Prioritizing Out-of-Care Case Investigations in King County, Washington

Daniel Cockson

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2023

Committee:

Dr. Julia Dombrowski (Chair)
Dr. Brandon Guthrie (Member)

Program Authorized to Offer Degree:

Epidemiology

University of Washington

Abstract

Prioritizing Out-of-Care Case Investigations in King County, Washington

Daniel Cockson

Chair of the Supervisory Committee:
Dr. Julie Dombrowski
Department of Medicine - Allergy and Infectious Diseases

Background

Health departments need to investigate cases of people with HIV who appear to be out of care but often have insufficient resources to investigate all cases and need a way to prioritize investigations.

Methods

In this retrospective cohort analysis, we used classification and regression tree (CART) methodology to develop and validate a decision algorithm for indicating which HIV cases need investigation. The goal is that this algorithm could be used prospectively to confirm out-of-care status and offer assistance relinking to care. The data utilized is from Public Health – Seattle King County's Comprehensive HIV/AIDS Database (CHARD) which is used to manage HIV case investigations. A "priority" designation is applied to investigations where 1.) the individual was confirmed to be out of care and had been successfully contacted or 2.) the individual was confirmed to be out of care and could not be contacted. We considered 20 potential predictors for priority designation relevant to patient demographics, laboratory result patterns, and reported investigation characteristics. We compared the test characteristics of an optimized algorithm and simple algorithm.

Results

During 01/2018 – 12/2022, 4,311 HIV cases were referred for further investigation. Across the validation data, the optimized and simplified algorithms correctly identified 79.9% and 81.3% of priority investigations, respectively. The optimized algorithm had the lowest specificity at 88.0% and the simplified algorithm had the highest specificity at 89.0%. Models did not perform significantly worse across gender, age, and racial ethnic strata except for when applied to Non-Hispanic Asians (poor positive predictive value) and Non-Hispanic Native Hawaiians/Pacific Islanders (poor negative predictive value).

Conclusions

We found that the performance of two algorithms (one optimized and one simplified) developed with CART were effective in identifying non-priority investigations and could be used prospectively to triage case investigations.

1. Introduction

Among people with diagnosed HIV (PWH) in King County (KC), 87% are virally suppressed, a higher level than the U.S. overall (**1**). Most individuals who are not virally suppressed need higher-intensity care due to mental illness, substance use, and other psychosocial barriers **(1-4**). Treatment and engagement for these PWH must address social needs and minimize structural barriers.

Toward that goal, Public Health – Seattle & King County has implemented a "Data to Care" (D2C) program to reach people beyond the walls of the clinic (**13-15**). This approach involves using health department surveillance data to identify people who are out-of-care and offer them assistance re-engaging in care. This program links major public health case registries with laboratory testing sites, jails in King County, and major emergency rooms/hospitals (**8-12**). Epidemiologists and case workers utilize these data to perform case investigations and re-engagement programming for PWH. These interventions can range from high-intensity (direct placement into a low-barrier care setting) to low-intensity (helping make a clinic appointment) depending on each individual's needs. For those needing high-intensity intervention, low-barrier care has been shown to increase viral suppression among PWH facing complex barriers by offering walk-in access to sexual health services, HIV care, social, mental health, and substance use services (**1, 3, 5-7**). The records of cases identified and the results of investigations and relinkage assistance are recorded in a data system specific to the local D2C program.

At present, the resources available to health departments cannot adequately support the number of cases that currently need investigation and/or could benefit from intervention. Case prioritization needs to be optimized to help triage investigations. Most cases identified for investigation do not actually represent a person who is out-of-care. More often, people have moved out of the area. Moreover, most contact attempts are unsuccessful; just under half of contact attempts from the health

department result in contact with an out-of-care individual. Health departments need a way to prioritize cases identified for D2C in order to focus resources on those most likely to represent a person who is truly out-of-care and contactable. Very little guidance is available to health departments to guide case prioritization. Often, the residual case load is simply neglected or investigations are limited to pre-defined at-risk subpopulations (**16-18**).

Data science methodologies can help guide prioritization and focus resources on those individuals who are in greatest need of and will most benefit from improved engagement in HIV and behavioral health-related services. Specifically, machine learning is particularly useful as it is not constrained by pre-specified hypotheses and can uncover unobserved patterns in data. Given these features, we constructed and compared classification and regression trees (CART) to find algorithms that reliably identify “priority investigations” and improve workflow (i.e. closing cases without investigation or contact attempt). CART methods are particularly useful in decision making scenarios as the graphical output is more easily interpretable than traditional regression models.

The goal of this study was to construct and evaluate the performance of two CART algorithms on identifying cases that require further investigation. The first model was an optimized algorithm that maximized the number of investigations correctly identified as priority. The second model was simplified for maximum efficiency. The two models had their performance compared to each other and to the actual investigation result found in investigation data provided by Public Health – Seattle & King County.

2. Methods

Design and Study Population

This was a secondary analysis using data prospectively collected from the cohort of people with HIV who appeared to be out-of-care or viremic in King County from 2018 through the end of 2022. These results were collected in the Comprehensive HIV/AIDS Database (CHARD), a locally developed database used to manage case assignments, investigations, and final disposition results. This data information system links public health HIV case registries, laboratory information/reporting, jail records, and emergency department utilization from King County (**8-12**). Individuals were categorized as out-of-care if there was a gap in reported laboratory results or were viremic (indicating untreated HIV). Specifically, cases identified for investigation met 3 criteria: 1) previously reported to have HIV based on inclusion in the local Enhanced HIV/AIDS Reporting System (eHARS), 2) no CD4 or VL results reported to PHSKC in the recent past (threshold varied from 12 – 18 months) or a VL > 1000 at the time of last report (viremic), 3) not known to have died or moved to another jurisdiction in another surveillance system. The study population was all individuals diagnosed with HIV who lived in King County from 2018-2022 and were identified as potentially out-of-care or virally unsuppressed at least once.

PHSKC has validated the completeness of its surveillance data, and the case investigations are done record-by-record to determine the true status of the individuals who appear to be out-of-care. Some subjects were identified as potentially out-of-care/virally unsuppressed multiple times over the study's period of investigation.

Outcome Definition

The analysis outcome was whether a case referred for investigation was a priority or non-priority investigation. We defined the outcome based on the results of the investigation. For this study, the term "priority" is used to indicate investigations that warrant prioritization because they are more likely to lead to an individual is truly out-of-care or viremic. A priority investigation was identified if 1) the

individual was confirmed to be out-of-care and had been successfully contacted or 2) the individual was confirmed to be out-of-care and could not be contacted. A non-priority investigation was identified as someone who was 1) confirmed to be in-care/virally suppressed or 2) not able to be located or 3) no longer residing in King County or 4) dead or 5) incarcerated.

Predictors

Predictor variables were identified on the basis of having a plausible relationship to engagement in HIV care or the likelihood of being in the area and identifiable. The variables were chosen based on data availability, expert opinion, and relevant scientific literature. Injection drug use and stimulant drug use at HIV diagnosis were included in the model as there is a strong relationship between substance use and retention in HIV care (**1-5, 14, 19-20**). Other factors that are associated with retention in care and were included in the model are age, sex at birth, gender, and race/ethnicity (**1-5, 19-20**). These were measured at the time of investigation. In accordance with CDC guidelines, risk of HIV transmission was categorized at the time of investigation as men who have sex with men (MSM), injection drug users (IDU), MSM and IDU, heterosexual contact, and perinatal transmission. Measurements since HIV diagnosis documenting housing status were included as being unstably housed has been linked to lower viral suppression rates, longer linkage to care, and lower retention rates (**1-5, 21**). Reasons for investigation were included in the model and were categorized as viremia (VL < 500), missing laboratory results (> 1 year since last lab), or missed pick-up for antiretroviral HIV drugs. The latter was collected through a D2C partnership between certain pharmacies in King County and PHSKC.

Additionally, a social vulnerability index was used based on the social and economic risk index (SERI) that was used by Public Health – Seattle King County during the COVID-19 pandemic to guide contact tracing for vulnerable populations (**22**). The index includes inputs from factors that are

associated with challenges to accessing healthcare and health information, historic and ongoing systemic racism and discrimination, and low income. The index was developed using 2018 census data and was regionally matched to the individuals last known residence.

Laboratory result patterns were also included in the model. Viral load at investigation, CD4 count at investigation, CD4 nadir, time between last laboratory report and investigation, time since diagnosis and investigation, time to reach viral suppression after HIV diagnosis, and documented duration of viremia have all been documented as being associated with HIV care retention (**1-5, 23-24**). Additionally, these laboratory reports can help determine the likelihood of an individual still residing in King County and being contactable (**9-10**).

The results of past case investigations were also utilized to predict priority designation. Specifically, the number of previous investigations and the designation of those investigations could be related to the likelihood that someone is truly out-of-care/viremic, in King County, and contactable. The method of case identification was also included in the model, (**12, 21**) including sexually transmitted infection (STI) partner services investigations, hospitals and emergency departments, and jail booking data.

Data Analysis

This analysis used classification and regression trees (CART) to identify predictors that are associated with the outcome of interest. This supervised machine learning method utilizes nonparametric, empirical statistical methods to output predictive models (**25-27**). The output forms a tree with progressive binary splits composed of similar groups categorized by the outcome of interest. When a split occurs, CART will utilize the predictor that best differentiates between individuals with or without the primary outcome of interest. A predictor can be used multiple times in the creation of the tree or not at all.

CART methods have been used successfully in medical research to categorize and predict the occurrence of events (**28-29**). The output of CART is easier to graphically interpret than traditional regression models, making them ideal for scientific communication and decision making. Additionally, this method was chosen as it does not make assumptions about underlying distributions and allows for complex interactions between variables.

For this analysis, we fit the models to the Public Health – Seattle King County data. The data were randomly split into a development and validation set. We used the development dataset to perform a CART analysis with predictors of interest. The fitted models developed were applied to the validation set to predict the response variable. The complexity parameter value was chosen to be 0.001 in order to build a very deep initial model. Given that public health practice often requires balancing accuracy and feasibility, we investigated if there was a complexity parameter that resulted in a smaller tree that better minimized the cross-validated error rate. Various complexity parameters were plotted against the size of the tree, and the pruned complexity parameter was identified to be 0.002. The model was then pruned, and both the full and pruned models were assessed based on their accuracy (overall correct classification of true positives and true negatives), sensitivity, specificity, and positive and negative predictive values.

Bagging was employed to see if the model's predictive abilities could be improved. This method bootstraps the training sets and constructs classification trees that are then averaged to get the resulting predictions. These trees are not pruned which leads to high variance, but low bias (**26**).

Bias in the predictions was investigated by examining the difference between actual investigation designation and predicted investigation designation across all three models stratified by gender, race and ethnicity, and age categories. Positive and negative predictive values were also compared.

This study was approved by the University of Washington Human Subjects Division. Data cleaning, model development, and analysis were conducted in R Software (version 4.3.0).

3. Results:

Characteristics of patients and investigations

The study sample totaled 3,159 PWH with a mean age of 49 ± 12.4 years old (Table 1). The majority of individual's sex at birth was male (85.7%) and most were cisgender (98.5%). 47.5% of participants identified as non-Hispanic white with the next largest group identifying as non-Hispanic Black or African American (22.7%). Of individuals with complete data, 537 (18.1%) reported injection drug use at diagnosis, and 273 (8.6%) reported stimulant drug use at diagnosis. 5.2% of participants reported being unstably housed at diagnosis. At HIV diagnosis, 1944 (61.5%) had their risk of transmission identified as men who have sex with men (MSM). Each individual is associated with at least one case investigation documenting relevant laboratory results and characteristics over the course of the study period.

There are 4,311 investigations in the database with some individuals being represented more than once. 1,364 (31.6%) of investigations had priority designations (Table 2). Most priority investigations (41.6%) had 1 to 6 months between referral and the last reported laboratory result. CD4 counts for priority investigations are often greater than 500 cells/mm^3 (40.0%) with a CD4 nadir of less than 200 cells/mm^3 (49.5%). Across priority investigations, the majority of viral load counts were greater than 1,000 copies per mL (50.3%). For priority designations with complete data on time spent viremic,

the majority of investigations (23.5%) had 1 to 6 months of documented viremia at the time of identification for investigation. Viremia was the most common reason for investigation (47.7%) with no recent laboratory results being the next most common reason (31.6%). Investigations designated as priority were more likely to have more than one previous investigation than non-priority investigation (35.6% vs. 24.8% respectively). Additionally, those previous investigations were more likely to have a priority designation (25.0% vs. 1.7%)

Outputs of CART models

The pruned CART model identified having a viral load of 200-1,000 copies per mL or greater than 1,000 copies per mL as the most significant indicator of priority for investigation (Figure 1). 32% of cases in this category met criteria for being a priority investigation. The second most predictive factor for priority outcomes was having one or more previous investigations that resulted in a priority outcome (67% priority). The third most predictive factor of priority designation was having no previous investigations at the time of referral with 60% being identified as priority investigations. There were six other variables included in the full tree (Figure 2) but they were ultimately not included in the final pruned tree as they did not significantly reduce the overall relative error of the model.

Test characteristics

Performance of the models was assessed utilizing the validation data set. The accuracy, sensitivity, specificity, positive predictive value, and negative predictive value were calculated for each model (Table 3). All models performed similarly to each other. Sensitivity was higher for the pruned model at 0.645. The bagged model had marginally higher specificity (0.893) and positive predictive value (0.735). The negative predictive value was equal for both the pruned and bagged model at 0.844. The area under the receiver operating characteristic curve (AUROC) was similar across the three models

(Figure 3). The full model had the highest AUROC at 0.804 and the pruned model had the lowest AUROC at 0.779.

Comparing performance across relevant demographic strata, both the pruned and full model had the two lowest PPVs for individuals not identifying as cisgender and for Non-Hispanic Asians (Table 4). The bagged model had the two lowest PPVs for those identifying as Non-Hispanic Asian or Hispanic/Latino. All models had the poorest NPV for those identifying as Non-Hispanic Native Hawaiian/Pacific Islander.

4. Discussion

In this retrospective analysis of health department case investigations, we found that the performance of two algorithms (one optimized and one simplified) developed with CART were effective in predicting non-priority investigations. For the optimized model, variables most associated with predicting non-priority investigation designation included viral load at referral, outcome of previous investigations, number of previous investigations, time since last laboratory result at time of referral, CD4 nadir at referral, time to achieve viral suppression, CD4 count at referral, age, and transmission risk categorization. For the simplified model, variables most associated with predicting non-priority investigation designation included viral load at referral, outcome of previous investigations, and number of previous investigations. The simplified model outperformed the optimized model marginally.

Given that public health practice often requires balancing accuracy and feasibility, applying the simplified algorithm could reduce the number of cases for investigation. According to this simplified CART model, cases are most likely to not require investigation if there is a low viral load reported at referral, there are no priority designations in a previous investigation (also indicating previous

investigations), or if there are no previous investigations. Additional predictors did not benefit the model's predictive ability giving confidence that the simplified algorithm appropriately identifies cases not in need of further investigation.

These findings are consistent with current literature. Viral loads are often more predictive of an individual needing investigation than gap in laboratory results for two main reasons. First, most individuals with gaps in laboratory results usually have moved away from the reporting region (**30-33**). Second, most people with viremia in the era of contemporary HIV treatment are off their medications. Investigations into viremic cases often yielded more public health benefit than cases with gaps in labs (**13**).

This work has important strengths. First, the data utilized in this analysis has been validated for completeness, and the case investigations were done record-by-record to acquire the true status of the individuals who appear to be out-of-care. Second, the models are visual and easy to comprehend, making them usable immediately in clinical practice. Third, the models provided offer variations in their accuracy and simplicity. Investigators and clinicians can choose which models best fit a specific setting or compare performance across models. Finally, these models provide guidance regarding investigation prioritization in a setting where there previously was very little.

This study also has notable limitations. First, as an initial investigation, the selected input variables were limited by data availability. There may be important variables that would improve prediction but were not included in the models. Second, there might be additional temporal relations among variables across multiple years of case investigations. This study did not aim to uncover those relationships. Third, the given predictive models should be further evaluated by PWH and clinical experts. To address these issues, future studies should (1) integrate mental health status markers

and improving assessment of housing status over time after the point of initial diagnosis, (2) include other variables that may arise as HIV surveillance practices include and link data, and (3) evaluate the equity and usefulness of the identified algorithm for surveillance practice.

Public health departments have little to no evidence to guide decisions about how to efficiently prioritize HIV case investigations. These algorithms (once validated) could be employed to help triage/identify cases needing investigation first. By streamlining the investigation process, investigations that would most likely result in a priority designation could be reached sooner in the workflow. For those investigations that are most likely to result in non-priority designations, investigations would occur based on resource availability. Thus, individuals who are more likely to be out-of-care can be reached sooner and connected with re-linkage services.

In conclusion, we evaluated two triage algorithms that can be implemented in public HIV surveillance settings to improve investigation efficiency and priority case identification. The simplified designation criteria of viral load at referral, outcomes of prior investigations, and number of prior investigations can have a high specificity for identifying cases that do not need further investigation and outreach.

Table 1: Sources of predictors used in models

Predictors	Source	Notes
Demographics		
Age	HIV case report & investigation	
Sex at birth	HIV case report & investigation	
Gender	HIV case report & investigation	
Race/Ethnicity	HIV case report & investigation	
Injection drug use	HIV partner services interview	Self-identified, in past year, at the time of HIV diagnosis.
Stimulant drug use	HIV partner services interview	Self-identified, in past year, at the time of HIV diagnosis.
Social Determinants of Health		
Social vulnerability index	Calculated using residence data connected with HIV case or laboratory report	Factors in the SERI score selected based on association with challenges accessing healthcare and health information (born outside the U.S.), historic and ongoing systemic racism and discrimination (people of color), and low income [indicated by homelessness/unstable housing or residence in a low-income neighborhood (>20% of households with an income level of <200% of the federal poverty limit)]
Homelessness	Residence data connected with HIV case report	At time of HIV diagnosis or any updated address in surveillance database
Laboratory result patterns		
CD4 nadir	Laboratory reporting to health department	Lowest CD4 count on record
CD4 count	Laboratory reporting to health department	
Viral load	Laboratory reporting to health department	
Time since last lab	Calculated using dates of laboratory reporting	Viremia defined as viral load >200
Duration of viremia	Calculated using dates and results of laboratory reporting	
Case investigation characteristics		
Time since HIV diagnosis	HIV case report & investigation	
Time to reach suppression	Laboratory reporting to health department	Time between date of HIV diagnosis and first VL<200
Number of investigations	Data to Care database (CHARD)	Categorized sum of the number of case investigations an individual previously at the most recent investigation (1 or more investigation(s) previously, no previous investigations).
Priority outcome in investigations	Data to Care database (CHARD)	Categorized sum of the number of case investigations designated as a priority an individual had previously at the most recent investigation date (more than 1 investigation designated priority previously, 1 investigation designated priority previously, 0 investigations designated priority previously).
Hospital/ED identification	Data to Care database (CHARD)	Identified for investigation based on an automated algorithm at the time of emergency department visit or hospitalization
Jail identification	Data to Care database (CHARD)	Identified for investigation based on daily matching of jail booking records and HIV surveillance
STI diagnosis identification	Data to Care database (CHARD)	Identified for investigation based on matching of STI case report and HIV surveillance
Transmission risk categorization	HIV case report & investigation	Defined at the time of diagnosis according to CDC definition

Table 2: Characteristics of individuals within sample

	Overall (N=3159)
Age categories (yrs)	
18-24	31 (1.0%)
25-34	402 (12.7%)
35-44	804 (25.5%)
45-54	771 (24.4%)
55-64	815 (25.8%)
65+	336 (10.6%)
Sex at birth	
Male	2706 (85.7%)
Female	453 (14.3%)
Gender	
Other	47 (1.5%)
Cisgender	3112 (98.5%)
Race/Ethnicity	
Hispanic/Latino	529 (16.7%)
Non-Hispanic AIAN	24 (0.8%)
Non-Hispanic Asian	115 (3.6%)
Non-Hispanic Black or African American	716 (22.7%)
Non-Hispanic NHPI	13 (0.4%)
Non-Hispanic White	1500 (47.5%)
Non-Hispanic Multi-race	262 (8.3%)
Time to viral suppression	
< 1 month	60 (1.9%)
1 month - 6 months	610 (19.3%)
6 months - 1 year	277 (8.8%)
1 year - 2 years	295 (9.3%)
2 years - 3 years	202 (6.4%)
3 years - 4 years	190 (6.0%)
4 years - 5 years	162 (5.1%)
> 5 years	1191 (37.7%)
Missing	172 (5.4%)
Unstably housed at diagnosis	
Yes	164 (5.2%)
No	2995 (94.8%)
Injection drug use at diagnosis	
Yes	573 (18.1%)
No	1175 (37.2%)
Missing	1411 (44.7%)
Stimulant drug use at diagnosis	
Yes	273 (8.6%)
No	124 (3.9%)
Missing	2762 (87.4%)

Table 2: Characteristics of individuals within sample

	Overall (N=3159)
Transmission risk categorization at diagnosis	
MSM	1944 (61.5%)
IDU	191 (6.0%)
MSM & IDU	388 (12.3%)
Heterosexual	347 (11.0%)
Perinatal	39 (1.2%)
Missing	250 (7.9%)

Table 3: Characteristics of cases within sample

	Priority investigation (N=1364)	Not priority investigation (N=2947)	Overall (N=4311)
Homeless at case identification			
Yes	186 (13.6%)	254 (8.6%)	440 (10.2%)
No	1178 (86.4%)	2693 (91.4%)	3871 (89.8%)
Social Vulnerability Index			
Mean (SD)	0.555 (0.302)	0.526 (0.305)	0.535 (0.304)
Median [Min, Max]	0.600 [0, 1.00]	0.500 [0, 1.00]	0.600 [0, 1.00]
Missing	147 (10.8%)	322 (10.9%)	469 (10.9%)
Years between referral and diagnosis			
Mean (SD)	12.6 (8.41)	13.7 (8.56)	13.3 (8.53)
Median [Min, Max]	12.0 [0, 39.0]	13.0 [0, 39.0]	12.0 [0, 39.0]
Missing	0 (0%)	8 (0.3%)	8 (0.2%)
Time since last lab			
< 1 month	96 (7.0%)	37 (1.3%)	133 (3.1%)
1 month - 6 months	568 (41.6%)	508 (17.2%)	1076 (25.0%)
6 months - 1 year	183 (13.4%)	196 (6.7%)	379 (8.8%)
1 year - 2 years	356 (26.1%)	1907 (64.7%)	2263 (52.5%)
2 years - 3 years	63 (4.6%)	119 (4.0%)	182 (4.2%)
3 years - 4 years	29 (2.1%)	67 (2.3%)	96 (2.2%)
4 years - 5 years	14 (1.0%)	29 (1.0%)	43 (1.0%)
> 5 years	40 (2.9%)	54 (1.8%)	94 (2.2%)
Missing	15 (1.1%)	30 (1.0%)	45 (1.0%)
CD4 count at case identification			
< 200	298 (21.8%)	205 (7.0%)	503 (11.7%)
200 - 500	494 (36.2%)	826 (28.0%)	1320 (30.6%)
> 500	545 (40.0%)	1857 (63.0%)	2402 (55.7%)
Missing	27 (2.0%)	59 (2.0%)	86 (2.0%)
CD4 nadir at case identification			
< 200	675 (49.5%)	1129 (38.3%)	1804 (41.8%)
200 - 500	474 (34.8%)	1109 (37.6%)	1583 (36.7%)
> 500	188 (13.8%)	650 (22.1%)	838 (19.4%)
Missing	27 (2.0%)	59 (2.0%)	86 (2.0%)
Viral load at case identification			
< 200	575 (42.2%)	2536 (86.1%)	3111 (72.2%)
200 - 1000	74 (5.4%)	55 (1.9%)	129 (3.0%)
> 1000	686 (50.3%)	313 (10.6%)	999 (23.2%)
Missing	29 (2.1%)	43 (1.5%)	72 (1.7%)
Time spent viremic at case identification			
< 1 month	30 (2.2%)	8 (0.3%)	38 (0.9%)
1 month - 6 months	321 (23.5%)	94 (3.2%)	415 (9.6%)
6 months - 1 year	123 (9.0%)	57 (1.9%)	180 (4.2%)
1 year - 2 years	115 (8.4%)	78 (2.6%)	193 (4.5%)
2 years - 3 years	38 (2.8%)	22 (0.7%)	60 (1.4%)

Table 3: Characteristics of cases within sample

	Priority investigation (N=1364)	Not priority investigation (N=2947)	Overall (N=4311)
3 years - 4 years	29 (2.1%)	19 (0.6%)	48 (1.1%)
4 years - 5 years	15 (1.1%)	5 (0.2%)	20 (0.5%)
> 5 years	60 (4.4%)	48 (1.6%)	108 (2.5%)
Missing	633 (46.4%)	2616 (88.8%)	3249 (75.4%)
Reason for referral			
missed ARV pick-up	282 (20.7%)	690 (23.4%)	972 (22.5%)
no recent labs	431 (31.6%)	2132 (72.3%)	2563 (59.5%)
viremia	651 (47.7%)	125 (4.2%)	776 (18.0%)
Number of previous investigations at identification			
One or more previous investigation	486 (35.6%)	731 (24.8%)	1217 (28.2%)
No previous investigations	878 (64.4%)	2216 (75.2%)	3094 (71.8%)
Number of previous priority investigations at identification			
None designated priority	145 (10.6%)	680 (23.1%)	825 (19.1%)
One or more designated priority	341 (25.0%)	51 (1.7%)	392 (9.1%)
Missing	878 (64.4%)	2216 (75.2%)	3094 (71.8%)
One or more identification from STI diagnosis in the past year			
Yes	127 (9.3%)	123 (4.2%)	250 (5.8%)
No	1203 (88.2%)	2798 (94.9%)	4001 (92.8%)
Missing	34 (2.5%)	26 (0.9%)	60 (1.4%)
One or more identification from STI diagnosis in the past 2 years			
Yes	156 (11.4%)	221 (7.5%)	377 (8.7%)
No	1155 (84.7%)	2639 (89.5%)	3794 (88.0%)
Missing	53 (3.9%)	87 (3.0%)	140 (3.2%)
One or more identification from jail booking in the past year			
Yes	60 (4.4%)	58 (2.0%)	118 (2.7%)
No	1284 (94.1%)	2869 (97.4%)	4153 (96.3%)
Missing	20 (1.5%)	20 (0.7%)	40 (0.9%)
One or more identification from jail booking in the past 2 years			
Yes	68 (5.0%)	86 (2.9%)	154 (3.6%)
No	1250 (91.6%)	2819 (95.7%)	4069 (94.4%)
Missing	46 (3.4%)	42 (1.4%)	88 (2.0%)
One or more identification from hospital encounter in the past year			
Yes	91 (6.7%)	146 (5.0%)	237 (5.5%)
No	1135 (83.2%)	2671 (90.6%)	3806 (88.3%)
Missing	138 (10.1%)	130 (4.4%)	268 (6.2%)
One or more identification from hospital encounter in the past 2 years			
Yes	95 (7.0%)	156 (5.3%)	251 (5.8%)
No	1121 (82.2%)	2633 (89.3%)	3754 (87.1%)
Missing	148 (10.9%)	158 (5.4%)	306 (7.1%)

Figure 1: Pruned CART Tree

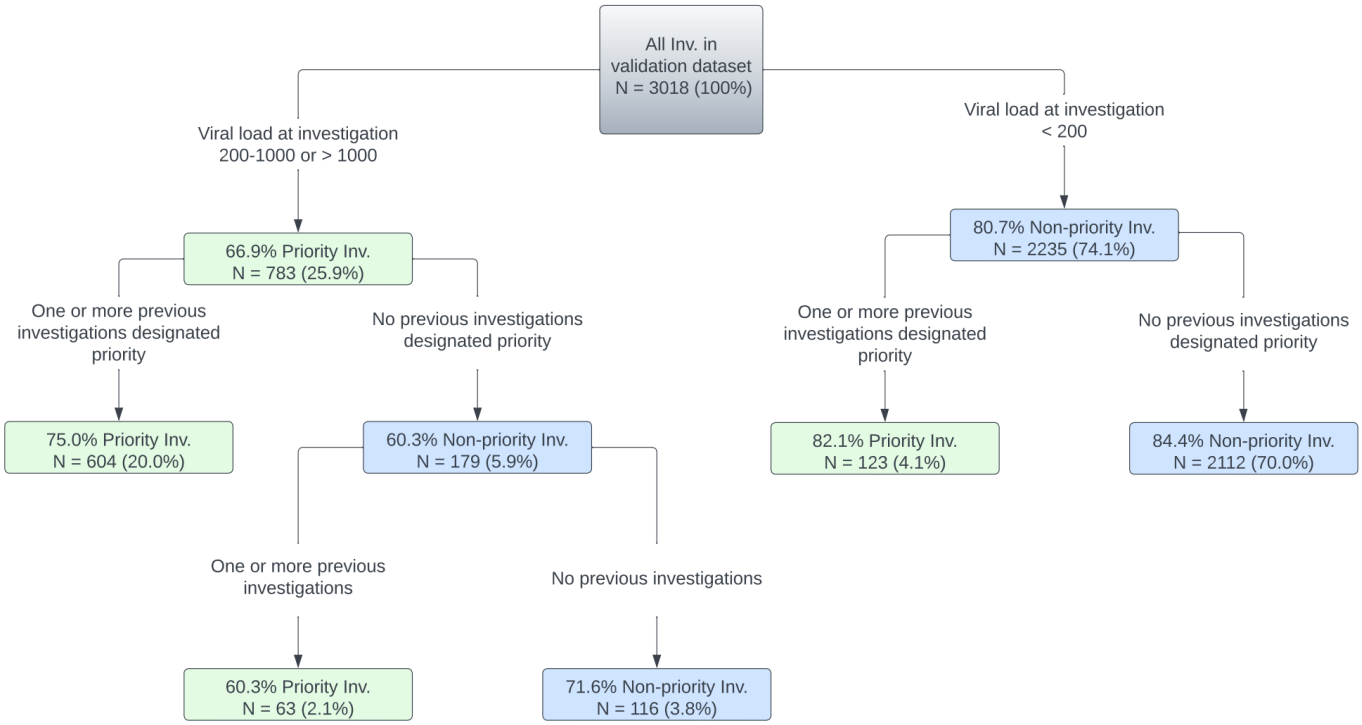


Figure 2: Full CART Tree

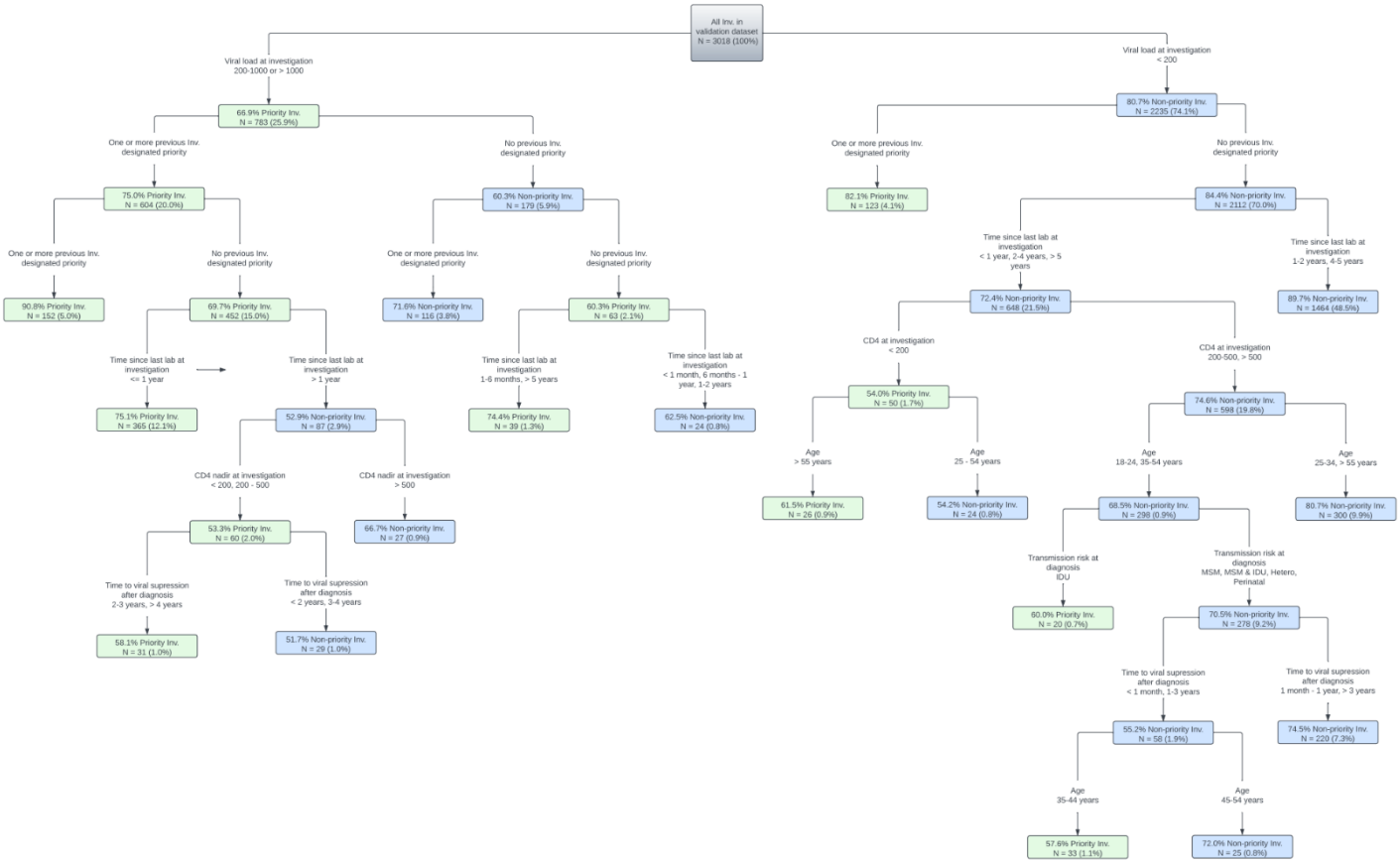


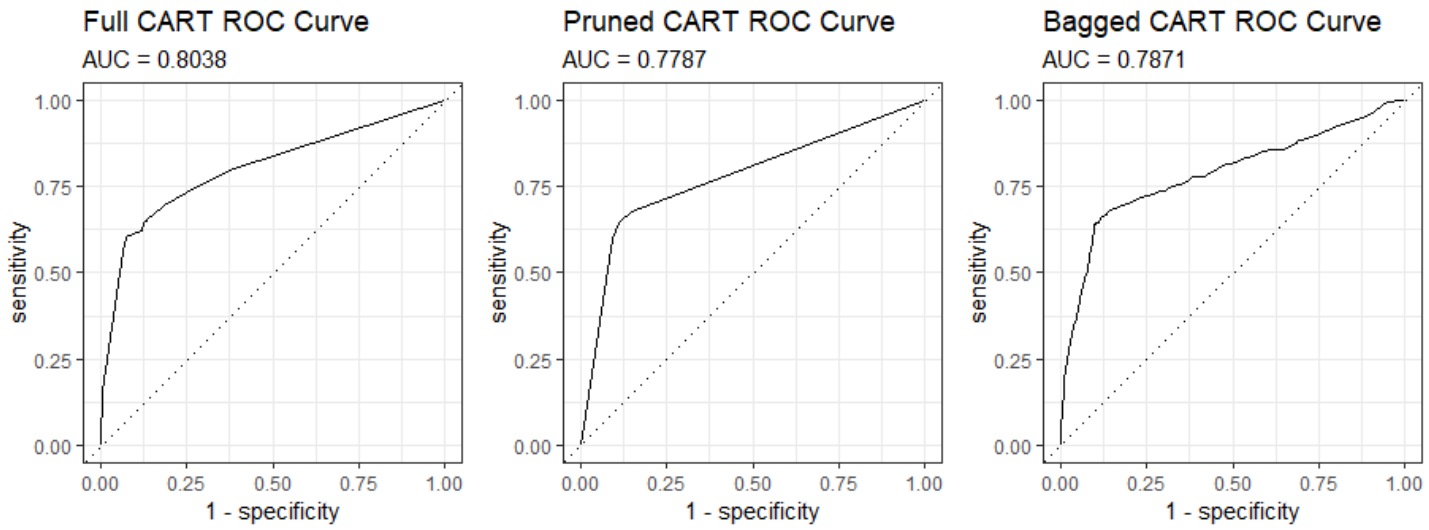
Table 4: Sensitivity analysis across three models conducted on validation set

Model	Accuracy	Sensitivity	Specificity	PPV	NPV
Full Tree	0.799	0.623	0.880	0.706	0.835
Pruned Tree	0.813	0.645	0.890	0.731	0.844
Bagged Trees	0.814	0.643	0.893	0.735	0.844

Table 5: Stratified model performance

		Full Model			Pruned Model			Bagged Model		
	Assigned priority investigations	PPV	NPV	Predicted priority investigations	PPV	NPV	Predicted priority investigations	PPV	NPV	Predicted priority investigations
Total	1364	0.750	0.841	1150	0.744	0.839	1151	0.744	0.839	1150
Sex at birth										
Female	221	0.799	0.843	194	0.816	0.844	190	0.820	0.845	189
Male	1143	0.740	0.840	956	0.729	0.838	961	0.730	0.838	960
Gender										
Cisgender	1348	0.750	0.840	1133	0.745	0.838	1136	0.745	0.838	1137
Other	16	0.706	0.918	17	0.667	0.882	15	0.714	0.885	14
Age categories (yrs)										
18-24	16	0.818	0.731	11	0.833	0.760	12	0.833	0.760	12
25-34	176	0.736	0.861	159	0.833	0.760	12	0.833	0.760	12
35-44	420	0.730	0.832	403	0.743	0.819	377	0.738	0.820	381
45-54	342	0.762	0.818	260	0.754	0.828	276	0.751	0.821	269
55-64	301	0.761	0.858	243	0.759	0.853	237	0.766	0.860	244
65+	109	0.797	0.855	74	0.800	0.848	70	0.800	0.848	70
Race/Ethnicity										
Hispanic/Latino	214	0.709	0.855	196	0.702	0.862	205	0.698	0.861	205
Non-Hispanic AIAN	14	0.833	0.818	12	0.818	0.783	11	0.818	0.783	11
Non-Hispanic Asian	38	0.680	0.835	25	0.667	0.828	24	0.640	0.827	25
Non-Hispanic Black or African American	358	0.784	0.819	292	0.764	0.818	301	0.766	0.821	304
Non-Hispanic NHPI	11	0.778	0.692	9	0.875	0.714	8	0.875	0.714	8
Non-Hispanic White	612	0.744	0.842	503	0.737	0.839	499	0.743	0.838	494
Non-Hispanic Multi-race	117	0.761	0.878	113	0.796	0.867	103	0.794	0.864	102

Figure 3: Model ROC curves



5. References

1. Public Health – Seattle & King County and Washington State Department of Health. HIV/AIDS Epidemiology Report and Community Profile, 2022.
2. Dombrowski JC, Galagan SR, Ramchandani M, Dhanireddy S, Harrington RD, Moore A, Hara K, Eastment M, Golden MR. HIV Care for Patients With Complex Needs: A Controlled Evaluation of a Walk In, Incentivized Care Model. Open Forum Infect Dis. 2019;6(7):ofz294. PMCID: PMC6641789
3. Dombrowski JC, Ramchandani M, Dhanireddy S, Harrington RD, Moore A, Golden MR. The Max Clinic: Medical Care Designed to Engage the Hardest-to-Reach Persons Living with HIV in Seattle and King County, Washington. AIDS Patient Care STDS. 2018;32(4):149-56. PMCID: PMC5905858.
4. Hood JE, Buskin SE, Golden MR, Glick SN, Banta-Green C, Dombrowski JC. The Changing Burden of HIV Attributable to Methamphetamine Among Men Who Have Sex with Men in King County, Washington. AIDS Patient Care STDS. 2018;32(6):223-33. PMID: 29851502
5. Dombrowski JC, Hughes JP, Buskin SE, Bennett A, Katz D, Fleming M, Nunez A, Golden MR. A Cluster Randomized Evaluation of a Health Department Data to Care Intervention Designed to Increase Engagement in HIV Care and Antiretroviral Use. Sex Transm Dis. 2017. PubMed PMID: 29465679.
6. Dombrowski JC, Simoni JM, Katz DA, Golden MR. Barriers to HIV Care and Treatment Among Participants in a Public Health HIV Care Relinkage Program. AIDS Patient Care STDS. 2015;29(5):279- 87. PMCID: 4410545.
7. Dombrowski JC, Carey JW, Craw JA, Pitts N, Freeman A, Golden MR, Bertolli J. Patient and Provider Perspectives on the Development of a Health Department "Data to Care" Program: a Qualitative Study. BMC Public Health. 2016;16:491. PMCID: PMC4901404.
8. Public Health – Seattle & King County, Washington State Department of Health, EHE & Non-EHE Partner Organizations, HIV Prevention & Care Community Members, New Voices. Plan to Support Ending the HIV Epidemic in King County
9. Centers for Disease Control and Prevention (CDC). Compendium of Evidence-Based Interventions and Best Practices for HIV Prevention
10. AIDS United (Awarding Agency: HRSA). Request for Proposals: Using Innovative Intervention Strategies to Improve Health Outcomes Among People with HIV, Implementation Site Application
11. Hood JE, Katz DA, Bennett AB, Buskin SE, Dombrowski JC, Hawes SE, Golden MR. Integrating HIV Surveillance and Field Services: Data Quality and Care Continuum in King County, Washington, 2010- 2015. Am J Public Health. 2017;107(12):1938-43. PMID: 29048962.
12. Dombrowski JC, Buskin SE, Bennett A, Thiede H, Golden MR. Use of Multiple Data Sources and Individual Case Investigation to Refine Surveillance-Based Estimates of the HIV Care Continuum. J Acquir Immune Defic Syndr. Epub 2014/08/21. PMID: 25140904.
13. Avoundjian T, Dombrowski JC, Golden MR, Hughes JP, Guthrie BL, Baseman J, Sadinle M. Comparing Methods for Record Linkage for Public Health Action: Matching Algorithm Validation Study. JMIR public health and surveillance. 2020;6(2):e15917. PMCID: PMC7226047.

14. Avoundjian T, Golden MR, Ramchandani M, Guthrie BL, Hughes JP, Baseman JG, Dombrowski JC. Evaluation of an Emergency Department and Hospital-Based Data Exchange to Improve HIV Care Engagement and Viral Suppression. Sex Transm Dis. 2020;47:535-540. PMID: 32404856.
15. 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. PMCID: PMC5388943.
16. Sweeney P, Hoyte T, Mulatu MS, Bickham J, Brantley AD, Hicks C, McGoy SL, Morrison M, Rhodes A, Yerkes L, Burgess S, Fridge J, Wendell D. Implementing a Data to Care Strategy to Improve Health Outcomes for People With HIV: A Report From the Care and Prevention in the United States
17. Mokotoff ED, Green Ruth K, Benbow N, Sweeney P, Nelson Sapiano T, McNaghten AD. Data to Care: Lessons Learned From Delivering Technical Assistance to 20 Health Departments. J Acquir Immune Defic Syndr. 2019;82 Suppl 1:S74-S9. PMID: 31425400.
18. Hart-Malloy R, Rajulu DT, Johnson MC, Shrestha T, Spencer EC, Anderson BJ, Tesoriero JM. Cross Jurisdictional Data to Care: Lessons Learned in New York State and Florida. J Acquir Immune Defic Syndr. 2019;82 Suppl 1:S42-S6. PMID: 31425394
19. Tobias CR, Cunningham W, Cabral HD, Cunningham CO, Eldred L, Naar-King S, et al. Living with HIV but without medical care: barriers to engagement. *AIDS Patient Care STDs.* 2007;21:426–34
20. Bulsara SM, Wainberg ML, Newton-John TRO. Predictors of Adult Retention in HIV Care: A Systematic Review. AIDS Behav. 2018;22(3):752-64. PMCID: PMC5476508
21. Rajabiun, Serena et al. “The Influence of Housing Status on the HIV Continuum of Care: Results From a Multisite Study of Patient Navigation Models to Build a Medical Home for People Living With HIV Experiencing Homelessness.” *American journal of public health* vol. 108,S7 (2018): S539-S545.
22. Public Health – Seattle & King County. Social and Economic Inequalities and COVID-19 Outcomes [cited 2022 Mar 25]. Available at: https://kingcounty.gov/depts/health/covid-19/data/inequities.aspx
23. Mugavero, Michael J et al. “Early retention in HIV care and viral load suppression: implications for a test and treat approach to HIV prevention.” *Journal of acquired immune deficiency syndromes (1999)* vol. 59,1 (2012): 86-93.
24. Hall HI, Tang T, Westfall AO, Mugavero MJ. HIV care visits and time to viral suppression, 19 U.S. jurisdictions, and implications for treatment, prevention and the national HIV/AIDS strategy. PLoS One. 2013 Dec 31;8(12):e84318. doi: 10.1371/journal.pone.0084318. PMID: 24391937; PMCID: PMC3877252
25. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods.* 2009;14(4):323–348.
26. Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison with Logistic Regression. Ann Behav Med. 2003;26(3):172-81. PubMed PMID: 14644693.
27. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning: with applications in R. New York :Springer

28. Chambers LC, Manhart LE, Katz DA, Golden MR, Barbee LA, Dombrowski JC. Comparison of Algorithms to Triage Patients to Express Care in a Sexually Transmitted Disease Clinic. Sex Transm Dis. 2018;45(10):696-702. PMCID: PMC6133713.
29. Podgorelec V, Kokol P, Stiglic B, Rozman L. Decision trees: An overview and their use in medicine. *Journal of Medical Systems.* 2002;26(5):445–63.
30. Dombrowski JC, Bove J, Roscoe JC, Harvill J, Firth CL, Khormooji S, Carr J, Choi P, Smith C, Schafer SD, Golden MR; Northwest Health Department Centers for AIDS Research (CFAR) Consortium. "Out of Care" HIV Case Investigations: A Collaborative Analysis Across 6 States in the Northwest US. J Acquir Immune Defic Syndr. 2017 Feb 1;74 Suppl 2(Suppl 2):S81-S87. doi: 10.1097/QAI.0000000000001237. PMID: 28079717; PMCID: PMC5234689.
31. Xia Q, Braunstein SL, Wiewel EW, et al. Persons Living with HIV in the United States: Fewer Than We Thought. *J Acquir Immune Defic Syndr.* 2016;72:552–7.
32. Dombrowski JC, Buskin SE, Bennett A, et al. Use of multiple data sources and individual case investigation to refine surveillance-based estimates of the HIV care continuum. *J Acquir Immune Defic Syndr.* 2014;67:323–30.
33. Dombrowski JC, Kent JB, Buskin SE, et al. Population-based metrics for the timing of HIV diagnosis, engagement in HIV care, and virologic suppression. *AIDS.* 2012;26:77–86.