PLOS ONE

RESEARCH ARTICLE

Methodology for linking Ryan White HIV/AIDS Program Services Report (RSR) client level data over multiple years

Julia Zhu[1]*, Miranda Fanning[1], Laura Sheehan[2], Kerry Grace Morrissey[3], Stan Legum[3], Sigurd Hermansen[3]

1 Health Resources and Services Administration, HIV/AIDS Bureau, Division of Policy and Data, Rockville, Maryland, United States of America, **2** Accenture Federal Services LLC, Arlington, Virginia, United States of America, **3** Westat Inc., Rockville, Maryland, United States of America

* xzhu@hrsa.gov, xiao828@gmail.com

OPEN ACCESS

Citation: Zhu J, Fanning M, Sheehan L, Morrissey KG, Legum S, Hermansen S (2020) Methodology for linking Ryan White HIV/AIDS Program Services Report (RSR) client level data over multiple years. PLoS ONE 15(8): e0237635. https://doi.org/10.1371/journal.pone.0237635

Editor: Rachel M. Presti, Washington University in Saint Louis, UNITED STATES

Received: October 3, 2019

Accepted: July 28, 2020

Published: August 21, 2020

Data Availability Statement: The Ryan White HIV/AIDS Program Services Report (RSR) client level data are not available due to client privacy and confidentiality. HRSA makes aggregate data available in the Ryan White HIV/AIDS Program Annual Client-Level Data Report: https://hab.hrsa.gov/data/data-reports. Requests for aggregate data can be made to: RWHAPdata@hrsa.gov.

Funding: This project was funded by the HIV/AIDS Bureau, Health Resources and Services Administration, under Contract No.

Abstract

Background

The Health Resources and Services Administration's (HRSA), HIV/AIDS Bureau (HAB) is responsible for leading the nation's efforts to provide health care, medications, and support services to low-income people living with HIV through the Ryan White HIV/AIDS Program (RWHAP). The RWHAP funds and coordinates with cities, states, and local community-based organizations to deliver efficient and effective HIV care, treatment, and support services for over half a million vulnerable people living with HIV (PLWH) and their families in the United States. The annual RWHAP Services Report (RSR) is an important source of information for monitoring RWHAP's progress towards National HIV/AIDS Strategy goals. Since 2010, HRSA HAB has used the annual client-level RSR data to monitor program-related outcomes, conduct program evaluations, understand service provision, and conduct extensive analysis on disparities in viral suppression and retention in HIV care. HRSA HAB receives annual RSR submissions from RWHAP recipients and sub-recipients. However, the de-identified nature of the data limits HRSA HAB's ability to expand beyond year-to-year analyses and conduct additional analyses to evaluate outcomes for clients who are seen in multiple years. The current paper describes the development and validation of a method to link RSR client-level records across multiple data years.

Methods and findings

Using seven RSR reporting years of data (2010 to 2016), we applied a Fellegi-Sunter (F-S) linkage model that used client demographic characteristics and their providers' geographic locations to calculate matching weights for each record pair based on estimated agreement and disagreement conditional probabilities across RSR years. To validate our methodology, we conducted an internal sample review and external validation to assess the level of accuracy of the linkage, and the extent to which the linked data set corresponds accurately to clinical records of individual clients. The linkage result yielded 70 to 80 percent year-to-year client carry-over rate over seven years of the RSR data; 96 percent linkage ratio from the internal sample review and 79.9 to 94.2 percent of provider network client carry- over rate per year from the external validation.

HAB55_C_6249. The funder provided support in the form of salaries for authors [JZ and MF], but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section. Co-authors KM, SL, SH, and LS are employed by the commercial enterprises funded under this contract. KM, SL, and SH are employed by Westat (www.westat.com), while LS is a former employee of Accenture Federal Services (AFS; www.accenture.com). Both companies provided support in the form of salary for the co-authors (KM, SL, SH, and LS) but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: Co-authors KM, SL and SH are employed by Westat and LS is a former employee of Accenture Federal Systems. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

Conclusions

This methodology addresses a gap in data analysis capabilities by allowing HRSA HAB to link RWHAP clients across reporting years. Despite weak identifying information and lack of continuity of service reporting, the longitudinal linkage improves HRSA HAB's ability to evaluate the patterns of viral suppression and monitor service utilization over time for individuals who receive services in multiple years. These analyses will support future analytic activities in understanding the impact and outcomes of the RWHAP, and will assist HRSA HAB in monitoring progress toward meeting National HIV/AIDS Strategy goals. For those looking for ways to assess health services data, the F-S unsupervised method combining weak identifying attributes and geographic proximity offers practical solutions to the problem of linking de-identified information about individuals across multiple years and improving longitudinal research.

Introduction

The Health Resources and Services Administration's (HRSA), HIV/AIDS Bureau (HAB) is responsible for leading the nation's efforts to provide health care, medications, and support services to low-income people living with HIV through the Ryan White HIV/AIDS Program (RWHAP) [1]. The RWHAP funds and coordinates with cities, states, and local community-based organizations to deliver efficient and effective HIV care, treatment, and support services for over half a million vulnerable people living with HIV (PLWH) and their families in the United States [2, 3]. As a condition of this funding, RWHAP grant recipients and providers are required to report annual data on clients served, services provided, and expenditures to HRSA HAB. HRSA HAB uses these data to monitor program-related health outcomes, evaluate program activities, track service utilization, and assess disparities in viral suppression and retention in HIV care.

RWHAP-funded recipients and service providers submit annual data in the RWHAP Services Report (RSR) -, a client-level data system that captures information on the characteristics of RWHAP grant recipients, service providers, and clients. Because HRSA HAB is statutorily prohibited from the collection of Personal Identifiable Information (PII), such as name, date of birth and social security number, client-level data are de-identified by RWHAP recipients and/or service providers prior to submission [4]. Clients within the same service provider or provider's network are uniquely identified using a forty- or forty-one character encrypted Unique Client Identifier (eUCI40 or eUCI41).

The eUCI40 is an encrypted client identification code generated from the following elements: the first and third characters of a client's first and last names, full date of birth, and gender code. This encrypted ID is a string with length 40; thus, it is called the eUCI40. If two or more clients have identical eUCI40s, service providers will add a character at the end of the eUCI40s to create 41-digit eUCIs (eUCI41) to distinguish between clients (see < https://targethiv.org/sites/default/files/file-upload/resources/eUCI_Application_User_Guide_Dec_2014.pdf>). The choice of this additional character is not universal across providers; each eUCI40 or eUCI41 is only unique for clients within the same service provider and year. Unfortunately, this process creates analytic challenges in the RSR data for HRSA HAB because it introduces the potential for obtaining duplicate records across providers and years as well as limits the ability to link client data across time.

HAB previously developed a single-year de-identification duplication methodology to uniquely identify clients within a single year across providers and provider networks. The

approach uses probabilistic matching to assess the likelihood that two clients are the same, given equal values on common data elements (i.e., race, ethnicity, and housing status). Records determined to belong to the same client are provided with a 2-digit suffix, such as "00", "01", "02", that replaces the provider assigned character at the end of the eUCI40 (i.e., eUCI42). De-duplication of RSR data is performed independently each year by HRSA HAB. As a result, the eUCI42 cannot be used to identify the same client across years. This inability to link data across years limits HRSA HAB's ability to assess longitudinal trends in HIV viral suppression and retention in care as well as understand the factors associated with these outcomes.

To address this limitation, HRSA HAB sought to develop and validate a method to link RSR client-level records across multiple data years. In this paper, we describe the development of a longitudinal record linkage methodology for RSR data and provide evidence for the validity of this approach. We also discuss the implications for using this methodology in other settings.

Methods

Ethics statement

The current analyses involved the analysis of existing de-identified data for which investigators cannot link back to participants and, as such, is exempt from human subjects review (U.S. Department of Health and Human Services (HHS) regulation 45 CFR 46.101(b)).

Selection of a longitudinal data linkage algorithm

To identify the most appropriate data linkage methodology related to the health care field and the study of people living with HIV (PLWH), we conducted a literature review of over 400 peer-reviewed articles, reports and publications from four major scientific research databases on the web (PubMed, PubMed Central, Google Scholar, and Lex Jansen). The inclusion criteria limited searches to: 1) publications since 2000, except for key or highly relevant publications published prior to 2000; 2) English language publications; 3) record linkage/matching methods and strategies, data integration, object identification in statistics, computer science, medicine, database management, and web technology search terms; and, 4) natural key linkage of disparate data sources and methodologies aimed at identifying and de-duplicating data drawn from the PLWH and other special populations.

The literature on data linkage methods (i.e., "data integration" and "entity resolution") includes many examples of the standard F-S probabilistic linkage model [5], its extensions (e.g., Expectation-Maximization), latent class, and statistical/machine learning (support vector machines, random forests, and deep learning) classifiers. For the HRSA HAB longitudinal linkage, only weak (in fact, mostly de-identified) attributes of persons were available for training a classifier.

We divided the prospective data linkage methods into two groups: supervised, requiring a substantial "truth set" of known client record links and non-links; and unsupervised, without that requirement. Based on the literature review, we selected the unsupervised F-S model because the methodology generally provides robust and reliable linkage results in the absence of a truth set, and it is the methodology commonly cited in the literature reviewed [5–47]. In addition, the use of the F-S unsupervised method has proven successful in contexts similar to that of HRSA HAB longitudinal linkage [6].

Model overview

The F-S model is a probabilistic approach to solving record linkage problems based on conditional probabilities. Although it comes from an earlier era, it has much in common with more recent classical and Bayesian statistical and machine learning classifiers in that it employs an

ensemble of a weak classifiers and prior distributions of true and false signals from classifiers. This method classifies all record pairs that come from two (or more) data sources into three independent and mutually exclusive groups: true matches, non-matches, and uncertain matches. Record pairs are classified based on a conditional probability model that calculates both the likelihood ratio and the match score that the record pairs represent the same individual based on the summarized and weighted level of agreement or disagreement between values of selected variables. Weights represent the estimated frequency of a match of a variable's values given a true match divided by the frequency of a true match given a non-match of the variable's values. In the absence of more precise conditional probabilities, a model with a weight of 0.9/0.1 for all variables will work well enough.

Under the F-S Model, the classification decision involves setting two threshold weights: an upper threshold above which a record pair is classified as a match, and a lower threshold below which a record pair is classified as a non-match. The choice of these selection thresholds aims to minimize both the linkage errors and the number of pairs with an indeterminate status between the two thresholds [5]. In practice, many applications use a single threshold for classification to achieve the desired error levels [7–8]. Our linkage methodology adopted the single threshold for classification. Pairs at or above the classification threshold were declared as a match; those below the classification threshold were declared as a non-match. In a longitudinal setting, we began with the eUCI40 linkage results for a base year (2010) and extended those to the next year, and then to each subsequent year. This progression linked records with similar attributes and proximity, and it de-linked records with differences in similarity and proximity.

Matching variable selection

The matching variables consisted of a set of personal attributes and geographic codes of service providers' addresses that were used to link client records in record comparisons in the F-S model. The selection criteria of matching variables from the RSR data included consistency and availability of each variable. An exploratory review of personal attributes that might serve to link client records preceded the final selection of variables for use in record comparisons. Within eUCI40 groups, client attributes carried over from one visit to a provider to the next visit, within and across years, and appeared consistent with eUCI40s assigned by that provider. Within the same geographic area, we also found that personal attributes within an eUCI40 group linked clients across different providers. Further, sequences of visits to providers by clients with very similar attributes showed potential for linking records belonging to a single client. In some instances, sequences of visits by clients ceased in one geographic area and appeared to resume in another area. The sequences of the eUCI40 IDs by provider showed continuous records in many cases, and gaps and breaks in others. Increases in continuity of the eUCI40 across years indicated correct linkage. Contemporaneous sequences of eUCI40 across years in different but proximate providers for a client provided even stronger evidence of correct linkage. These patterns suggest complementary roles of sets of identifying personal attributes and of geographical location in linking clients longitudinally.

A redeeming virtue compensating for weakly identifying personal attributes in RSR data turned out to be the relatively small number of RSR records in a typical eUCI40 group. Because different providers used the same rules to create the eUCI40 for a client, the eUCI40 groups carried over from provider to provider. Even though different clients could have the same eUCI40, the eUCI40 confined most clients to groups (or blocks) within which geographic and personal attributes could potentially distinguish one client from another.

The matching variables that were initially considered were race, ethnicity, housing status, poverty level, HIV status, HIV risk factors, provider, enrollment status, transgender status,

geographic unit, first service year, HIV diagnosis year, death date, and first ambulatory care year. After examining each of the selected variables, we removed the variables that had a high rate of nonresponse, invalid values, or with a low information value for distinguishing true matches from non-matches. We also excluded variables that contributed to the eUCI40 generation (date of birth gender, and transgender status). In addition, we added a new variable Provider Core Based Statistical Area (PCBSA) [48] by geocoding the providers' address and mapping the provider location to a metropolitan statistical area (MSA) [49]. Because RSR data not include client addresses, we used the PCBSA variable as a proxy for client location. The variables ultimately selected for RSR linkage were race, ethnicity, housing status, poverty status, HIV risk factor, HIV/AIDS status, provider ID, and provider service location: state and Core Base Statistical Area (CBSA) [49] codes.

Each matching variable carried a different weight in contributing evidence of distinguishing records. Some had more discriminative value than others. For example, agreement on provider, provider's CBSA location code and state code contributed more evidence of a match than agreement on characteristics such as ethnicity or race.

Matching algorithm

We first created pairs of client records from years 2010 (*A2010*) and 2011 (*B2011*) RSR data sets. The algorithm forms pairs of records in blocks (groups) sharing the same eUCI40s. The initial comparison was made on $NA2010i * NB2011i$ number of pairs within each block in year 2010 data set (*A2010*) and 2011 data set (*B2011*), where *NA2010i* is the number of records in a block in data set *A2010*, *and NB2011i* is the number of records in the same block i in data set *B2011*.

Then the F-S model was applied to the linkage algorithm to calculate the matching weights for each variable based on the estimated agreement conditional probabilities (m) and the disagreement conditional probabilities (u). A starting estimate of m could be 0.9 for a name match in two records of the same person, and a u estimate of 0.05 for a name match between two different persons. The agreement weight was calculated as (log $[m/u]$) and the disagreement weight was calculated as (log $[(1\text{-}m)/(1\text{-}u)]$) for each matching variable. Agreement weights were generally positive and disagreement weights were generally negative.

For each record pair, the match score was calculated by adding the agreement weights for each matching variable that was in agreement and the disagreement weights for each matching variable that was in disagreement. Record pairs with agreement on multiple matching variables will have large positive match scores; record pairs with disagreement on most matching variables will have negative match scores. The matching score represents the likelihood of records belonging to the same individual, given the agreement or disagreement on the set of matching variables. The above approach for estimating the linkage parameters was an iterative refitting process. Each cycle merged an annual data set to the previous data set and formed a combined data set in the longitudinal series.

The formal F-S model specifies a comparison vector of variables v [1,J] and weights **w**[1,J;k] from data i ∈{1,2} having latent probabilities of k = $\{m_j,u_j\}$ if there is a match or non-match of v_{*j}, where match score s =

$$\sum_v (c_{ij}) \times w_{jk=m} + (1 - c_{ij}) \times w_{ij=m}$$

$$where\ c_{ij} = 1\ if\ (v_{1j} = v_{2j}) \cup 0\ if\ (v_{1j} \neq v_{2j})$$

A threshold score determines the assignments of pairs of records to matched or unmatched classes. The selection of a threshold has a key role in data linkage of two data sets. All analyses were performed in SAS version 9.3 (SAS Institute Inc., Cary, North Carolina, USA).

Threshold selection

We selected the threshold for classification based on the distributions of match scores for true matches and non-matches from the pilot test on 2015 RSR data sets. We plotted both the observed frequencies and simulated frequencies to evaluate the estimated linkage error at the selection threshold and to minimize errors. Fig 1 shows the cumulative distribution functions (cdf) of the matched weights for the matched and unmatched pairs in Cycle 2015. The first plot shows the cdf for all matched pairs; the second plot shows the best pairs retained in 1:1 matching. Using the threshold at a match weight score of 1, the 2010–2015 pilot test yielded a 60.5 percent (695,985 record pairs) agreement on all matching variables of the matched pairs within eUCI40s. For individual match variables, there was 96 percent agreement for race, 99 percent agreement for the risk and provider MSA variables, and greater than 80 percent agreement for the other variables.

Validation

We conducted sample review and external validation to identify potential errors, as well as to assess the level of accuracy of the linkage and the extent to which the linked data set

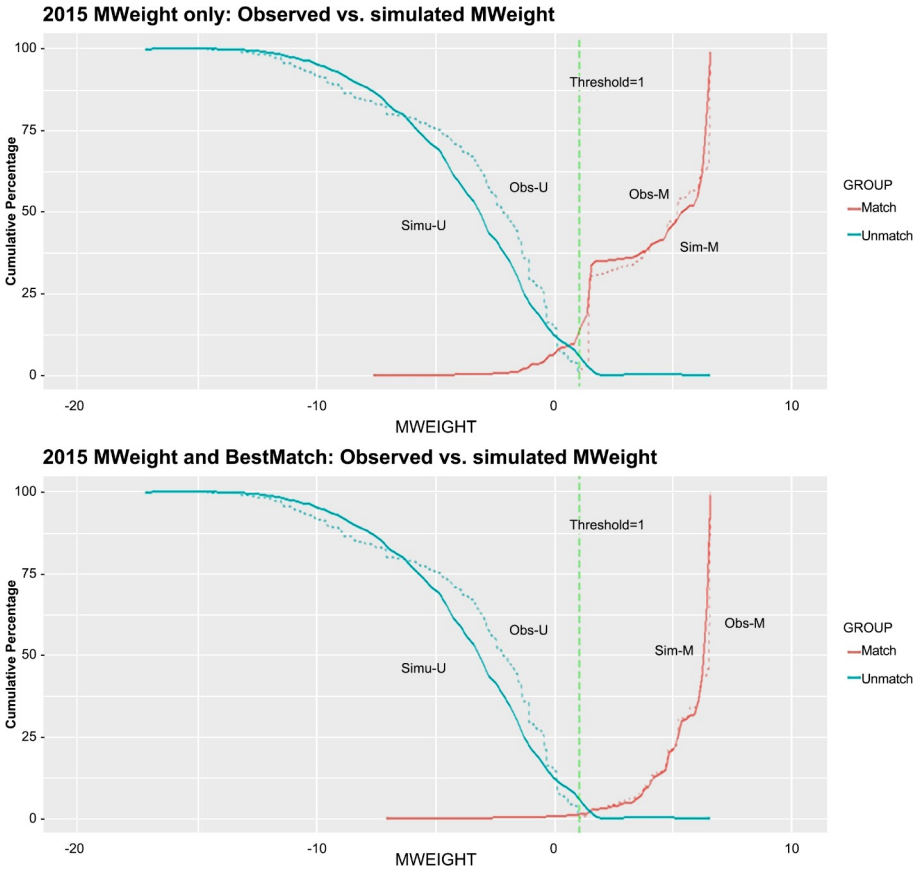

Fig 1. Cumulative distribution functions of the final match weight, 2015.

https://doi.org/10.1371/journal.pone.0237635.g001

corresponds to what was collected and captured within RWHAP clinics. In addition, we assessed evidence of improvements in the accuracy, coverage, and continuity in the longitudinal database.

Sample review. We reviewed a stratified client record sample from the RSR linked data set to assess the accuracy of the longitudinal record linkage. Our review included all client reports within each sampled eUCI40 and variables for matching and other support variables (e.g. the client enrollment status). To select our sample, we created stratification groups of eUCI40s that appeared in the RSR data for two, three, four, five, or six years. We then selected ten (10) eUCI40s from each stratification group, yielding a total sample of 50 eUCI40s. Each eUCI40 could be associated with multiple client records within and across years. If two or more of these records represented the same client, then the records were linked. This sample review only covered the 2010–2015 interval because only 2010–2010 RSR data were available when this review was conducted.

Comparison with an external data source. We also examined the extent to which our linked RSR records corresponded to records collected in RWHAP clinics. To accomplish this task, we used data from the HIV Research Network (HIVRN). HIVRN is a network of HIV care providers from across the United States who provide timely demographic, clinical, and health services costs and utilization data on people with HIV from a consortium of adult and pediatric clinics across the United States [50]. HIVRN data are close to an ideal data set for comparison because most of the HIVRN providers are funded by RWHAP and all clients in the network were uniquely identified across providers and years. We obtained aggregated provider-level data from the eleven service providers that are both in the HIVRN health care network and are funded through RWHAP. The final data for comparison contain information on the number of clients within the service provider each year, and the number of new clients within the service provider in both the current and prior year for calendar years 2010 through 2015. Since we only obtained the 2010 to 2015 aggregated provider-level data from the HIVRN, the validations only covered the 2010–2015 interval.

We used the HIVRN-RWHAP linkage ratio across all eleven service providers as a measure of comparison between these two data sets. The ratio is defined as:

$$\frac{\%\text{ of RWHAP clients found in both the current and previous analysis year}}{\%\text{ of HIVRN clients found in both the current and previous analysis year}}$$

A client was deemed "year-to-year linked" if they received care in both the previous year and the current year. The denominator of "% of clients found in both the current and previous analysis year" was the number of clients who received care in the previous analysis year. The rate of carryover from year to year, though not an explicit part of threshold selection or external validation methods, had a crucial role in both as a reality check. Linkage resulting in too low or too high a carryover rate would imply increases or decreases in numbers of clients that would be inconsistent with overall service levels and payments to providers.

Results

RSR linkage

Table 1 shows the results of the longitudinal linkage across the seven cycles from 2010–2016. Seventy percent of clients who appeared in 2010 also appeared in 2011. Seventy-four percent of clients who appeared in 2012 also appeared in either 2010 or 2011, or in both. In the year 2011, 30 percent of clients had not been seen in 2010, and in 2012, 26 percent of clients had not been seen in 2010 and 2011. In 2011, about 30 percent of clients had appeared in 2010 but were not found in 2011.

Table 1. RSR linkage results by year.

	# Clients Appearing in Current Year and Previous Years*	# of Clients Appearing in Current Year but not Previous Years**	# of Clients Appearing in Previous Year but Not Current Year***	# of Unique Clients Within Current Year
2010	-	-	-	
2011	385,753 (70%)	168,893 (30%)	170,422 (31%)	554,646
2012	395,152 (74%)	141,067 (26%)	176,263 (32%)	536,219
2013	394,801 (75%)	129,874 (25%)	168,726 (31%)	524,675
2014	407,267 (80%)	104,947 (20%)	146,650 (28%)	512,214
2015	423,625 (79%)	110,127 (21%)	128,138 (25%)	533,752
2016	437,456 (79%)	114,108 (21%)	127,271 (24%)	551,564

*The second column is the total number of clients who appeared in the current year and earlier years.

**The third column in this table is the total number of clients who appeared in the previous year but not in the current year.

***The fourth column shows the total of unique clients per year.

https://doi.org/10.1371/journal.pone.0237635.t001

Table 2 provides information on the overall number of RWHAP clients who were linked using the longitudinal linkage. Overall, 39.5 percent of clients only appeared in one year; 49.6 percent of clients appeared in multiple years, but not all seven years; and 10.9 percent of clients appeared in all seven years.

The total number of clients who appeared in multiple years was 802,686, representing about 60 percent of the total number of clients over the seven-year study.

Validation

Sample review. The longitudinal linkage worked well among all sampled cases except for two cases in the 2014 cycle. Forty-eight of the fifty (96 percent) eUCI40 blocks had correct assignments of client records. Following annual deduplication in 2014, two cases previously matched on identifiers that included geographic unit codes failed to link after collection of the client location ended due to stricter privacy constraints.

Comparison to HIVRN data. Table 3 shows the 2010–2015 year-to-year HIVRN-RWHAP linkage ratio for the eleven matched service providers. The overall year-to-year client carry over rate for the HIVRN sites was about 84 percent, while the annual RSR longitudinal client carry over rate fluctuated and improved between 2010–2015 (from 67.7 to 78.6 percent). The range of the year-to-year HIVRN-RWHAP linkage ratio was from 79.9 to 94.2 percent. The RWHAP and RWHAP/HIVRN Ratio have slight upward trends after the first year of linkage.

Discussion

In the current paper, we describe the development of a longitudinal record linkage methodology for use with de-identified longitudinal data on PLWH. This methodology addresses a previous gap in data analytic capabilities by allowing HRSA HAB to link RWHAP clients across reporting years. Initial validation efforts suggest that this proposed methodology is strong in

Table 2. RSR overall linkage rate (2010–2016).

	# of Clients	Percent of Clients
Clients only appearing in one year (2010–2016)	**524,286**	**39.5**
Clients appearing in multiple years	**802,686**	**60.5**
Clients appearing in multiple years, but not all seven years	657,869	49.6
Clients appearing in all seven years	144,817	10.9

https://doi.org/10.1371/journal.pone.0237635.t002

Table 3. Year-to-year HIVRN-RWHAP linkage ratio from 11 matched service providers, 2010–2015.

Year	HIVRN % Linked*	RWHAP % Linked**	RWHAP-HIVRN Ratio
2010–2011	82.8	75.9	91.6
2011–2012	84.7	67.7	79.9
2012–2013	84.2	67.8	80.5
2013–2014	84.7	77.4	91.4
2014–2015	83.4	78.6	94.2

*_HIVRN % Linked_ refers to the percent of HIV clients found in both analysis years, according to the provider's records.

**_RWHAP % Linked_ refers to the percent of HIV clients found in both analysis years, according to the F-S linkage methodology executed on anonymized data supplied to RWHAP by providers.

https://doi.org/10.1371/journal.pone.0237635.t003

its ability to identify the same client across multiple reporting years. This longitudinal linkage methodology also will allow HRSA HAB to better evaluate patterns of viral suppression and service utilization over time. These types of analyses will support future activities to better understand the impact of the RWHAP on HIV/AIDS outcomes and will assist HRSA HAB in monitoring progress toward meeting National HIV/AIDS Strategy goals.

Although there were some noted differences in the comparison between HIVRN and RWHAP longitudinal linkage rates, these differences were not unexpected. The de-identified nature of the RSR data prevented a direct linkage of patients receiving care in the HIVRN network compared to RWHAP clients. Instead, we relied on cross sectional information about the RWHAP funded service providers in the HIVRN network. Some patients who received care at the HIVRN sites may not have been eligible to receive RWHAP services or did not receive RWHAP-funded services in the year. As such, these clients would not be reported in the RSR, but still would have been included in the HIVRN data set. Despite these differences, however, the improvements in the critical carryover rate from one year to another among RWHAP clients (Table 3) indicates improvements in longitudinal linkage during the 2010–2015 interval. Some of the difference between the HIVRN and RWHAP linkage rates do not appear to be due to RWHAP linkage failures alone. While over 50 percent of clients in the 2010–2016 longitudinal database appear consecutively in time spans of at least three years, the remainder appear in sequences with gaps that suggest changes in the RWHAP client or provider populations.

There are some important limitations to our approach that should be noted. Because RSR data do not include any client identification information (i.e., client first and last name, date of birth, and social security number), the quality of the linkage relies heavily on the quality of the eUCI40 and completeness of the client level data submitted to HRSA HAB annually by each provider. A client may receive multiple eUCI40s if the client's identifying information was recorded differently across providers or across years. If so, this client's records would never be matched using our algorithm. This is also true for our within-year linkages. Additionally, our linkage methodology is more likely to miss clients who moved across states and CBSA and were served by an entirely new list of providers. Unfortunately, the extent to which this linkage methodology may miss clients who have moved to other geographic areas is unknown. If attributes of these individuals are associated with lower rates of viral suppression or access to care, subsequent results from analyses of these outcomes using RSR data could be biased. Nevertheless, the proportion of clients who fall into this category could be small. While many factors associated with HIV risk (e.g., homelessness) are linked to high mobility, studies suggest that residential mobility, particularly among the poor, are often short distances as opposed to moving out of state [51–54].

The development of this longitudinal linkage methodology has important implications in the fields of public health and health services research [51]. The linkage of multiple administrative and/or electronic health records can provide valuable opportunities to answer research questions that are not possible with single data sources alone. Nevertheless, there are often data sharing restrictions between and across the organizations that collect these data, particularly when data are contained in medical records. Even if there are no such restrictions on sharing data, laws and regulations, such as the Social Security Number Fraud Prevention Act of 2017, Privacy Act of 1974, HIPPA, Health Insurance Portability and Accountability Act, prohibit the sharing of personal identifiers that can be used to link the data. As a result, there is an important need for alternative methods to facilitate accurate linkages of health records in the absence of personal identifiers.

Supporting information

S1 Appendix.
(DOCX)

Acknowledgments

Shelita Merchant, (HRSA); Marianne Winglee, Westat (retired); Jay Clark, Westat; Jane Xue, Westat; Yu Sun (NCHS); Ann Truelove, Westat; and Kelly Gebo, HIV Research Network (HIVRN).

Author Contributions

Conceptualization: Sigurd Hermansen.

Formal analysis: Julia Zhu, Sigurd Hermansen.

Investigation: Julia Zhu.

Methodology: Sigurd Hermansen.

Project administration: Julia Zhu, Sigurd Hermansen.

Resources: Stan Legum.

Supervision: Julia Zhu.

Validation: Julia Zhu, Laura Sheehan.

Writing – original draft: Julia Zhu.

Writing – review & editing: Julia Zhu, Miranda Fanning, Kerry Grace Morrissey, Stan Legum, Sigurd Hermansen.

References

1. The Ryan White HIV/AIDS Treatment Extension Act of 2009—Title XXVI of the Public Health Service Act, as amended—the Ryan White HIV/AIDS Program Legislation. (2015). https://hab.hrsa.gov/about-ryan-white-hivaids-program/ryan-white-hivaids-program-legislation
2. Centers for Disease Control and Prevention. HIV Surveillance Report, 2016. Vol.28. Published November 2017. Available from: www.cdc.gov/hiv/library/reports/hiv-surveillance.html.
3. Health Resources and Services Administration. Ryan White HIV/AIDS Program Annual Client-Level Data Report 2016. Published November 2017. http://hab.hrsa.gov/data/data-reports. Accessed January 4, 2018.
4. Health Insurance Portability and Accountability Act (HIPAA), Privacy Rule, Pub. L. No. 104–191, 42 U. S.C. § 1320d-2 note. (2000). Available from: http://www.hhs.gov/ocr/privacy/index.html

5. Fellegi IP, Sunter AB. A Theory for Record Linkage. Journal of the American Statistical Association. 1969; 64(328):1183–210.
6. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health services research and data linkages: issues, methods, and directions for the future. Health services research. 2010; 45(5p2):1468–88.
7. Winkler WE, Yancey W, Porter EH. Fast record linkage of very large files in support of decennial and administrative records projects. In Proceedings of the Section on Survey Research Methods, American Statistical Association. 2010:2120–30.
8. Winkler WE. Overview of record linkage and current research directions. In Bureau of the Census. 2006. Available from https://www.census.gov/srd/papers/pdf/rrs2006-02.pdf.
9. Baxter R, Christen P, Churches T. A comparison of fast blocking methods for record linkage. Proc ACM SIGKDD'03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation; Washington, DC. 2003:25–27
10. Borg A, Sariyar M. RecordLinkage: Record linkage in R. R package version 0.4–8. 2015. Available from http://crantastic.org/packages/RecordLinkage/versions/40837.
11. Christen P. A survey of indexing techniques for scalable record linkage and deduplication. IEEE transactions on knowledge and data engineering. 2011; 24(9):1537–55.
12. Churches T, Christen P. Some methods for blindfolded record linkage. BMC Medical Informatics and Decision Making. 2004; 4(1):9.
13. Daggy JK, Xu H, Hui SL, Gamache RE, Grannis SJ. A practical approach for incorporating dependence among fields in probabilistic record linkage. BMC Med Inform Decis Mak. 2013; 13:97. https://doi.org/10.1186/1472-6947-13-97 PMID: 24001000
14. DuVall SL, Kerber RA, Thomas A. Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators. Journal of biomedical informatics. 2010; 43(1):24–30. https://doi.org/10.1016/j.jbi.2009.08.004 PMID: 19683070
15. Elfeky MG, Verykios VS, Elmagarmid AK, Ghanem TM, Huwait AR. Record linkage: a machine learning approach, a toolbox, and a digital government web service. 2003. Available from http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=2572&context=cstech.
16. Elmagarmid AK, Ipeirotis PG, Verykios VS, Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering. 2007; 19(1):1–16.
17. Fair M. Generalized record linkage system–Statistics Canada's record linkage software. Austrian Journal of Statistics. 2004; 33(1&2):37–53.
18. Harron K, Goldstein H, Dibben C. Methodological developments in data linkage. John Wiley & Sons; 2015.
19. Hermansen SW, Leitzmann MF, Schatzkin A. The impact on National Death Index ascertainment of limiting submissions to Social Security Administration Death Master File matches in epidemiologic studies of mortality. Am J Epidemiol. 2009; 169(7):901–8. https://doi.org/10.1093/aje/kwn404 PMID: 19251755
20. Hernández MA, Stolfo SJ. Real-world data is dirty: Data cleansing and the merge/purge problem. Data mining and knowledge discovery. 1998; 2(1):9–37.
21. Imel L, Mule VT Jr. Implementing a Bayesian Approach to Record Linkage. SAS Conference Proceedings. Savannah, GA: SouthEast SAS Users Group. 2015: 8.
22. Jaro MA. Probabilistic linkage of large public health data files. Statistics in medicine. 1995; 14(5–7):491–8. https://doi.org/10.1002/sim.4780140510 PMID: 7792443
23. Larsen MD, Zhao Y. A study of factors affecting record linkage in federal statistical databases. In Federal Committee on Statistical Methodology 2012 Research Conference, Washington, DC. 2012. Available from https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/Larsen_2012FCSM_X-B.pdf
24. Leulescu A, Agafitei M. Statistical matching: a model based approach for data integration. Eurostat-Methodologies and Working papers. 2013:10–2.
25. McNeill N, Kardes H, Borthwick A. Dynamic record blocking: efficient linking of massive databases in mapreduce. In Proceedings of the 10th International Workshop on Quality in Databases (QDB) 2012.
26. Michelson M, Knoblock CA. Learning blocking schemes for record linkage. Proceedings of the National Conference on Artificial Intelligence (AAAI); London; AAAI Press. 2006.
27. Mitchell TM. Machine learning. McGraw Hill: Burr Ridge, IL. 1997.
28. National Center for Health Statistics. National Death Index user's guide. Hyattsville, MD. 2013:1–61. Available from http://www.cdc.gov/nchs/data/ndi/NDI_Users_Guide.pdf.
29. Newcombe HB, Kennedy JM. Record linkage: making maximum use of the discriminating power of identifying information. Commun ACM. 1962; 5(11):563–6.

30. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. Science. 1959; 130(3381):954–9. https://doi.org/10.1126/science.130.3381.954 PMID: 14426783

31. Organisation for Economic Co-operation and Development. Glossary of statistical terms. 2006. Retrieved from https://stats.oecd.org/glossary/detail.asp?ID=3103

32. Randall SM, Ferrante AM, Boyd JH, Semmens JB. The effect of data cleaning on record linkage quality. BMC Medical Informatics and Decision Making. 2013; 13(1):64.

33. Sadinle M, Fienberg SE. A generalized Fellegi–Sunter framework for multiple record linkage with application to homicide record systems. Journal of the American Statistical Association. 2013; 108 (502):385–97.

34. Sariyar M, Borg A, Pommerening K. Missing values in deduplication of electronic patient data. Journal of the American Medical Informatics Association. 2012; 19(e1):e76–82. https://doi.org/10.1136/amiajnl-2011-000461 PMID: 22003173

35. Schnell R. Privacy-preserving record linkage and privacy-preserving blocking for large files with cryptographic keys using multibit trees. JSM Proceedings, Survey Research Methods Section. 2013:187–94.

36. Schnell R, Center GRL. Efficient private record linkage of very large datasets. Proceedings of the 59th World Statistics Congress (WSC). 2013:1862–67.

37. Steorts RC, Ventura SL, Sadinle M, Fienberg SE. A comparison of blocking methods for record linkage. In International Conference on Privacy in Statistical Databases; Springer, Cham. 2014:253–68.

38. Nations United. Handbook of Vital Statistics Systems and Methods. Legal, Organisational and Technical Aspects (United Nations Studies in Methods, Glossary, Series F, No. 35). New York. 1991; 1.

39. van Hest NA, Smit F, Baars HW, De Vries G, De Haas PE, Westenend PJ, et al. Completeness of notification of tuberculosis in The Netherlands: how reliable is record-linkage and capture-recapture analysis? Epidemiology and infection. 2007; 135(6):1021–9. https://doi.org/10.1017/S0950268806007540 PMID: 17156496

40. Wagner D, Lane M. The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software. Center for Economic Studies, US Census Bureau; 2014. Available from https://ideas.repec.org/p/cen/cpaper/2014-01.html.

41. Wilson DR. Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage. The 2011 International Joint Conference on Neural Networks; IEEE. 2011:9–14.

42. Yancey WE. BigMatch: A program for extracting probable matches from a large file for record linkage. Computing. 2004. Available from: https://www.census.gov/srd/papers/pdf/rrc2002-01.pdf. https://doi.org/10.1109/MCISE.2004.1267606

43. Yancey WE. BigMatch: A program for extracting probable matches from a large file for record linkage. Computing. 2007. Available from: https://www.census.gov/srd/papers/pdf/rrc2007-01.pdf. https://doi.org/10.1109/MCSE.2007.74

44. Yancey WE. An adaptive string comparator for record linkage. Statistics. Washington, DC. 2004. Available from https://www.census.gov/srd/papers/pdf/rrs2004-02.pdf.

45. Yancey WE. Bigmatch: A program for large-scale record linkage. In Proceedings of the American Statistical Association, Section on Survey Research Methods [CD-ROM] 2004:4652–55.

46. Yancey WE. Improving EM algorithm estimates for record linkage parameters. Washington, DC. 2004. Available from https://www.census.gov/srd/papers/pdf/rrs2004-01.pdf.

47. Winkler WE. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. US Census Bureau: Washington, DC. 1991.

48. Management and Budget (OMB). [image on internet]. 2019 [updated 2019 Jan 10; cited 2019 April 1]. Available from: https://www.census.gov/programs-surveys/metro-micro/about/omb-standards.html.

49. Management and Budget (OMB). [image on internet]. 2019 [updated 2018 Jan 17; cited 2019 April 1]. Available from: https://www.census.gov/programs-surveys/metro-micro/about/omb-standards.html.

50. HIV Research Network. About. [image on internet]. 2019 [updated 2018 Jan 17; cited 2019 April 1]. Available from: https://cfar.globalhealth.harvard.edu/hiv-research-network-hivrn.

51. Marks G, Patel U, Stirratt MJ, Mugavero MJ, Mathews WC, Giordano TP, Crepaz N, Gardner LI, Grossman C, Davila J, Sullivan M. Single viral load measurements overestimate stable viral suppression among HIV patients in care: clinical and public health implications. Journal of acquired immune deficiency syndromes (1999). 2016; 73(2):205.

52. Parker RD, Dykema S. The reality of homeless mobility and implications for improving care. Journal of community health. 2013; 38(4):685–9. https://doi.org/10.1007/s10900-013-9664-2 PMID: 23494281

53. Cooke TJ. Residential mobility of the poor and the growth of poverty in inner-ring suburbs. Urban Geography. 2010; 31(2):179–93.

54. Fitchen JM. On the Edge of Homelessness: Rural Poverty and Housing Insecurity 1. Rural Sociology. 1992; 57(2):173–93.