**Supplemental Figure 1.** Distribution of pairwise genetic distances within and between individuals in the internal validation population. Vertical lines indicate the 95th and 5th percentiles of intra-individual and inter-individual genetic distances, respectively, at 0.0193 and 0.0444. Calculations between individuals included only the first available sequence for each individual. A cutoff of 1.5% was chosen for cluster identification because (1) 1.5% included >90% of intra-individual sequences and <0.015% of inter-individual sequences in this dataset [Supplemental Figure 1], making it extremely unlikely that a pair of randomly selected, unrelated sequences would cluster; (2) 1.5% is the standard used by many in the field and is recommended by the CDC for cluster detection by health departments;1─4 and (3) a prior evaluation of within host evolution demonstrated that HIV sequences from mono-infected participants showed <1% divergence over time.5

****

**Supplemental Figure 2.** Development, selection, internal, and external validation of a predictive model for HIV transmission cluster growth in NC.****

**Supplemental Figure 3.** A. Visualization of member status of 1835 individuals in 352 established clusters in the internal validation population. Individuals are grouped by putative HIV transmission cluster membership. Dark blue circles indicate sequences sampled prior to baseline (baseline cluster members), light blue circles indicate sequences sampled following baseline from new diagnoses (new cluster members), and gray circles indicate sequences sampled following baseline from prior diagnoses (hidden cluster members). B. Distribution of baseline cluster size among growing and non-growing clusters in the internal validation population.

****

**Supplemental Figure 4.** Comparison of ROC-AUC curves for the full model and candidate reduced models, including predictors retained in ≥20%, ≥30%, ≥40%, ≥50%, and ≥70% of final models across 100 bootstrapped samples. ROC-AUC values shown are not adjuted for optimism.

****

**Supplemental Table 1.** Evaluation of candidate predictors via bivariable analyses with cluster growth, collinearity assessments, coding evaluation, and backward elimination.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Predictor**  | **Final Coding** | **Definition** | **Point of removal from consideration** | **In full model** | **In final model\*** |
| Cluster size | Continuous | Number of individuals in the cluster with sequences sampled prior to baseline. aOR represents a one-member increase in baseline cluster size | Retained in final model - Wald Chi-square p-value ≤ 0.10 in 93/100 bootstrapped samples | Yes | Yes |
| **Temporal** |  |  |  |  |  |
| Years since a diagnosis | Binomial | Number of years from the last HIV diagnosis among individuals in the cluster to baseline. Binomial: ≤1, >1. >1 is referent for aOR | Retained in final model - Wald Chi-square p-value ≤ 0.10 in 91/100 bootstrapped samples | Yes | Yes |
| Cluster age: diagnoses | Categorical | Number of years from the first HIV diagnosis among individuals in the cluster to baseline. Categorical: <4, 4-8, 8-12, ≥12 | Removed fifth in backward elimination - Wald Chi-square p-value ≤ 0.10 in 17/100 bootstrapped samples | Yes | No |
| Cluster age: sequences | Continuous | Number of years from the first sampled HIV sequence among individuals in the cluster to baseline  | Removed third in backward elimination - Wald Chi-square p-value ≤ 0.10 in 15/100 bootstrapped samples  | Yes | No |
| **Demographic** |  |  |  |  |  |
| Median age | Continuous | Median age in years of individuals in the cluster as of baseline. aOR represents a ten year increase in median age  | Retained in final model - Wald Chi-square p-value ≤ 0.10 in 73/100 bootstrapped samples | Yes | Yes |
| Percent PWID | Binomial | Percentage of individuals in the cluster by baseline who identified as PWID or MSM and PWID: Binomial: 0%, >0% | Removed sixth in backward elimination - Wald Chi-square p-value ≤ 0.10 in 21/100 bootstrapped samples | Yes | No |
| Percent MSM | Continuous | Percentage of individuals in the cluster by baseline who identified as MSM or MSM and PWID | Removed eighth in backward elimination - Wald Chi-square p-value ≤ 0.10 in 30/100 bootstrapped samples | Yes | No |
| Percent male | Continuous | Percentage of individuals in the cluster by baseline who identified as male | Removed in preliminary assessments due to Pearson correlation coefficient > 0.7 in collinearity assessments with percent MSM | No | No |
| Percent Black, non-Hispanic | Binomial | Percentage of individuals in the cluster by baseline who identified as Black, non-Hispanic: Binomial: <100%, 100% | Removed seventh in backward elimination - Wald Chi-square p-value ≤ 0.10 in 25/100 bootstrapped samples | Yes | No |
| Percent in dominant geographic region | Binomial | Percentage of individuals in the cluster by baseline who reported living in the dominant region for that cluster at the time of diagnosis. Binomial: <100%, 100% | Removed second in backward elimination - Wald Chi-square p-value ≤ 0.10 in 14/100 bootstrapped samples  | Yes | No |
| **Clinical** |  |  |  |  |  |
| Percent with HIV viremia | Categorical | Percentage of individuals in the cluster by baseline with detectable viral loads (≥1000 copies/mL) at the most recent care visit during the one year prior to baseline, or with no viral load available during that year. Categorical: 0%, 0 < - 25%, 25% < - 50%, ≥50.0%. 0% is referent for aORs | Retained in final model - Wald Chi-square p-value ≤ 0.10 in 81/100 bootstrapped samples | Yes | Yes |
| Percent acute | Binomial | Any cluster members diagnosed during the two years prior to baseline who were diagnosed during acute or recent HIV infection. NC STAT currently defines acute HIV infection (AHI) as a reactive fourth generation antigen/antibody combination test with a negative or indeterminate multispot assay and a detectable HIV RNA. Recent HIV infection (RHI) is treated as a subset of AHI and is defined by NC STAT as a repeatedly reactive HIV-1 antibody test with verification of a negative test in the past three months and/or symptoms compatible with AHI in the three months prior to diagnosis. Binomial: 0%, >0% | Removed fourth in backward elimination – Wald Chi-square p-value ≤ 0.16 in 66/100 bootstrapped samples | Yes | No |
| Median time to HIV care entry | Continuous | Median number of years from diagnosis to the first care visit (date of first CD4, viral load, or sequence) among individuals in the cluster by baseline. aOR represents a one year increase in time to care entry | Retained in final model - Wald Chi-square p-value ≤ 0.10 in 55/100 bootstrapped samples | Yes | Yes |
| Percent in HIV care  | Binomial | Percentage of individuals in the cluster by baseline who had any HIV care visits (CD4, VL, or sequence date) during the one year prior to baseline. Excluded those who died prior to baseline from calculations. Binomial: <100%, 100% | Removed first in backward elimination - Wald Chi-square p-value ≤ 0.10 in 10/100 bootstrapped samples | Yes | No |
| Percent with prior/ prevalent STI | N/A | Percentage of individuals in the cluster by baseline who reported a prior or prevalent sexually transmitted infection | Excluded due to substantial missingness (64.2%) | No | No |
| **Contact tracing** |  |  |  |  |  |
| Percent interviewed by DIS | Binomial | Percentage of individuals in the cluster by baseline who were interviewed by Disease Intervention Specialists (DIS). Binomial: 0 - < 66.67%, ≥66.67% | Removed ninth in backward elimination - Wald Chi-square p-value ≤ 0.10 in 49/100 bootstrapped samples | Yes | No |
| Percent with no named contacts | Binomial | Percentage of individuals in the cluster by baseline who did not report any identifiable contacts at DIS interview: Binomial: ≤50%, >50%. ≤50% is referent for aOR | Retained in final model - Wald Chi-square p-value ≤ 0.10 in 84/100 bootstrapped samples | Yes | Yes |
| Percent meeting sex partners online | N/A | Percentage of individuals in the cluster by baseline who reported meeting sex partners online | Excluded due to substantial missingness (64.8%) | No | No |

aCandidate predictors included in the final predictive model are highlighted in gray.

**Supplemental Table 2.** Retained predictors across 100, 200, and 500 bootstrapped samples.

|  |  |  |  |
| --- | --- | --- | --- |
| **Candidate Predictor** | **100 Samples (N, %)** | **200 Samples (N, %)** | **500 Samples (N, %)** |
| Cluster size | 93 (93.0%) | 188 (94.0%) | 470 (94.0%) |
| Percent with HIV viremia | 81 (81.0%) | 165 (82.5%) | 391 (78.2%) |
| Percent acute/recent | 16 (16.0%) | 29 (14.5%) |  88 (17.6%) |
| Median time to HIV care entry | 55 (55.0%) | 110 (55.0%) | 287 (57.4%) |
| Percent in HIV care  | 10 (10.0%) | 17 (8.5%) | 58 (11.6%) |
| Years since a diagnosis | 91 (91.0%) | 184 (92.0%) |  450 (90.0%) |
| Percent interviewed by DIS | 49 (49.0%) | 99 (49.5%) | 253 (50.6%) |
| Cluster age: diagnoses | 17 (17.0%) | 31 (15.5%) | 79 (15.8%) |
| Median age | 73 (73.0%) | 156 (78.0%) | 397 (79.4%) |
| Percent PWID | 21 (21.0%) | 34 (17.0%) | 72 (14.4%) |
| Percent MSM | 30 (30.0%) | 55 (27.5%) | 125 (25.0%) |
| Percent Black, non-Hispanic | 25 (25.0%) | 47 (23.5%) | 106 (21.2%) |
| Percent with no named contacts | 84 (84.0%) | 172 (86.0%) | 431 (86.2%) |
| Percent in dominant geographic region | 14 (14.0%) | 30 (15.0%) | 81 (16.2%) |
| Cluster age: sequences | 15 (15.0%) | 36 (18.0%) | 89 (17.8%) |

**Supplemental Table 3.** Evaluation of goodness of fit and predictive ability of candidate models in the internal and external validation datasets.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Candidate model** | **Predictors dropped from prior model** | **Chi-square p-value** | **ROC-AUC**b | **AIC** | **BIC** |
| **Internal Validation** |  |  |  |  |  |
| Full model |  | 0.939 | 0.840 | 322.86 | 392.40 |
| All predictors in ≥20% of bootstrapped samples | % in care, % dominant region, % cluster age by sequence dates, % acute, cluster age by diagnosis dates | 0.690 | 0.839 | 313.94 | 364.16 |
| All predictors in ≥30% of bootstrapped samples | % PWID, % Black | 0.695 | 0.839 | 310.48 | 352.98 |
| All predictors in ≥40% of bootstrapped samples | % MSM | 0.778 | 0.837 | 309.16 | 347.79 |
| All predictors in ≥50% of bootstrapped samples | % interviewed by DIS | 0.200 | 0.835 | 309.70 | 344.47 |
| All predictors in ≥70% of bootstrapped samples | Median time to care entry | 0.619 | 0.826 | 313.74 | 344.65 |
| **External Validation** |  |  |  |  |  |
| Final model from development & internal validation | Includes: cluster size, years since a diagnosis, % with no named contacts, % with HIV viremia, median age, and median time to HIV care entry | 0.100 | 0.831 |  |  |

aFinal predictive model is highlighted in gray; bROC-AUC values in the internal validation sample are not adjusted for optimism

**Supplemental Table 4.** Predictor adjusted odds ratios in the internal and temporal external validation populations of clusters for the final model in the main analysis and a sensitivity analysis excluding dyads at baseline.

|  |  |  |
| --- | --- | --- |
|  | Main analysis | Excluding dyads |
| Predictor | Internal OR (95% CI) | External OR (95% CI) | Internal OR (95% CI) | External OR (95% CI) |
| Cluster size at baseline | 1.17 (1.08, 1.28) | 1.15 (1.08, 1.23) | 1.12 (1.03, 1.22) | 1.12 (1.04, 1.20) |
| Median years to care entry | 0.85 (0.74, 0.98) | 0.93 (0.80, 1.09) | 0.86 (0.72, 1.02) | 0.94 (0.75, 1.17) |
| Median age at baseline | 0.67 (0.49, 0.93) | 0.67 (0.48, 0.93) | 0.83 (0.55, 1.25) | 0.49 (0.31, 0.79) |
| >50% with no named contacts | 2.13 (1.16, 3.91) | 1.28 (0.71, 2.31) | 2.01 (0.97, 4.13) | 1.63 (0.79, 3.39) |
| Years since a diagnosis | 2.69 (1.44, 5.03) | 4.48 (2.51, 8.01) | 2.70 (1.32, 5.54) | 3.98 (1.97, 8.03) |
| Viremia |  |  |  |  |
|  0 < % ≤ 25 with HIV viremia | 2.03 (0.52, 7.88) | 3.46 (1.19, 10.06) | 0.53 (0.11, 2.50) | 4.20 (0.96, 18.40) |
|  25 < % < 50 with HIV viremia | 4.31 (1.49, 12.45) | 1.49 (0.61, 3.66) | 1.42 (0.37, 5.51) | 2.22 (0.54, 9.16) |
|  ≥ 50% with HIV viremia  | 2.74 (0.93, 8.09) | 2.71 (1.08, 6.81) | 0.83 (0.55, 1.25) | 3.78 (0.92, 15.58) |

**SUPPLEMENTAL REFERENCES**

1. Chaillon A, Essat A, Frange P, et al. Spatiotemporal Dynamics of HIV-1 Transmission in France (1999-2014) and Impact of Targeted Prevention Strategies. Retrovirology **2017**; 14.

2. Aldous JL, Pond SK, Poon A, et al. Characterizing HIV Transmission Networks Across the United States. Clin Infect Dis **2012**; 55: 1135─1143.

3. Fisher M, Pao D, Brown AE, et al. Determinants of HIV-1 transmission in men who have sex with men: a combined clinical, epidemiological and phylogenetic approach. AIDS **2010**; 24: 1739-1747.

4. Division of HIV/AIDS Prevention. Detecting and Responding to HIV Transmission Clusters: A Guide for Health Departments. Centers for Disease Control and Prevention. **2018**. <https://www.cdc.gov/hiv/pdf/funding/announcements/ps18-1802/CDC-HIV-PS18-1802-AttachmentE-Detecting-Investigating-and-Responding-to-HIV-Transmission-Clusters.pdf>.

5. Hightower GK, May SJ, Perez-Santiago J, Pacold ME, Wagner GA, Little SJ, et al. HIV-1 Clade B pol Evolution following Primary Infection. PLoS One **2013**; 8: e68188.