cyber security risks.

3. Patterns: There is a lack of analytic algorithm techniques to automatically detect data patterns and alert the stakeholders early or in real time.

4. Privacy: There is a significant tension between security and privacy. Who trusts the “watchers"? Who watches the watchers?

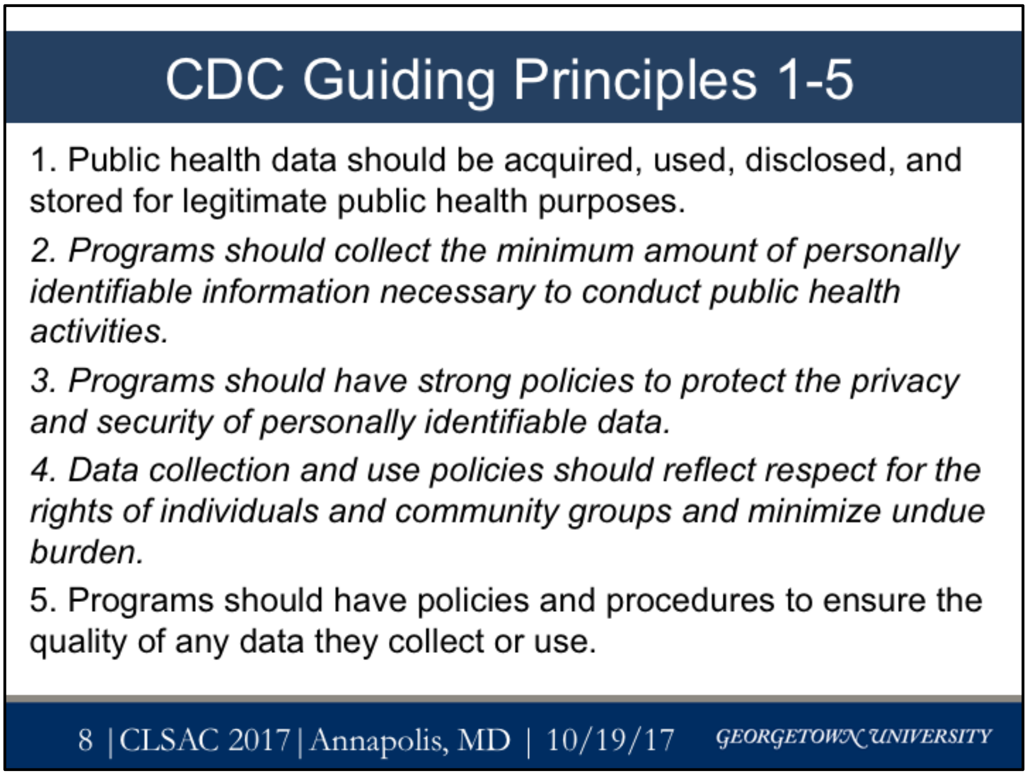
5. Politics: What’s in it for me? Policies: What policies govern how data can be collected, stored, and used? How can work progress when those policies vary widely from one jurisdiction to another?

6. Personnel: Even if there is a willingness to share data, there may be a lack of staff, time, or technical expertise – or all three – at the local level.



The Role of the CDC

- No federal laws regulating HIV/AIDS surveillance; laws set at the state level (vary widely today)
- Pre-April 2008 state data collection practices varied widely
- CDC Electronic HIV/AIDS Reporting System (eHARS)
- All 50 U.S. states, DC, and the territories submit scrubbed data to eHARS
- CDC Routine Interstate Duplicate Review (RIDR) process
- 2011 CDC guidelines for public health departments to collect, store, share, and use HIV/AIDS data in 2011



CDC Guiding Principles 1-5

1. Public health data should be acquired, used, disclosed, and stored for legitimate public health purposes.

2. Programs should collect the minimum amount of personally identifiable information necessary to conduct public health activities.

3. Programs should have strong policies to protect the privacy and security of personally identifiable data.

4. Data collection and use policies should reflect respect for the rights of individuals and community groups and minimize undue burden.

5. Programs should have policies and procedures to ensure the quality of any data they collect or use.

8 | CLSAC 2017 | Annapolis, MD | 10/19/17 GEORGETOWN UNIVERSITY

To ensure data security and confidentiality, the Centers for Disease Control and Prevention (CDC) has published 10 guiding principles for collection, storing, sharing, and using public health data [CDC 2011, CDC 2014].

Here I highlight the guiding principles that are especially relevant to our discussion today on privacy, policy, and ethics.

Guiding principle number 2:

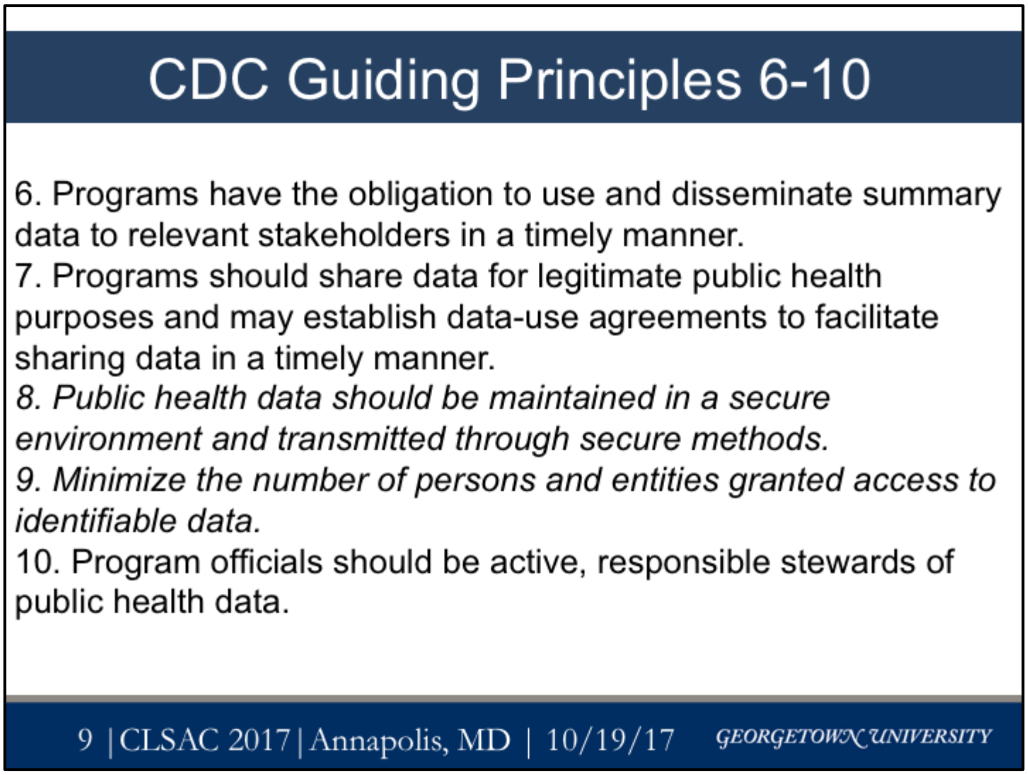
2. Programs should collect the minimum amount of personally identifiable information necessary to conduct public health activities.

Principle number 3:

3. Programs should have strong policies to protect the privacy and security of personally identifiable data.

And principle number 4:

4**.** Data collection and use policies should respect respect for the rights of individuals and community groups and minimize undue burden.



CDC Guiding Principles 6-10

6. Programs have the obligation to use and disseminate summary data to relevant stakeholders in a timely manner.
7. Programs should share data for legitimate public health purposes and may establish data-use agreements to facilitate sharing data in a timely manner.
8. Public health data should be maintained in a secure environment and transmitted through secure methods.
9. Minimize the number of persons and entities granted access to identifiable data.
10. Program officials should be active, responsible stewards of public health data.

9 | CLSAC 2017 | Annapolis, MD | 10/19/17 GEORGETOWN UNIVERSITY

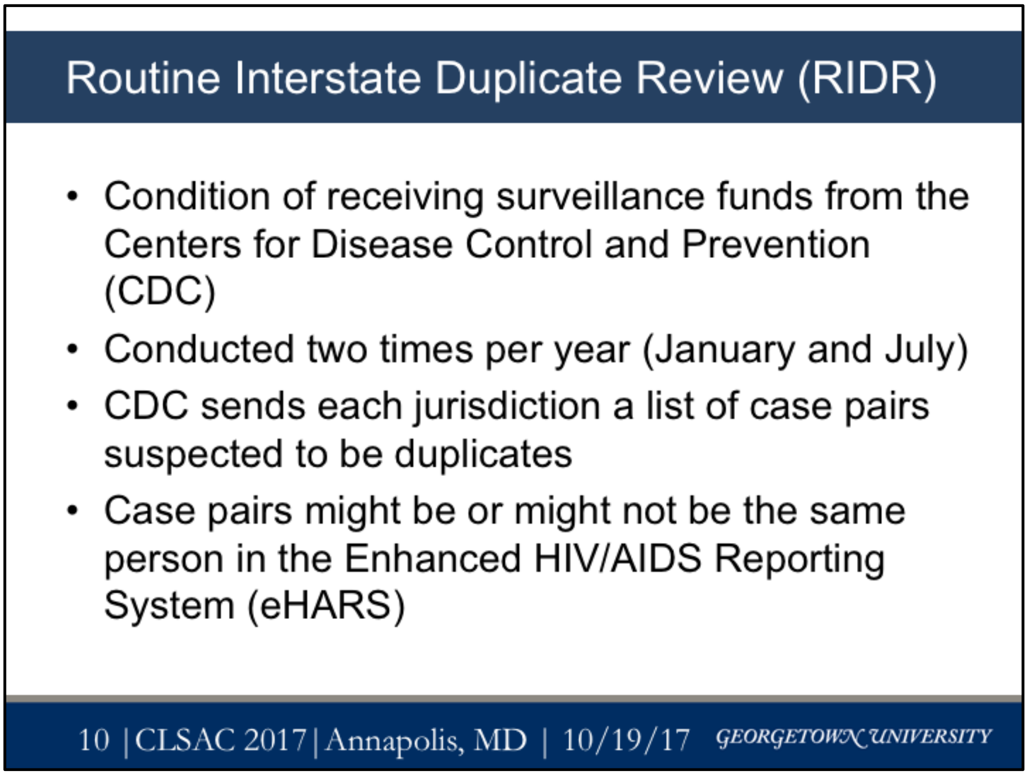
The Centers for Disease Control and Prevention (CDC) has published ten guiding principles for public health data collection, storage, sharing, and use in order to ensure security and confidentiality (CDC 2011, CDC 2014).

Number 8:

Public health data should be maintained in a secure environment and transmitted through secure methods.

Number 9:

Minimize the number of persons and entities granted access to identifiable data.



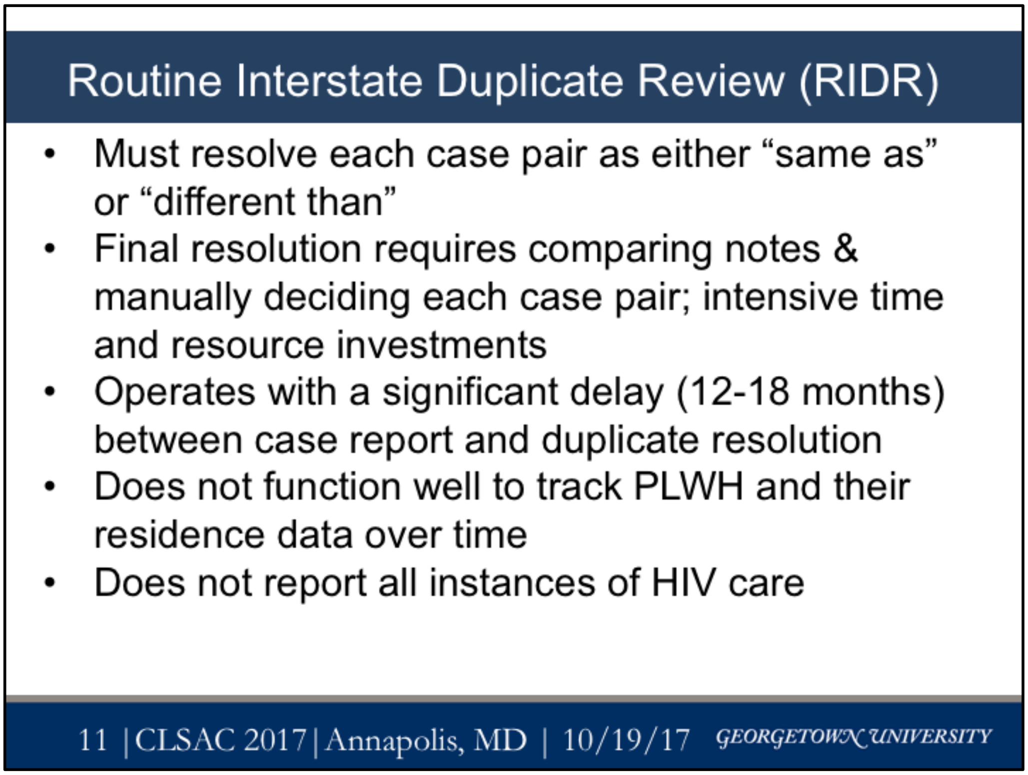
Routine Interstate Duplicate Review (RIDR)

- Condition of receiving surveillance funds from the Centers for Disease Control and Prevention (CDC)
- Conducted two times per year (January and July)
- CDC sends each jurisdiction a list of case pairs suspected to be duplicates
- Case pairs might be or might not be the same person in the Enhanced HIV/AIDS Reporting System (eHARS)

10 | CLSAC 2017 | Annapolis, MD | 10/19/17 GEORGETOWN UNIVERSITY

[Public health jurisdictions expend much resources tracking PLWH across the continuum of care. The Centers for Disease Control and Prevention (CDC) directs public health departments to conduct the Routine Interstate Duplicate Review (RIDR) two times per year. Conducting the RIDR review is a condition of receiving surveillance funds. The Council of State and Territorial Epidemiologists (CSTE) recommends that public health departments communicate and collaborate to identify and resolve duplicates.

In January and July of every year, CDC sends each jurisdiction a list of case pairs suspected to be duplicates; that is, PLWH case pairs who may or may not be the same person in the Enhanced HIV/AIDS Reporting System (eHARS) archives of specific jurisdictions.



Routine Interstate Duplicate Review (RIDR)

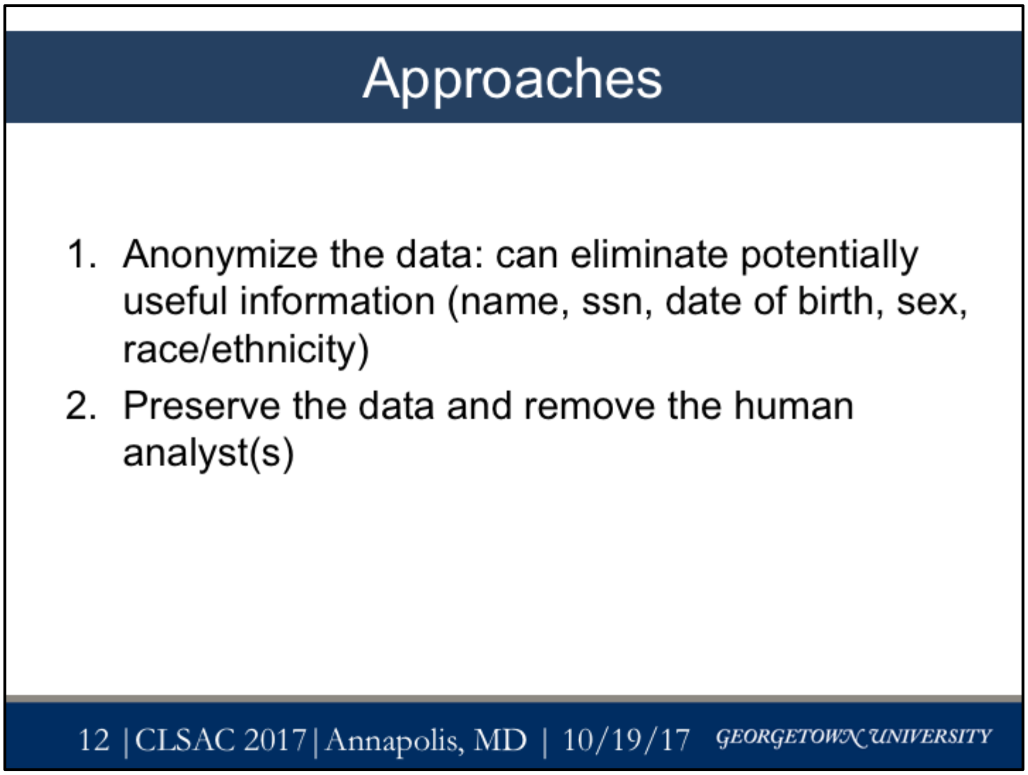
- Must resolve each case pair as either “same as” or “different than”
- Final resolution requires comparing notes & manually deciding each case pair; intensive time and resource investments
- Operates with a significant delay (12-18 months) between case report and duplicate resolution
- Does not function well to track PLWH and their residence data over time
- Does not report all instances of HIV care

On the basis of available information from various sources, two jurisdiction must resolve each of their case pairs. For each case pair, they must decide if the two people are the “same as” or “different than” each other.

Traditionally, final resolution of case pairs requires staff from jurisdictions to call each other on the telephone to compare notes and manually decide the “same as” or “different than” status of each specific case pair.

Thus, RIDR requires intensive time and resource investments, and operates with a significant delay (12-18 months) between case report and duplicate resolution. RIDR also does not function well to track PLWH and their residence data over time, or to report all instances of HIV care.

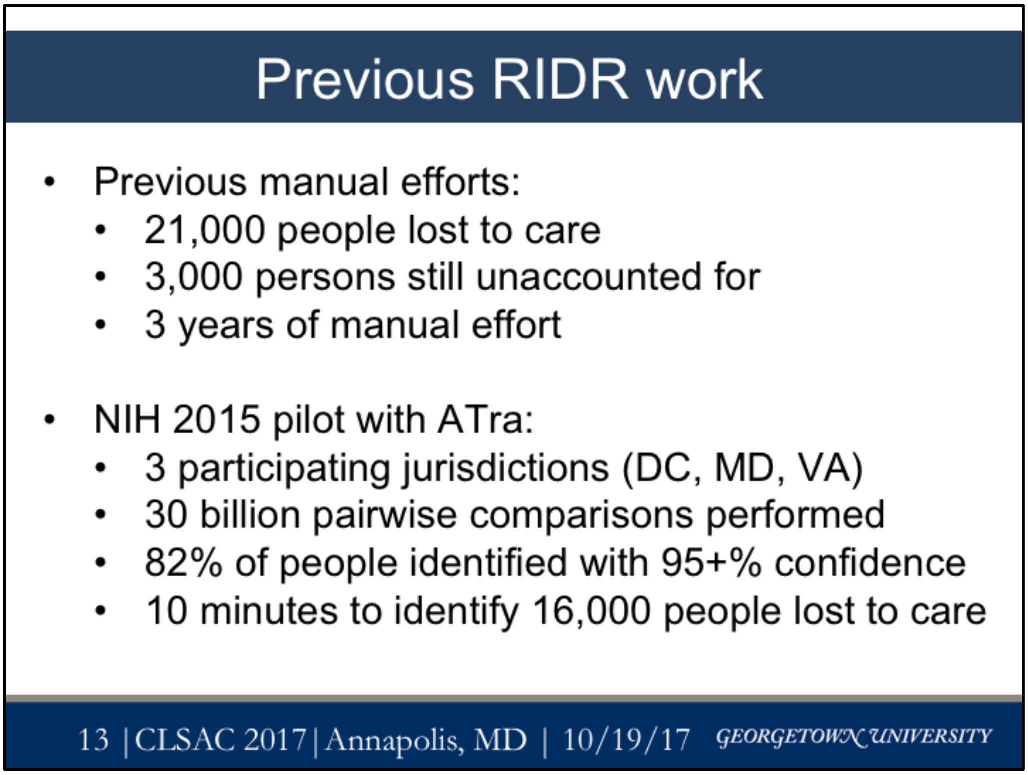
In August 2016, the CDC commissioned a special project involving HIV surveillance teams from 9 jurisdictions and Georgetown University to evaluate the efficacy of a privacy device (or “Black Box”) with a matching and computational algorithm (“ATra$^{\text{TM}}$”) in overcoming these deficiencies in the RIDR process.



Approaches

1. Anonymize the data: can eliminate potentially useful information (name, ssn, date of birth, sex, race/ethnicity)
2. Preserve the data and remove the human analyst(s)

For (2), by using a “black box” matching algorithm

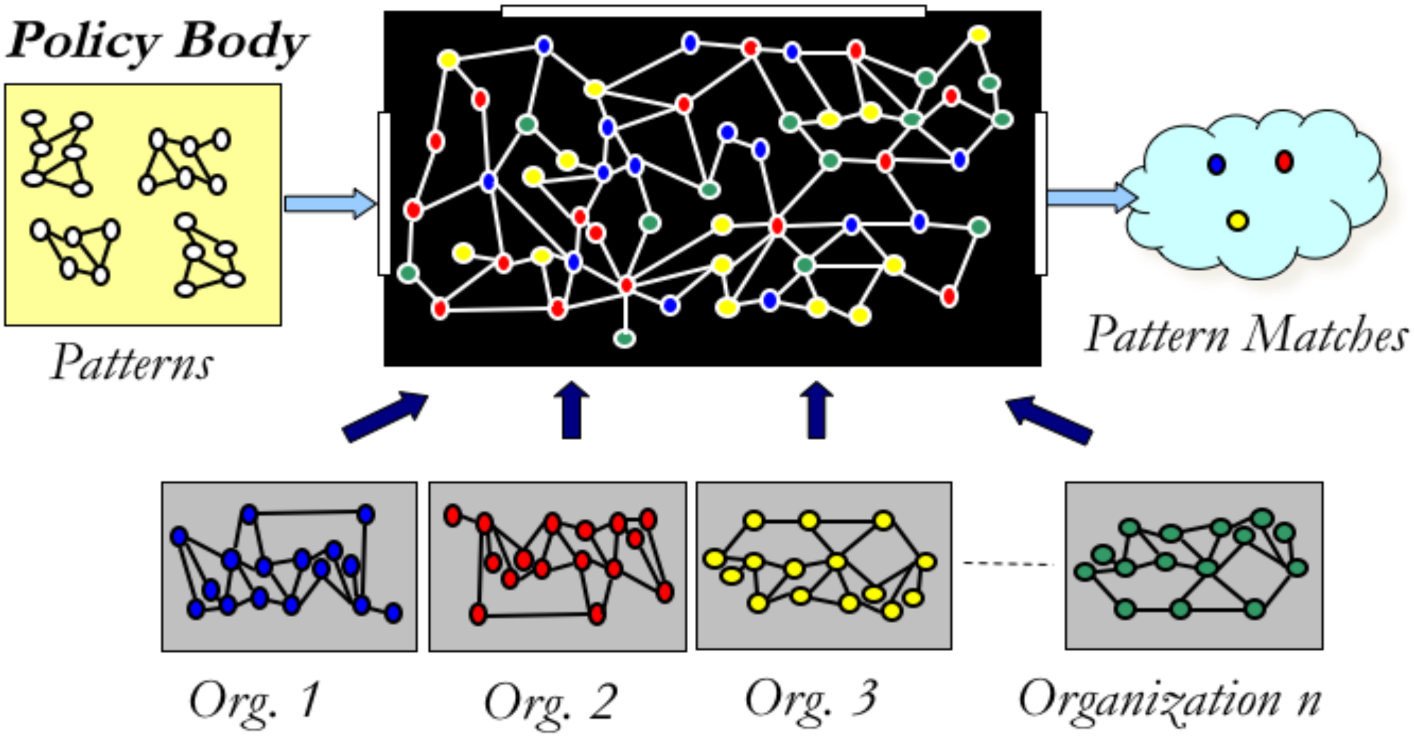


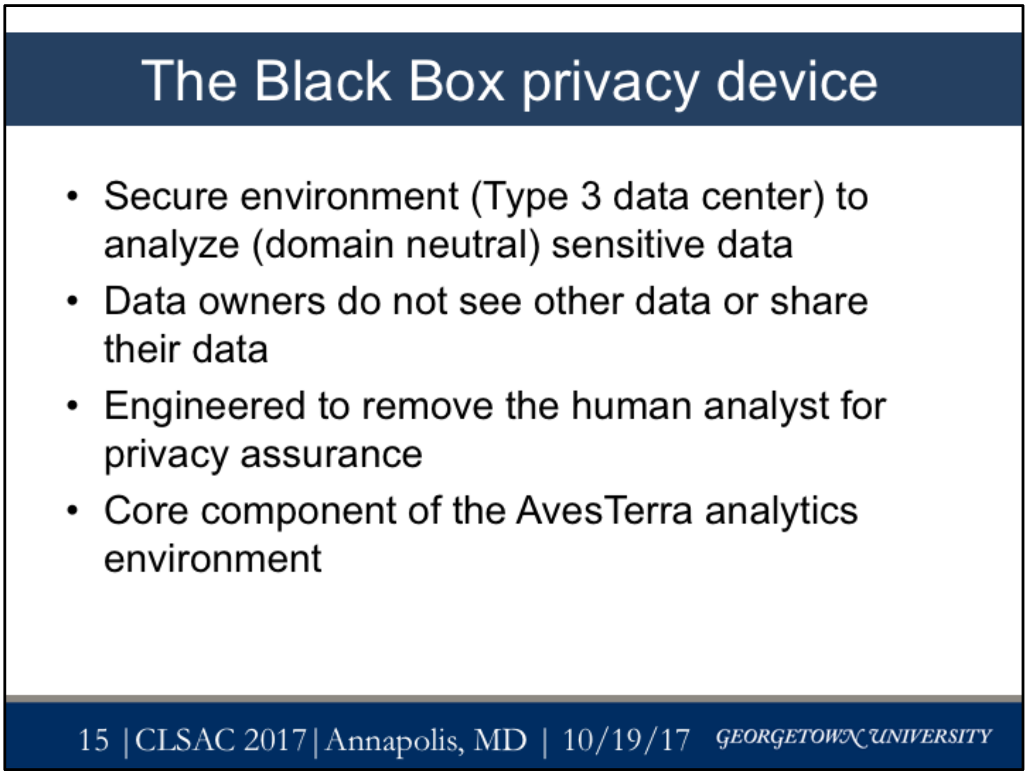
Previous RIDR work

- Previous manual efforts:
  - 21,000 people lost to care
  - 3,000 persons still unaccounted for
  - 3 years of manual effort

- NIH 2015 pilot with ATra:
  - 3 participating jurisdictions (DC, MD, VA)
  - 30 billion pairwise comparisons performed
  - 82% of people identified with 95+% confidence
  - 10 minutes to identify 16,000 people lost to care

The ATra privacy assurance concept





The Black Box privacy device

- Secure environment (Type 3 data center) to analyze (domain neutral) sensitive data
- Data owners do not see other data or share their data
- Engineered to remove the human analyst for privacy assurance
- Core component of the AvesTerra analytics environment

15 | CLSAC 2017 | Annapolis, MD | 10/19/17 GEORGETOWN UNIVERSITY

Type 3 – "best commercial practices". For more details, see [Smart 2016]

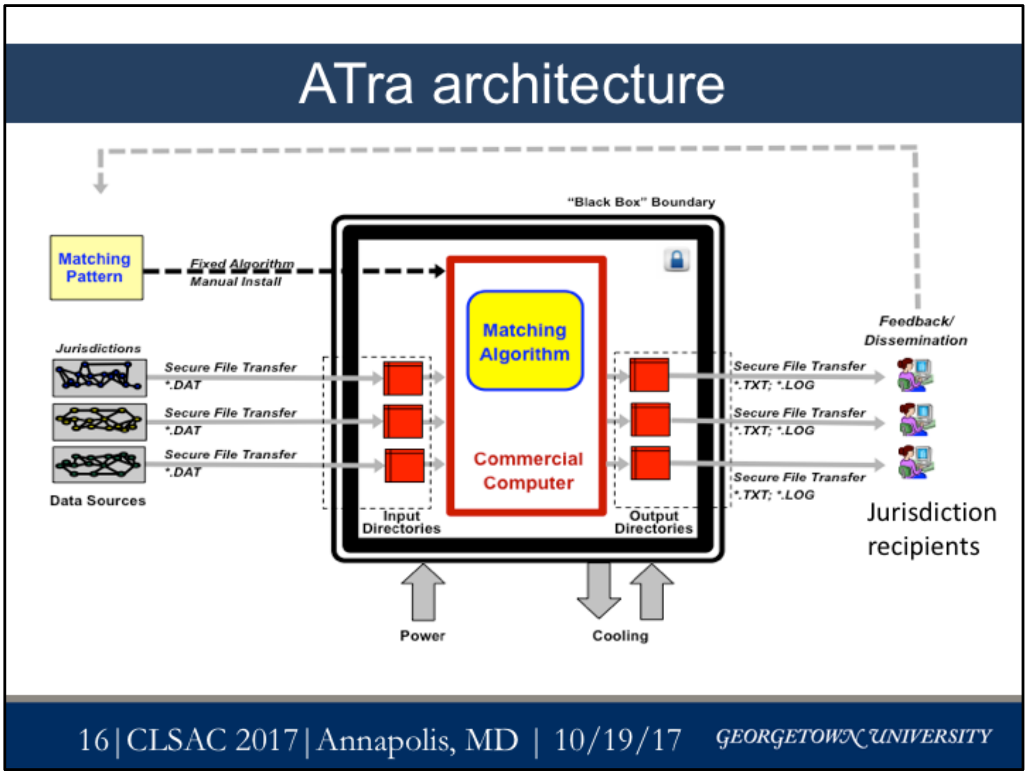
Domain neutral – The Black Box is domain neutral. It can analyze ANY sensitive data, not just public health data or HIV/AIDS data.

Data owners do not see data from other data owners. Their data is not shared with other data owners.

ATra is a methodology that creates a secure environment when analyzing sensitive data.

Unlike other methods that seek to anonymize data, ATra can be engineered to remove the human analyst from the data analysis for privacy assurance.

ATra is a core component of the AvesTerra distributed, large-scale, analytics environment

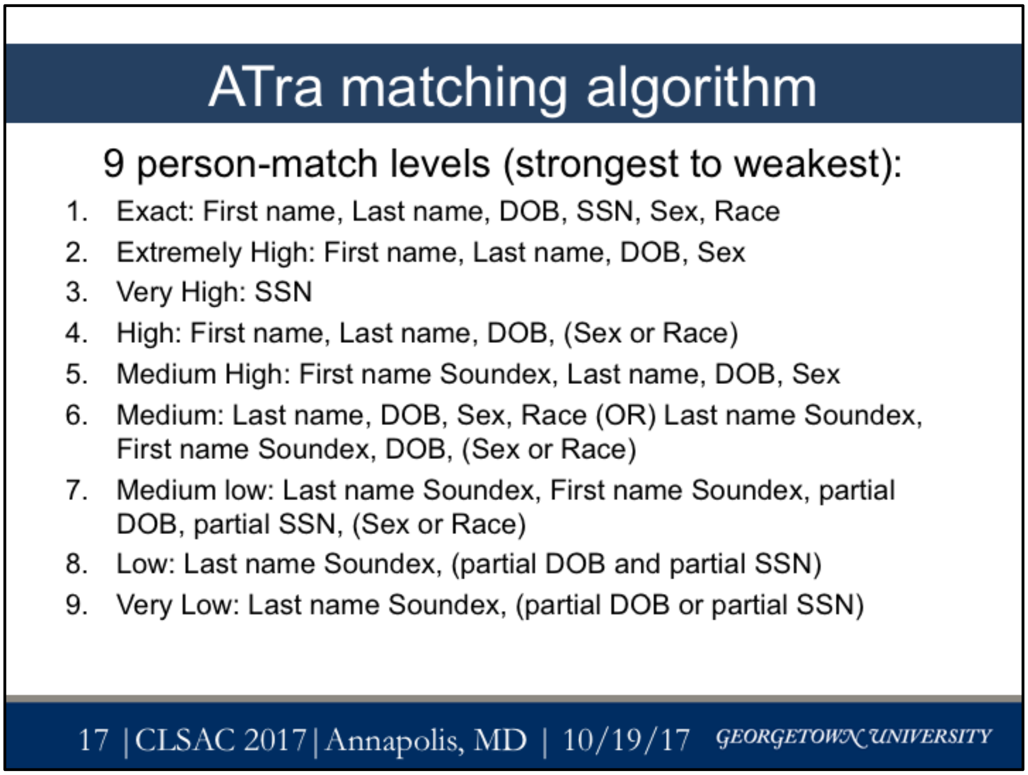


ATra architecture

16 | CLSAC 2017 | Annapolis, MD | 10/19/17 GEORGETOWN UNIVERSITY

Five general design principles informed both the NIH and the CDC Black Box projects.

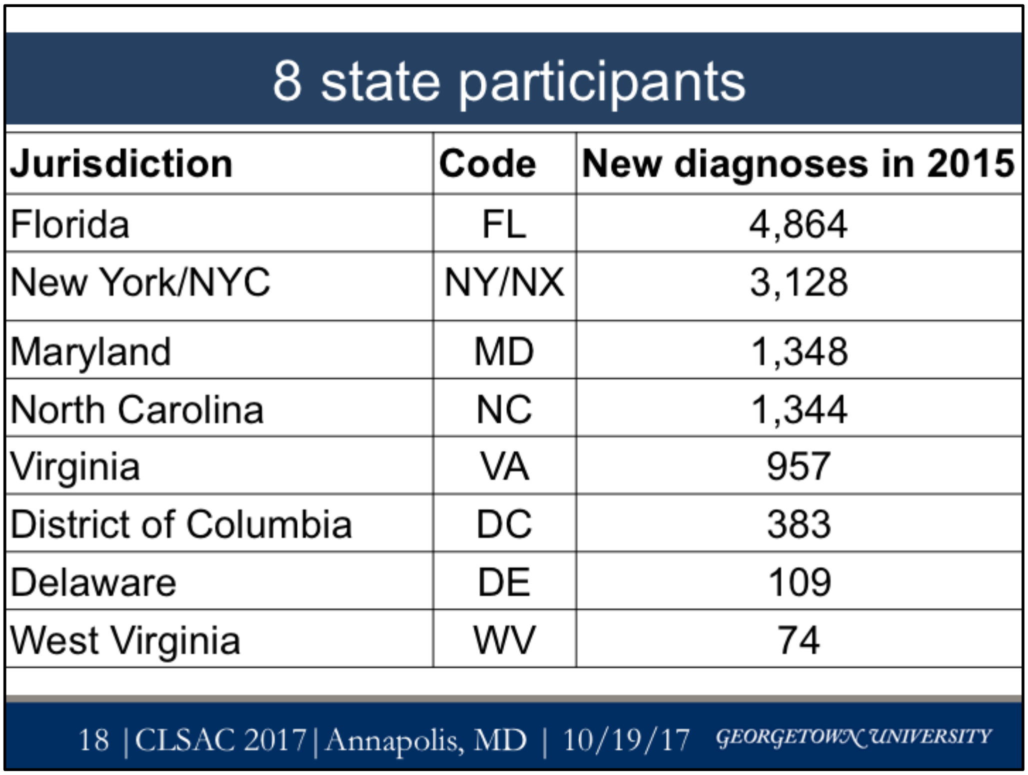
1. The computation platform resides in a contained space with varying layers of perimeter protection depending on the location
2. Stripped computation platform includes only those features necessary to run and support data processing through the algorithm.
3. Only the jurisdictions have network access to the Black Box and only into their individual, dedicated directories.
4. Incoming data is securely erased from the input directory upon being accepted by the algorithm for processing. The incoming data resides temporarily only in volatile memory .
5. Power dies to the Black Box if someone tries to gain physical access, thus wiping all data in the directories or volatile memory.



ATra matching algorithm

9 person-match levels (strongest to weakest):

1. Exact: First name, Last name, DOB, SSN, Sex, Race
2. Extremely High: First name, Last name, DOB, Sex
3. Very High: SSN
4. High: First name, Last name, DOB, (Sex or Race)
5. Medium High: First name Soundex, Last name, DOB, Sex
6. Medium: Last name, DOB, Sex, Race (OR) Last name Soundex, First name Soundex, DOB, (Sex or Race)
7. Medium low: Last name Soundex, First name Soundex, partial DOB, partial SSN, (Sex or Race)
8. Low: Last name Soundex, (partial DOB and partial SSN)
9. Very Low: Last name Soundex, (partial DOB or partial SSN)



8 state participants

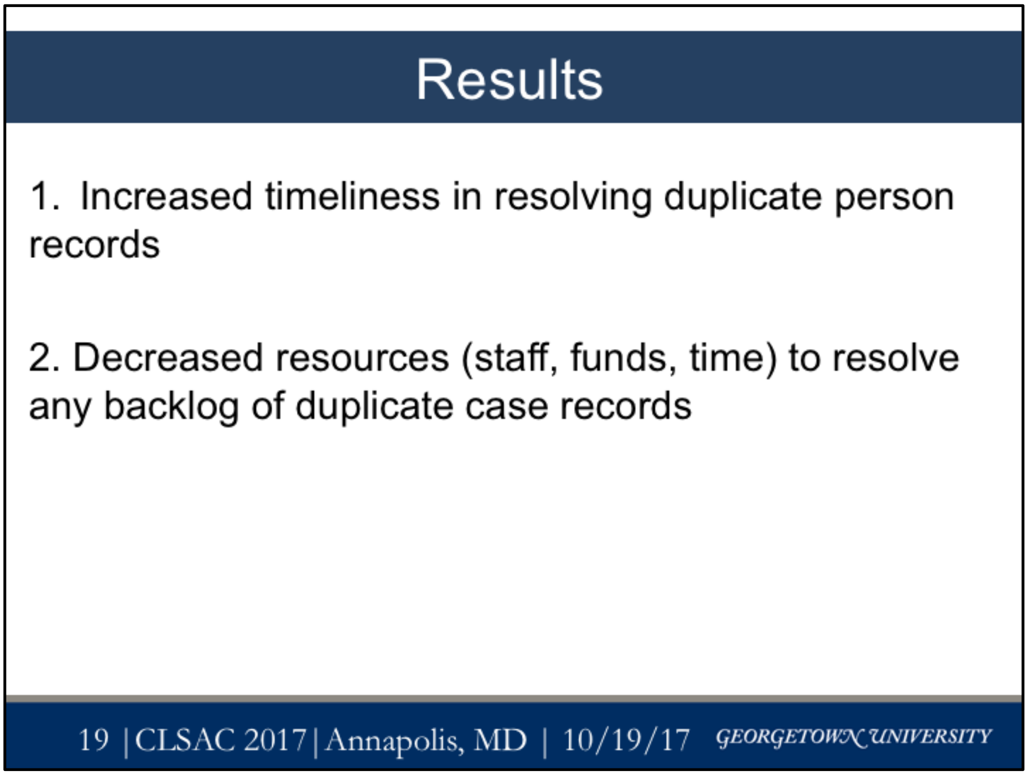
Jurisdiction	Code	New diagnoses in 2015
Florida	FL	4,864
New York/NYC	NY/NX	3,128
Maryland	MD	1,348
North Carolina	NC	1,344
Virginia	VA	957
District of Columbia	DC	383
Delaware	DE	109
West Virginia	WV	74

18 |CLSAC 2017|Annapolis, MD | 10/19/17 GEORGETOWN UNIVERSITY

The figures in Column are from [CDC 2016a].

Florida had the highest number of new diagnoses of HIV/AIDS in 2015 of any U.S. jurisdiction. The large number of PLWH are in the South region of the United, which includes Florida

For more statistics on HIV/AIDS, see [CDC 2016a], [CDC 2017a] and [CDC 2017b]



Results

1. Increased timeliness in resolving duplicate person records

2. Decreased resources (staff, funds, time) to resolve any backlog of duplicate case records

19 |CLSAC 2017|Annapolis, MD | 10/19/17 GEORGETOWN UNIVERSITY

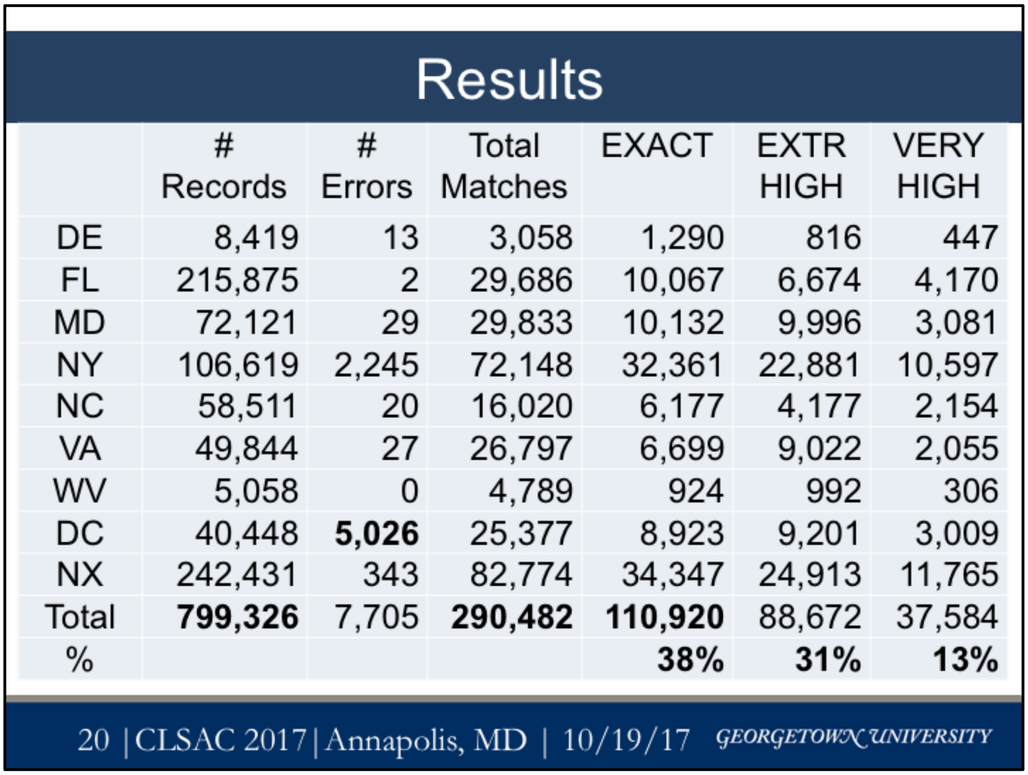
The CDC identified five measures of benefit that were expected from the RIDR Black Box process.

Measure 1: Timeliness.

The amount of time to resolve the EXACT matched case pairs through phone calls was reduced to almost zero. The data is to the CDC and to the jurisdictions much earlier than by the conventional method. Not having to spend time on the EXACT matches also enables the jurisdictions to begin and finish work earlier on all other case pairs on the RIDR list, and, potentially, makes the entire RIDR process more timely and complete.

Measure 2: Decreased resources. Resources include the number of staff, funds, and time to resolve any backlog of duplicate case records for the eHARS RIDR process. “Minutes saved” was used as a surrogate measure of all resources.

We found that the reduction in resources saved by the Black Box varies depending on the level of match accepted for direct importation. For EXACT matches, the estimated time that was saved was approximately 135 hours of calling time. The 126 hours is actual phone time. It does not include the additional time necessary to arrange the calls.



Results

	# Records	# Errors	Total Matches	EXACT	EXTR HIGH	VERY HIGH
DE	8,419	13	3,058	1,290	816	447
FL	215,875	2	29,686	10,067	6,674	4,170
MD	72,121	29	29,833	10,132	9,996	3,081
NY	106,619	2,245	72,148	32,361	22,881	10,597
NC	58,511	20	16,020	6,177	4,177	2,154
VA	49,844	27	26,797	6,699	9,022	2,055
WV	5,058	0	4,789	924	992	306
DC	40,448	**5,026**	25,377	8,923	9,201	3,009
NX	242,431	343	82,774	34,347	24,913	11,765
Total	**799,326**	7,705	**290,482**	**110,920**	88,672	37,584
%				**38%**	**31%**	**13%**

20 | CLSAC 2017 | Annapolis, MD | 10/19/17 GEORGETOWN UNIVERSITY

The total number of records that were uploaded was 799,326.

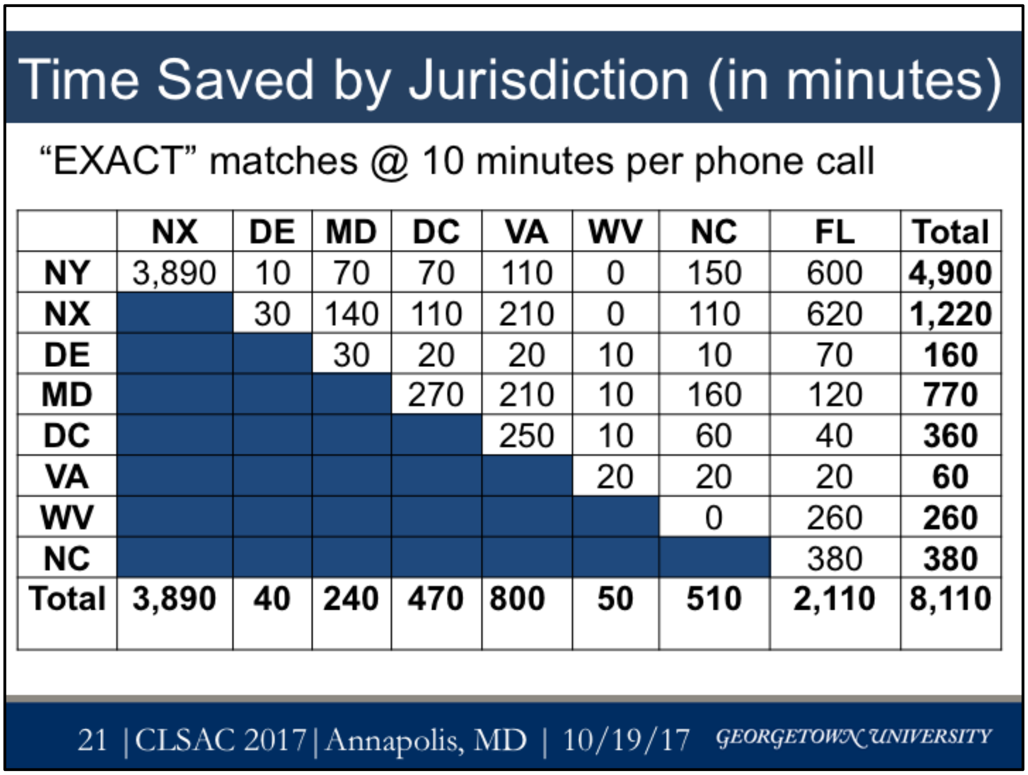
Some of the records could not be used because of errors, such as required data fields that were missing. Those records with errors numbered 7,705.

DC had a large error rate in their data because of missing names (first names and last names) in their early data collection. This goes back to the slide that we saw earlier that speaks to the wide variety of data collection practices, especially before April 2008.

The total number of records uploaded to the ATra system were just under 800,000 – 791,621 to be precise.

Of those approximately 800,000 records, approximately 290,000 of the records – 36.7% -- were matched at some level.

Of the total matches, 82% of the matches occurred at the EXACT level (38%), EXTREMELY HIGH level (31%), or the VERY HIGH level (13%).



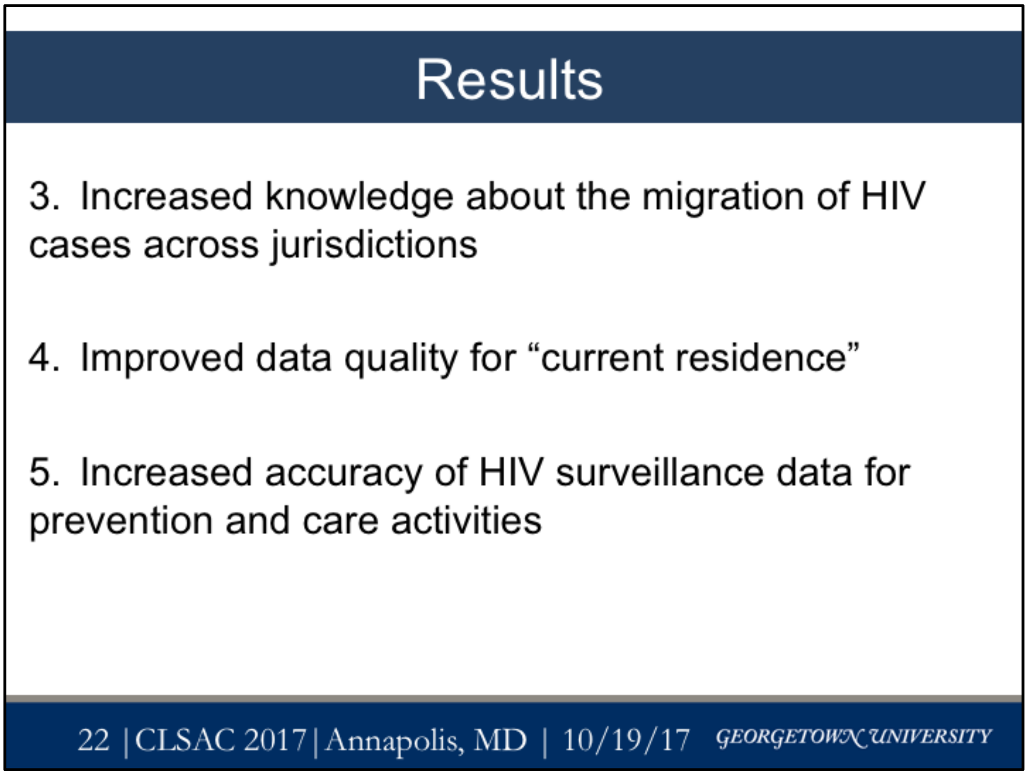
Time Saved by Jurisdiction (in minutes)

"EXACT" matches @ 10 minutes per phone call

	NX	DE	MD	DC	VA	WV	NC	FL	Total
NY	3,890	10	70	70	110	0	150	600	**4,900**
NX		30	140	110	210	0	110	620	**1,220**
DE			30	20	20	10	10	70	**160**
MD				270	210	10	160	120	**770**
DC					250	10	60	40	**360**
VA						20	20	20	**60**
WV							0	260	**260**
NC								380	**380**
Total	**3,890**	**40**	**240**	**470**	**800**	**50**	**510**	**2,110**	**8,110**

21 | CLSAC 2017 | Annapolis, MD | 10/19/17 GEORGETOWN UNIVERSITY

135 total hours



Results

3. Increased knowledge about the migration of HIV cases across jurisdictions

4. Improved data quality for “current residence”

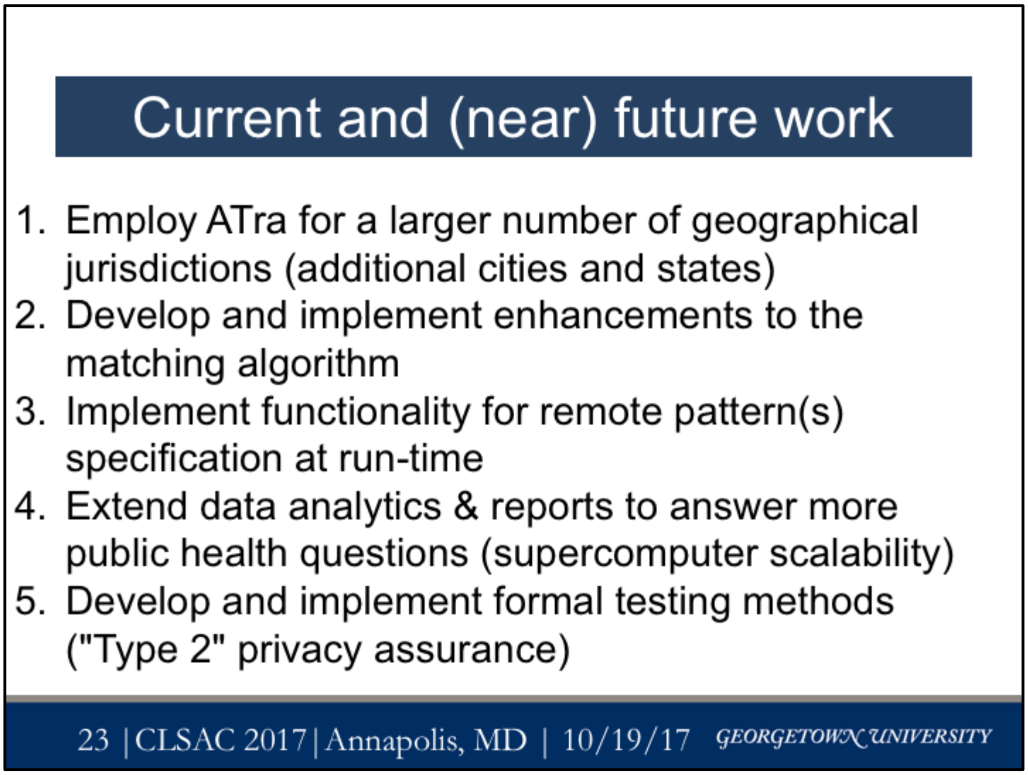
5. Increased accuracy of HIV surveillance data for prevention and care activities

22 | CLSAC 2017 | Annapolis, MD | 10/19/17 GEORGETOWN UNIVERSITY

Measure 3: Knowledge about migration. In an effort to provide a tool for exploration of migration information implicit in the Ridealong variables, Dr. J.C. Smart coded the algorithm to produce a report that tracks match totals for each ZIP code based on where individuals currently are living, were living at HIV disease, were living at HIV infection diagnosis, and were living at AIDS diagnosis. Each jurisdiction received this report as part of the results of the Black Box match on August 4, 2017.

Measure 4: Improved quality for case 'current residence' data. All jurisdictions received data from other jurisdictions about the “current residence” of matched case pairs for EXACT, EXTREMELY HIGH, and VERY HIGH matches.

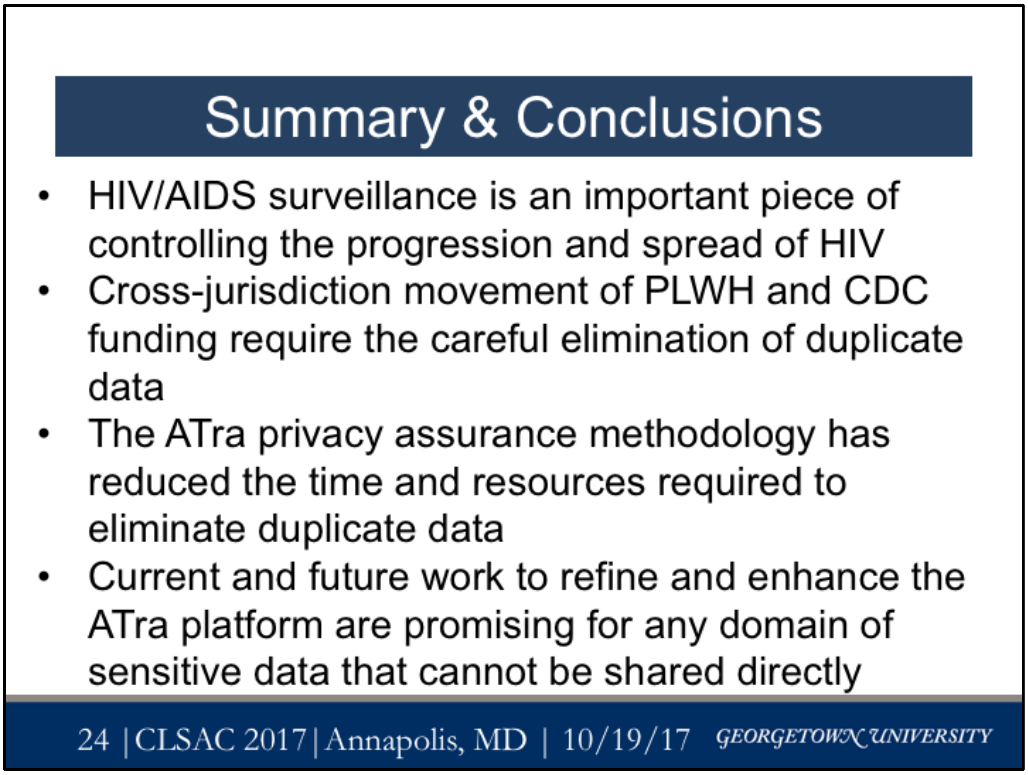
Measure 5: Increased accuracy of HIV surveillance data for prevention and care activities. The extensive exchange of data about persons in the jurisdictions’ eHARS archive offers great potential for increased accuracy of HIV surveillance data. The ultimate value of this substantial data sharing, however, depends upon creating the organizational means and conditions for participating jurisdictions to continue their collaboration that began during this project. In a previous study [Ocampo 2016], the jurisdictions of the District of Columbia, Maryland and Virginia developed routine processes of sharing HIV surveillance data through a variety of measures, including data exchanges, ongoing contacts, and face-to-face meetings on HIV surveillance topics.



Current and (near) future work

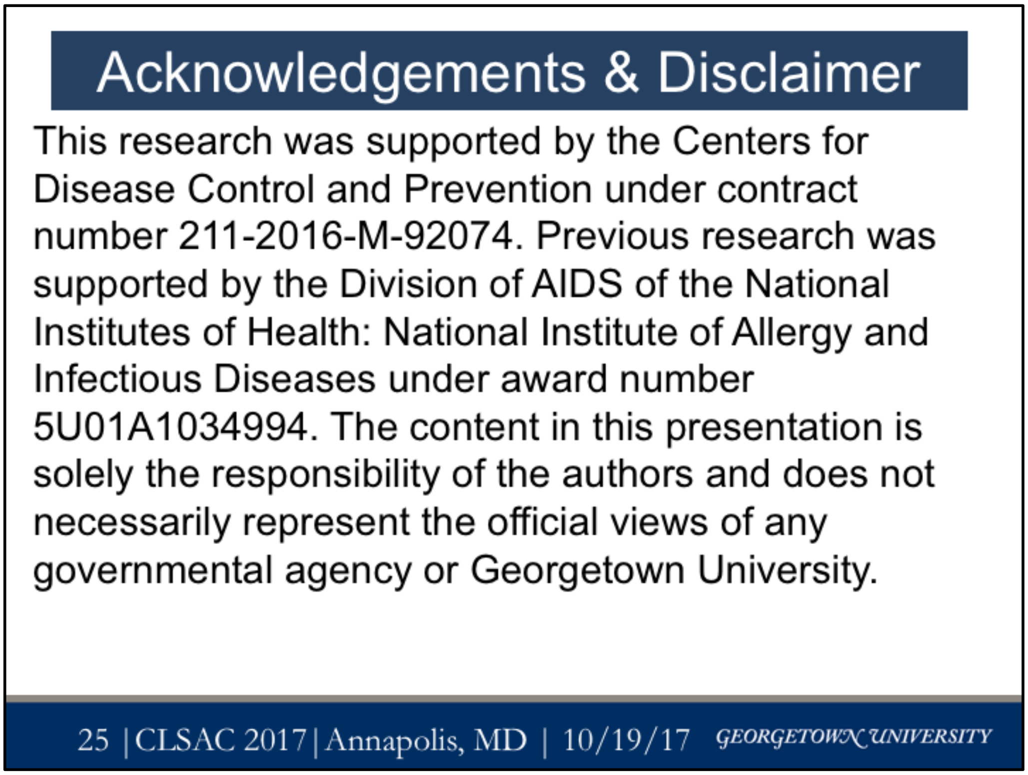
1. Employ ATra for a larger number of geographical jurisdictions (additional cities and states)
2. Develop and implement enhancements to the matching algorithm
3. Implement functionality for remote pattern(s) specification at run-time
4. Extend data analytics & reports to answer more public health questions (supercomputer scalability)
5. Develop and implement formal testing methods ("Type 2" privacy assurance)

Type 2 privacy assurance – for more details, see [Smart 2016]



Summary & Conclusions

- HIV/AIDS surveillance is an important piece of controlling the progression and spread of HIV
- Cross-jurisdiction movement of PLWH and CDC funding require the careful elimination of duplicate data
- The ATra privacy assurance methodology has reduced the time and resources required to eliminate duplicate data
- Current and future work to refine and enhance the ATra platform are promising for any domain of sensitive data that cannot be shared directly



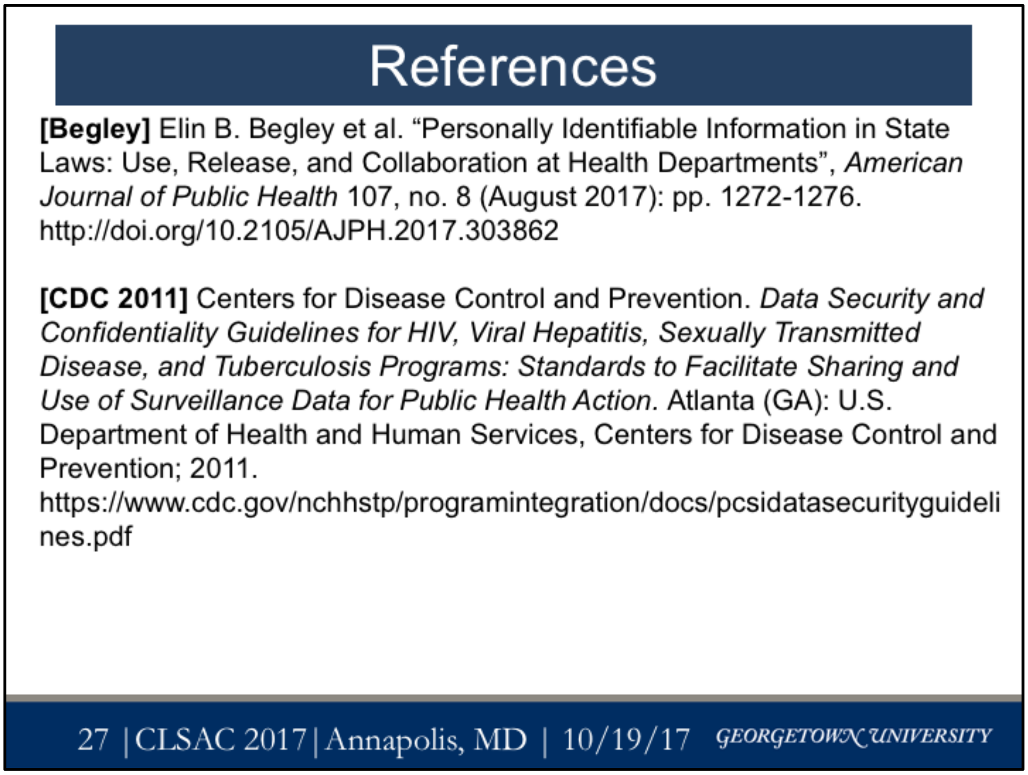
Acknowledgements & Disclaimer

This research was supported by the Centers for Disease Control and Prevention under contract number 211-2016-M-92074. Previous research was supported by the Division of AIDS of the National Institutes of Health: National Institute of Allergy and Infectious Diseases under award number 5U01A1034994. The content in this presentation is solely the responsibility of the authors and does not necessarily represent the official views of any governmental agency or Georgetown University.

Contacts

Dr. J. C. Smart
AvesTerra Chief Scientist and Research Professor
Office of the Senior Vice President for Research
Georgetown University, Washington, DC
smart@georgetown.edu

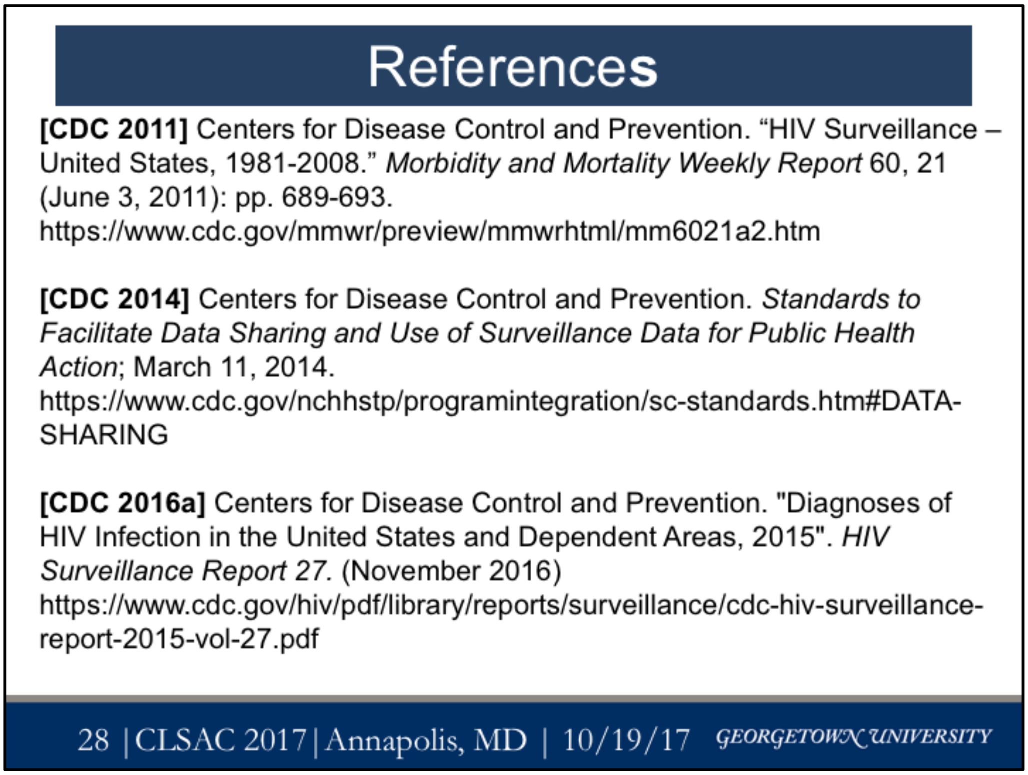
Presenter:
Dr. Helen Karn
Research Specialist – Computational Sciences
Office of the Senior Vice President for Research
Georgetown University, Washington, DC
karnh@georgetown.edu



References

[Begley] Elin B. Begley et al. “Personally Identifiable Information in State Laws: Use, Release, and Collaboration at Health Departments”, *American Journal of Public Health* 107, no. 8 (August 2017): pp. 1272-1276. http://doi.org/10.2105/AJPH.2017.303862

[CDC 2011] Centers for Disease Control and Prevention. *Data Security and Confidentiality Guidelines for HIV, Viral Hepatitis, Sexually Transmitted Disease, and Tuberculosis Programs: Standards to Facilitate Sharing and Use of Surveillance Data for Public Health Action.* Atlanta (GA): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention; 2011. https://www.cdc.gov/nchhstp/programintegration/docs/pcsidatasecurityguidelines.pdf

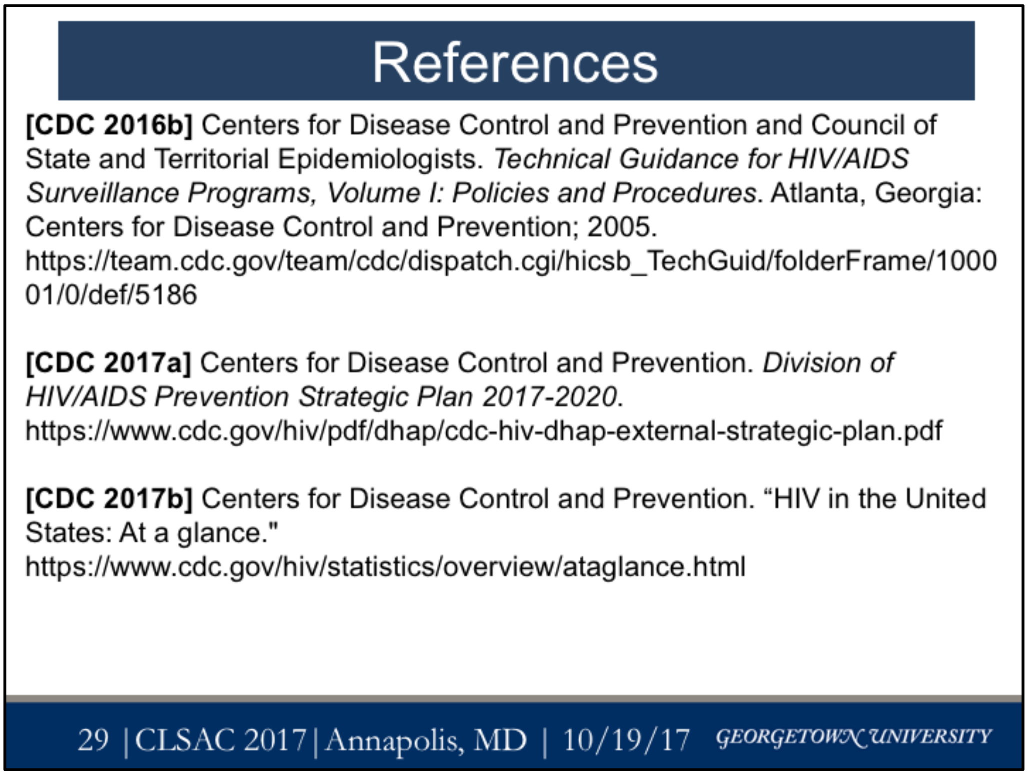


References

[CDC 2011] Centers for Disease Control and Prevention. “HIV Surveillance – United States, 1981-2008.” *Morbidity and Mortality Weekly Report* 60, 21 (June 3, 2011): pp. 689-693. https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6021a2.htm

[CDC 2014] Centers for Disease Control and Prevention. *Standards to Facilitate Data Sharing and Use of Surveillance Data for Public Health Action*; March 11, 2014. https://www.cdc.gov/nchhstp/programintegration/sc-standards.htm#DATA-SHARING

[CDC 2016a] Centers for Disease Control and Prevention. "Diagnoses of HIV Infection in the United States and Dependent Areas, 2015". *HIV Surveillance Report 27.* (November 2016) https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-report-2015-vol-27.pdf

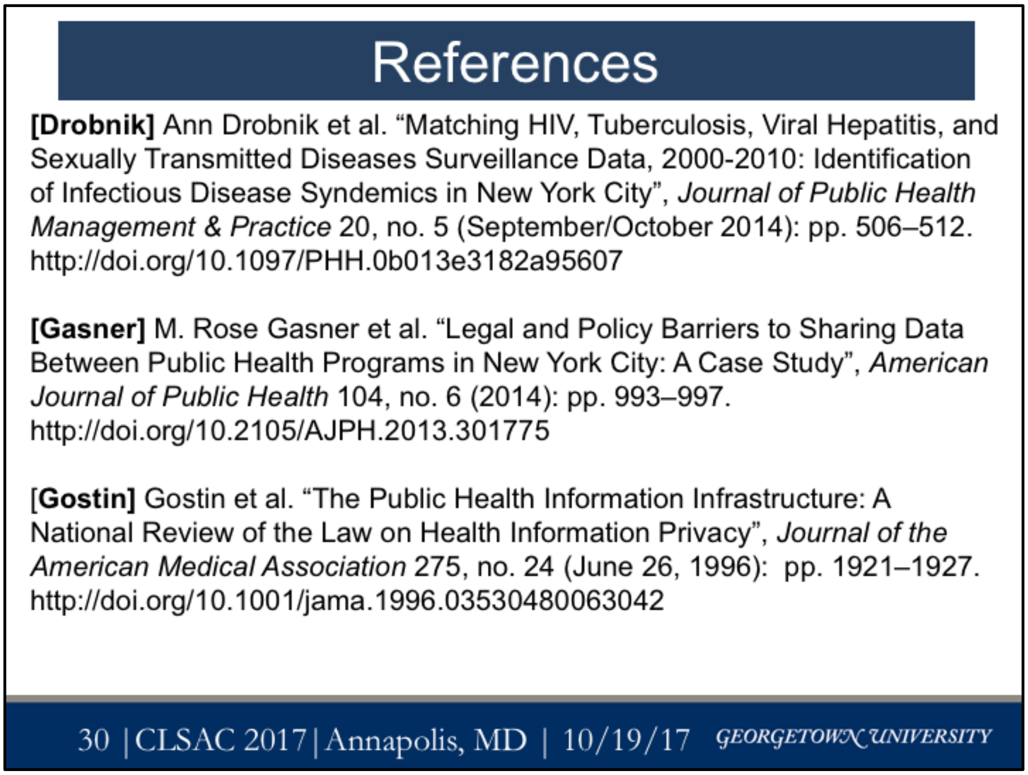


References

[CDC 2016b] Centers for Disease Control and Prevention and Council of State and Territorial Epidemiologists. *Technical Guidance for HIV/AIDS Surveillance Programs, Volume I: Policies and Procedures*. Atlanta, Georgia: Centers for Disease Control and Prevention; 2005. https://team.cdc.gov/team/cdc/dispatch.cgi/hicsb_TechGuid/folderFrame/100001/0/def/5186

[CDC 2017a] Centers for Disease Control and Prevention. *Division of HIV/AIDS Prevention Strategic Plan 2017-2020*. https://www.cdc.gov/hiv/pdf/dhap/cdc-hiv-dhap-external-strategic-plan.pdf

[CDC 2017b] Centers for Disease Control and Prevention. “HIV in the United States: At a glance." https://www.cdc.gov/hiv/statistics/overview/ataglance.html



References

[Drobnik] Ann Drobnik et al. “Matching HIV, Tuberculosis, Viral Hepatitis, and Sexually Transmitted Diseases Surveillance Data, 2000-2010: Identification of Infectious Disease Syndemics in New York City”, *Journal of Public Health Management & Practice* 20, no. 5 (September/October 2014): pp. 506–512. http://doi.org/10.1097/PHH.0b013e3182a95607

[Gasner] M. Rose Gasner et al. “Legal and Policy Barriers to Sharing Data Between Public Health Programs in New York City: A Case Study”, *American Journal of Public Health* 104, no. 6 (2014): pp. 993–997. http://doi.org/10.2105/AJPH.2013.301775

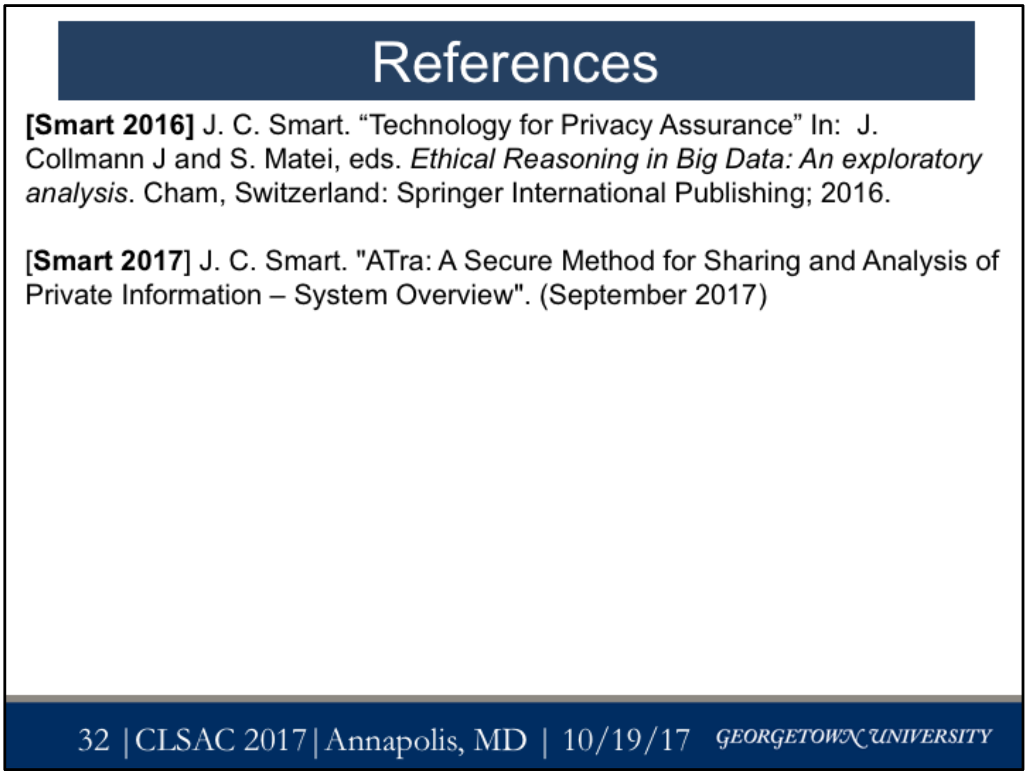
[**Gostin]** Gostin et al. “The Public Health Information Infrastructure: A National Review of the Law on Health Information Privacy”, *Journal of the American Medical Association* 275, no. 24 (June 26, 1996): pp. 1921–1927. http://doi.org/10.1001/jama.1996.03530480063042

References

[GU OSVPR 2017] Georgetown University Office of the Senior Vice President for Research et al. “Privacy Data Sharing Tool to Support De-duplication of Cases in the National HIV Surveillance System (NHSS)”, Final Report submitted to the Center for Disease Control for Contract #211-2016-M-92074

[Ocampo 2016] Joanne Michelle Ocampo et al. “Improving HIV Surveillance Data for Public Health Action in Washington, DC: A Novel Multiorganizational Data-Sharing Method”, *JMIR Public Health Surveillance* 2, no. 1 (e3). DOI: 10.2196/publichealth.5317
http://publichealth.jmir.org/2016/1/e3/

[O’Connor & Matthews] Jean O’Connor & Gene Matthews. “Informational Privacy, Public Health, and State Laws.”, *American Journal of Public Health* 101, no. 10 (October 2011): pp. 1845–1850.
http://doi.org/10.2105/AJPH.2011.300206



References

[Smart 2016] J. C. Smart. “Technology for Privacy Assurance” In: J. Collmann J and S. Matei, eds. *Ethical Reasoning in Big Data: An exploratory analysis*. Cham, Switzerland: Springer International Publishing; 2016.

[**Smart 2017**] J. C. Smart. "ATra: A Secure Method for Sharing and Analysis of Private Information – System Overview". (September 2017)