### Title

Use of molecular HIV surveillance data and predictive modeling to prioritize persons for transmission-reduction interventions.




### Journal

AIDS, 34(3)

### ISSN

0269-9370



### Authors

Xia, Qiang
Wertheim, Joel O
Braunstein, Sarah L
et al.



### Publication Date

2020-03-01

### DOI

10.1097/qad.0000000000002452

Peer reviewed

HHS Public Access

Author manuscript

*AIDS*. Author manuscript; available in PMC 2021 November 11.

Published in final edited form as:

*AIDS*. 2020 March 01; 34(3): 459–467. doi:10.1097/QAD.0000000000002452.

# Use of molecular HIV surveillance data and predictive modeling to prioritize persons for transmission-reduction interventions

**Qiang Xia**[1], **Joel O. Wertheim**[2], **Sarah L. Braunstein**[1], **Kavita Misra**[1], **Chi-Chi Udeagu**[1], **Lucia V. Torian**[1]

[1]HIV Epidemiology and Field Services Program, Bureau of HIV Prevention and Control, New York City Department of Health and Mental Hygiene, Queens, NY

[2]Department of Medicine, University of California San Diego, San Diego, CA

## Abstract

**Background:** To develop a predictive model to prioritize persons with a transmissible HIV viral load for transmission-reduction interventions.

**Methods:** New York City (NYC) HIV molecular surveillance data from 2010–2013 were used to build a model to predict the probability that the partial *pol* gene of the virus of a person with a transmissible HIV viral load (>1,500 copies/mL) would be genetically similar to that of a person with a new HIV infection (diagnosis at stage 0 or 1 according to the revised Centers for Disease Control and Prevention classification system). Data from 2013–2016 were then used to validate the model and compare it with five other selection strategies that can be used to prioritize persons for transmission-reduction interventions.

**Results:** A total of 10,609 persons living with HIV (PLWH) were included in the development dataset, and 8,257 were included in the validation dataset. Among the six selection strategies, the predictive model had the highest area under the receiver operating characteristic curve (AUC) (0.86, 95% confidence interval [CI]: 0.84, 0.88), followed by the "Young men who have sex with men (MSM)" (0.79, 95% CI: 0.77, 0.82), "MSM with high viral loads" (0.74, 95% CI: 0.72, 0.76), "Random sample of MSM" (0.73, 95% CI: 0.71, 0.76), "Persons with high viral loads" (0.56, 95% CI: 0.54, 0.59), and "Random sample" (0.50, 95% CI: 0.48, 0.53) strategies.

**Conclusions:** Jurisdictions should consider applying predictive modeling to prioritize persons with a transmissible viral load for transmission-reduction interventions and to evaluate its feasibility and effectiveness.

### Keywords

HIV; molecular epidemiology; surveillance; transmission; predictive modeling

In the first two decades of the HIV/AIDS epidemic in the United States, HIV prevention programs primarily focused their efforts on HIV-negative persons at high risk for infection.[1] Focusing only on persons who are HIV-negative undermines the effectiveness of HIV

Correspondence to: Qiang Xia, MD, MPH, HIV Epidemiology and Field Services Program, Bureau of HIV, New York City Department of Health and Mental Hygiene, 42-09 28th St., Queens, NY 11101; qxia@health.nyc.gov.

prevention, as it overlooks the population that is the source of onward transmission—HIV-positive individuals. To maximize reductions in HIV transmission, the Centers for Disease Control and Prevention (CDC) in 2003 recommended incorporating HIV prevention services into the medical care of persons living with HIV (PLWH).[2]

Lower plasma HIV viral load is associated with lower risk of HIV transmission.[3, 4] In 2011, the landmark HIV Prevention Trials Network (HPTN) 052 study among sero-discordant couples, with the majority (97%) being heterosexual, reported that early initiation of antiretroviral treatment (ART) by the HIV-positive partner reduced the risk of transmission to the HIV-negative partner by 96% compared to delayed treatment.[5] Two follow-up studies, the PARTNER (Partners of People on ART—A New Evaluation of the Risks) study among heterosexuals and men who have sex with men (MSM) and the Opposites Attract study among MSM, have found zero transmissions between sero-discordant couples when the HIV-positive partner was on treatment and had an undetectable viral load.[6, 7]

The findings from these studies support the strategy of treatment as prevention, i.e., that treating PLWH with ART to prevent HIV transmission be included as a key component of HIV prevention programs. The Panel on Antiretroviral Guidelines for Adults and Adolescents now recommends immediate initiation of ART for all people living with HIV, regardless of CD4 count.[8] Despite these recommendations, some patients do not initiate ART due to a host of individual- and structural-level factors, whereas others may take ART but are unable to achieve an undetectable viral load due to non-adherence or drug resistance, putting them at risk of transmitting HIV to their negative partners.[9–11]

One approach to reducing HIV transmission among PLWH, is to identify persons with a transmissible viral load and assist them to achieve viral suppression. This approach is constrained by limited resources, because the number of PLWH with a transmissible viral load at any given time is usually larger than a state or local HIV program can manage. Therefore, a prioritization strategy is needed to identify those at the highest risk of transmitting HIV. Programs have already preferentially selected some sub-populations for intervention, including MSM, Black and Latino people, and persons with high HIV viral loads, including those with acute HIV infection, co-infected with sexually transmitted infections (STIs), or belonging to a “recent and rapid” transmission cluster,[3, 12–14] but there is no systematic way to select individuals for intervention while simultaneously considering multiple factors, e.g., race/ethnicity, transmission risk, age, and viral load, in order to improve the specificity of the targeting strategy and the effectiveness of the intervention.

In the United States, CDC supports local jurisdictions to conduct Molecular HIV Surveillance (MHS), which collects, reports, and analyzes HIV genetic sequences generated during HIV drug resistance testing. It has been used as a tool to identify and respond to PLWH who have one or more viral genetic connections within networks containing recent HIV diagnoses.[15] The aim of this analysis is to use MHS data and predictive modeling to demonstrate a method that can be used to prioritize PLWH for intervention to reduce HIV transmission.

# METHODS

## Data source

The data source was the New York City (NYC) HIV surveillance registry. AIDS diagnoses have been reportable in New York State since 1981, and HIV diagnoses have been reportable since 2000. All CD4 counts, viral loads, and nucleotide sequences obtained for genotypic analyses have been reported to the registry since June 1, 2005. As of December 31, 2017, the registry contained a cumulative total of more than 240,000 cases (both living and deceased) and more than 10 million laboratory tests. In 2017, 2,157 people were diagnosed with HIV in NYC and there were about 90,500 PLWH, of whom 7% did not know their HIV-positive status.[16]

## Analysis population

Separating data into development and validation datasets is a common initial step when building a predictive model.[17] For this analysis, PLWH who had HIV sequence data in the NYC registry and were 13 years of age or older and living in NYC at the end of 2010 with a transmissible viral load, defined as >1,500 copies/mL, were included in the development dataset; PLWH who had HIV sequence data and were 13 years of age or older and living in NYC at the end of 2013 with a transmissible viral load were included in the validation dataset.[18] Patients who met the following criteria were included in both development and validation datasets: 1) 13 years of age or older by December 31, 2010, 2) diagnosed with HIV by December 31, 2010, 3) alive by December 31, 2013, and 4) viral load >1,500 copies/mL at both times: the end of 2010 and the end of 2013.

## Definition of new HIV infection

CDC classifies HIV diagnoses by stage based on patient's CD4 count at diagnosis and AIDS-defining opportunistic illness. Early infection, defined as a documented negative HIV test within 6 months prior to diagnosis, is classified as stage 0, regardless of CD4 count; CD4 count ≥500 cells/$mm^3$ is classified as stage 1; CD4 count between 200 and 499 cells/$mm^3$ is classified as stage 2; and CD4 count <200 cells/$mm^3$ or an AIDS-defining opportunistic illness (e.g., Kaposi's sarcoma, pneumocystis pneumonia, and tuberculosis) regardless of CD4 count, is classified as stage 3.[19, 20] Using patient's stage information, we defined an individual to have a new HIV infection if he/she acquired HIV through a non-perinatal route and was diagnosed with a stage 0 or 1.

## Sequence analysis

To determine whether the partial *pol* sequence of a PLWH was genetically similar to that of a person with a new HIV infection, we used a CDC funded computational tool, Secure HIV-TRACE (HIV TRAnsmission Cluster Engine), following a procedure described previously.[14, 21–23]

When we ran HIV-TRACE on the development dataset to determine whether the partial *pol* sequence of a PLWH at the end of 2010 was linked, i.e., genetically similar, to that of at least one person with a new HIV infection in NYC in 2011–2013, we included the last sequence from each PLWH at the end of 2010 and the first sequence from each HIV

case diagnosed in 2011–2013. First, all sequences were aligned to the HXB2 reference sequence (coordinates: 2253–3869) using an extension of the Smith-Waterman algorithm.[24] Next, HIV-TRACE calculated the pairwise Tamura-Nei 93 (TN93) genetic distance among all sequences, using an ambiguity fraction of 0.015 (i.e., genetic distance between ambiguous nucleotides were resolved only when the sequence contained ≤1.5% ambiguous nucleotides).[25] A viral genetic distance ≤0.015 substitution/site between a PLWH and a new infection was considered evidence of similarity, i.e., a link.

When we ran HIV-TRACE on the validation dataset to determine whether the viral sequence of a PLWH at the end of 2013 was genetically linked to that of at least one person with a new HIV infection in NYC in 2014–2016, the same procedure was followed but with different sequence data—the last sequence from each PLWH at the end of 2013 and the first sequence from each HIV case diagnosed in 2014–2016.

## Outcome variable

In the development dataset, we included an outcome variable indicating whether a patient's viral sequence was genetically linked to that of at least one new HIV infection diagnosed in NYC in the next three calendar years, i.e., 2011–2013; in the validation dataset, we included an outcome variable indicating whether a patient's virus was genetically linked to that of at least one new infection in NYC in the next three calendar years, i.e., 2014–2016.

## Model development and validation

We developed our predictive model following the guidelines for Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD),[26] and using the multivariate adaptive regression splines (MARS) algorithm in the Classification And Regression Training (CARET) package in R.[27] The candidate variables included sex, transgender status, race/ethnicity, age at diagnosis, current age, transmission category, year of diagnosis, ever diagnosed with AIDS, nadir CD4 count in the last three years, the highest $\log_{10}$ viral load in the last three years, and the $\log_{10}$ last viral load. The final model from the development dataset included the following variables: year of diagnosis, current age, ever diagnosed with AIDS, age at diagnosis, the highest $\log_{10}$ viral load in the last three years, and nadir CD4 count in the last three years.

The predictive model was tested with the validation dataset and compared with five other selection strategies that can be used to select PLWH with a transmissible viral load for transmission-reduction interventions: 1) random sample, 2) persons with high viral loads (sorting PLWHs' last viral load in descending order and selecting patients with the highest viral loads), 3) random sample of MSM, 4) MSM with high viral loads (sorting MSMs' last viral load in descending order and selecting patients with the highest viral loads), and 5) young MSM (sorting MSM by age in ascending order and selecting the youngest patients).

In addition to the area under the receiver operating characteristic curve (AUC) to assess the performance of these six selection strategies, we introduced a new measure, the viral genetic linkage rate, which was defined as the proportion of PLWH whose partial *pol* sequences were genetically linked to that of at least one new HIV infection diagnosed in the next three calendar years. The reason for introducing this new measure is that the

performance of each selection strategy also depends on the number of PLWH selected for the transmission-reduction interventions. Using the difference-in-difference analysis, we compared the viral genetic linkage rates by arbitrarily assuming that 250, 500, 750, and 1,000 PLWH, respectively, had been selected for the interventions.

### Sensitivity analysis

To assess the sensitivity of our findings, we repeated the above model-building process and analysis by changing the genetic distance threshold from 0.015 substitution/site to 0.005 substitution/site and limiting new diagnoses to acute HIV infections only.

## RESULTS

Of 10,609 PLWH living with a documented transmissible viral load in NYC at the end of 2010 and included in development dataset, two-thirds (66.6%) were men and one-third (33.4%) were women; over half (52.2%) were Black and one-third (35.8%) were Hispanic; 5.1% had viruses genetically linked to that of at least one new HIV infection. Of 8,257 PLWH living with a documented transmissible viral load in NYC at the end 2013 included in the validation dataset, the proportions were similar to those in the development dataset (Table 1).

Among the six selection strategies, the predictive model has the highest AUC (0.86, 95% confidence interval [CI]: 0.84, 0.88), followed by the “Young MSM” (0.79, 95% CI: 0.77, 0.82), “MSM with high viral load” (0.74, 95% CI: 0.72, 0.76), “Random sample of MSM” (0.73, 95% CI: 0.71, 0.76), “Persons with high viral load” (0.56, 95% CI: 0.54, 0.59), and “Random sample” (0.50, 95% CI: 0.48, 0.53) strategies (Figure 1).

Assuming that 500 PLWH with a transmissible viral load could have been selected for a transmission-reduction intervention, there would be striking differences in the characteristics of these PLWH selected by each selection strategy (Table 2). For example, the “Random sample” and “Persons with high viral loads” strategies selected 175 (35.0%) and 158 (31.6%) women, respectively, and the predictive model selected only 10 (2.0%) women. By definition, no women were selected by the three strategies that focused only on MSM.

In terms of age, the proportion of PLWH 45 years or older selected for intervention was 49.6% by the “Random sample,” 45.2% by the “Persons with high viral loads,” 32.6% by the “Random sample of MSM,” 32.4% by the “MSM with high viral loads,” 0% by the “Young MSM” strategy, and 0.2% by the predictive model.

Among 500 PLWH at the end of 2013 selected by the “Random sample” strategy, 17 of them were linked to at least one new HIV infection diagnosed in NYC in 2014–2016, with a genetic linkage rate of 3.4% (17/500). The genetic linkage rates were 5.8%, 7.6%, 12.0%, 23.2%, and 27.4%, respectively, for the “Persons with high viral loads,” “Random sample of MSM,” “MSM with high viral loads,” and “Young MSM” strategies, and the predictive model. The predictive model had the highest genetic linkage rate and was 8.06 (95% CI: 4.95, 13.13) times higher than the “Random sample” strategy.

Figure 2 shows the genetic linkage rate of each strategy by number of PLWH selected for transmission-reduction interventions. When the number of PLWH selected for interventions increased from 250 to 1,000, no changes in the genetic linkage rate were observed for the four strategies with the lowest rates, i.e., the "Random sample," "Persons with high viral loads," "Random sample of MSM," and "MSM with high viral loads" strategies. The genetic linkage rate for the top two strategies decreased, and the gap between them narrowed, as the number of PLWH selected for intervention increased. When 250 PLWH were selected for transmission-reduction interventions, the genetic linkage rates were 32.4% and 27.6%, respectively, for the predictive model and the "Young MSM" strategy, and the gap between them was 4.8 percentage points (95% CI: −3.5, 13.0); when 1,000 PLWH were selected, the genetic linkage rates were 22.2% and 20.0%, respectively, and the gap between them was 2.2 percentage points (95% CI: −1.5, 5.8), with a difference-in-difference of 2.6 percentage points (95% CI: −5.6, 10.8).

As expected, the sensitivity analysis results showed that the genetic linkage rate decreased as the genetic distance threshold decreased from 0.015 to 0.005 substitution/site and genetic linkages were limited to acute HIV infections (Table 3). However, the rate ratio stayed relatively stable, and the predictive model always had the highest ratio.

## DISCUSSION

The CDC recommends that jurisdictions include both HIV-negative persons and PLWH in interventions to reduce HIV transmission.[2] Since PLWH with an undetectable viral load cannot transmit HIV to their sexual partners,[6, 28] it would be more effective to focus on PLWH with a transmissible viral load. The majority of PLWH in the United States receive regular HIV care, including viral load monitoring.[29, 30] Therefore, it is not difficult to identify PLWH with a transmissible viral load in state and local surveillance systems. The challenge is how to prioritize them effectively. Using HIV sequence data, we developed a predictive model to select PLWH with a transmissible viral load for transmission-reduction interventions and have shown that the model performs better than all other selection strategies included in our analysis. The predictive model should perform even better when the program has more limited resources and must select fewer PLWH for transmission-reduction interventions.

Besides its better performance, the predictive model has a number of advantages. First, it can systematically select PLWH for interventions by considering multiple factors simultaneously, while other selection strategies focus on only one or two factors and may not prioritize individuals at the greatest risk for transmission.

Second, unlike other selection strategies, the predictive model does not limit PLWH with certain characteristics for intervention. For example, women and heterosexual men would not be selected by a selection strategy that focuses only on MSM, but they could be selected by the predictive model if other factors put them at a higher risk of transmitting HIV, such as young age, high viral loads, or history of drug use.[31, 32]

Third, HIV sequence data are used to build the predictive model, but after the predictive model is built, HIV sequence data are not needed for the model to identify PLWH for transmission-reduction interventions, although additional sequence data can be used to refine the model. At the end of 2013, there were 10,128 PLWH with a transmissible viral load in NYC, of whom 8,257 (81.5%) had HIV sequence data and 1871 (18.5%) did not. Applying the predictive model to select 500 PLWH for intervention, 375 (75.0%) would be selected from those with sequence data and 125 (25.0%) from those without. The higher probability of being selected from those without sequence data (125/1,871 = 6.7% vs. 375/8,257 = 4.5%) suggests that some factors, e.g., gender, age, race/ethnicity, and CD4 count at diagnosis,[33] that make PLWH less likely to be genotyped may also put them at a higher risk of transmitting HIV.

The predictive model also has limitations. First, our definition of new infection is based on the CD4 count at the time of diagnosis and there could be some misclassifications—new infections with a CD4 count at diagnosis lower than 500 copies/mL (false negatives) and established infections with a CD4 count at diagnosis higher than 500 copies/mL (false positives). To predict a new transmission, minimizing false positives is more important than minimizing false negatives. We were able to further minimize false positives by conducting a sensitivity analysis among people with an acute HIV infection and reached the same conclusion—the predictive model performed better than the other five selection strategies.

Second, we used genetic links to evaluate each selection strategy, and genetic links alone, particularly at this liberal distance threshold, cannot be used to represent direct transmissions. Although we are unable to confirm direct transmissions at the individual level, it is reasonable to conclude at the population level that a subpopulation with more genetic links represents more transmissions. Selecting such a subpopulation for intervention would be an effective way to prevent onward transmission.

Third, to keep the model simple, PLWH whose viruses were genetically linked to that of at least one new infection in the next three calendar years were all treated equally in the model building, despite the fact that some were linked to more than one new infection. Building the model this way should not affect our conclusion that the predictive model performs better than the other five selection strategies, because, 1) it was common (41.7%) for PLWH to link to more than one new infection, 2) it was also common (28.7%) for new infections to link to more than one PLWH, and 3) the same method was applied to all six selection strategies.

Fourth, HIV-TRACE requires a minimum of 500 nucleotides to calculate the genetic distance between two sequences. The sequences included in our analysis have a length between 669 and 1,600 nucleotides, with a median length of 1,212 (interquartile range [IQR]: 1,212, 1,497). Different lengths of sequences may have an impact on the distance calculation, but they should have little impact on our conclusion, because the same genetic distance calculation method was applied to all six selection strategies.

Fifth, the current predictive model can only be used to prioritize viremic PLWH who are in care, i.e., it cannot be used to prioritize out-of-care patients for transmission-reduction interventions because their viral load data are not available. However, since out-of-care

patients are eligible for re-engagement services by health departments that do “Data to Care” work, they have other opportunities to be selected for HIV interventions.[34]

Finally, the analysis was conducted in NYC, where HIV surveillance data may be more complete than other jurisdictions, including the high proportion of diagnosed infections, the high percentages of diagnosed PLWH entered in the registry, PLWH with sequence data, and PLWH with complete information on the variables included in the model. It is also possible that PLWH in NYC have different relative transmission rates by the categories we examined than those elsewhere. The purpose of this analysis is to demonstrate that predictive modeling using case and molecular surveillance data can be used to prioritize persons with a transmissible HIV viral load for transmission-reduction interventions. Jurisdictions should evaluate the quality and completeness of their data before using our method to build their own model for transmission-reduction interventions.

Using case and molecular surveillance data, we developed a predictive model to prioritize PLWH with a transmissible viral load for interventions to reduce onward transmission and found the model to perform better than other selection strategies. We suggest investigating applying this method in the real world to evaluate its feasibility and effectiveness. Before implementation, jurisdictions also need to consider the ethical implications of selecting persons with specific, readily identifiable characteristics, e.g., race/ethnicity, sexual orientation, and transmission risk, for targeted interventions, and evaluate its latent consequences, possibly in consultation with a community advisory board that is sensitive to community concerns about stigma and discrimination against PLWH. Both external and internalized stigma may drive people away from the services that they need.

## Acknowledgments:

The authors would like to thank Drs. Kent Sepkowitz, Oni Blackstock, Demetre Daskalakis, and James Hadler for their review and comments on this paper.

**Funding:**

This project was supported in part by a Cooperative Agreement with the Centers for Disease Control and Prevention (PS18-1802). JOW was supported in part by an NIH-NIAID K01 Career Development Award (K01AI110181) and an NIH-NIAID R01 (AI135992).

## References

1. Janssen RS, Holtgrave DR, Valdiserri RO, Shepherd M, Gayle HD, De Cock KM. The serostatus approach to fighting the HIV epidemic: prevention strategies for infected individuals. Am J Public Health 2001; 91(7):1019–1024. [PubMed: 11441723]
2. Centers for Disease Control and Prevention. Incorporating HIV prevention into the medical care of persons living with HIV: recommendations of CDC, the Health Resources and Services Administration, the National Institutes of Health, and the HIV Medicine Association of the Infectious Diseases Society of America. MMWR Morb Mortal Wkly Rep 2003; 52(RR-12):1–24. [PubMed: 12549898]
3. Attia S, Egger M, Muller M, Zwahlen M, Low N. Sexual transmission of HIV according to viral load and antiretroviral therapy: systematic review and meta-analysis. AIDS 2009; 23(11):1397–1404. [PubMed: 19381076]

4. Quinn TC, Wawer MJ, Sewankambo N, Serwadda D, Li C, Wabwire-Mangen F, et al. Viral load and heterosexual transmission of human immunodeficiency virus type 1. Rakai Project Study Group. N Engl J Med 2000; 342(13):921–929. [PubMed: 10738050]
5. Cohen MS, Chen YQ, McCauley M, Gamble T, Hosseinipour MC, Kumarasamy N, et al. Prevention of HIV-1 infection with early antiretroviral therapy. N Engl J Med 2011; 365(6):493–505. [PubMed: 21767103]
6. Rodger AJ, Cambiano V, Bruun T, Vernazza P, Collins S, van Lunzen J, et al. Sexual activity without condoms and risk of HIV transmission in serodifferent couples when the HIV-positive partner is using suppressive antiretroviral therapy. JAMA 2016; 316(2):171–181. [PubMed: 27404185]
7. Bavinton B, Grinsztejn B, Phanuphak N, Jin F, Zablotska I, Prestage G, et al. HIV treatment prevents HIV transmission in male serodiscordant couples in Australia, Thailand and Brazil. 9th IAS Conference on HIV Science (IAS 2017); Paris, France: 23–26 July, 2017.
8. The Panel on Antiretroviral Guidelines for Adults and Adolescents (2017). Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. The U.S. Department of Health and Human Services. Washington, DC.
9. 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 2014; 67(3):323–330. [PubMed: 25140904]
10. Xia Q, Lazar R, Bernard MA, McNamee P, Daskalakis DC, Torian LV, et al. New York City achieves the UNAIDS 90-90-90 targets for HIV-infected whites but not Latinos/Hispanics and blacks. J Acquir Immune Defic Syndr 2016; 73(3):e59–e61. [PubMed: 27763998]
11. Xia Q, Sun X, Wiewel EW, Torian LV. HIV prevalence and the prevalence of unsuppressed HIV in New York City, 2010–2014. J Acquir Immune Defic Syndr 2017; 75(2):143–147. [PubMed: 28207429]
12. Torian LV, Forgione LA. Young MSM at the leading edge of HIV in New York City: back to the future? J Acquir Immune Defic Syndr 2015; 68(4):e63–68. [PubMed: 25559598]
13. Pilcher CD, Tien HC, Eron JJ Jr., Vernazza PL, Leu SY, Stewart PW, et al. Brief but efficient: acute HIV infection and the sexual transmission of HIV. J Infect Dis 2004; 189(10):1785–1792. [PubMed: 15122514]
14. Oster AM, France AM, Panneer N, Banez Ocfemia MC, Campbell E, Dasgupta S, et al. Identifying clusters of recent and rapid HIV transmission through analysis of molecular surveillance data. J Acquir Immune Defic Syndr 2018; 79(5):543–550. [PubMed: 30222659]
15. Oster AM, Wertheim JO, Hernandez AL, Ocfemia MC, Saduvala N, Hall HI. Using molecular HIV surveillance data to understand transmission between subpopulations in the United States. J Acquir Immune Defic Syndr 2015; 70(4):444–451. [PubMed: 26302431]
16. HIV Epidemiology and Field Services Program (2018). HIV Surveillance Annual Report, 2017. New York City Department of Health and Mental Hygiene. New York, NY.
17. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: Springer; 2013.
18. Marks G, Gardner LI, Rose CE, Zinski A, Moore RD, Holman S, et al. Time above 1500 copies: a viral load measure for assessing transmission risk of HIV-positive patients in care. Aids 2015; 29(8):947–954. [PubMed: 25768835]
19. Selik RM, Mokotoff ED, Branson B, Owen SM, Whitmore S, Hall HI. Revised surveillance case definition for HIV infection--United States, 2014. MMWR Recomm Rep 2014; 63(RR-03):1–10.
20. Xia Q, Braunstein SL, Torian LV. Using the revised Centers for Disease Control and Prevention staging system to classify persons living with human immunodeficiency virus in New York City, 2011–2015. Sex Transm Dis 2017; 44(11):653–655. [PubMed: 28876316]
21. Kosakovsky Pond SL, Weaver S, Leigh Brown AJ, Wertheim JO. HIV-TRACE (Transmission Cluster Engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens. Mol Biol Evol 2018.
22. Wertheim JO, Leigh Brown AJ, Hepler NL, Mehta SR, Richman DD, Smith DM, et al. The global transmission network of HIV-1. J Infect Dis 2014; 209(2):304–313. [PubMed: 24151309]

23. Wertheim JO, Kosakovsky Pond SL, Forgione LA, Mehta SR, Murrell B, Shah S, et al. Social and genetic networks of HIV-1 transmission in New York City. PLoS Pathog 2017; 13(1):e1006000. [PubMed: 28068413]
24. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981; 147(1):195–197. [PubMed: 7265238]
25. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 1993; 10(3):512–526. [PubMed: 8336541]
26. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Ann Intern Med 2015; 162(10):735–736.
27. Kuhn M Building predictive models in R using the caret package. J Stat Softw 2008; 28(5):1–26. [PubMed: 27774042]
28. Vernazza P, Hirschel B, Bernasconi E, Flepp M. Les personnes séropositives ne souffrant d'aucune autre MST et suivant un traitement antirétroviral efficace ne transmettent pas le VIH par voie sexuelle. Bull Med Suissess 2008; 89(5):165–169.
29. Centers for Disease Control and Prevention. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data -- United States and 6 U.S. dependent areas -- 2015. HIV Surveillance Supplemental Report 2017; 22(2):1–63.
30. Xia Q, Kersanske LS, Wiewel EW, Braunstein SL, Shepard CW, Torian LV. Proportions of patients with HIV retained in care and virally suppressed in New York City and the United States: higher than we thought. J Acquir Immune Defic Syndr 2015; 68(3):351–358. [PubMed: 25501613]
31. Donnell D, Baeten JM, Kiarie J, Thomas KK, Stevens W, Cohen CR, et al. Heterosexual HIV-1 transmission after initiation of antiretroviral therapy: a prospective cohort analysis. Lancet 2010; 375(9731):2092–2098. [PubMed: 20537376]
32. Fideli US, Allen SA, Musonda R, Trask S, Hahn BH, Weiss H, et al. Virologic and immunologic determinants of heterosexual transmission of human immunodeficiency virus type 1 in Africa. AIDS Res Hum Retroviruses 2001; 17(10):901–910. [PubMed: 11461676]
33. Scott SN, Forgione LA, Torian LV. Racial disparities in baseline genotyping in the era of "ART for all" in New York City, 2006–2017 (Poster #1579). Conference on Retroviruses and Opportunistic Infections (CROI 2019); Seattle, WA: March 4–7, 2019.
34. Udeagu CC, Webster TR, Bocour A, Michel P, Shepard CW. Lost or just not following up: public health effort to re-engage HIV-infected persons lost to follow-up into HIV medical care. AIDS 2013; 27(14):2271–2279. [PubMed: 23669157]

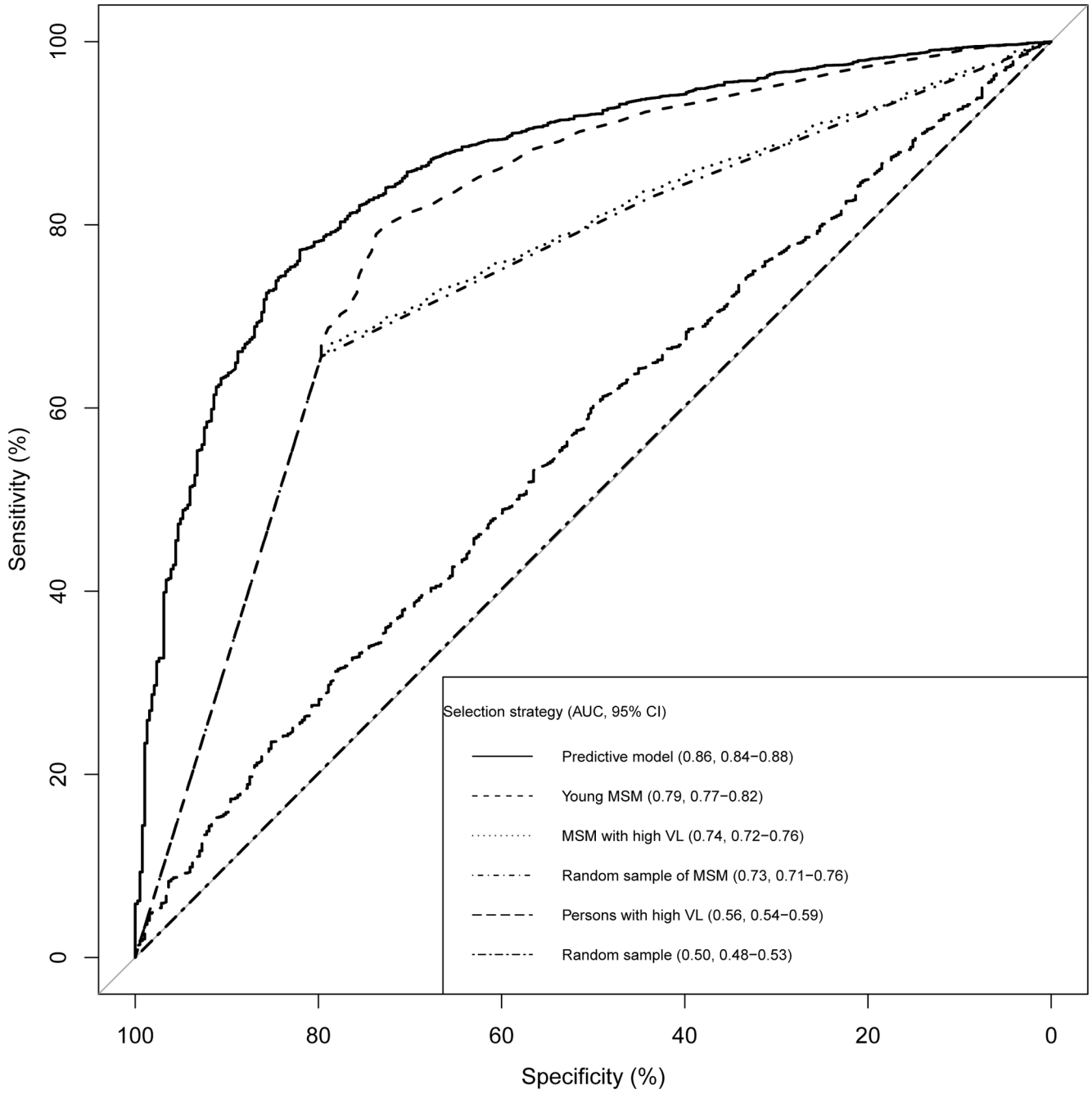


**Figure 1. Receiver operating characteristic curves and the area under the receiver operating characteristic curve (AUC) for the six selection strategies**

MSM, men who have sex with men; VL, viral load.

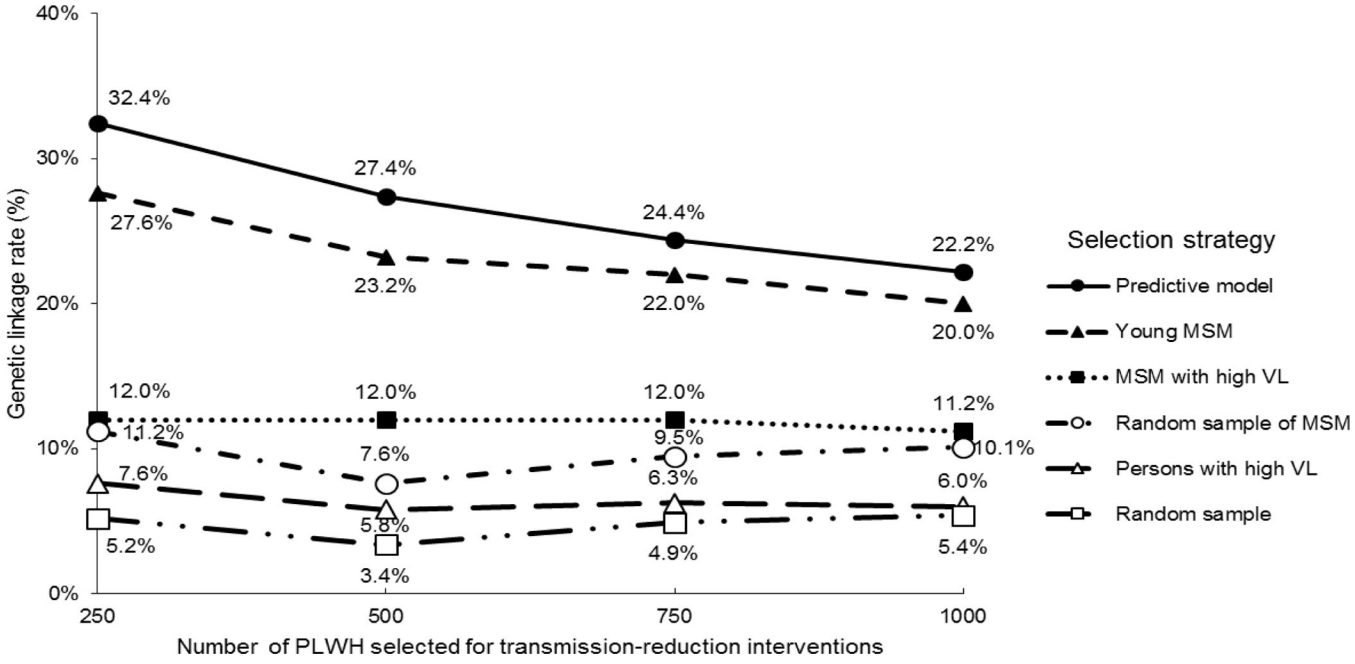


**Figure 2. Genetic linkage rate, by selection strategy and number of PLWH with a transmissible viral load (>1,500 copies/mL) selected for transmission-reduction interventions*†**

HIV, human immunodeficiency virus; MSM, men who have sex with men; PLWH, persons living with human immunodeficiency virus; VL, viral load.

*Genetic linkage rate was defined as the proportion of PLWH with a transmissible viral load (>1,500 copies/mL) at the end of 2013 whose viral sequence was genetically linked to that of at least one new HIV infection in 2014–2016.

†A new HIV infection was defined as a person diagnosed with HIV at stage 0 or 1 in New York City in 2014–2016. The stage of HIV infection was determined based on the revised Centers for Disease Control and Prevention (CDC) classification system.

*AIDS*. Author manuscript; available in PMC 2021 November 11.

**Table 1.**

Characteristics and genetic linkage in the development and validation datasets

| | Development dataset | | | | Validation dataset | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Linked to ≥1 new infection* | | | | Linked to ≥1 new infection* | |
| | N | Col % | n | Row % | N | Col % | n | Row % |
| Total | 10,609 | 100.0 | 538 | 5.1 | 8,257 | 100.0 | 384 | 4.7 |
| Sex | | | | | | | | |
| Male | 7,066 | 66.6 | 480 | 6.8 | 5,485 | 66.4 | 353 | 6.4 |
| Female | 3,543 | 33.4 | 58 | 1.6 | 2,772 | 33.6 | 31 | 1.1 |
| Race/ethnicity | | | | | | | | |
| Black | 5,540 | 52.2 | 218 | 3.9 | 4,555 | 55.2 | 165 | 3.6 |
| Hispanic | 3,803 | 35.8 | 199 | 5.2 | 2,800 | 33.9 | 133 | 4.8 |
| White | 1,099 | 10.4 | 107 | 9.7 | 746 | 9.0 | 61 | 8.2 |
| API | 118 | 1.1 | 11 | 9.3 | 106 | 1.3 | 19 | 17.9 |
| Other | 49 | 0.5 | 3 | 6.1 | 50 | 0.6 | 6 | 12.0 |
| Age | | | | | | | | |
| 13–24 | 931 | 8.8 | 150 | 16.1 | 802 | 9.7 | 104 | 13.0 |
| 25–44 | 4,628 | 43.6 | 329 | 7.1 | 3,378 | 40.9 | 234 | 6.9 |
| 45–64 | 4,850 | 45.7 | 59 | 1.2 | 3,827 | 46.3 | 46 | 1.2 |
| 65+ | 200 | 1.9 | 0 | 0.0 | 250 | 3.0 | 0 | 0.0 |
| Transmission risk | | | | | | | | |
| MSM | 3,228 | 30.4 | 389 | 12.1 | 2,715 | 32.9 | 296 | 10.9 |
| IDU | 2,010 | 18.9 | 17 | 0.8 | 1,283 | 15.5 | 9 | 0.7 |
| MSM-IDU | 454 | 4.3 | 15 | 3.3 | 298 | 3.6 | 10 | 3.4 |
| Heterosexual | 2,541 | 24.0 | 55 | 2.2 | 2,039 | 24.7 | 35 | 1.7 |
| Perinatal | 398 | 3.8 | 5 | 1.3 | 409 | 5.0 | 4 | 1.0 |
| Unknown | 1,978 | 18.6 | 57 | 2.9 | 1,513 | 18.3 | 30 | 2.0 |
| Year of diagnosis | | | | | | | | |
| Pre-1991 | 1,279 | 12.1 | 6 | 0.5 | 838 | 10.1 | 3 | 0.4 |
| 1991–1995 | 1,501 | 14.1 | 12 | 0.8 | 974 | 11.8 | 7 | 0.7 |
| 1996–2000 | 2,640 | 24.9 | 24 | 0.9 | 1,856 | 22.5 | 17 | 0.9 |

| | Development dataset | | | | Validation dataset | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Linked to ≥1 new infection* | | | | Linked to ≥1 new infection* | |
| | N | Col % | n | Row % | N | Col % | n | Row % |
| 2001–2005 | 2,400 | 22.6 | 67 | 2.8 | 1,761 | 21.3 | 30 | 1.7 |
| 2006–2010 | 2,789 | 26.3 | 429 | 15.4 | 1,646 | 19.9 | 114 | 6.9 |
| 2011–2013 | — | — | — | — | 1,182 | 14.3 | 213 | 18.0 |
| Nadir CD4 count in the last 3 years ( cells/mm$^3$) | | | | | | | | |
| 0–199 | 5,184 | 48.9 | 90 | 1.7 | 4,062 | 49.2 | 87 | 2.1 |
| 200–349 | 2,645 | 24.9 | 153 | 5.8 | 1,906 | 23.1 | 105 | 5.5 |
| 350–499 | 1,702 | 16.0 | 160 | 9.4 | 1,263 | 15.3 | 81 | 6.4 |
| 500+ | 1,049 | 9.9 | 132 | 12.6 | 987 | 12.0 | 106 | 10.7 |
| Unknown | 29 | 0.3 | 3 | 10.3 | 39 | 0.5 | 5 | 12.8 |
| Last viral load (copies/mL) | | | | | | | | |
| 1,500–9,999 | 3,528 | 33.3 | 161 | 4.6 | 2,422 | 29.3 | 81 | 3.3 |
| 10,000–99,999 | 4,939 | 46.6 | 273 | 5.5 | 4,125 | 50.0 | 206 | 5.0 |
| 100,000–999,999 | 2,016 | 19.0 | 100 | 5.0 | 1,585 | 19.2 | 85 | 5.4 |
| 1,000,000+ | 126 | 1.2 | 4 | 3.2 | 125 | 1.5 | 12 | 9.6 |

API, Asian/Pacific Islander; HIV, human immunodeficiency virus; IDU, injection drug users; MSM, men who have sex with men.

*A new HIV infection was defined as a person diagnosed with HIV at stage 0 or 1 in New York City in the next three calendar years (2011–2013 in the development dataset and 2014–2016 in the validation dataset). The stage of HIV infection was determined based on the revised Centers for Disease Control and Prevention (CDC) classification system.

*AIDS*. Author manuscript; available in PMC 2021 November 11.

**Table 2.**

Characteristics of persons with a transmissible HIV viral load (>1,500 copies/mL) in New York City at the end of 2013 who could have bee selected for transmission-reduction interventions, by selection strategy

| | Random sample | Persons with high VL | Random sample of MSM | MSM with high VL | Young MSM | Predictive model |
|---|---|---|---|---|---|---|
| Total | 500 | 500 | 500 | 500 | 500 | 500 |
| Sex | | | | | | |
| Men | 325 | 342 | 500 | 500 | 500 | 490 |
| Women | 175 | 158 | 0 | 0 | 0 | 10 |
| Race/ethnicity | | | | | | |
| Black | 286 | 268 | 235 | 212 | 257 | 196 |
| Hispanic | 165 | 168 | 163 | 170 | 191 | 200 |
| White | 40 | 54 | 87 | 103 | 31 | 73 |
| API | 5 | 8 | 13 | 10 | 13 | 19 |
| Other | 4 | 2 | 2 | 5 | 8 | 12 |
| Age | | | | | | |
| 13–24 | 47 | 49 | 63 | 54 | 347 | 244 |
| 25–44 | 205 | 225 | 274 | 284 | 153 | 255 |
| 45–64 | 228 | 216 | 157 | 154 | 0 | 1 |
| 65+ | 20 | 10 | 6 | 8 | 0 | 0 |
| Transmission risk | | | | | | |
| MSM | 159 | 174 | 447 | 451 | 480 | 415 |
| IDU | 86 | 85 | 0 | 0 | 0 | 5 |
| MSM-IDU | 16 | 13 | 53 | 49 | 20 | 15 |
| Heterosexual | 117 | 122 | 0 | 0 | 0 | 29 |
| Perinatal | 27 | 21 | 0 | 0 | 0 | 0 |
| Unknown | 95 | 85 | 0 | 0 | 0 | 36 |
| Year of diagnosis | | | | | | |
| Pre-1991 | 52 | 43 | 42 | 27 | 0 | 0 |
| 1991–1995 | 59 | 53 | 27 | 32 | 0 | 0 |
| 1996–2000 | 114 | 109 | 78 | 86 | 0 | 0 |
| 2001–2005 | 118 | 103 | 99 | 100 | 14 | 0 |

| | Random sample | Persons with high VL | Random sample of MSM | MSM with high VL | Young MSM | Predictive model |
|---|---|---|---|---|---|---|
| 2006–2010 | 92 | 96 | 132 | 125 | 194 | 9 |
| 2011–2013 | 65 | 96 | 122 | 130 | 292 | 491 |
| Last viral load in 2013 (copies/mL) | | | | | | |
| 1,500–9,999 | 165 | 0 | 143 | 0 | 101 | 96 |
| 10,000–99,999 | 245 | 0 | 248 | 0 | 288 | 296 |
| 100,000–999,999 | 86 | 375 | 98 | 451 | 103 | 92 |
| 1,000,000+ | 4 | 125 | 11 | 49 | 8 | 16 |
| Genetically linked to ≥1 new HIV infection in 2014–2016 * | | | | | | |
| No | 483 | 471 | 462 | 440 | 384 | 363 |
| Yes | 17 | 29 | 38 | 60 | 116 | 137 |
| Genetic linkage rate (%) † | 17/500 = 3.4% | 29/500 = 5.8% | 38/500 = 7.6% | 60/500 = 12.0% | 116/500 = 23.2% | 137/500 = 27.4% |
| Rate ratio | 1.00 | 1.71 | 2.24 | 3.53 | 6.82 | 8.06 |
| 95% CI | — | 0.95, 3.06 | 1.28, 3.91 | 2.09, 5.96 | 4.17, 11.18 | 4.95, 13.13 |

API, Asian/Pacific Islander; CI, confidence interval; HIV, human immunodeficiency virus; IDU, injection drug users; MSM, men who have sex with men; VL, viral load.

* A new HIV infection was defined as a person diagnosed with HIV at stage 0 or 1 in New York City in 2014–2016. The stage of HIV infection was determined based on the revised Centers for Disease Control and Prevention (CDC) classification system.

† Genetic linkage rate was defined as the proportion of PLWH with a transmittable viral load (>1,500 copies/mL) at the end of 2013 whose viral sequence was genetically linked to that of at least one new HIV infection in 2014–2016.

*AIDS*. Author manuscript; available in PMC 2021 November 11.

**Table 3.**

Sensitivity analysis results—number of PLWH with a transmissible viral load (>1,500 copies/mL) whose viral sequence was genetically linked to that of at least one new HIV infection or acute HIV infection in the next three calendar years, by selection strategy, genetic distance, and stage of new HIV infection [*,†]

| | Random sample | Persons with high VL | Random sample of MSM | MSM with high VL | Young MSM | Predictive model |
|---|---|---|---|---|---|---|
| Total number of persons selected for a transmission-reduction intervention (N) | 500 | 500 | 500 | 500 | 500 | 500 |
| Genetically linked to ≥1 new HIV infection within a 1.5% genetic distance [*] | | | | | | |
| n | 17 | 29 | 38 | 60 | 116 | 137 |
| Genetic linkage rate (%)[§] | 17/500 = 3.4% | 29/500 = 5.8% | 38/500 = 7.6% | 60/500 = 12.0% | 116/500 = 23.2% | 137/500 = 27.4% |
| Rate ratio | 1.00 | 1.71 | 2.24 | 3.53 | 6.82 | 8.06 |
| 95% CI | — | 0.95, 3.06 | 1.28, 3.91 | 2.09, 5.96 | 4.17, 11.18 | 4.95, 13.13 |
| Genetically linked to ≥1 new HIV infection within a 0.5% genetic distance [*] | | | | | | |
| n | 7 | 9 | 15 | 22 | 31 | 57 |
| Genetic linkage rate (%)[§] | 7/500 = 1.4% | 9/500 = 1.8% | 15/500 = 3.0% | 22/500 = 4.4% | 31/500 = 6.2% | 57/500 = 11.4% |
| Rate ratio | 1.00 | 1.29 | 2.14 | 3.14 | 4.43 | 8.14 |
| 95% CI | — | 0.48, 3.43 | 0.88, 5.21 | 1.36, 7.29 | 1.97, 9.96 | 3.75, 17.68 |
| Genetically linked to ≥1 acute HIV infection within a 1.5% genetic distance [†] | | | | | | |
| n | 6 | 15 | 17 | 31 | 43 | 58 |
| Genetic linkage rate (%)[§] | 6/500 = 1.2% | 15/500 = 3.0% | 17/500 = 3.4% | 31/500 = 6.2% | 43/500 = 8.6% | 58/500 = 11.6% |
| Rate ratio | 1.00 | 2.50 | 2.83 | 5.17 | 7.17 | 9.67 |
| 95% CI | — | 0.98, 6.39 | 1.13, 7.13 | 2.18, 12.27 | 3.08, 16.68 | 4.21, 22.20 |
| Genetically linked to ≥1 acute HIV infection within a 0.5% genetic distance [†] | | | | | | |
| n | 2 | 5 | 7 | 12 | 12 | 17 |
| Genetic linkage rate (%)[§] | 2/500 = 0.4% | 15/500 = 1.0% | 7/500 = 1.4% | 12/500 = 2.4% | 12/500 = 2.4% | 17/500 = 3.4% |
| Rate ratio | 1.00 | 2.50 | 3.50 | 6.00 | 6.00 | 8.50 |
| 95% CI | — | 0.49, 12.82 | 0.73, 16.76 | 1.35, 26.67 | 1.35, 26.67 | 1.97, 36.59 |

CI, confidence interval; HIV, human immunodeficiency virus; PLWH, persons living with HIV; VL, viral load.

[*] A new HIV infection was defined as a person diagnosed with HIV at stage 0 or 1 in New York City in 2014–2016. The stage of HIV infection was determined based on the revised Centers for Disease Control and Prevention (CDC) classification system.

[†] An acute HIV infection was defined as a person diagnosed with stage 0 HIV infection in New York City in 2014–2016. The stage of HIV infection was determined based on the revised CDC classification system.

[§] Genetic linkage rate was defined as the proportion of PLWH with a transmissible viral load (>1,500 copies/mL) at the end of 2013 whose viral sequence was genetically linked to that of at least one new HIV infection or acute HIV infection diagnosed in 2014–2016.