PUBLIC
HEALTH
REPORTS

Public Health Reports
1–8
© 2022, Association of Schools and
Programs of Public Health
All rights reserved.
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00333549211066171
journals.sagepub.com/home/phr
SAGE

Enhancing the ATra Black Box Matching Algorithm: Use of All Names for Deduplication Across Jurisdictions

Auntré D. Hamp, MEd, MPH[1,2]; Helen E. Karn, PhD[1];
Frances Y. Kwon, MPH[1]; Anne Rhodes, PhD[1];
James Carrier, MPH[3]; Reshma Bhattacharjee, MBBS, MS, MPH[3];
Colin Flynn, MPH[3]; Trevor Hsu, MPH[3]; John McNeice, MPH[4];
Bridget J. Anderson, PhD[5]; Joyce Chicoine, BS[6],
Jessica Fridge, MPH[7]; Justice King, MPH[7];
Garret R. Lum, MPH[8]; Tej Mishra, MPH[8];
Alisa Kang, MA[9]; and J.C. Smart, PhD[1,10]

Abstract

Objectives: Achieving accurate, timely, and complete HIV surveillance data is complicated in the United States by migration and care seeking across jurisdictional boundaries. To address these issues, public health entities use the ATra Black Box—a secure, electronic, privacy-assuring system developed by Georgetown University—to identify and confirm potential duplicate case records, exchange data, and perform other analytics to improve the quality of data in the Enhanced HIV/AIDS Reporting System (eHARS). We aimed to evaluate the ability of 2 ATra software algorithms to identify potential duplicate case-pairs across 6 jurisdictions for people living with diagnosed HIV.

Methods: We implemented 2 matching algorithms for identifying potential duplicate case-pairs in ATra software. The Single Name Matching Algorithm examines only 1 name for a person, whereas the All Names Matching Algorithm examines all names in eHARS for a person. Six public health jurisdictions used the algorithms. We compared outputs for the overall number of potential matches and changes in matching level.

Results: The All Names Matching Algorithm found more matches than the Single Name Matching Algorithm and increased levels of match. The All Names Matching Algorithm identified 9070 (4.5%) more duplicate matches than the Single Name Matching Algorithm (n = 198 828) and increased the total number of matches at the exact through high levels by 15.4% (from 167 156 to 192 932; n = 25 776).

Conclusions: HIV data quality across multiple jurisdictions can be improved by using all known first and last names of people living with diagnosed HIV that match with eHARS rather than using only 1 first and last name.

Keywords
HIV, surveillance, data sharing, deduplication

Several initiatives focus on reducing new HIV infections and ensuring people living with diagnosed HIV (PWDH) are in care and achieve viral suppression, including the UNAIDS 90-90-90 plan, the Ending the HIV Epidemic initiative, and the National HIV Strategic Plan.[1-3] These plans track progress through metrics such as linkage and retention in HIV care and viral load suppression. These metrics use data from the Centers for Disease Control and Prevention's (CDC's) National HIV Surveillance System (NHSS). The NHSS collects data from public health agencies to accurately assess HIV disease prevalence and incidence at the national and jurisdictional levels and serves as the foundation for programmatic interventions including Data to Care.[4,5]

As the use of HIV surveillance data becomes a more integral part of ongoing patient engagement strategies, accurate, complete, and timely data are needed.[6] Because of strict policies and regulations about data security and confidentiality, US public health jurisdictions do not routinely share person-level HIV surveillance data, outside the biannual case-by-case manual review process, the Routine Interstate Duplicate Review, and a new 5-year process, the Cumulative Interstate Duplicate Review.

The Enhanced HIV/AIDS Reporting System (eHARS) is a browser-based surveillance application developed by CDC that public health departments use to collect, report, manage, and analyze data on PWDH.[7] Each funded jurisdiction maintains its own eHARS; as such, routine, real-time sharing of data across jurisdictions does not happen consistently. Although eHARS data are sent to CDC monthly, they contain a limited number of identifiers and do not contain names or dates of birth.

The movement of people across jurisdictional boundaries poses challenges for a system designed to follow and support access to care for PWDH.[8-10] Several efforts have been deployed to assist in the deduplication of HIV cases across the NHSS, including regional data exchange exemplified in the metropolitan District of Columbia region, Maryland, and Virginia and the use of technology tools such as the ATra Black Box by these and other jurisdictions to deduplicate their eHARS case records and enhance the data quality of their case surveillance records.

In 2016, the health departments of the District of Columbia, Maryland, and Virginia began a collaborative relationship with Georgetown University to enhance ATra Black Box technology with funding from the National Institutes of Health.[11] The ATra Black Box is an electronic privacy-assuring system that allows for the secure and streamlined exchange of data between public health jurisdictions.[12] The ATra Black Box was highly effective in the National Institutes of Health pilot project in identifying potential duplicate records of people living in and accessing care in the District of Columbia, Maryland, and Virginia jurisdictions. Subsequently, the Louisiana and New York State departments of health joined these 3 areas in sharing data. Each jurisdiction executed a contractual agreement with Georgetown University to systematically address HIV surveillance data quality in the ATra Black Box. The New York State Department of Health submitted records from 2 installations of eHARS: 1 each for New York State and New York City. Among the system's core features are algorithms for matching potential duplicate case-pairs for ≥2 jurisdictions.[13] Georgetown University developed these algorithms in close consultation with users in public health jurisdictions, and the algorithms have been validated by those jurisdictions.[11]

eHARS is a document-based surveillance system that allows all documents to be stored and retained electronically in their original format. Because jurisdictions receive surveillance documents from many types of sources, different names and/or spellings are often recorded. eHARS includes an alias table that contains all iterations of identifying information received for the person, including multiple names (eg, legal names, alternate spellings of first and last names, alias names, nicknames), dates of birth, social security numbers (SSNs), and other identifiers. The Document View in eHARS comprises all laboratory and patient identification documents for each person. The Person View provides 1 summary record for each person, derived from all entered records for that person and using a hierarchy to determine data elements with multiple entries in the Document View.[14]

Initial runs of the ATra Black Box used only Person View data to match between jurisdictions, because the Person View data are most often used for manual deduplication processes and the tracking of aliases varies among jurisdictions. The jurisdictions proposed an enhancement to the Black Box that would incorporate all first and last names of each case record that are contained in the eHARS alias table, which pulls data from the Document View. This enhancement to the Black Box is the All Names Matching Algorithm. The goals of this enhancement were to increase the ability of jurisdictions to deduplicate case records by exchanging additional name information and to increase the yield and accuracy of identified case matches. The previous algorithm, which examined only the Person View name, is termed the Single Name Matching Algorithm. The objective of this evaluation was to assess differences in the identification of duplicate case-pairs between the All Names Matching Algorithm and the Single Name Matching Algorithm among participating jurisdictions.

Methods

Both matching algorithms contain a set of rules called *match levels* (Table 1). The match levels indicate confidence that 2 eHARS records from 2 different jurisdictions belong to the same person. The highest match level (exact) expresses the greatest certainty that case records belong to the same person. The lowest match level (low) expresses the least certainty that the pair of records belongs to the same person—the only data

[1] Office of the Senior Vice President for Research, Georgetown University, Washington, DC, USA
[2] Center for Global Health Practice and Impact, Georgetown University, Washington, DC, USA
[3] Center for HIV Surveillance, Epidemiology and Evaluation, Maryland Department of Health, Baltimore, MD, USA
[4] HIV Surveillance Program, Virginia Department of Health, Richmond, VA, USA
[5] Center for Community Health, New York State Department of Health, Albany, NY, USA
[6] Bureau of HIV/AIDS Epidemiology, New York State Department of Health, Albany, NY, USA
[7] STD/HIV/Hepatitis Program, Louisiana Department of Health, New Orleans, LA, USA
[8] HIV/AIDS, Hepatitis, STD and TB Administration, District of Columbia Department of Health, Washington, DC, USA
[9] University Information Systems, Georgetown University, Washington, DC, USA
[10] Department of Computer Science, Georgetown University, Washington, DC, USA

Corresponding Author:
Anne Rhodes, PhD, Georgetown University, Office of the Senior Vice President for Research, Healy Hall, Ste 200, 37th and O Streets NW, Washington, DC 20007, USA.
Email: anne.rhodes@georgetown.edu

Table 1. Comparison of the ATra Black Box Single Name Matching Algorithm and All Names Matching Algorithm, December 2019[a]

Match level	Match rule	Name variables[b]	Additional match variables
Single Name Matching Algorithm			
Exact	Exact	PV last name and PV first name	and DOB and SSN and sex
Extremely high	Extremely high	PV last name and PV first name	and DOB and partial SSN and sex
Very high	Very high	PV last name and PV first name	and DOB and SSN
High	High-1	PV last name and PV first name	and DOB and sex
	High-2	PV last name and PV first name	and DOB and partial SSN
Medium high	Medium high	PV last name and PV first name soundex	and DOB and sex
Medium	Medium-1	PV last name	and DOB and sex
	Medium-2	PV last name soundex and PV first name soundex	and DOB and sex
	Medium-3	PV last name soundex and PV first name soundex	and DOB
Medium low	Medium low-1	PV last name soundex and PV first name soundex	and partial DOB and partial SSN and sex
	Medium low-2	PV last name soundex and PV first name soundex	and partial DOB and partial SSN
Low	Low-1	PV last name soundex	and partial DOB and partial SSN and sex
	Low-2	PV last name soundex	and partial DOB and partial SSN
All Names Matching Algorithm			
Exact	Exact-1	PV last name and PV first name	and DOB and SSN and sex
	Exact-2[c]	PV last name and PV first name = other name (OR) other name = other name	and DOB and SSN and sex
Extremely high	Extremely high-1	PV last name and PV first name	and DOB and partial SSN and sex
	Extremely high-2[c]	PV last name and PV first name = other name (OR) other name = other name	and DOB and partial SSN and sex
Very high	Very high-1	PV last name and PV first name	and DOB and SSN
	Very high-2[c]	PV last name and PV first name = other name (OR) other name = other name	and DOB and SSN
High	High-1	PV last name and PV first name	and DOB and sex
	High-2[c]	PV last name and PV first name = other name (OR) other name = other name	and DOB and sex
	High-3	PV last name and PV first name	and DOB and partial SSN
	High-4[c]	PV last name and PV first name = other name (OR) other name = other name	and DOB and partial SSN
	High-5[c]	PV last name	and DOB and SSN
	High-6[c]	PV first name	and DOB and SSN and sex
	High-7[c]	NA	DOB and SSN

Abbreviations: DOB, date of birth; NA, not applicable; PV, Person View; SSN, social security number.

[a] ATra Black Box is a secure, electronic, privacy-assuring system developed by Georgetown University to identify and confirm potential duplicate case records, exchange data, and perform other analytics to improve the quality of data in the Enhanced HIV/AIDS Reporting System.[7,14]

[b] The Person View (PV) provides 1 summary record for each person, derived from all entered records for that person. The soundex value is a 4-character alpha-numeric string that represents how the name is pronounced in English. Soundex values are calculated by a soundex algorithm and consist of the first letter of the name followed by 3 digits (0-9).

[c] Match rules that were added to the All Names Matching Algorithm and that do not exist in the Single Name Matching Algorithm. For match levels low through medium high, no changes were made.

element that they have in common is the soundex value of the Person View last name. (The soundex value is a 4-character alpha-numeric string that represents how the name is pronounced in English. Soundex values are calculated by a soundex algorithm and consist of the first letter of the name followed by 3 digits [0-9]. For example, the soundex value for the name William is W452; the soundex value for the name Smith is S530.) If a pair of case records matches on a rule, they appear in a detailed report that lists the case record identifiers, the match level, and additional data elements. A description of the validation of the match levels is available elsewhere.[11] The Georgetown University Institutional Review Board determined this research to be exempt from institutional review board review.

The Single Name Matching Algorithm considers only 1 first name and 1 last name for each eHARS case record from the Person View table. We refer here to the eHARS Person View first name as Person View First Name and the eHARS Person View last name as Person View Last Name.

The All Names Matching Algorithm considers all names that are stored in eHARS for each case record and contains new rules at the higher match levels for matching on individual components of the name: first name only, last name only, or no name component. For matching purposes in the All

Table 2. Three possible match scenarios (using fabricated data and fictional names) in the ATra Black Box All Names Matching Algorithm, December 2019[a]

Match scenario[b]	Jurisdiction 1	Jurisdiction 2	Social security number	Date of birth	Birth sex
Scenario A					
PV Name	Mary Baker	Mary Jane Baker	012-34-5678	19260601	Female
Other Name[c]	Mary Jane Baker	M. Baker	012-34-5678	19260601	Female
Scenario B					
PV Name[d]	John Doe	John Earl Do, Jr	xxx-xx-6789	19241001	Male
Other Name	Johnny Doe	John Doe	123-45-6789	19241001	Male
Scenario C					
PV Name	Jane Smith	Jane Jones	234-56-7890	19150407	Male
Other Name[e]	Janet Jones	Janet Jones	234-56-7890	1915xx07	Male

[a] ATra Black Box is a secure, electronic, privacy-assuring system developed by Georgetown University to identify and confirm potential duplicate case records, exchange data, and perform other analytics to improve the quality of data in the Enhanced HIV/AIDS Reporting System.[7,14]

[b] The Person View (PV) provides 1 summary record for each person, derived from all entered records for that person. The combination of a non-PV first name and non-PV last name that appear together in an Enhanced HIV/AIDS Reporting System document is called the Other Name.

[c] Other Name (jurisdiction 1) matches PV Name (jurisdiction 2).

[d] PV Name (jurisdiction 1) matches Other Name (jurisdiction 2).

[e] Other Name (jurisdiction 1) matches Other Name (jurisdiction 2).

Names Matching Algorithm, the person's first name and last name must appear together on at least 1 document in the jurisdiction's eHARS database. The combination of a non–Person View First Name and non–Person View Last Name that appear together in an eHARS document is called the Other Name.

Jurisdictions extracted Other Name records from eHARS via a SAS program written by one of the authors (J.C.). Overall, 98.9% of the total eHARS records (333 426 of 337 281) had 1 to 5 Other Names per case record (District of Columbia, 98.3%; Louisiana, 99.2%; Maryland, 98.6%; New York State, 98.8%; Virginia, 99.4%). The maximum number of Other Names per case record ranged from 15 to 56 per jurisdiction. The most frequent number of Other Names by jurisdiction was 1 (District of Columbia, 71.6%; Louisiana, 66.7%; Maryland, 72.5%; New York State, 58.7%; Virginia, 76.6%). New York City was unable to submit data on the number of Other Names per case record.

Jurisdictions uploaded a file containing all Other Names in addition to the regular eHARS file. To determine if a pair of case records is a match, the All Names Matching Algorithm checks the Person View First Name and Person View Last Name to see if they are identical. If the names and other matching level variables are the same, the algorithm determines the match level, just as it does in the Single Name Matching Algorithm. If the pairs of Person View names are not the same but other variables *do* match (such as SSN and/or date of birth), the All Names Matching Algorithm then compares each jurisdiction's set of Other Names for a possible match. We used fabricated data and fictional names to illustrate 3 possible combinations of name matching between 2 jurisdictions (Table 2).

For quality assurance and validation purposes, we created input files of sample test data for each jurisdiction. Each test file contained randomly generated values as well as a predetermined number of handcrafted match pairs prepared by one of the authors (F.K.). Output reports from the validation run with the test files were examined to verify that the match pairs were identified correctly by the Black Box, according to the rules of each matching algorithm.

The higher-confidence match level rules (exact through high) include more match variables than the lower-confidence match level rules (medium high through low). The higher-confidence match levels also require all variables to contain complete values, such as 9 valid digits for the SSN and 8 valid digits for the date of birth. The lower-confidence match levels allow partially missing values for SSN and date of birth. Both algorithms do not allow any match variable to contain a blank value or an unknown value to match another record.

To examine the differences between the 2 algorithms, we compared the number of total matches by jurisdiction and the number of matches by match level. We assessed changes in match levels between the algorithms. Specifically, we assessed the number of matches that moved to a higher-confidence match level resulting from use of the All Names Matching Algorithm.

Results

Overall Number of eHARS Records Uploaded and Total Matches

In December 2019, jurisdictions uploaded a total of 584 290 eHARS case records (District of Columbia, 41 908; Maryland, 77 086; Virginia, 54 256; Louisiana, 49 294; New York State, 115 284; and New York City, 246 462) into the ATra Black Box during separate runs with the Single Name Matching Algorithm and the All Names Matching Algorithm (Table 3). The same number of records were uploaded successfully during both ATra Black Box runs for all jurisdictions except New York City. In the run with the All Names Matching Algorithm,

Table 3. Number of HIV case surveillance records matched by jurisdiction (exact through high match levels) from the ATra Black Box Single Name Matching Algorithm and the All Names Matching Algorithm, December 2019[a]

		No. of matches		
Jurisdiction[b]	No. of records loaded[b]	Single Name Matching Algorithm	All Names Matching Algorithm	Total no. of new matches (% increase) in All Names Matching Algorithm
District of Columbia	41 908	24 374	25 485	1111 (4.6)
Maryland	77 086	24 512	25 730	1218 (5.0)
Virginia	54 256	17 439	18 246	807 (4.6)
Louisiana	49 294	3418	3549	131 (3.8)
New York State	115 284	62 772	65 618	2846 (4.5)
New York City	246 462[b]	66 313	69 270	2957 (4.5)
Total	584 290	198 828	207 898	9070 (4.6)

[a] ATra Black Box is a secure, electronic, privacy-assuring system developed by Georgetown University to identify and confirm potential duplicate case records, exchange data, and perform other analytics to improve the quality of data in the Enhanced HIV/AIDS Reporting System.[7,14]
[b] Jurisdictions had the same number of records uploaded during both Black Box runs except New York City, which had 242 547 records loaded during the run with the Single Name Matching Algorithm and 246 462 records loaded during the run with the All Names Matching Algorithm.

New York City uploaded 246 462 total records to the ATra Black Box. Because of an uploading issue, 3915 fewer records were uploaded from New York City during the run with the Single Name Matching Algorithm.

For the run with the Single Name Matching Algorithm, the ATra Black Box identified a total of 198 828 (of 580 375; 34.3%) case-pairs in all jurisdictions across all match levels: exact (61.3%), extremely high (1.5%), very high (0.6%), high (20.7%), medium high (3.1%), medium (12.7%), medium low (0.1%), and low (0.1%).

With the All Names Matching Algorithm, the ATra Black Box identified 207 898 (of 584 290; 35.6%) case-pairs in all jurisdictions (Table 3) across all match levels: exact (65.6%), extremely high (1.8%), very high (0.6%), high (24.8%), medium high (0.3%), medium (6.7%), medium low (0.1%), and low (<0.1%). More than two-thirds (67.4%) of case-pairs identified by matching on all names stored in eHARS were at the exact or extremely high levels.

New Case-Pairs Identified by the All Names Matching Algorithm

The All Names Matching Algorithm added a total of 9070 new case-pairs; these had not been previously identified as matches by the Single Name Matching Algorithm. From the All Names Matching Algorithm, jurisdictions saw an average increase of 4.6% in the total number of matches when compared with the Single Name Matching Algorithm. New York State and New York City had the highest number of matches added by the All Names Matching Algorithm. Maryland had the highest percentage increase in matches (5.0%), followed by the District of Columbia and Virginia (both 4.6%) (Table 3).

The introduction of the All Names Matching Algorithm resulted in a total of 4578 new exact-level case-pairs (Table 4). Half of the newly matched case-pairs identified by the ATra Black Box were at the exact level, and more than one-third of the newly identified matches were at the high-2 match level, which matched names using the alias table names, along with date of birth and sex. One jurisdiction reported that 20.7% (n = 252/1217) of the new matches found by the All Names Matching Algorithm had not previously been identified by any deduplication processes, including the Routine Interstate Duplicate Review and the Cumulative Interstate Duplicate Review.

Shift of Case-Pairs to the Higher Match Levels

In addition to the 9070 new matches that the ATra Black Box identified using the All Names Matching Algorithm, 16 706 records moved from a low or medium match level to a high level or better in the All Names Matching Algorithm (Table 5). Most (59.2%) shifting case-pairs in all jurisdictions moved to the exact level. New York State and New York City had the highest number of matches move to higher match levels (5219 and 5498, respectively) with the All Names Matching Algorithm. We found an 11.9% (from 121 912 to 136 374; difference = 14 462) increase in the number of exact matches between jurisdictions and a 15.4% (from 167 156 to 192 932; difference = 25 776) increase overall in matches at the exact through high match levels in the All Names Matching Algorithm.

By matching on all names stored in eHARS, 14 462 (56.1% of all improved matches) case-pairs moved from not matched or lower match levels in the Single Name Matching Algorithm to the exact level. A total of 956 case-pairs that previously were unmatched or matched at lower levels were matched at the extremely high or very high match levels, and 10 358 case-pairs moved from a no match or lower level to 1 of the 5 high match levels.

Lessons Learned

As public health jurisdictions focus efforts on reducing new HIV infections and ensuring PWDH are virally suppressed, data

Table 4. Number of matches, by match level and jurisdiction, identified by the ATra Black Box All Names Matching Algorithm that were not identified by the Single Name Matching Algorithm, December 2019

	Match level,[b] no. (%) of total new matches in jurisdiction				
Jurisdiction	**Exact-2**	**Extremely high-2**	**Very high-2**	**High-2, high-4, high-5, high-6, high-7**	**Total no. of new matches in All Names Matching Algorithm**
District of Columbia	449 (40.4)	86 (7.7)	13 (1.2)	563 (50.7)	1111
Maryland	541 (44.4)	93 (7.6)	12 (1.0)	572 (47.0)	1218
Virginia	337 (41.8)	54 (6.7)	6 (0.7)	410 (50.8)	807
Louisiana	63 (48.1)	2 (1.5)	0	66 (50.4)	131
New York State	1579 (55.5)	11 (0.4)	44 (1.5)	1212 (42.6)	2846
New York City	1609 (54.4)	8 (0.3)	47 (1.6)	1293 (43.7)	2957
Total	4578 (50.5)	254 (2.8)	122 (1.3)	4116 (45.4)	9070

[a] ATra Black Box is a secure, electronic, privacy-assuring system developed by Georgetown University to identify and confirm potential duplicate case records, exchange data, and perform other analytics to improve the quality of data in the Enhanced HIV/AIDS Reporting System.[7,14]
[b] Table 1 provides details on matching levels.

Table 5. Number of matches, by match level and jurisdiction, that the ATra Black Box All Names Matching Algorithm moved to a higher match level, December 2019[a]

	Match level,[b] no. (%)				
Jurisdiction	**Exact**	**Extremely high**	**Very high**	**High**	**Total no. of matches moved to higher level**
District of Columbia	1083 (49.7)	187 (8.6)	3 (0.1)	906 (41.6)	2179
Maryland	1058 (50.9)	163 (7.9)	4 (0.2)	852 (41.0)	2077
Virginia	670 (45.6)	125 (8.5)	1 (0.1)	675 (45.9)	1471
Louisiana	142 (54.2)	4 (1.5)	0	116 (44.3)	262
New York State	3434 (65.8)	17 (0.3)	28 (0.5)	1740 (33.3)	5219
New York City	3497 (63.6)	20 (0.4)	28 (0.5)	1953 (35.5)	5498
Total	9884 (59.2)	516 (3.1)	64 (0.4)	6242 (37.4)	16 706

[a] ATra Black Box is a secure, electronic, privacy-assuring system developed by Georgetown University to identify and confirm potential duplicate case records, exchange data, and perform other analytics to improve the quality of data in the Enhanced HIV/AIDS Reporting System.[7,11]
[b] Table 1 provides details on matching levels.

quality is critical in implementing appropriate interventions.[15] Timely, accurate, and complete data on PWDH are critical for public health personnel to reach and engage PWDH in activities including Data to Care and Ending the HIV Epidemic. The ATra Black Box has provided substantial assistance in identifying the current residence and care status of PWDH through deduplication, which, in turn, improves the quality of HIV surveillance data in NHSS.[13]

The ATra Black Box is used by more than 30 jurisdictions to assist in the deduplication of records in the NHSS through a cooperative agreement funded by CDC.[16] These jurisdictions use the original matching algorithm that was created in 2017. As exemplified in the results presented here for 6 jurisdictions, improvements in the yield and accuracy of duplicate case-pairs would be expected at the national level if the All Names Matching Algorithm were implemented in the national Black Box project.

In our assessment, for the All Names Matching Algorithm, the total number of exact matches increased by 11.9% (n = 14 462) compared with the Single Name Matching Algorithm, whereas the total number of matches at the high through exact levels increased by 15.4% (n = 25 776). This increase is important, because matches at the high through exact levels have been validated as true matches.[11] Each participating jurisdiction found 3.8% to 5.0% (all jurisdiction total = 9070 match pairs) of their matches using the All Names Matching Algorithm that were never previously identified as matches by the ATra Black Box.

The Single Name Matching Algorithm, which was validated during multiple runs of the ATra Black Box, enabled jurisdictions to accurately and securely match people across jurisdictional boundaries. The evaluation of the All Names Matching Algorithm presented here builds on that foundation. The increased yield and higher match levels found in the All Names Matching Algorithm for 6 jurisdictions suggest that integrating the All Names Matching Algorithm into the national-level Black Box project would greatly increase the number of high-quality matches. This improvement would enable public health personnel to have more complete and accurate data on PWDH living in their jurisdiction. The information provided in the output reports after each run of the ATra Black Box allows jurisdictions to better target limited resources to PWDH who are living in their jurisdictions. Jurisdictions could use their scarce resources to reach out to PWDH not in care to provide support in linking them to a care provider.

Although the All Names Matching Algorithm increases the number of high-confidence matches between jurisdictions, the deterministic algorithms used by the ATra Black Box have some general limitations. A recent study found that probabilistic algorithms detected more matches than the original ATra Black Box matching algorithm when matching surveillance data for HIV and sexually transmitted infections.[17] For public health jurisdictions, an important factor in matching is ensuring that health data are correctly associating a person, because these data are used for many purposes, including locating people out of care. Future work should assess false match rates across algorithms and examine the staff time burden on jurisdictions to implement different algorithms, because public health staff personnel and resources are limited.

Using multiple names to match people is also important in the larger context of public health, because people may have different names in various health data systems. For example, a recent study found that using alias names was important in ascertaining death rates among young people involved in the justice system.[18] The enhancement to the matching algorithm of the ATra Black Box presented here demonstrates that including all names of PWDH in a jurisdiction's eHARS database improved the number and match confidence levels of identified duplicate case records. Additional jurisdictions would benefit from using the All Names Matching Algorithm in the ATra Black Box in their deduplication efforts. Use of the All Names Matching Algorithm would improve the quality of epidemiological data that are needed to track the health outcomes of PWDH and engage people in HIV medical care. It would also improve the quality of the national HIV surveillance data used for measuring progress toward ending the HIV epidemic.

Acknowledgments

The authors acknowledge Spiros Dimolitsas, Mary Beth Fargo, and Peter Luger in the Office of the Senior Vice President for Research at Georgetown University for their leadership and assistance in developing and executing the data-sharing agreements with the participating departments of health, which made possible the work described in this article. ATra Black Box and ATra are registered trademarks of Georgetown University; all rights reserved.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Anne Rhodes, PhD https://orcid.org/0000-0002-0301-2432
Bridget J. Anderson, PhD https://orcid.org/0000-0002-1233-3780

References

1. Joint United Nations Programme on HIV/AIDS. 90-90-90: an ambitious treatment target to help end the AIDS epidemic. 2014. Accessed March 15, 2021. https://www.unaids.org/en/resources/documents/2017/90-90-90
2. Centers for Disease Control and Prevention. Ending the HIV epidemic in the U.S. Accessed October 14, 2020. https://www.cdc.gov/endhiv/index.html
3. US Department of Health and Human Services. *HIV National Strategic Plan: A Roadmap to End the Epidemic for the United States 2021-2025*. 2021. Accessed March 15, 2021. https://files.hiv.gov/s3fs-public/HIV-National-Strategic-Plan-2021-2025.pdf
4. Centers for Disease Control and Prevention. Diagnoses of HIV infection in the United States and dependent areas, 2018 (updated). *HIV Surveill Rep*. 2020;31:1-119.
5. Centers for Disease Control and Prevention. Estimated HIV incidence and prevalence in the United States, 2014-2018. *HIV Surveill Suppl Rep*. 2020;25(1):1-78.
6. 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. *Milbank Q*. 2013;91(3):558-603. doi:10.1111/milq.12018
7. Cohen SM, Gray KM, Ocfemia MC, Johnson AS, Hall HI. The status of the National HIV Surveillance System, United States, 2013. *Public Health Rep*. 2014;129(4):335-341. doi:10.1177/003335491412900408
8. Buskin SE, Kent JB, Dombrowski JC, Golden MR. Migration distorts surveillance estimates of engagement in care: results of public health investigations of persons who appear to be out of HIV care. *Sex Transm Dis*. 2014;41(1):35-40. doi:10.1097/OLQ.0000000000000072
9. Gill MJ, Krentz HB. Unappreciated epidemiology: the churn effect in a regional HIV care programme. *Int J STD AIDS*. 2009;20(8):540-544. doi:10.1258/ijsa.2008.008422
10. Hamp AD, Doshi RK, Lum GR, Allston A. 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;4(3):e62. doi:10.2196/publichealth.9800
11. Ocampo JMF, Smart JC, Allston A, et al. Improving HIV surveillance data for public health action in Washington, DC: a novel multiorganizational data-sharing method. *JMIR Public Health Surveill*. 2016;2(1):e3. doi:10.2196/publichealth.5317
12. Smart JC. Technology for privacy assurance. In: Collmann J, Matei S, eds. *Ethical Reasoning in Big Data: An Exploratory Analysis*. Springer International Publishing; 2016:93-114.
13. Ocampo JMF, Hamp A, Rhodes A, et al. Improving HIV surveillance data by using the ATra Black Box System to assist regional deduplication activities. *J Acquir Immune Defic Syndr*. 2019;82(suppl 1):S13-S19. doi:10.1097/QAI.0000000000002090

14. Council of State and Territorial Epidemiologists. *HIV Surveillance Training Manual*. 2005. Accessed March 15, 2021. https://www.cste.org/members/group.aspx?id=87601
15. Lesko C, Sampson L, Miller W, et al. Measuring the HIV care continuum using public health surveillance data in the United States. *J Acquir Immune Defic Syndr*. 2015;70(5):489-494. doi:10.1097/QAI.0000000000000788
16. Centers for Disease Control and Prevention. Secure data sharing tool to support de-duplication of cases in the National HIV Surveillance System (NHSS): CDC-RFA-PS18-1805. Accessed April 1, 2021. https://govtribe.com/file/government-file/cdc-rfa-ps18-1805-final-dot-pdf
17. Avoundjian T, Dombrowski JC, Golden MR, et al. Comparing methods for record linkage for public health action: matching algorithm validation study. *JMIR Public Health Surveill*. 2020;6(2):e15917. doi:10.2196/15917
18. Tibble H, Law HD, Spittal MJ, et al. The importance of including aliases in data linkage with vulnerable populations. *BMC Med Res Methodol*. 2018;18(1):76. doi:10.1186/s12874-018-0536-4