The Journal of Infectious Diseases

RESEARCH ARTICLE

Quantifying and predicting ongoing Human Immunodeficiency Virus Type 1 (HIV-1) transmission dynamics in Switzerland using a distance-based clustering approach

Marco Labarile[11,2], Tom Loosli[1,2], Marius Zeeb[1,2], Katharina Kusejko[1,2], Michael Huber[2], Hans H. Hirsch[3,4], Matthieu Perreau[5], Alban Ramette[6], Sabine Yerly[7], Matthias Cavassini[8], Manuel Battegay[4], Andri Rauch[9], Alexandra Calmy[7], Julia Notter[10], Enos Bernasconi[11], Christoph Fux[12], Huldrych F. Günthard[¶1,2], Chloé Pasin[¶1,2], Roger D. Kouyos[¶1,2], and the Swiss HIV Cohort Study*

[1]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, 8091 Zurich, Switzerland [2]Institute of Medical Virology, University of Zurich, 8057 Zurich, Switzerland, [3]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, University of Basel, 4031 Basel, Switzerland [4]Transplantation and Clinical Virology, Department of Biomedicine, University of Basel, 4009 Basel, Switzerland [5]Division of Immunology and Allergy, Lausanne University Hospital, University of Lausanne, 1011 Lausanne, Switzerland [6]Institute for Infectious Diseases, University of Bern, 3001 Bern, Switzerland [7]Laboratory of Virology and Division of Infectious Diseases, Geneva University Hospital, University of Geneva, 1205 Geneva, Switzerland [8]Division of Infectious Diseases, Lausanne University Hospital, 1011 Lausanne, Switzerland [9]Department of Infectious Diseases, Bern University Hospital, University of Bern, 3010 Bern, Switzerland [10]Division of Infectious Diseases, Cantonal Hospital St. Gallen, 9007 St. Gallen, Switzerland [11]Division of Infectious

---

[¶] These authors contributed equally to the manuscript

***** Membership of the Swiss HIV Cohort Study is listed in the Acknowledgments

Corresponding author: Prof. Dr. Roger D. Kouyos, +41 44 255 36 10, roger.kouyos@uzh.ch

Alternate corresponding author: Marco Labarile, +41 79 668 52 20, marco.labarile@usz.ch

DOI: 10.1093/infdis/jiac457

Diseases, Regional Hospital Lugano, 6900 Lugano, Switzerland [12]Department of Infectious Diseases, Kantonsspital Aarau, 5001 Aarau, Switzerland

Background. Despite effective prevention approaches, ongoing HIV-1 transmission remains a public health concern indicating a need for identifying its drivers.

Methods. We combine a network-based clustering method using evolutionary distances between viral sequences with statistical learning approaches to investigate the dynamics of HIV-1 transmission in the Swiss HIV Cohort Study and to predict the drivers of ongoing transmission.

Results. We find that only a minority of clusters and patients acquire links to new infections between 2007 and 2020. While the growth of clusters and the probability of individual patients acquiring new links in the transmission network was associated with epidemiological, behavioral and virological predictors, the strength of these associations decreased substantially when adjusting for network characteristics. Thus, these network characteristics can capture major heterogeneities beyond classical epidemiological parameters. When modeling the probability of a newly diagnosed patient being linked with future infections, we found that the best predictive performance (median AUC_{ROC}=0.77) was achieved by models including characteristics of the network as predictors and that models excluding them performed substantially worse (median AUC_{ROC}=0.54).

Conclusions. These results highlight the utility of molecular epidemiology-based network approaches for analysing and predicting ongoing HIV-1-transmission dynamics. This approach may serve for real-time prospective assessment of HIV-1-transmission.

Keywords: HIV transmission dynamics, cluster analysis, distance-based clustering

BACKGROUND

Since the peak of the global Human Immunodeficiency Virus (HIV-1) epidemic in the mid-1990s, the worldwide HIV-1-incidence has been declining [1] and continuous efforts to diagnose, counsel and treat people living with HIV-1 (PLWH) have contributed to this trend. Large-scale efforts have been undertaken to reach the ambitious UNAIDS 90-90-90 goal consisting of 90% of all PLWH knowing of their status, 90% of those receiving antiretroviral treatment (ART), and 90% of PLWH on ART having undetectable plasma viral load by the year 2020 [2], although this goal was ultimately not met in the majority of countries. Other studies have shown that the strategy of treatment as prevention is effective at lowering the number of new HIV-1-infections and that the use of pre-exposure prophylaxis for at-risk patients further compounds this effect [3–6]. However, transmission of HIV-1 is still ongoing, with an estimated 1.5 million new infections occurring in 2020 [7]. To achieve the goal of stopping the spread of HIV-1, there is a need to better characterise patients who contribute to transmission and to develop approaches that can be applied to inform targeted preventive measures [8].

Approaches from molecular epidemiology have provided major insights into different aspects of HIV-1 transmission, both on the level of individual cases and large-scale transmission networks [9–14]. In this context, viral sequences are typically used to build phylogenetic trees, based on the rationale that sequences from patients who belong to the same transmission chain share a common ancestor and hence form subtrees [15]. While these approaches have led to important insights into HIV-1 transmission, they have limitations in the context of assessing the temporal growth of transmission chains and for potential real-time uses of molecular epidemiology to identify foci of ongoing HIV-1 transmission. These limitations include computation time and the sensitivity of phylogenetic trees to the addition of new sequence data. Indeed, updating the phylogenetic tree by adding new sequences over time can alter the topology of the resulting tree: the newly defined clusters are not always strict supersets of the clusters found at a previous timepoint, which complicates the identification and long-term analysis of some clusters. For example, in [4], we were able to quantify growth rates of clusters of men that have sex with men (MSM) in Switzerland using a phylogenetic clustering approach. However, as clustering patterns were not always consistent across years, this analysis had to be restricted to those clusters showing a minimal overlap with the pre-existing clusters from the previous year, thereby potentially causing selection bias (excluding unstable clusters) and complicating the study of long-term cluster growth.

Evolutionary distance-based networks have been proposed and used as an alternative to infer transmission chains [16–18]. While this approach ignores the inference of common ancestry inherent in phylogenetic trees, it results in more robust clusters that are less sensitive to the addition of new patients. This substantially reduces computation time, particularly in the context of repeatedly updating clustering analyses in real-time. Moreover, distance-based clustering was non-inferior to phylogenetic approaches both in simulation studies and empirical assessment of the overlap between clusters and contact networks [19,20]. Finally, the distance network approaches are more directly amenable to the application of tools and metrics from network science. We therefore adapted one such distance-based clustering mechanism implemented in HIV-TRACE [16] and combined it with statistical learning approaches in the context of the Swiss HIV Cohort Study (SHCS) to assess its ability to analyse cluster growth dynamics in the Swiss HIV-1 epidemic, as well as the predictive capabilities that can be achieved in this framework.

METHODS

Swiss hiv cohort study

The Swiss HIV Cohort Study (SHCS) is a prospective multi-center cohort study that covers an estimated 53% of all HIV-1 diagnoses ever issued and an estimated 80% of all HIV-1-positive MSM in Switzerland [21]. The 13299 analysed sequences were obtained from genotypic drug resistance tests that have been performed up to 2020-12-31, each sequence being the earliest

available of the corresponding patient. The SHCS was approved by the local ethical committees of the participating centers, and written informed consent was obtained from all participants.

Sequences and clustering

HIV-1-pol sequences from resistance tests of patients enrolled in the SHCS were used in the clustering, which was performed using HIV-TRACE [16]. HIV-TRACE performs pairwise sequence alignment, during which insertions and deletions relative to the reference are discarded, pairwise distance calculation based on the Tamura-Nei 93 distance criterion [22], and subsequent clustering of sequences with a genetic distance smaller than 0.01. The genetic distance threshold was chosen with the aim of maximising the number of clusters and therefore the resolution of our analyses, while staying within the realm of previously applied thresholds in studies on HIV-*pol* [19,23,24] (Supplementary Figure 1). A comparison of cluster growth rates for genetic distance thresholds between 0.005 and 0.015 is shown in Supplementary Table 1. The *pol*-gene of the reference HIV-1 genome HXB2 (GenBank accession number K03455.1) was used as the reference sequence. For the purposes of this study, a node represents a single HIV-1-*pol* sequence and therefore one patient, while a cluster is defined as a connected component in the network produced by this method. By this definition, a cluster has at least two members and an unconnected node is not part of any cluster.

Data management and cluster analysis

Data management was performed in two steps: first, we parsed the outputs of HIV-1-TRACE, calculated statistics of interest (Table 1) for clusters and nodes and linked the sequences to patient information from the SHCS, using Python 3.8. Graph-theoretical analysis of the clusters was performed using igraph 0.8.3 for Python [25]. Further analyses and visualizations were performed using R version 4.0.5 [26] and the package ggplot [27].

Cluster- and node-level growth modeling

We used Poisson regression to model the number of nodes acquired in each cluster from 2014 to 2017 and assess factors associated with cluster growth. One considered factor was the past cluster growth (defined as the change in cluster size from 2011 to 2014). We used logistic regression to model the acquisition of new links of individual nodes within the first three years of being enrolled in the cohort (with a binary outcome variable). We included variables that have been found to be predictive of cluster growth or clustering in similar work [4,19,28], such as age, sex, CD4-cell count, virus load and transmission risk factor.

More details on the models and included variables can be found in the supplementary material.

Cross-validation

To predict whether a node will acquire a new link in the following three years, we compared logistic regressions and classification random forests [29] built on several subsets of variables.

For each set of predictors, a logistic regression model and a random forest model were trained on the same training data. To supplement the sets of predictors manually chosen, we also employed a method of automated variable selection implemented in Variable Selection Using Random Forests (VSURF) [30].

For more detailed information on the methods employed, please refer to the methods section in the supplementary material.

RESULTS

Analysing a total of 13299 sequences with the distance-based clustering algorithm yielded a total of 998 clusters that were highly robust over the observed timeframe, making it possible to assess the dynamics of the clusters and their constituent nodes over the 13 year-long period (Supplementary Figures 2-4).

Out of the 13299 included sequences, 4074 (30.6%) clustered with at least one other sequence at the time of sampling. At the last observed timepoint (2020-12-31), 5415 (40.7%) of all sequences were linked to at least one other sequence. We found that although IV-drug users represented 2572 (19.3%) of the total number of sequences, they constituted 29% of the clustered sequences (Table 2). On the other hand, patients in the heterosexual acquisition risk category represented 4782 (36%) of the total number of sequences but only 25% of the clustered sequences, indicating potentially less frequent transmission in this subpopulation ($p_{\chi 2} < 0.001$).

Most clusters had less than 10 nodes at the end of 2020 (943/998, 94.5%). The largest identified cluster contained 1577 nodes, 43.7% of which were categorized as IV-drug users, and 24.0% of which were categorized as heterosexuals.

The obtained clusters exhibit a large heterogeneity in terms of composition, size, and growth patterns (Figures 1, 2; Supplementary Figures 4-7). Of 575 clusters identified up to 2007-12-31, only 134 (23.3%) gained any new nodes in the following 13 years, of which only 33 (5.7%) gained 5 or more new nodes (Figure 2a). Despite the small fraction of clusters that gained 5 or more nodes, they accounted for 443 (70.9%) of all 625 nodes that were gained by all 575 clusters collectively. The clusters that gained 5 or more nodes were disproportionately MSM clusters (27 of 33, 81.8%). We found a strong correlation between cluster size in 2007 and number of new nodes acquired until 2020 (Spearman's $r = 0.72$, $p < 0.001$, Figure 2a). Similarly, of the 9308 identified patients in the SHCS as of 2007, only 1079 (11.6%) gained links to new sequences up to the year 2020 (Figure 2b). Most patients that acquired new links only gained very few: Only 206 (2.2%) gained links to three or more new sequences, and they accounted for 1103 (50.4%) of the total 2190 new links over the studied period.

When modeling cluster growth using Poisson regression (with $\log_{10}$ cluster size in the year 2014 as an offset), we found that past growth of a cluster was a good predictor for future growth

(Figure 3a; adjusted incidence rate ratio, aIRR, [95%-CIs], 5.11 [2.62, 9.95] and 11.03 [6.44, 18.88] for past cluster growth of 2-3 and >=4, respectively). Besides past growth, no other variable yielded a statistically significant estimate in the multivariable model. Clusters with older individuals had lower growth rates in the univariable models (aIRR 0.62 [0.42, 0.91] and 0.17 [0.11, 0.29] for median ages of 40-49 and >=50, respectively), as did clusters made up of mostly heterosexuals (aIRR 0.38 [0.24, 0.60]), clusters with more than 90% virally suppressed patients (aIRR 0.48 [0.37, 0.62]), and clusters with high rates of condom use with occasional partners (aIRR 0.15 [0.07, 0.31]). On the other hand, clusters with more patients using non-IV drugs had significantly higher growth rates in the univariable model (aIRR 2.18 [1.41, 3.37]). This indicates that the effect of these variables can be captured by including past cluster growth as a proxy for behavioral and demographic risk factors. Similar results were obtained when restricting the analysis to clusters where MSM was the most common acquisition risk category (Supplementary Figure 8) and when varying the timepoint (Supplementary Figures 9-12)

To quantify the relevant factors of growth at the individual node level, i.e., a node's risk of acquiring new links over time, we specified a logistic regression model where we used a similar set of variables for predicting the addition of new links to a given node within three years of being sequenced (Figure 3b). Node degree had a significant effect on the outcome, with larger node degrees being associated with higher probabilities to gain new links (Odds ratios, OR, [95%-Cis], 2.41 [1.94, 3.00], 4.98 [3.98, 6.24], 11.35 [8.34, 15.45] for node degree 1, 2-4 and >=5, respectively). Accordingly, removing node-degree from the regression model led to a significantly worse model fit (Likelihood ratio test: $p < 0.001$). In other words, the growth of the network occurs by preferential attachment, meaning more connected nodes acquire more new links, which also explains the approximately scale-free pattern observed for the degree distribution of the whole network (Supplementary Figure 13). Besides node degree, several epidemiological and virological factors were associated with acquisition of new links: Patients between 40 and 49 years old were at a significantly lower risk than younger patients (OR 0.52 [0.42, 0.64]), as were IV-drug users compared to MSM (OR 0.70 [0.54, 0.89]). Viral loads above 10000 copies/ml were associated with the gain of new links (OR 1.35 [1.05, 1.74]), as were CD4-cell counts above 300 cells/µL (OR 1.59 [1.31, 1.93]) and inconsistent condom use with occasional partners (OR 1.37 [1.13, 1.67]). Restricting the analysis to MSM patients yielded similar results (Supplementary Figure 14), as did adding the enrolment year as a linear effect (Supplementary Figure 15) and random subsampling of 75% or 50% of the available sequences (Supplementary Figures 16, 17).

Model comparison

We trained multiple models using five different sets of predictors (specified in Table 3 and Supplementary Table 2) with the goal of identifying the best model for predicting whether a certain node is going to acquire a link to a new node within three years. To assess the performance of these models, we performed a 10-fold cross-validation and compared the median AUCs of the ROC-curves based on the model predictions.

Among models with preselected predictor sets (Table 3), models that used both network and patient characteristics yielded the most accurate predictions (Figure 4, compare *Mix* and *Cluster* predictor sets with *Patient*). Random forests and logistic regression models performed similarly in all cases except one. Notably, restricting the set of predictors to demographical and clinical variables (*Patient* predictor set) resulted in a large drop in accuracy: From the *Mix* to the *Patient* predictor set, the median AUC decreased from 0.78 to 0.67 for the logistic regression and from 0.76 to 0.55 for the random forest. On the other hand, restricting the set of predictors to variables pertaining to the topological characteristics of clusters and nodes (model *Cluster*) did not decrease accuracy to the same degree, as the median AUC was 0.76 both for the logistic regression and the random forest. Accordingly, variables with the highest variable importance in the *Mix* random forest model were cluster characteristics, namely node degree, cluster past growth and cluster size (Supplementary Figure 18).

Additionally, we identified two more subsets of predictors using Variable Selection Using Random Forests (VSURF) [30]. From a mix of demographical, clinical and cluster topology-related predictors VSURF repeatedly selected only the latter category of variables (Supplementary Table 2). The performance of random forests based on the variables selected by VSURF was similar to the *Mix* and *Cluster* models, with median AUCs of 0.77 and 0.74 for VSURF_Interpretation and VSURF_Prediction, respectively (Figure 4). ROC curves of each model are displayed in Supplementary Figure 19.

DISCUSSION

Here we combined the distance-based clustering method HIV-Trace [16] with longitudinal cohort data and statistical learning approaches to analyse cluster growth dynamics in the Swiss HIV Cohort study. In concordance with previous work [4], we found that, in the timespan from 2007 to 2020, only a minority of the HIV-1-clusters in Switzerland were growing. Similarly, only a small fraction of patients enrolled up to the year 2007 have formed any new links, which would be an indication of onward transmission of HIV-1. Consistent with earlier work, [4], we found that the fraction of virally suppressed patients and behavioral risk factors were predictive of cluster growth. When adjusting for network characteristics however, these associations were no longer statistically significant, suggesting that part of the information provided by the aforementioned variables can be captured by the characteristics of the network.

When modeling the risk of acquiring new links on the patient-level, we found that viral loads of more than 10000 copies/ml were associated with a high risk of gaining links, adding to the evidence that suppressing viral loads is essential for HIV-1 prevention [5,31–34]. Additionally, we observed a subgroup of MSM with a sudden burst of growth early in the studied period, indicating that this subgroup or their undiagnosed or HIV-1-negative contacts might benefit from targeted preventive efforts (Figure 2b, Supplementary Figures 4,5).

As has been partly shown previously in a US-based study [35], we find that cluster size and its previous growth activity is predictive of future growth. When comparing demographical, clinical and behavioral variables with network-based variables, we observed a significant improvement in the predictive capacity of both cluster-level and patient-level growth models when network-based variables were added as predictors. Keeping in mind the goal of prospectively analysing the state of HIV-1-epidemics, these variables derived from the network topology provided a substantial increase in predictive accuracy that should not be ignored. The predictive power of past cluster growth and the node degree, the small fraction of active clusters and patients, and the degree distributions observed in the clusters also suggest some degree of preferential attachment being responsible in the generation of the clusters. This underlines the need for approaches that allow the precise and timely identification of foci of ongoing transmission for the sake of preventive action. The predictive models established in this work could thus form the basis of such precision public health approaches to HIV-1 prevention.

One limitation of our analyses is that they depend on an ad-hoc choice of clustering threshold. Here, in line with previous work [19,36,37], we chose a threshold of 0.01. This conservative threshold maximizes the number of clusters in our cohort and thereby provides the best resolution for our analysis. This way, we avoid the two extremes of an unnecessarily strict threshold, which would fail to cluster sequences even if they correspond to real transmission pairs, and a too lenient threshold, which would combine even very different sequences into large uninformative clusters that don't reflect the underlying transmission network. In addition, a strict threshold was preferable in the case of the patient-based prediction models. We also conducted a sensitivity analysis showing that in this study, results were robust to the threshold choice. Another limitation is that we cannot establish individual transmission events between linked patients. Furthermore, the SHCS contains only part of the Swiss HIV-1-positive population, which means that the analysed clusters are missing patients that are not enrolled in the cohort. Consequently, the appearance of new links between patients of the SHCS can be caused by undiagnosed or otherwise not enrolled PLWH. However, with close to 21000 total patients and nearly 10000 patients under follow-up as of 2020, the SHCS is representative of the Swiss HIV-1-epidemic [38]. Another limitation is the use of the first sequence per patient only, which does not account for intra-patient evolution of the virus. Since there was only a single sequence available for most (63.0%) SHCS patients for whom a genotypic resistance test had been performed, this was a practical choice with the added benefit of maximizing the long-term robustness of the clusters generated by HIV-TRACE. Future extensions of this study could possibly take this intra-patient evolution into account, therefore more precisely modeling the real epidemic, though this is contingent on the availability of sequence data on a large amount of longitudinally sampled patients.

Despite these limitations, this study provides insight into the long-term dynamics of cluster growth of HIV-1 in Switzerland. It makes use of the densely sampled SHCS, representing a significant and representative part of the Swiss HIV-1-positive population. The clustering

method used makes longitudinal follow up on individual clusters feasible and opens the possibility of prospective analyses performed in real-time. Additionally, it demonstrates the importance of considering cluster-derived variables in addition to demographical and clinical variables when modeling cluster and individual growth dynamics.

In conclusion, we present new insights into the long-term dynamics of HIV-1-cluster growth including the value of using cluster-based variables in predicting future growth both on the level of clusters and individual patients in the Swiss HIV-1-epidemic.

FOOTNOTES

Acknowledgments We thank the participants of the Swiss HIV Cohort Study (SHCS); the physicians, and study nurses for excellent patient care; the resistance laboratories for high-quality genotyping drug resistance testing; the SHCS data center (A. Scherrer, K. Kusejko, J. Meier, Y. Schäfer, and O. Follonier) for excellent data management; and D. Perraudin and M. Amstad for administrative assistance. **Members of the Swiss HIV Cohort Study**: Abela I, Aebi-Popp K, Anagnostopoulos A, Battegay M, Bernasconi E, Braun DL, Bucher HC, Calmy A, Cavassini M, Ciuffi A, Dollenmaier G, Egger M, Elzi L, Fehr J, Fellay J, Furrer H, Fux CA, Günthard HF (President of the SHCS), Hachfeld A, Haerry D (deputy of "Positive Council"), Hasse B, Hirsch HH, Hoffmann M, Hösli I, Huber M, Kahlert CR (Chairman of the Mother & Child Substudy), Kaiser L, Keiser O, Klimkait T, Kouyos RD, Kovari H, Kusejko K (Head of Data Centre), Martinetti G, Martinez de Tejada B, Marzolini C, Metzner KJ, Müller N, Nemeth J, Nicca D, Paioni P, Pantaleo G, Perreau M, Rauch A (Chairman of the Scientific Board), Schmid P, Speck R, Stöckle M (Chairman of the Clinical and Laboratory Committee), Tarr P, Trkola A, Wandeler G, Yerly S.

Funding This work was supported by the Swiss National Science Foundation [33CS30_177499 to H. F. G.] in the framework of the Swiss HIV Cohort Study; and further supported by the Swiss National Science Foundation [324730B_179571, 310030_141067 to H. F. G. and 324730_207957, BSSGI0_155851 to R. D. K.]; the Yvonne-Jacob Foundation (to H. F. G.); the University of Zurich Clinical Research Priority Program for “Viral Infectious Disease, the Zurich Primary HIV Infection Cohort Study” (to H. F. G.); and an unrestricted research grant from Gilead Sciences to the SHCS Research Foundation. The data were gathered by the 5 Swiss university hospitals, 2 cantonal hospitals, 15 affiliated hospitals, and 36 private physicians (listed at [34]).

Potential conflicts of interest R. D. K. has received grants from the SNF and personal fees from Gilead Sciences, outside the submitted work. H. F. G. has received unrestricted research grants from Gilead Sciences and Roche; fees for data and safety monitoring board membership from Merck; and consulting/advisory board membership fees from Gilead Sciences, ViiV, Sandoz,

and Mepha. The institution of H. F. G. has received unrestricted educational grants from Gilead Sciences, MSD, ViiV, Sandoz, and Abbvie. All other authors report no conflicts.

References

1. Rehle TM, Hallett TB, Shisana O, et al. A Decline in New HIV Infections in South Africa: Estimating HIV Incidence from Three National HIV Surveys in 2002, 2005 and 2008. PLOS ONE. Public Library of Science; **2010**; 5(6):1–8.
2. HIV/AIDS (UNAIDS) JUNP on, others. 90-90-90: an ambitious treatment target to help end the AIDS epidemic. Geneva: UNAIDS; 2014 [Internet]. 2017 [cited 2022 Jun 11]. Available from: https://www.unaids.org/en/resources/documents/2017/90-90-90
3. Kusejko K, Marzel A, Hampel B, et al. Quantifying the drivers of HIV transmission and prevention in men who have sex with men: a population model-based analysis in Switzerland. HIV Med. **2018**; 19(10):688–697.
4. Bachmann N, Kusejko K, Nguyen H, et al. Phylogenetic Cluster Analysis Identifies Virological and Behavioral Drivers of Human Immunodeficiency Virus Transmission in Men Who Have Sex With Men. Clin Infect Dis. **2021**; 72(12):2175–2183.
5. Cohen MS, Chen YQ, McCauley M, et al. Prevention of HIV-1 Infection with Early Antiretroviral Therapy. N Engl J Med. **2011**; 365(6):493–505.
6. Celum C, Baeten J. PrEP for HIV Prevention: Evidence, Global Scale-up, and Emerging Options. Cell Host Microbe. **2020**; 27(4):502–506.
7. UNAIDS. Global HIV & AIDS statistics—Fact sheet [Internet]. UNAIDS Geneva, Switzerland; 2021 [cited 2022 Jun 11]. Available from: https://www.unaids.org/en/resources/fact-sheet
8. Dennis AM, Herbeck JT, Brown AL, et al. Phylogenetic studies of transmission dynamics in generalized HIV epidemics: an essential tool where the burden is greatest? J Acquir Immune Defic Syndr 1999. **2014**; 67(2):181–195.
9. Sivay MV, Hudelson SE, Wang J, et al. HIV-1 diversity among young women in rural South Africa: HPTN 068. PloS One. **2018**; 13(7):e0198999.
10. Castro-Nallar E, Pérez-Losada M, Burton GF, Crandall KA. The evolution of HIV: Inferences using phylogenetics. Mol Phylogenet Evol. **2012**; 62(2):777–792.
11. Oster AM, France AM, Mermin J. Molecular Epidemiology and the Transformation of HIV Prevention. JAMA. **2018**; 319(16):1657–1658.
12. Grabowski MK, Herbeck JT, Poon AFY. Genetic Cluster Analysis for HIV Prevention. Curr HIV/AIDS Rep. **2018**; 15(2):182–189.
13. Beloukas A, Psarris A, Giannelou P, Kostaki E, Hatzakis A, Paraskevis D. Molecular epidemiology of HIV-1 infection in Europe: An overview. Infect Genet Evol. **2016**; 46:180–189.
14. Peeters M, Jung M, Ayouba A. The origin and molecular epidemiology of HIV. Expert Rev Anti Infect Ther. **2013**; 11(9):885–896.
15. Hassan AS, Pybus OG, Sanders EJ, Albert J, Esbjörnsson J. Defining HIV-1 transmission clusters based on sequence data. AIDS. **2017**; 31(9):1211–1222.
16. 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. Shapiro B, editor. Mol Biol Evol. **2018**; 35(7):1812–1819.

17. Xia Q, Wertheim JO, Braunstein SL, Misra K, Udeagu C-C, Torian LV. Use of molecular HIV surveillance data and predictive modeling to prioritize persons for transmission-reduction interventions. AIDS [Internet]. **2020**; 34(3). Available from: https://journals.lww.com/aidsonline/Fulltext/2020/03010/Use_of_molecular_HIV_surveillance_data_and.13.aspx
18. Oster AM, France AM, Panneer N, et al. Identifying Clusters of Recent and Rapid HIV Transmission Through Analysis of Molecular Surveillance Data. J Acquir Immune Defic Syndr 1999. **2018**; 79(5):543–550.
19. Wertheim JO, Kosakovsky Pond SL, Forgione LA, et al. Social and Genetic Networks of HIV-1 Transmission in New York City. Bonhoeffer S, editor. PLOS Pathog. **2017**; 13(1):e1006000.
20. Villandre L, Stephens DA, Labbe A, et al. Assessment of Overlap of Phylogenetic Transmission Clusters and Communities in Simple Sexual Contact Networks: Applications to HIV-1. PLOS ONE. Public Library of Science; **2016**; 11(2):1–18.
21. Scherrer AU, Traytel A, Braun DL, et al. Cohort Profile Update: The Swiss HIV Cohort Study (SHCS). Int J Epidemiol [Internet]. **2021**; . Available from: https://doi.org/10.1093/ije/dyab141
22. 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.
23. Fujimoto K, Bahl J, Wertheim JO, et al. Methodological synthesis of Bayesian phylodynamics, HIV-TRACE, and GEE: HIV-1 transmission epidemiology in a racially/ethnically diverse Southern U.S. context. Sci Rep. Nature Publishing Group; **2021**; 11(1):3325.
24. Chato C, Kalish ML, Poon AFY. Public health in genetic spaces: a statistical framework to optimize cluster-based outbreak detection. Virus Evol. **2020**; 6(1):veaa011.
25. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. **2006**; Complex Systems:1695.
26. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/
27. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org
28. Wertheim JO, Murrell B, Mehta SR, et al. Growth of HIV-1 Molecular Transmission Clusters in New York City. J Infect Dis. **2018**; 218(12):1943–1953.
29. Breiman L. Random Forests. Mach Learn. **2001**; 45(1):5–32.
30. Genuer R, Poggi J-M, Malot CT-. VSURF: An R Package for Variable Selection Using Random Forests. R J. **2015**; 7(2):19–33.
31. Marzel A, Shilaih M, Yang W-L, et al. HIV-1 Transmission During Recent Infection and During Treatment Interruptions as Major Drivers of New Infections in the Swiss HIV Cohort Study. Clin Infect Dis. **2016**; 62(1):115–122.
32. Attia S, Egger M, Müller 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.
33. Quinn TC, Wawer MJ, Sewankambo N, et al. Viral Load and Heterosexual Transmission of Human Immunodeficiency Virus Type 1. N Engl J Med. Massachusetts Medical Society; **2000**; 342(13):921–929.
34. Rodger AJ, Cambiano V, Bruun T, et al. Risk of HIV transmission through condomless sex in serodifferent gay couples with the HIV-positive partner taking suppressive antiretroviral therapy

(PARTNER): final results of a multicentre, prospective, observational study. The Lancet. Elsevier; **2019**; 393(10189):2428–2438.
35. Billock RM, Powers KA, Pasquale DK, et al. Prediction of HIV Transmission Cluster Growth With Statewide Surveillance Data. JAIDS J Acquir Immune Defic Syndr. **2019**; 80(2):152–159.
36. Smith DM, May SJ, Tweeten S, et al. A public health model for the molecular surveillance of HIV transmission in San Diego, California. AIDS. **2009**; 23(2):225–232.
37. Poon AFY, Joy JB, Woods CK, et al. The Impact of Clinical, Demographic and Risk Factors on Rates of HIV Transmission: A Population-based Phylogenetic Analysis in British Columbia, Canada. J Infect Dis. **2015**; 211(6):926–935.
38. Shilaih M, Marzel A, Yang WL, et al. Genotypic Resistance Tests Sequences Reveal the Role of Marginalized Populations in HIV-1 Transmission in Switzerland. Sci Rep. **2016**; 6(1):27580.

Table 1: Network properties that were calculated for clusters and nodes.

Variable	Object	Description
Node degree	Node	Number of links
Past node growth	Node	Number of links gained over the past three years
Future node growth	Node	Number of links gained over the next three years
Closeness	Node	$(n - 1) / (\Sigma_i^n p_i)$, where n is the number of nodes in the cluster and p_i is the shortest path from the node of interest to node i
Betweenness	Node	Number of shortest paths between each pair of nodes in the cluster that pass through the node of interest
Cluster size	Cluster	Number of nodes in the cluster
Past cluster growth	Cluster	Number of nodes gained over the past three years
Future cluster growth	Cluster	Number of nodes gained over the next three years
Density	Cluster	$m / (n * (n - 1) / 2)$, where m is the total number of links and n is the total number of nodes in the cluster
Transitivity	Cluster	Probability of two neighbors of the same node being linked directly
Median degree	Cluster	Median of all node degrees in the cluster
Median distance	Cluster	Median Tamura-Nei 93-distance of all the links in the cluster
Median closeness	Cluster	Median of the node closenesses
Median betweenness	Cluster	Median of the node betweennesses

Table 2: Characteristics of the patients whose HIV-*pol* sequences were used in the analysis.

	Clustered[a]	**Not clustered**[a]	**All**
Age[a]	years	years	Years
Mean (SD)	37.3 (9.73)	38.7 (10.7)	38.3 (10.4)
Median [Q1, Q3]	36 [31, 42]	37 [31, 45]	37 [31, 44]
Sex	N (%)	N (%)	N (%)
Female	899 (22.1%)	2844 (30.8%)	3743 (28.1%)
Male	3175 (77.9%)	6381 (69.2%)	9556 (71.9%)
Acquisition risk group			
MSM	1751 (43.0%)	3588 (38.9%)	5339 (40.1%)
Heterosexuals	1017 (25.0%)	3765 (40.8%)	4782 (36.0%)
IV-drug users	1180 (29.0%)	1392 (15.1%)	2572 (19.3%)
unknown	126 (3.1%)	480 (5.2%)	606 (4.6%)
RNA concentration[b]	copies/ml	copies/ml	copies/ml
Median [Q1, Q3]	27162 [3345, 110023]	15900 [790, 87078]	19605 [1260, 95938]
Missing	686 (16.8%)	2054 (22.3%)	2740 (20.6%)
CD4-cell count[b]	cells/µl	cells/µl	cells/µl
Median [Min, Max]	379 [222, 568]	340 [180, 527]	350 [191, 540]
Missing	68 (1.7%)	184 (2.0%)	252 (1.9%)
Total	N = 4074	N = 9225	N = 13299

[a] when the sample for the genotypic resistance test was taken

[b] at the follow up visit closest to the sampling for the genotypic resistance test

MSM: Men who have sex with men

Table 3: Predictor sets used in the model comparison. Mix, Cluster and Patient are predictor sets with mixed, only cluster-based predictors and only demographical and clinical predictors, respectively. The model comparison further included two predictor sets generated with automatic variable selection algorithm and which are described in the Supplementary Table 2.

Predictor set	Network-based predictors	Demographical predictors	Clinical predictors
Mix	node degree, past cluster growth, cluster size	acquisition risk group, registration center, age, sex	RNA-concentration
Cluster	node degree, past cluster growth, cluster size, median closeness in the cluster, node closeness, cluster density, median distance in the cluster	None	None
Patient	None	acquisition risk group, registration center, age, sex	RNA-concentration

Figure 1: Graph-representations of 4 different clusters (For a larger selection of clusters, see Supplementary Figure 7). Each node represents a single patient. Two linked nodes are patients whose HIV-pol-sequences have a Tamura Nei 93-genetic distance of less than or equal to 0.01. The sample year refers to the year when the sample for the genotypic drug resistance test was taken. A small amount of noise was added to the coordinates of each node for better readibility of clusters with many overlapping links. Abbreviations: MSM, men who have sex with men

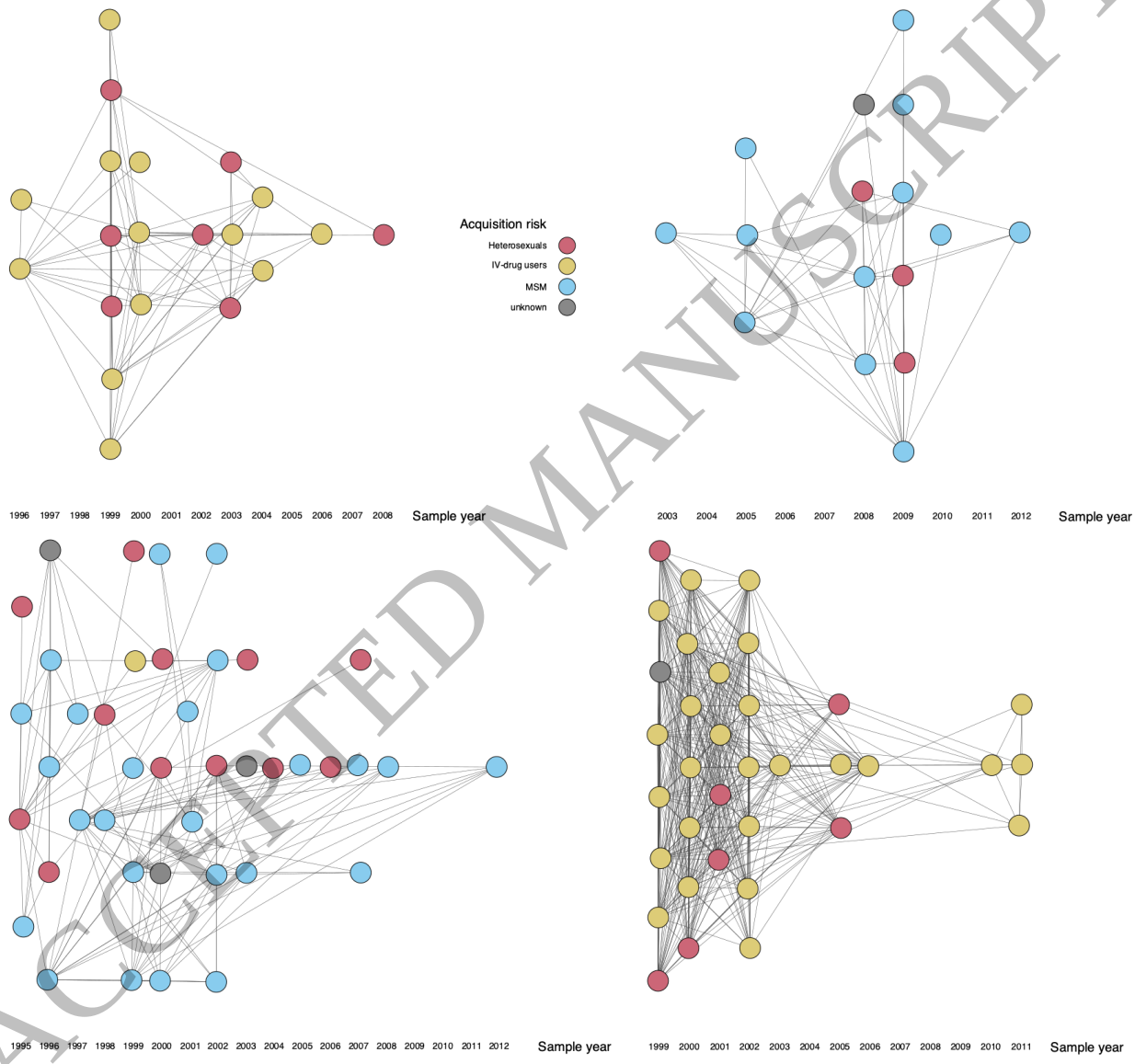

Figure 2: (a) Cluster growth from 2007-12-31 to 2020-12-31 as a function of cluster size in 2007. Clusters where the most common acquisition risk group did not constitute >50% of all members were assigned to a combined (hyphenated) category consisting of the two most common risk groups in alphabetical order. (b) Node growth from 2007-12-31 to 2020-12-31 as a function of node degree, i.e., the number of links, in 2007. Abbreviations: Het, Heterosexuals; IDU, Intravenous drug users; MSM, men who have sex with men

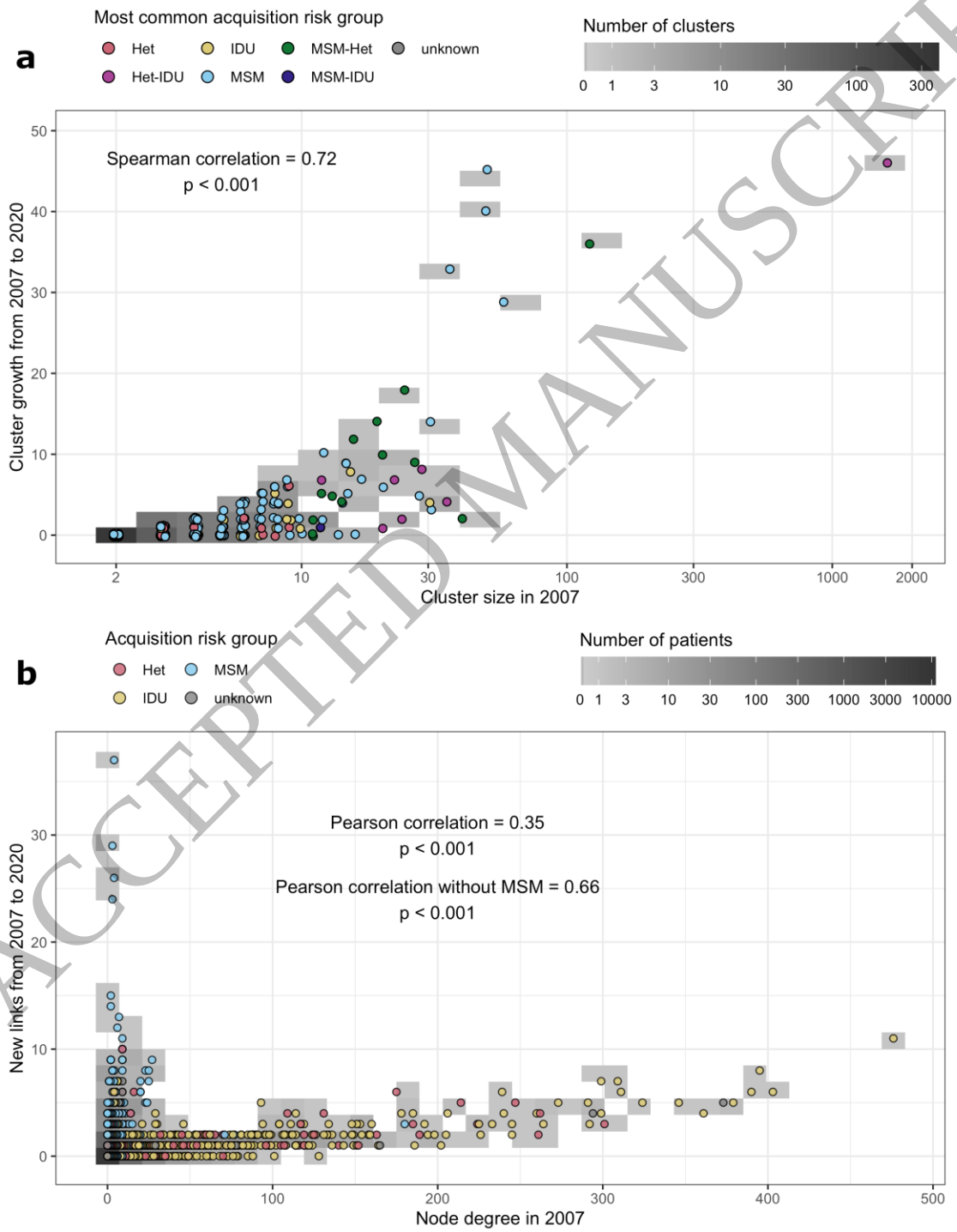

Figure 3: (a) Factors associated with the number of nodes gained by a cluster over the span of 3 years, as assessed by a Poisson regression model. Parameters estimated and 95%-confidence interval from univariable (red) and multivariable (blue) regressions are represented. Past cluster growth was calculated as the number of nodes gained from 2011-12-31 to 2014-12-31, and future cluster growth was calculated as the number of nodes gained from 2014-12-31 to 2017-12-31. (b) Factors associated with the gain of new links for a node within 3 years, as modeled by a logistic regression. Odds ratio and 95% confidence interval from univariable (red) and multivariable models (blue) are represented. The outcome was a binary variable based on the number of links gained in the first 3 years after the date of the genotypic resistance test.

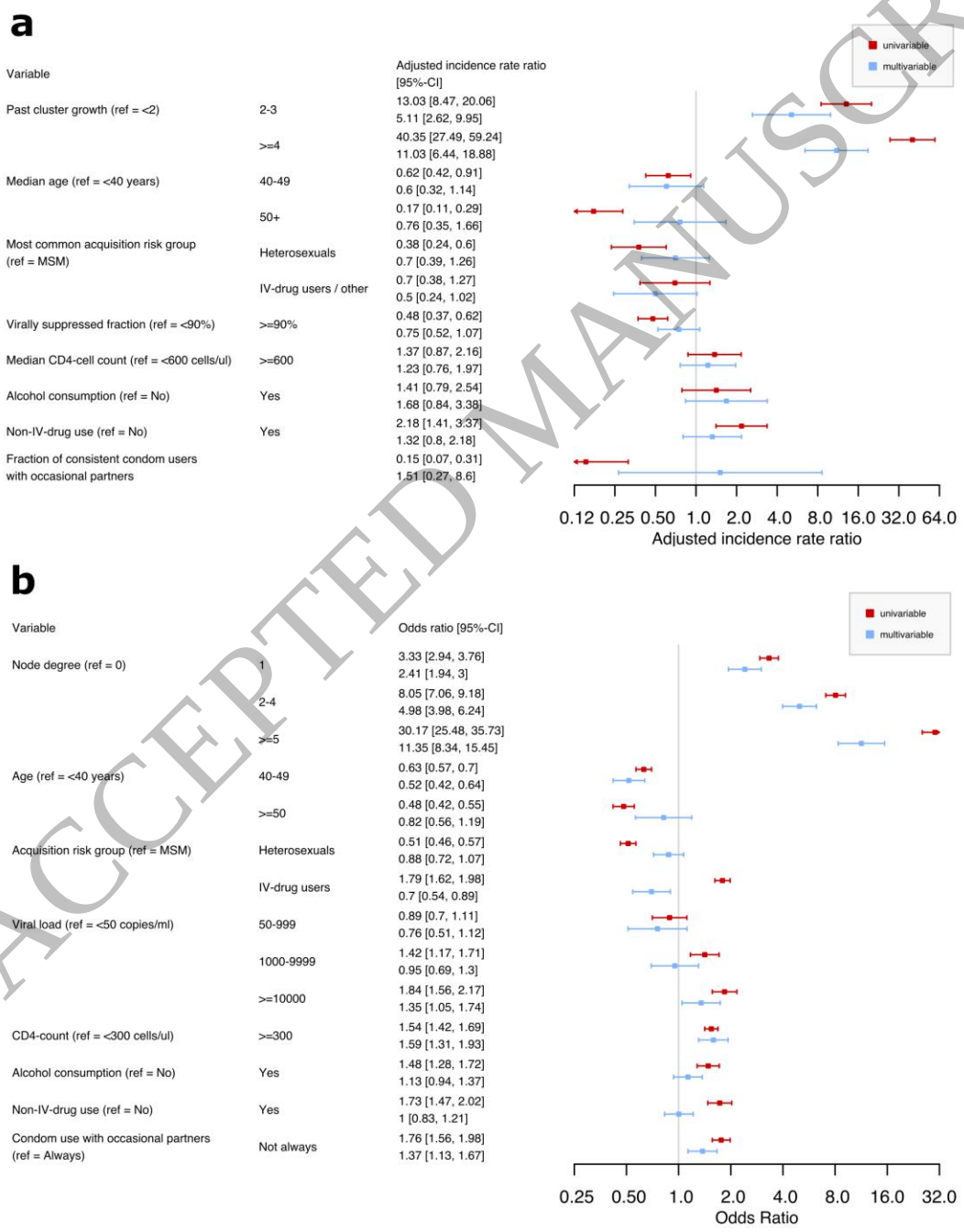

Figure 4: Comparison of the predictive abilities of 12 different classification models. These models are based on the combination of five different sets of predictors (described in Table 3 and Supplementary Table 2) with two different modeling methods: logistic regression and random forest, colored red and blue respectively. Each one of these combinations was assessed in a 10-fold cross validation. Predictive ability was assessed by comparing the areas-under-the curve (AUCs) of the receiver-operator-characteristic (ROC)-curves of the models.