BMC Public Health

RESEARCH

Open Access

Evaluating the utility of public-facing jail registers to inform public health practice, Washington state 2023

Steven Erly[1,2*], Richard J. Lechtenberg[3], Vasiliki Georgoulas-Sherry[4], Anna Berzkalns[3], Allison Moore[3], Julia C. Dombrowski[3,5] and Jennifer E. Balkus[2,3]

Abstract

Background Data on criminal justice system involvement can support public health efforts in ways that have been recognized for decades, but data protections and jurisdictional boundaries can make data sharing difficult. In many jurisdictions, carceral facilities are required to publish lists of currently incarcerated individuals. Automated collection of these lists may be one way for public health to access this information. The purpose of this project was to evaluate the availability, completeness, and utility of carceral data collected from public-facing registers in Washington State.

Methods Program staff at the Washington State Department of Health catalogued the websites of all carceral facilities in Washington State and identified what information was available about currently incarcerated individuals. This information was downloaded daily from 1/1/2023 to 12/31/2023 using R software. The completeness of this data was compared relative to the Washington State Jail Booking and Release System (JBRS) during the same time frame. To evaluate the utility of the scraped data (which may contain only partial identifiers) for record linkage, we performed a set of simulated linkages between two external datasets with a known relationship (King County Jail bookings and a surveillance list of people living with HIV who may be out of care). We applied a simple match algorithm to copies of these datasets that had been reduced to match the different combinations of identifier variables available in the public data (full names and ages, partial names, etc.) We compared the sensitivity and positive predictive value (PPV) of the algorithm applied to the reduced datasets and calculated an estimate for the entire dataset weighted by incarcerated population size.

Results At the time of the project, 61 of 71 facilities in Washington State published information about current inmates. 100% of these 61 published names of inmates, 33% age or date of birth, and 13% other identifiers. We collected data from 58 facilities over the project span. 89% of individuals in JBRS were present in the daily scraped data and 95% of individuals in JBRS who were incarcerated for more than 24 h. We estimated that the collected data had 87.7% sensitivity of and 88.8% PPV in linkages with HIV registries.

Conclusions Public facing carceral data in Washington State constitute a data source with high completeness and adequate information for record linkage.

*Correspondence:
Steven Erly
steven.erly@doh.wa.gov

Full list of author information is available at the end of the article

Keywords Jail, Linkage, Public health practice

Introduction

Although it has long been recognized that data on criminal legal involvement can inform public health efforts, data protections and jurisdictional boundaries have hindered effective data sharing [1–3]. Prisons and jails disproportionately house marginalized and underserved populations and can present an opportunity for screening, care linkage, and treatment of infectious and non-infectious disease [4–7]. At the same time, carceral institutions tend to be cautious about data sharing and a 2020 report by the RAND corporation found that data sharing efforts "often are hindered by misconceptions about protected data, a lack of trust in outside organizations, and a lack of confidence in the jail's own ability to manage the risks involved." [8].

Public facing carceral registers may be a source of information that can bridge this gap. In the United States, state prisons and local jails generally publish lists of their current population [9]. These provide some information on the incarcerated population and have been used with modest success to facilitate care linkage for HIV and syphilis and characterize the population entering and exiting the carceral system, but there are several logistical challenges to their widespread use [10–12]. The information contained in these rosters is generally scattered across multiple websites and formats and it can be labor intensive to search them manually [13]. The information contained in a given jail roster is also transient; once a person leaves a carceral facility their name is removed from the roster and their information is unavailable. Finally, it is unclear how complete and timely the data contained in these rosters are. Regulations that create public carceral rosters are controlled at the state and local level, and the requirements and their enforcement may vary [13]. In many cases, the information released may consist of only a list of names with little other context; the value of this limited set of identifiers is uncertain.

One way to increase the utility of public-facing carceral rosters for improving public health services is to use automated computer programs to collect, organize, and temporarily store data. This type of procedure is known colloquially as "web scraping" or "data mining" and involves writing individual programming scripts to download and parse the data on a regular basis. This has been used on a wide variety of health and non-health related topics such as collecting spatial data and identifying organ trade networks [14, 15]. If successful, this would yield a data source that public health could use for activities such as locating individuals who would benefit from relinkage to care. Using Washington State as an example, we describe the availability of public-facing carceral data, evaluate its completeness, and quantify its utility for record linkage using HIV surveillance data. A more complete understanding of the quality of this data will allow public health agencies to consider its value and interpret the results of its use.

Methods

This was a descriptive analysis of data contained in public-facing carceral registers in Washington State between 1/1/2023 and 12/31/2023 and a simulation of linkages using this data.

Program staff at Washington State Department of Health (WA DOH) generated a list of all carceral institutions in Washington State. Under Washington State legal code, every prison and jail institution is required to publish a list of current residents [16]. Each carceral institution website was reviewed to confirm the availability of the information. If no roster was available for an institution's website, the information was requested via email.

Descriptive statistics were used to summarize data availability, including number of facilities with available data, the jurisdictions represented (State, County, City, or Tribal), types of identifiers available (Full Name [e.g. "John Smith"], Partitioned Name [e.g. "First name "John", Last name "Smith"], Age, Date of Birth, and Other Identifiers), and register structure (HTML, API, PDF, and email report). Values weighted by the median daily population of each facility were calculated to express the data availability as a proportion of the incarcerated population of Washington.

For each institution, a web scraping script was developed using R software. The contents of these scripts depended on the structure of the report. HTML rosters were parsed via the HTTR package, JavaScript rosters were parsed via Rvest, and PDF rosters were parsed via PDFTools. Scripts were scheduled to run daily at 6 am and updated as website structure or location changed. The number of days that each program ran successfully was also summarized.

Comparison to state data sources

The dataset generated through web scraping scripts ("scraping dataset") was compared to information from the Washington Statewide Jail Booking and Reporting system (JBRS). JBRS is a database of booking and release records reported to the Washington Association of Sheriffs and Police Chiefs by 46 city and county jails in Washington State [17]. WA DOH received a one-time dataset of the names and dates of incarceration for people who were booked in Washington State in 2023 and matched this to the incarcerations in the scraping

dataset based on first name, last name, and facility of incarceration. For each facility, the percent of individuals in JBRS who were recorded in the scraping dataset overall and for the days data were successfully scraped were calculated.

Evaluation of match utility

To evaluate the usefulness of the limited identifiers contained in the scraped dataset for record linkage, a set of simulated linkages was performed between two external datasets with a known relationship: a list of bookings at two King County Jails between 5/3/2018 and 4/23/2024 and the population of people living with HIV in King County. King County Correctional Facility is the largest single-county jail in Washington State. These data were generated as part of an on-going jail-based HIV relinkage effort ("JBLink") undertaken by the health department. Details about the source of these data have been previously reported [12]. In the simulation, we applied a simple match algorithm to copies of these datasets that had been reduced to match the different combinations of identifier variables available in the public data (supplemental material 1).

Sensitivity and positive predictive value (Clopper-Pearson 95% confidence intervals) of this algorithm were calculated relative to the known matches in the dataset for each combination of identifiers found in the scraping dataset. An average sensitivity and positive predictive value for the scraped dataset was computed, weighted by the proportion of the observed incarcerated population represented with each combination of identifiers. Finally, as a benchmark, sensitivity and positive predictive value were calculated using the fastLink algorithm that underlies the JBLink project [12]. The JBLink project uses a modification of the fastLink algorithm that adds a deterministic match on first and last names and DOB, regardless of the match probability estimated by fastLink. A fastLink threshold of 95% was used until September of 2021; following this a threshold of 87.5% was used.

It has been demonstrated that linkage processes have biases with respect to race and ethnicity [18]. Many commonly used linkage processes are structured around Western name standards (i.e. first, middle, and last name) and do not account for cultural differences in name structures. A sensitivity analysis was performed repeating the analysis stratified by the race and ethnicity of the individuals in the linked datasets to understand how these processes might contribute to racial and ethnic disparities if used in a public health setting.

Results

In 2023, data were successfully collected from 58 of the 61 facilities with complete public rosters (Table 1). The remaining three facilities were not identified until after the project period, and thus were not included in this analysis. Over the year, the scraping programs downloaded the public facing rosters 91% of days. However, the frequency varied by facility type with successful downloads from Tribal jails 74% of days, followed by prisons (83%), county jails (92%), and city jails (97%). Reasons for non-collection included the data not being posted, host website unavailability, website structure changes, or an outage of the computer performing the data collection.

Comparison to state data sources

There were 8 city jails and 25 county jails represented in both our scraping and the JBRS datasets. Of the 47,404 unique people in JBRS during 2023, 32,778 (85%) were present in our scraping dataset. On days when scraping was completed, 38,458 of 43,278 (89%) were present in the scraping dataset. This was higher in county jails (90%) than city (81%) and from lists accessed via API or HTML download (both 89%) rather than PDF or emailed reports (82%). The 4,820 individuals who were not present in the scraping dataset had shorter median time incarcerated (Median < 1 day; interquartile range [IQR] < 1–2 days)1st compared to those who were present in both data systems (median 2 days; IQR 1–12 days). Of those in JBRS who could be identified as being incarcerated for more than 24 h ($n = 8,314$), 95% were present in the scraping dataset.

Evaluation of match utility

Between 5/3/2018 and 4/23/2024 there were 1,779 instances of a person living with HIV and reported in the HIV registry being booked in King County Jails. Simple matching algorithms were implemented using different sets of matching variables (Table 2), with the highest degree of matching (sensitivity) seen when the match variables included unpartitioned name and date of birth (97%) and lowest with partitioned name only or unpartitioned name only (85% for both). PPV was highest (96%) when date of birth or age were present along with name vs. not (77%). The JBLink benchmark had a sensitivity of 90% and PPV of 96%. The weighted average estimate for sensitivity and PPV for all the data available in Washington was 87.7% and 88.8%, respectively. Full results are available in Table 2.

Sensitivity was relatively consistent across racial/ethnic categories; however it was lower among people who identified as Hispanic or Latino/a/x (sensitivity range 74–95% versus 85–97% for the entire population; Table 3).

Discussion

In this evaluation we found that public-facing carceral data in Washington State constitute a data source with high completeness and adequate information for record

Table 1 Characteristics of publicly available jail and prison rosters in Washington state as of December 31, 2023

Variable	Number of Facilities	% of Facilities[a]	% of Population[b]
Availability			
Full Availability	61	-	-
Partial Availability	4[c]	-	-
Inaccessible	6[d]	-	-
Jurisdictions Represented[e]			
State	12	19%	50%
County	39	72%	40%
City	8	13%	10%
Tribe	2	3%	1%
Available Data			
Full Name	61	100%	100%
Partitioned Name	8	11%	6%
Age	16	26%	59%
Date of Birth	4	7%	3%
Other Identifiers[f]	9	13%	4%
Access Method			
HTML Download	21	34%	27%
API Access	34	53%	72%
PDF Download	5	10%	1%
Emailed Report	1	2%	< 1%

[a]Percent of facilities successfully accessed

[b]Weighted percentage based on average daily facility population from scraped data

[c]Two facilities only have booking information accessible. Two facilities have registries that are not updated regularly

[d]Two small tribal facilities and four small county jails. In 2022, the tribal facilities represented tribes with a total of 31,910 members and the counties had a total population of 59,204. The total population of Washington in 2022 was 7,785,786 [19].

[e]The following sections summarize the 61 facilities whose rosters were fully available

[f]Other identifiers include race, ethnicity, gender, and city of residence

Table 2 Sensitivity and positive predictive value of linkages using different sets of matching variables, King County jail and HIV out of care program 5/3/2018 to 4/23/2024[a]

Match Variables	Matches missed	Total matched	True matches	False matches	Sensitivity	PPV
Unpartitioned name + Date of Birth	62	1787	1717	70	97% (96–97%)	96% (95–97%)
Partitioned name + Age	190	1654	1589	65	89% (88–91%)	96% (95–97%)
Unpartitioned name + Age	191	1653	1588	65	89% (88–91%)	96% (95–97%)
Partitioned name only	267	1953	1512	441	85% (83–87%)	77% (76–79%)
Unpartitioned name only	268	1954	1511	443	85% (83–87%)	77% (75–79%)
JBLink Algorithm (Benchmark)	178	1660	1601	59	90% (88–91%)	96% (95–97%)

[a]Sensitivity and positive predictive value were calculated by applying a basic match algorithm to link a dataset of jail bookings in King County to King County's subset of the statewide HIV registry

linkage with public health data. Over the course of a calendar year, we were able to collect data on an estimated 85% of individuals who were incarcerated in Washington and 95% of those who were incarcerated for more than 24 h. The number of identifiers in the data varied by facility, but yielded a high sensitivity and positive predictive value when linked to a population of people living with HIV and, and a higher sensitivity even than a probabilistic linkage using a larger set of identifiers. Taken together, this suggests that this public data can be a valuable source of information for public health programs seeking to locate individuals for public health purposes, such as finding named partners of individuals diagnosed with syphilis so that they can be offered testing or coordination with release planners for people living with HIV who have special care needs, and for better understanding the role of criminal justice involvement in health disparities. This finding is supported by successful projects to use these data in Washington [11, 20].

The data provided in these rosters do not require data sharing agreements to access and do not require additional effort on the part of the carceral facilities to share. Although this makes it convenient to access the data, it may circumvent important collaboration processes that are necessary for successful implementation of jail-based relinkage services. For example, testing and treatment of infectious conditions in jail would most likely require disclosing some amount of personal health information to

Table 3 Sensitivity of linkages using different sets of matching variables by race/ethnicity, King County jail and HIV out of care program 5/3/2018 to 4/23/2024[a]

Match variables	Total	AI/AN	Asian	Black	Hispanic	NHOPI	White	Multiple
Unpartitioned name + Date of Birth	97% (97%–97%)	100% (85–100%)	100% (87–100%)	97% (95–98%)	95% (92–97%)	100% (66–100%)	96% (94–97%)	99% (96–100%)
Partitioned name + Age	89% (88-91%)	91% (72–99%)	93% (76–99%)	91% (89–94%)	76% (68–79%)	100% (66–100%)	92% (89–94%)	91% (86–95%)
Unpartitioned name + Age	89% (88-91%)	91% (72–99%)	93% (76–99%)	91% (89–94%)	76% (70–81%)	100% (66–100%)	92% (90–94%)	91% (86–95%)
Partitioned name only	85% (83-87%)	91% (72–99%)	89% (71–98%)	88% (85–91%)	74% (68–79%)	89% (52–100%)	87% (84–89%)	72% (76–87%)
Unpartitioned name only	85% (83-87%)	91% (72–99%)	89% (71–98%)	88% (85–91%)	74% (68–79%)	89% (62–100%)	87% (84–89%)	82% (76–87%)
JBLink (Benchmark)	90% (88–91%)	100% (85–100%)	89% (71–98%)	91% (89–94%)	80% (75–85%)	100% (66–100%)	92% (90–94%)	91% (86–95%)

[a]Sensitivity was calculated by applying a basic match algorithm to link a dataset of jail bookings in King County and HIV Cases Assigned to King County

jail staff. There would need to be guardrails to how this information is shared, which may require the formal data sharing processes that the scraping initially avoided. In general, public health should not be sharing health information with correctional staff, but this type of information can be used to facilitate communication with health and social services staff within the jails or prisons or to identify opportunities to talk directly with affected individuals. There are also outstanding ethical questions about privacy and the role that carceral data should play in research and public health practice, though a workshop in North Carolina found that stakeholders were largely supportive of using public jail data to support HIV relinkage work [13, 21]. Finally, the lower rate of data availability from tribal institutions and the lower sensitivity for Hispanic/Latino/a/x individuals suggests that programs developed from this scraped data may not be equitable with respect to ethnicity.

This evaluation has several limitations. JBRS contains only a subset of carceral facilities in Washington State and was not developed as an evaluation tool. We were also unable to fully evaluate the completion of data from state prisons and tribal facilities as there is no equivalent system to JBRS for these facilities. We also do not have any insight as to why some individuals were not represented in the publicly available rosters or what biases these may create. We hypothesize that jails representing smaller jurisdictions (e.g. city rather than county) or with simpler reporting mechanisms (e.g. pdf or email rather than API or HTML) may have fewer resources to dedicate to ensuring data quality. The accuracy of record linkages will depend on the extent to which the populations being linked overlap. People living with HIV who are not receiving routine medical care may have a higher likelihood of criminal-legal involvement than other populations; the utility of record linkages for other public health topics may be lower [10].

The United States has the highest incarceration rate of any country in the world, and the health of incarcerated communities is inseparable from public health as a whole. As many marginalized populations face increased burden of comorbid conditions, scraping public jail rosters can be a way for health departments to understand how services are reaching these populations and facilitate care linkage efforts [11, 12].

Appendix 1

Description of linkage algorithm

The linkage algorithm used the following information to determine whether a name was a match: date of birth, year of birth, middle name, and first name. We also calculated scores for the frequency of a person's first and last names as the base 2 logarithm of the name's frequency

in the 2010 census. Names that didn't appear in the 2010 census were given the maximum possible value. After blocking on first and last name soundex, a pair of names were considered a match if any of the following were true:

1) Match of full date of birth.
2) Match of year of birth, middle initial, first name, last name.
3) Match of year of birth, first name, last name, the first or last name have a frequency score > 12, and no middle name.
4) Match middle name, first name, last name, the first or last name have a frequency score > 10, and no date or year of birth.
5) Match middle name, first name, last name, the first or last name have a frequency score > 12.6, and no date or year of birth.
6) Match first name, match last name, the first or last name have a frequency score > 19, no middle name, no date or year of birth.

This algorithm was developed for linking syphilis cases to the scraped jail data in a separate project. The original algorithm included criteria for matching county of residence to county of incarceration which were not useful in the context of this analysis (as the individuals in our simulation generally resided and were incarcerated in King County).

Acknowledgements
Not applicable.

Authors' contributions
SE, JCD, RJL, and JEB conceived of the evaluation. SE and JEB performed the analysis. VGS and SE curated and provided data for the evaluation. AB, AM, and JCD provided subject matter expertise of the topic. SE drafted the manuscript. All authors read and approved the final manuscript.

Funding
There was no funding received for this work.

Data Availability
The analytic datasets used and/or analysed during the current study are available from the corresponding author on reasonable request. Access to JBRS data is restricted under Washington State law and only available from the Washington Association of Sheriffs and Police Chiefs.

Declarations

Ethics approval and consent to participate
This project was performed as program evaluation per The Washington State Agency Policy on Protection of Human Research Subjects and is exempt from IRB review.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Washington State Department of Health, Office of Infectious Disease, Olympia, USA
[2]Department of Epidemiology, University of Washington, Seattle, USA
[3]Public Health Seattle & King County, HIV/STI/HCV Program, Seattle, USA
[4]Office of Financial Management, Public Safety Policy and Research Center, Olympia, USA
[5]Division of Allergy and Infectious Diseases, University of Washington School of Medicine, Seattle, USA

Received: 18 March 2025 / Accepted: 17 June 2025
Published online: 08 August 2025

References

1. Glaser JB. Correctional health care: A public health opportunity. Ann Intern Med. 1993;118(2):139.
2. Glowalla G, Subbian V. Data Sharing Between Jail and Community Health Systems: Missing Links and Lessons for Re-Entry Success. In: Otero P, Scott P, Martin SZ, Huesing E, editors. Studies in Health Technology and Informatics. IOS Press; 2022. Available from: https://doi.org/10.3233/SHTI220029. Cited 2024 Apr 22.
3. Binswanger IA, Maruschak LM, Mueller SR, Stern MF, Kinner SA. Principles to guide National data collection on the health of persons in the criminal justice system. Public Health Rep. 2019;134(1suppl):S34–45.
4. Eastment MC, Toren KG, Strick L, Buskin SE, Golden MR, Dombrowski JC. Jail booking as an occasion for HIV care reengagement: A Surveillance-Based study. Am J Public Health. 2017;107(5):717–23.
5. Masarone M, Caruso R, Aglitti A, Izzo C, De Matteis G, Attianese MR, et al. Hepatitis C virus infection in jail: Difficult-to-reach, not to-treat. Results of a point-of-care screening and treatment program. Dig Liver Disease. 2020;52(5):541–6.
6. Rhodes T. Risk environments and drug harms: A social science for harm reduction approach. Int J Drug Policy. 2009;20(3):193–201.
7. Martin MS, Crocker AG, Potter BK, Wells GA, Grace RM, Colman I. Mental health screening and differences in access to care among prisoners. Can J Psychiatry. 2018;63(10):692–700.
8. Russo J, Vermeer M, Woods D, Jackson B. Data-Informed Jails: Challenges and Opportunities. Santa Monica, CA: RAND Corporation; 2020. Available from: https://www.rand.org/pubs/research_reports/RRA108-1.html. Cited 2024 May 2.
9. Cornwall D. Prisoner Locator Tools from State Agency Databases. GODORT; 2023. Available from: https://godort.libguides.com/prisonerdbs.. Cited 2024 Apr 22.
10. Shook-Sa BE, Hudgens MG, Kavee AL, Rosen DL. Estimating the number of persons with HIV in jails via web scraping and record linkage. J R Stat Soc Ser Stat Soc. 2022;185(Suppl 2):S270–87.
11. Haecker K. Real-Time linkage of public jail registries and STI surveillance data for partner services. Oral Presentation presented at: STI Engage; 2024. Atlanta, GA.
12. Avoundjian T, Dombrowski JC, Golden MR, Hughes JP, Guthrie BL, Baseman J, et al. Comparing methods for record linkage for public health action: matching algorithm validation study. JMIR Public Health Surveill. 2020;6(2):e15917.
13. Rennie S, Buchbinder M, Juengst E, Brinkley-Rubinstein L, Blue C, Rosen DL. Scraping the web for public health gains: ethical considerations from a 'big data' research project on HIV and incarceration. Public Health Ethics. 2020;13(1):111–21.
14. Galvez-Hernandez P, Gonzalez-Viana A, Gonzalez-de Paz L, Shankardass K, Muntaner C. Generating contextual variables from web-Based data for health research: tutorial on web scraping, text mining, and Spatial overlay analysis. JMIR Public Health Surveill. 2024;10:e50379.
15. Li MH, Siddique AB, Wilson B, Patel A, El-Amine H, Koizumi N. Identifying kidney trade networks using web scraping data. BMJ Glob Health. 2022;7(9):e009803.
16. Jail register. open to the public—Records confidential—Exception. RCW Jan 1, 1988. Available from: https://apps.leg.wa.gov/rcw/default.aspx?cite=70.48.100

17. Washington Association of Sheriffs and Police Chiefs. Jail Booking and Reporting System (JBRS). Washington Association of Sheriffs and Police Chiefs. Available from: https://www.waspc.org/jail-booking-reporting-system-jbrs-. Cited 2024 May 11.
18. Grath-Lone LM, Libuy N, Etoori D, Blackburn R, Gilbert R, Harron K. Ethnic bias in data linkage. Lancet Digit Health. 2021;3(6):e339.
19. US Census Bureau. 2018–2022 American Community Survey 5-Year Estimates. Available from: https://data.census.gov/table?q=B01003. Cited 2024 Jul 9.
20. Avoundjian T, Golden MR, Guthrie BL, Hughes JP, Baseman J, Jones-Vanderleest JG, et al. Integration of jail booking and HIV surveillance data to facilitate care coordination. J Correctional Health Care. 2024;30(6):406–13.
21. Juengst E, Buchbinder M, Blue C, Rennie S, Brinkley-Rubinstein L, Rosen DL. Improving the continuity of care for people living with HIV experiencing incarceration in North Carolina jails: stakeholder perspectives. N C Med J. 2022;83(5):382–8.

Publisher's Note