Published in final edited form as:
J Acquir Immune Defic Syndr. 2019 September 01; 82(Suppl 1): S13–S19. doi:10.1097/QAI.0000000000002090.

[*]**Corresponding Author:** Auntré Hamp, Address: 2115 Wisconsin Avenue NW, Suite 603, Washington DC, 20007, Fax: (202) 687-9109.

Author Contributions

Joanne Michelle F. Ocampo[1a/d], contributed to study design, led manuscript coordination, manuscript development/edits.
Auntré Hamp[1,5], contributed to study design, implementation, manuscript development, and is the corresponding author.
Anne Rhodes[2], contributed to study design and implementation, and manuscript development.
J. C. Smart[1b], contributed to study design and implementation, and manuscript development.
Raghu Pemmaraju[1c], contributed to study design and implementation, and manuscript development.
Karalee Poschman[3], contributed to study design and implementation, and manuscript development.
Kristen L. Hess[4], contributed to study design and implementation, and manuscript development.
Reshma Bhattacharjee[6], contributed to study design and implementation, and manuscript development.
Colin Flynn[6], contributed to study design and implementation, and manuscript development.
Bridget J. Anderson[7], contributed to study design and implementation, and manuscript development.
James E. Dowling[8], contributed to study design and implementation, and manuscript development.
Fred Maccormack[8], contributed to study design and implementation, and manuscript development.
Rupali Doshi[5], contributed to study design and implementation, and manuscript development.
Garret Lum[5], contributed to study design and implementation, and manuscript development.
Lorene Maddox[3], contributed to study design and implementation, and manuscript development.
Brenda Moncur[7], contributed to study design and implementation, and manuscript development.
John E. Barnhart[9], contributed to study design and implementation, and manuscript development.
Jason Maxwell[9], contributed to study design and implementation, and manuscript development.
Sahithi Boggavarapu Aurand[2], contributed to study design and implementation, and manuscript development.
Vicki Hogan[10], contributed to study design and implementation.
David Wills[10], contributed to study design and implementation, and manuscript development.
Stacy Prowell[11], contributed to study design and implementation, and manuscript development.
Seble G. Kassaye[1d], contributed to study design and implementation, and manuscript development.
Helen E. Karn[1a], contributed to manuscript development.
Benjamin T. Laffoon[4], contributed to study design and implementation, and manuscript development.
Jeff Collmann[1a] led study design and implementation as Principal Investigator and contributed to manuscript development.

Disclaimers

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.

Notice: This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Conflict of interest

Joanne Michelle F. Ocampo at the time of manuscript preparation and publication was also a part-time employee with the Norwegian Institute of Public Health. This work was only associated with her capacity as a Georgetown University employee and not with the Norwegian Government, and she declares no conflict of interest.
Auntré Hamp at the time of manuscript preparation and publication was employed with Georgetown University, but during the study period was employed with DC Department of Health, and declares no conflict of interest.
Anne Rhodes declares no conflict of interest.
JC Smart declares no conflict of interest.
Raghu Pemmaraju declares no conflict of interest.
Karalee Poschman declares no conflict of interest.
Kristen L. Hess declares no conflict of interest.
Reshma Bhattacharjee declares no conflict of interest.
Colin Flynn declares no conflict of interest.
Bridget J. Anderson declares no conflict of interest.
James E. Dowling declares no conflict of interest.
Fred Maccormack declares no conflict of interest.
Rupali Doshi, declares no conflict of interest.
Garret Lum declares no conflict of interest.
Lorene Maddox declares no conflict of interest.
Brenda Moncur, declares no conflict of interest.
John E Barnhart declares no conflict of interest.
Jason Maxwell declares no conflict of interest.
Sahithi Boggavarapu Aurand declares no conflict of interest.
Vicki Hogan declares no conflict of interest.
David Wills declares no conflict of interest.
Stacy Prowell declares no conflict of interest.

Improving HIV surveillance data by using the ATra Black Box System to assist regional deduplication activities

Joanne Michelle F. Ocampo[1a,d], **Auntré Hamp**[1a,5,*], **Anne Rhodes**[2], **J. C. Smart**[1b], **Raghu Pemmaraju**[1c], **Karalee Poschman**[3,4], **Kristen L. Hess**[4], **Reshma Bhattacharjee**[6], **Colin Flynn**[6], **Bridget J. Anderson**[7], **James E. Dowling**[8], **Fred Maccormack**[8], **Rupali Doshi**[5,12], **Garret Lum**[5], **Lorene Maddox**[3], **Brenda Moncur**[7], **John E. Barnhart**[9], **Jason Maxwell**[9], **Sahithi Boggavarapu Aurand**[2], **Vicki Hogan**[10], **David Wills**[10], **Stacy Prowell**[11], **Seble G. Kassaye**[1d], **Helen E. Karn**[1a], **Benjamin T. Laffoon**[4], **Jeff Collmann**[1a]

[1a]Georgetown University, Office of the Senior Vice President for Research,

[1b]Georgetown University, Department of Computer Science,

[1c]Georgetown University, University Information Systems,

[1d]Georgetown University, Department of Medicine, Division of Infectious Diseases,

[2]Virginia, Department of Health,

[3]Florida, Department of Health,

[4]Centers for Disease Control and Prevention,

[5]District of Columbia, Department of Health,

[6]Maryland, Department of Health,

[7]New York State Department of Health,

[8]Delaware Division of Public Health,

[9]North Carolina, Department of Health,

[10]West Virginia Department of Health and Human Resources, Bureau for Public Health,

[11]Oak Ridge National Laboratory,

[12]The George Washington University

Abstract

Background—Focused attention on Data to Care underlines the importance of high quality HIV surveillance data. This study identified the number of total duplicate and exact duplicate HIV case records in nine separate Enhanced HIV/AIDS Reporting System (eHARS) databases reported by eight jurisdictions, and compared this approach to traditional Routine Interstate Duplicate Review (RIDR) resolution.

Seble G. Kassaye declares no conflict of interest.
Helen E. Karn declares no conflict of interest.
Benjamin T. Laffoon declares no conflict of interest.
Jeff Collmann was retired from his faculty position at Georgetown University throughout project planning, execution, and analysis as well as preparation and publication of this manuscript and declares no conflict of interest.

Methods—This study used the ATra Black Box System and six eHARS variables for matching case records across jurisdictions: Last Name, First Name, Date of Birth (DOB), Sex assigned at birth (Birth Sex), Social Security Number (SSN), and Race/Ethnicity, plus four system-calculated values (First Name Soundex, Last Name Soundex, Partial DOB, Partial SSN).

Results—In approximately 11 hours, this study matched 290,482 cases from 799,326 uploaded records, including 55,460 exact case pairs. Top case pair overlaps were between NYC and NYS (51%), DC and MD (10%), and FL and NYC (6%), followed closely by FL and NYS (4%), FL and NC (3%), DC and VA (3%), and MD and VA (3%). Jurisdictions estimated that they realized a combined 135 labor hours in time efficiency by using this approach compared with manual methods previously used for interstate duplication resolution.

Discussion—This approach discovered exact matches that were not previously identified. It also decreased time spent resolving duplicated case records across jurisdictions while improving accuracy and completeness of HIV surveillance data in support of public health program policies. Future uses of this approach should consider standardized protocols for post-processing eHARS data.

Keywords

HIV surveillance; Data to Care; data quality; deduplication; case pair resolution; ATra Black Box System

Background

Critical public health tasks to improve population-level health outcomes for persons with HIV (PWH) include early diagnosis of HIV, rapid linkage to HIV care, and treatment with antiretroviral medications to achieve viral suppression.[1,2] However, for public health departments, it remains challenging to achieve optimal levels of these goals in part due to the difficulty in accurately measuring this spectrum, otherwise known as the HIV Care Continuum.[3] In the United States, interstate migration and differences in state and local public health reporting laws and interpretations among jurisdictions regarding data sharing and privacy, challenge accurate measurements of the HIV Care Continuum, which in turn, affect the public health outreach and intervention that depend on these data.[4]

Data to Care[5] is a public health strategy that aims to use HIV surveillance data to identify individuals with diagnosed HIV who are not in care, link or reengage them to care, and support the HIV Care Continuum.[6] The Data to Care strategy relies on accurate data, and in particular, current residential address, vital status, and care status, which are collected in HIV surveillance systems at state/local health departments. A key characteristic of a well-functioning surveillance system and its data quality is its ability to link records on the same person across different jurisdictions to minimize duplicate records of reports/cases. In the United States, much of this information is collected through deduplication activities among jurisdictions. The Centers for Disease Control and Prevention (CDC) coordinates the Routine Interstate Duplicate Review (RIDR). This is a bi-annual process to identify and resolve duplicate cases in the Enhanced HIV/AIDS Reporting System (eHARS) across public health jurisdictions, for whom this process is a condition for receiving

CDC surveillance funds.[7,8] CDC identifies records suspected of being duplicate reports on the same individual using a CDC-developed matching algorithm. CDC then provides jurisdictions with lists of suspected duplicate records for them to review, discuss, and agree upon a resolution (‘same as’ or ‘different than’) during resource intensive telephone case conferencing between jurisdictions. Currently, RIDR operates with an estimated 12-month time lag between case reporting and duplicate resolution, as the process involves extensive manual follow-up for case pair resolution across jurisdictions.

In 2015, the health departments of the District of Columbia (DC), Maryland (MD), and Virginia (VA) with Georgetown University used a novel privacy-assuring data technology—the ATra Black Box System — to identify 21,472 eHARS potential duplicates from 161,343 case records across the three public health jurisdictions in a computational processing time of 21 minutes and 58 seconds.[9] This previous study showed significant eHARS case record overlap across jurisdictions in the DC metropolitan area, reflecting persons’ interactions with health systems reporting to different public health departments across these jurisdictional borders. It also gave jurisdictions the opportunity to improve accuracy of their data by identifying additional cases that were actually still in care but living out of jurisdiction and those who were deceased.

The study detailed here sought to examine the public health utility of using the ATra Black Box System in an expanded geographic area to determine its potential role in improving efficiency of case-pair identification and determine the improvements in overall quality of HIV surveillance data across participating jurisdictions.

Study objectives

Our study objective was to use the ATra Black Box System approach for the District of Columbia (DC); Delaware (DE); Florida (FL); Maryland (MD), North Carolina (NC); New York State (NYS), including data from New York City (NYC); Virginia (VA), and; West-Virginia (WV) to: 1) identify the overall number of duplicate case records in eHARS across jurisdictions; 2) identify the number of exact duplicate case records in eHARS across jurisdictions; and, 3) compare this approach to traditional RIDR resolution by estimating time efficiency realized and assess congruence with the July 2017 RIDR process.

Methods

A governing body

This highly collaborative technical approach was contingent upon first establishing a governing body that determined the hypothesis in question and met regularly to discuss and implement study activities. This body included regional representatives from jurisdictional sites and study partners, all of whom received legal clearance to participate in this study – a process that involved productive dialogue between participating organizations. This governing body consisted of members from public health jurisdictions (DC, DE, FL, MD, NC, NYS, VA, WV), and members from Georgetown University (GU), CDC, and Oak Ridge National Laboratory (ORNL). Guided by the public health jurisdictions’ need, this body reached consensus in selecting the analytical question to query via this data technology: what is the total number and nature of duplicate HIV case records across participating jurisdictions along the East Coast corridor?

Data privacy and ethics

Among the jurisdictions on the United States East Coast that were offered the opportunity to participate, eight (DC, DE, FL, MD, NC, NYS, VA, and WV) agreed to participate in this effort. The jurisdictions and GU, with support from ORNL drafted, agreed upon and signed Data Sharing Agreements and a Data Security and Confidentiality Procedures Manual following CDC's standard format for such documents.[10] In NYS, the state and New York City (NYC) maintain separate HIV surveillance databases, but NYC reports cases to NYS for duplicate resolution purposes on a weekly basis. As a participating jurisdiction in this effort, NYS submitted their eHARS data as well as NYC's, making a total of nine eHARS data sets included in the final match. Due to the privacy-centered engineering design and technical approach of this study, which prohibits any person from seeing the data once in the ATra Black Box system and disallows long-term permanent storage of data, the GU Institutional Review Board (IRB) determined that the study was exempt from review.

Data technology

The ATra Black Box System approach has been described previously.[9] Briefly, the ATra Black Box System has a physically protected server with extremely high privacy assurance that was located at a secure Data Center in Virginia. Once closed, no one was able to inspect its contents, including the system administrators or the software developers. This server had no external connections to any device other than a power source. It saved data in temporary memory for data matching[9,11], and was programmed with manual and automatic mechanisms for cleaning out memory in the event of non-authorized access. The ATra Black Box System was available only to participating jurisdictions through designated encrypted Virtual Private Network (VPN) links. Encryption techniques were in compliance with the Advanced Encryption Standard (AES) to protect the highly sensitive public health HIV data during transit between the jurisdiction and the ATra Black Box System. For this study, the system securely processed eHARS data uploaded directly from each jurisdiction without permanently storing the data. Each jurisdiction was assigned a single, unique, dedicated directory on the server. The jurisdictions uploaded tab-delimited data files to their assigned directories. The jurisdictions prepared input files using a SAS program that was written to combine demographic, geographic, HIV diagnostic and laboratory information from each jurisdiction's eHARS database. Each jurisdiction downloaded output reports from their individually assigned subdirectory upon match completion. Each jurisdiction received information pertinent to only their jurisdiction, including the results of match runs, a real-time log, an error report, a case-by-case match report with values of additional variables for the three highest match categories, eHARS-importable files, match totals, grand totals, and matches by zip code.

Information Technology system coordination and system testing

Information technology (IT) staff from all collaborating jurisdictions and GU collaborated to enable their health department staff to securely upload their eHARS data file for matching and reporting of results. IT and HIV surveillance staff from all jurisdictions became sponsored users of GU's system, and received unique logins, passwords, and VPN access for the duration of this project. All jurisdictions first participated in an End-to-End test of the match process. The End-to-End test served several purposes, including confirming the communication channels between all jurisdictions and the ATra Black Box System server, testing operational efficiencies and technical processes including uploading of the data, monitoring the error logs in real time, downloading the reports, and testing the matching algorithm using a set of nine synthetic data sets (one for each jurisdiction plus one for NYC) provided by CDC. After correcting a minor logic error and running a second test, the system successfully passed all aspects of the End-to-End test as indicated by the precise reproduction of a master list of expected results.

Matching variables and levels

This system used ten matching variables: Last Name, First Name, Date of Birth (DOB), sex assigned at birth (Birth Sex), Social Security Number (SSN), Race/Ethnicity, First Name Soundex, Last Name Soundex, Partial DOB, and Partial SSN. The values for these six matching variables were retrieved from eHARS using the SAS program: Last Name, First Name, DOB, Birth Sex, SSN, and Race. The Black Box calculated these four matching variables: First Name Soundex, Last Name Soundex, Partial DOB, and Partial SSN. Five of the ten matching variables were required to be present in the input data record in order for the record to be processed and matched: Last Name, First Name, Date of Birth, Birth Sex, and Race. Three additional variables were required to be in the input data record, but were not used in the matching process: Stateno, Vital Status, and Transmission Category. The output displayed the number of matched individual HIV case records and the number of matched case pairs (i.e., ≥ two case records representing one unique person with HIV) for each jurisdiction's report files. In this study, the same individual could belong to one or more case pair(s) if they matched across more than two jurisdictions.

There were ten levels of matching confidence:

- Exact: last name and first name and DOB and SSN and Birth sex and race,
- Extremely High: last name and first name and DOB and Birth sex,
- Very High: SSN,
- High: last name and first name and DOB and (Birth sex or race),
- Medium High: (last name and first soundex and DOB and Birth sex) or (last name and first soundex and DOB and Birth sex),
- Medium: (last name and DOB and Birth sex and race) or (last soundex and first soundex and DOB and (Birth sex or race)),
- Medium Low: last soundex and first soundex and partial DOB and partial SSN and (Birth sex or race),
- Low: last soundex and (partial DOB and partial SSN) and (Birth sex or race),
- Very low: last soundex and (partial DOB or partial SSN)

These match levels were previously validated to assess the specificity of the matching algorithm in the Ocampo et al 2016 study.[9] Specificity differed by match level with case-pair matches at the exact level being validated as 100% true matches. The remainder of this paper focuses primarily on exact matches, since jurisdictions found the exact level acceptable for automatic eHARS import without further validation.

In addition, jurisdictions could upload up to 93 optional "ride-along" variables per individual. These variables represented data that are typically exchanged during the manual case resolution that occurs in the traditional RIDR process, including: HIV/AIDS case definition, state and county of residence at diagnosis of HIV/AIDS, current residential address information, laboratory test results associated with initial HIV disease diagnosis, and most recent HIV viral load, and CD4+ T lymphocyte count. Although not included in the matching algorithm, the ride-along variable data were included in the output for exact matches.

Comparing this method to traditional RIDR process

Estimating time efficiency realized—The governing body decided to use minutes per phone call per case pair resolution as an indirect measure of jurisdictional resources spent conducting aspects of the traditional RIDR resolution process, because RIDR resolution is typically conducted via phone to resolve batches of several case pairs. The time was estimated based on the typical amount of time to organize, conduct and document calls between jurisdictions to resolve specific case pairs. Jurisdictions estimated an average of five minutes per call per case with two persons (one from each jurisdiction) for an average of 10 minutes overall. This estimate did not account for variation among local conditions.

Congruence with July 2017 RIDR process—Prior to conducting the ATra Black Box System run, jurisdictions had previously received a CDC July 2017 RIDR list. The CDC July 2017 RIDR list was comprised of previously unresolved potential duplicates that were found in the eHARS system between January 1, 2017 and June 30, 2017. In order to assess the impact of the ATra Black Box System run on the RIDR resolution process, we checked if case-pair matches found by the ATra Black Box at the exact level were also present on the jurisdictions CDC July 2017 RIDR list. Additionally, we examined whether the ATra Black Box System found exact case-pair matches that were not present on the CDC July 2017 RIDR list that had not been previously unresolved in eHARS. We reviewed exact matches that did not appear on the CDC July 2017 RIDR list and those case pairs that were "previously resolved" through previous deduplication efforts and "not previously resolved" in eHARS.

Results

Overall number of duplicate records across jurisdictions

Jurisdictions uploaded a total of 799,326 eHARS case records (DC = 40,448; DE = 8,419; FL = 215,875; MD= 72,121; NC= 58,511; NYC = 242,431; NYS = 106,619; VA = 49,844; WV = 5,058), of which 7,705 (1%) were not uploaded successfully and were reported as errors (data not shown). A total of 290,482 (36%) eHARS records across these eight East

Coast public health jurisdictions matched across all levels: very low (8.9%), low (0.0%), medium low (0.0%), medium (8.1%), medium high (1.2%), high (0.2%), very high (12.9%), extremely high (30.5%), and exact (38.2%) (Table 1). Overall, close to 70% of matches were exact or extremely high.

Exact case pairs across jurisdictions

A total of 110,920 individual case records fell into the exact matching level (Table 1). These cases represent a total of 55,460 case pairs matched at the exact level. As shown in Table 2, the top three eHARS case pairs overlap were between NYC and NYS (51%), DC and MD (10%), and FL and NYC (6%), followed closely by FL and NYS (4%), FL and NC (3%), DC and VA (3%), and MD and VA (3%) (Table 2).

Congruence with July 2017 RIDR process

In July 2017, jurisdictions received their semi-annual RIDR case pair lists from CDC, these case pairs represented possible duplicates of new persons entered between January 1, 2017 and June 30, 2017. A total of 811 exact case pairs identified using this approach also appeared on the July 2017 RIDR lists for jurisdictions (Table 3).

Estimated time efficiency realized

This study estimated that jurisdictions realized approximately 8,110 minutes (or 135.2 labor hours) in time efficiency using this approach compared to aspects of the traditional RIDR resolution process. NYC and NYS conduct an automated intrastate deduplication process. Therefore, the time efficiency realized may be inflated by the large number of matches between NYC/NYS that would be resolved through other methods. The time efficiency realized, when not including the NYC/NYS matches, was approximately 4,220 minutes (or 70.3 labor hours).

Post-processing of results and the NYS case example

To describe the added value of using this approach, NYS examined the number of exact case pair matches that were not on the July 2017 RIDR list compared to other jurisdictions and found case pairs that were defined as “previously resolved” and “previously not resolved” in eHARS (Table 5). In the case of NYS, there were 2,371 case pairs matched as exact identified as “previously not resolved”.

Discussion

Main findings

Here, we demonstrated successfully using the ATra Black Box System to assist deduplication activities across jurisdictions along the United States East Coast corridor. This effort identified previously unidentified duplicates and likely helped realize time efficiency for resource-constrained public health jurisdictions. The highly collaborative public-private partnership between government, academic, and public health partners motivated jurisdictions to increase the frequency at which they directly communicate with each other about overlapping cases, and has thus improved cooperative activities among public health jurisdictions. Moreover, this study addressed the critically important arena of working together to more effectively use surveillance data while enhancing the privacy safeguards for sensitive public health data.

Monthly data transfers from jurisdictions to the CDC provide the National HIV Surveillance System (NHSS) with necessary information to track and monitor HIV across the nation. But by design, CDC does not have access to personal identifiers like First Name, Last Name, or SSN, and thus the variables are not available for deduplication purposes. The ATra Black Box System allows for automatic identification of matches, but without any person seeing or storing personally identifying information while in the matching process. This facilitates increased specificity in the identification and resolution of potential duplicates without compromising privacy.

This work can translate into improved Data to Care efforts reliant on surveillance data by providing more accurate, timely, and updated case data across jurisdictions. Activities related to Data to Care (i.e., linking surveillance data more closely to health care outcomes) underlie the need for the improvement of data quality in HIV surveillance. Enhanced data quality allows jurisdictions to better focus their valuable public health resources on cases in need of follow-up with confidence, and less so on individuals who have demonstrated continued engagement in care. Also, updated HIV surveillance data provides a better overview of HIV for public health planning purposes and has implications for funding public health efforts and health service delivery.

Public health implications

Identification of exact matches, especially those that were previously known to be duplicates, when accompanied by ride-along variable data, enabled jurisdictions to update their local eHARS case records with information from other jurisdictions for more complete and accurate information – complementing the conventional RIDR resolution approach. The ride-along variables also provide added value for future deduplication, such as SSN for positive identification, obtaining demographic characteristics, current address, and HIV transmission risk factors. For local public health jurisdictions, these data are critical for outbreak investigations, as well as epidemiologic analysis and reporting, which act to fine-tune policy and targeting prevention and control activities. An additional benefit of conducting this study was the identification of exact matches that were not previously resolved and not in the July 2017 RIDR list. Such case pairs that were previously unidentified exemplify cases that had not yet been distributed for resolution through the RIDR activity, leading to earlier identification of duplicates and hence improved accuracy of the surveillance data.

Our study suggests that jurisdictions with large seasonal migration or urban areas might stand to benefit the most from use of the ATra Black Box System given the complex nature of movement of people through cities and the need to clarify their interaction with the public health system across jurisdictional borders to ensure the most effective follow-up. This study made it clear that New York State had significant HIV surveillance data overlap with Florida.

The overlap of eHARS records identified here indicates people's interactions with health systems across jurisdictional borders, which is challenging when the system is designed with states independently conducting surveillance in isolation from other jurisdictions; and the data at the national level does not have sufficient detail for deduplication. This paper demonstrates a need to account for the mobility of people living with HIV in the United States, which leads people to engage with health care systems across different states (i.e., non-residence states). This has care implications, especially for public health departments allocating valuable resources to provide outreach, care, and support services to PWH. To reach optimal levels of each milestone in the HIV Care Continuum, and to provide a bridge between public health data and clinical care, there is a need to better understand and perhaps readjust our public health outreach to the dynamic nature of modern living. This study also presents significant benefits to jurisdictions' abilities to update their eHARS data (e.g., updating cases that may have never been identified for duplicate resolution in eHARS through other efforts.) Here, many cases that matched were not previously resolved in eHARS and also were not present on the most recent RIDR list, which presented an opportunity to improve the quality of local HIV surveillance data. Several reasons could explain why case-pairs had not been identified by previous deduplication procedures, including that, by design, identifying information is not available to CDC for traditional national level deduplication efforts, new cases which could possibly be on the upcoming RIDR list, cases from previous RIDR lists that were not resolved, and case-pairs that had been not included in previous deduplication efforts. In each of the potential scenarios, case-pair resolution enables jurisdictions to update and thus enhance their HIV surveillance data and better assist with national-level case pair deduplication.

Future work and study weaknesses

Evaluating time efficiency realized could be expanded to a more comprehensive cost-to-benefit analyses in future efforts. The initial time spent on legal clearances, setting up data sharing agreements, establishing secure IT-protocols between several organizations, and creating and tailoring the matching algorithm needs to be accounted for in future cost-to-benefit analyses. While jurisdictions will naturally spend time on initial set-up activities in the earlier years, we envision that they will spend less time on the same activities in subsequent years because of growing familiarity, experience and continued training with this approach. This work is expected to ultimately reduce the number of duplicated records existing across public health jurisdictions, but needs to be done on a regular basis for maximum efficiency. Future uses of this approach should also incorporate a streamlined mechanism to maximize efficiency and avoid error messages due to data aspects like lack of names or coded names. The location of the ATra Black Box System server may also affect future uses of this approach, and should be discussed with potential users.

In addition to NYS, other jurisdictions have reviewed using eHARS data after an ATra Black Box System run,[12] but to improve overall efficiency there must also be development of best practices for post-processing eHARS data across jurisdictions after matching case pairs using this approach. A single suite of software programs to be used by all participating jurisdictions are required to import matching datasets and ride-along variables back into eHARS, in order to realize the full potential for time and cost savings for this automated data deduplication process. As with other new approaches, there remains significant work, including finding the best methods for evaluating less than exact matches, resolving case pair conflicts, processing ride-along variables, and developing post-processing software.

Future work should detail this highly collaborative process to implement a novel approach to resolving case pairs in eHARS, especially given that some state privacy laws prevented other invited public health jurisdictions from participating in this study. Other uses may also consider further refining the sensitivity of the algorithm to detect more matches without losing specificity. Finally, it would be informative to learn how many case pairs overlap among jurisdictions in other geographic areas and consider how this could improve our understanding of HIV in the United States.

Conclusion

This study identified 290,482 potentially duplicated case records from 799,326 uploaded case records in nine separate eHARS data sets across eight participating jurisdictions, of which 55,460 were exact case pairs. An estimated 135 labor hours in time efficiency was realized using this process to identify duplicate case records in eHARS compared to the traditional process for CDC RIDR resolution. This privacy-centered deduplication process of eHARS records across multiple public health jurisdictions has the potential for improving the timeliness, accuracy, and completeness of nationwide HIV surveillance data. This may help to reduce potential case pairs on future CDC RIDR lists, strengthen collaborative relationships among jurisdictions, provide a fruitful cooperative platform between academia and government entities, and lead to the more efficient use of public health resources. To enhance the added value of using this approach, future applications should consider standardized protocols for post-processing duplicate eHARS data.

Acknowledgements

We would like to extend our gratitude to all administrative and other personnel who have helped to facilitate the success of this study.

Funding

This study was funded by Centers for Disease Control and Prevention (CDC) Contract #211-2016-M-92074. Analyses herein were made possible through the support of CDC PS18-1802: Integrated Human Immunodeficiency Virus Surveillance and Prevention Programs for Health Departments.

References

1. Sweeney P, Gardner LI, Buchacz K, et al. Shifting the paradigm: Using HIV surveillance data as a foundation for improving HIV care and preventing HIV infection. The Milbank Quarterly. 2013;91(3):558–603. doi:10.1111/milq.12018. [PubMed: 24028699]
2. National HIV/AIDS Strategy for the United States: Updated to 2020. URL: https://files.hiv.gov/s3fs-public/nhas-update.pdf [accessed: 2017-12-13]
3. Gardner EM, McLees MP, Steiner JF, et al. The spectrum of engagement in HIV care and its relevance to test-and-treat strategies for prevention of HIV infection. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America. 2011;52(6):793–800. doi:10.1093/cid/ciq243. [PubMed: 21367734]

4. Gill MJ, Krentz HB. Unappreciated epidemiology: The churn effect in a regional HIV care programme. Int J STD AIDS. 2009;20:540–544. DOI: 10.1258/ijsa.2008.008422 [PubMed: 19625584]
5. 2018: Data to Care, High Impact Prevention. Essential Elements. About Data to Care. URL: https://effectiveinterventions.cdc.gov/en/2018-design/data-to-care/group-1/data-to-care/data-to-care-essential-elements. [accessed: 2018-09-26]
6. 2018: Understanding the HIV Care Continuum. URL: https://www.cdc.gov/hiv/pdf/library/factsheets/cdc-hiv-care-continuum.pdf [accessed: 2018-09-26]
7. 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. URL: https://team.cdc.gov/team/cdc/dispatch.cgi/hicsb_TechGuid/folderFrame/100001/0/def/5186. [accessed: 2017-08-16]
8. Florida Health. Routine Interstate Duplicate Review Process. The CDC’s Routine Interstate Duplicate Review (RIDR) Process. URL: http://www.floridahealth.gov/diseases-and-conditions/aids/surveillance/routine-interstate-duplicate-review-process.html [accessed: 2018-09-26]
9. Ocampo JMF, Smart J, Allston A, et al. Improving HIV surveillance data for public health action in Washington, DC: A novel multiorganizational data-sharing method. Sanchez T, ed. JMIR Public Health and Surveillance. 2016;2(1):e3. doi:10.2196/publichealth.5317. [PubMed: 27227157]
10. 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. URL: https://www.cdc.gov/nchhstp/programintegration/docs/pcsidatasecurityguidelines.pdf [accessed: 2018-09-26]
11. Smart JC. “Technology for Privacy Assurance” in Ethical Reasoning in Big Data: An exploratory analysis. In: Collmann J, Matei S, editors. Ethical Reasoning in Big Data: an exploratory analysis. Cham, Switzerland: Springer International Publishing; Jun 30, 2016.
12. Hamp AD, Doshi RK, Lum GR, et al. Cross-jurisdictional data exchange impact on the estimation of the HIV population living in the District of Columbia: Evaluation study. JMIR Public Health Surveill. 2018 Aug 13;4(3):e62. doi: 10.2196/publichealth.9800. [PubMed: 30104182]

Table 1.

The total number of matched HIV case records and match levels across nine separate eHARS data sets.

Jurisdiction (no. of total uploads)	Exact	Extremely high	Very high	High	Medium high	Medium	Medium low	Low	Very low	Total matches
DC (40,448)	8,923	9,201	3,009	72	459	1,997	1	0	1,715	**25,377**
DE (8,419)	1,290	816	447	5	28	307	0	0	165	**3,058**
FL (215,875)	10,067	6,674	4,170	25	304	4,498	0	0	3,948	**29,686**
MD (72,121)	10,132	9,996	3,081	70	456	2,735	1	0	3,362	**29,833**
NC (58,511)	6,177	4,177	2,154	34	243	1,667	0	0	1,568	**16,020**
NYC (242,431)	34,347	24,913	11,765	106	860	6,265	0	0	4,518	**82,774**
NYS (106,619)	32,361	22,881	10,597	85	627	3,710	0	0	1,887	**72,148**
VA (49,844)	6,699	9,022	2,055	46	430	2,147	0	0	6,398	**26,797**
WV (5,058)	924	992	306	7	63	174	0	0	2,323	**4,789**
Total	**110,920**	**88,672**	**37,584**	**450**	**3,470**	**23,500**	**2**	**0**	**25,884**	**290,482**

Abbreviations: DC = District of Columbia, DE = Delaware, FL = Florida, MD = Maryland, NC = North Carolina, NYS = New York State, NYC = New York City, VA = Virginia, WV = West Virginia.

Table 2.

The number of exact case pairs (percentages) (total N=55,460, indicating the number of unique cases) by nine separate eHARS data sets.

	DC	DE	FL	MD	NC	NYC	NYS	VA	WV
DC	.	81 (0.15)	510 (0.92)	5,369 (9.68)	649 (1.17)	425 (0.77)	185 (0.33)	1585 (2.86)	119 (0.21)
DE	.	.	244 (0.44)	495 (0.89)	86 (0.16)	155 (0.28)	107 (0.19)	110 (0.20)	12 (0.02)
FL	.	.	.	918 (1.66)	1,756 (3.17)	3,139 (5.66)	2,255 (4.07)	1,046 (1.89)	199 (0.36)
MD	.	.	.	.	777 (1.40)	569 (1.03)	338 (0.61)	1,475 (2.66)	191 (0.34)
NC	.	.	.	.	.	951 (1.71)	634 (1.14)	1199 (2.16)	125 (0.23)
NYC	.	.	.	.	.	.	28,409 (51.22)	663 (1.20)	36 (0.06)
NYS	.	.	.	.	.	.	.	406 (0.73)	27 (0.05)
VA	.	.	.	.	.	.	.	.	215 (0.39)
WV	.	.	.	.	.	.	.	.	.

Abbreviations: DC = District of Columbia, DE = Delaware, FL = Florida, MD = Maryland, NC = North Carolina, NYS = New York State, NYC = New York City, VA = Virginia, WV = West Virginia.

Table 3:

The number of exact case pairs (total N= 811) also on July 2017 RIDR list across nine eHARS data sets.

	DC	DE	FL	MD	NC	NYC	NYS	VA	WV
DC	.	2	4	27	6	11	7	25	1
DE	.	.	7	3	1	3	1	2	1
FL	.	.	.	12	38	62	60	2	26
MD	.	.	.	.	16	14	7	21	1
NC	.	.	.	.	.	11	15	2	0
NYC	.	.	.	.	.	.	389	21	0
NYS	.	.	.	.	.	.	.	11	0
VA	.	.	.	.	.	.	.	.	2
WV	.	.	.	.	.	.	.	.	.

Abbreviations: DC = District of Columbia, DE = Delaware, FL = Florida, MD = Maryland, NC = North Carolina, NYS = New York State, NYC = New York City, VA = Virginia, WV = West Virginia.

Table 4.

Distribution of estimated time efficiency realized (in minutes; total N=8,110) for nine separate eHARS data sets.

	DC	DE	FL	MD	NC	NYC	NYS	VA	WV
DC	.	20	40	270	60	110	70	250	10
DE	.	.	70	30	10	30	10	20	10
FL	.	.	.	120	380	620	600	20	260
MD	.	.	.	.	160	140	70	210	10
NC	.	.	.	.	.	110	150	20	0
NYC	.	.	.	.	.	.	3,890	210	0
NYS	.	.	.	.	.	.	.	110	0
VA	.	.	.	.	.	.	.	.	20
WV	.	.	.	.	.	.	.	.	.

Abbreviations: DC = District of Columbia, DE = Delaware, FL = Florida, MD = Maryland, NC = North Carolina, NYS = New York State, NYC = New York City, VA = Virginia, WV = West Virginia.

Table 5:

New York State's exact matches across eight other eHARS data sets by "previously resolved/not previously resolved" in eHARS status.

			Not on July 2017 RIDR list	
Reporting Jurisdiction	**ATra Black Box System Exact Matches**	**Exact Matches on July 2017 RIDR List**	**Previously Resolved**	**Not Previously Resolved**
DE	107	1	55	51
DC	185	7	116	62
FL	2,255	60	1,357	838
MD	338	7	198	133
NYC	28,409	389	27,117	903
NC	634	15	381	238
VA	406	11	255	140
WV	27	0	21	6
Total	32,361	490	29,500	2,371

Abbreviations: DC = District of Columbia, DE = Delaware, FL = Florida, MD = Maryland, NC = North Carolina, NYS = New York State, NYC = New York City, VA = Virginia, WV = West Virginia.