Clusters of Sexual Behavior in Human Immunodeficiency Virus–positive Men Who Have Sex With Men Reveal Highly Dissimilar Time Trends

Luisa Salazar-Vizcaya,1,6 Katharina Kusejko,2,3 Axel J. Schmidt,15 Germán Carrillo-Montoya,6 Dunja Nicca,7 Gilles Wandeler,1 Dominique L. Braun,2,3 Jan Fehr,7 Katharine E. A. Darling,6 Enos Bernasconi,3 Patrick Schmid,6 Huldrych F. Günthard,2,3 Roger D. Kouyos,2,3 and Andri Rauch1; for the Swiss HIV Cohort Study

Background. Separately addressing specific groups of people who share patterns of behavioral change might increase the impact of behavioral interventions to prevent transmission of sexually transmitted infections. We propose a method based on machine learning to assist the identification of such groups among men who have sex with men (MSM).

Methods. By means of unsupervised learning, we inferred “behavioral clusters” based on the recognition of similarities and differences in longitudinal patterns of condomless anal intercourse with nonsteady partners (nsCAI) in the HIV Cohort Study over the last 18 years. We then used supervised learning to investigate whether sociodemographic variables could predict cluster membership.

Results. We identified 4 behavioral clusters. The largest behavioral cluster (cluster 1) contained 53% of the study population and displayed the most stable behavior. Cluster 3 (17% of the study population) displayed consistently increasing nsCAI. Sociodemographic variables were predictive for both of these clusters. The other 2 clusters displayed more drastic changes: nsCAI frequency in cluster 2 (20% of the study population) was initially similar to that in cluster 3 but accelerated in 2010. Cluster 4 (10% of the study population) had significantly lower estimates of nsCAI than all other clusters until 2017, when it increased drastically, reaching 85% by the end of the study period.

Conclusions. We identified highly dissimilar behavioral patterns across behavioral clusters, including drastic, atypical changes. The patterns suggest that the overall increase in the frequency of nsCAI is largely attributable to 2 clusters, accounting for a third of the population.

Keywords. sexual behavior; machine learning; men who have sex with men; HIV; condom.

Sexual behavior among men who have sex with men (MSM) is heterogeneous and dynamic [1–3]. Recent changes in sexual behavior have been associated with increases in sexually transmitted infections (STI) among MSM [4–8]. Yet potentially oversimplified assumptions on exposure to STI transmission are common [9–11]. Lack of understanding and misconceptions about choices regarding sexual encounters are likely to have limited the efficiency and efficacy of public health interventions, as well as the accuracy of mathematical modeling projections, because they may have failed to capture underlying drivers of risk taking and heterogeneity.

Separately addressing specific groups of people who share similar attitudes toward risk taking might increase the impact of interventions aimed at preventing exposure to STI and reducing their transmission. Recognizing such attitudes in individuals may be critical to design interventions that are more effective [12, 13]. For instance, previous modeling of individual decision making considering game theory found that emerging behaviors in populations are highly sensitive to the value/worth that individuals attribute to condomless sex with human immunodeficiency virus (HIV)-serodiscordant partners [14]. Yet assumptions regarding risk taking have often been based on individuals’ sociodemographic characteristics [15–17]. A priori categorizations are subject to stereotypical representations of people. Although these approaches sometimes succeed in identifying predictors of sexual behavior, it is less clear to what extent such characteristics drive decisions regarding sexual behavior.
informed consent had been obtained from all participants. Written consent was obtained from all participating institutions (Kantonale Ethikkommission Bern, Ethikkommission des Kantons St. Gallen, Comité départemental éthique des spécialités médicales et de médecine communautaire et de premier recours, Hôpitaux Cantonale de Genève, Kantonale Ethikkommission Zürich, Repubblica e Cantone Ticino—Comité Éthico Cantonale, Commission cantonale d'éthique de la recherche sur l'être humain, Canton de Vaud, Lausanne, Ethikkommission beider Basel for the SHCS and Kantonale Ethikkommission Zürich for the ZPHI). Written informed consent had been obtained from all participants.

METHODS
Systematic Data Collection on Sexual Behavior in the SHCS
The SHCS (www.shcs.ch) is a nationwide prospective cohort that routinely collects behavioral, laboratory, and clinical data from HIV-positive persons aged ≥16 years since 1988. Individual data are recorded at study entry and every 3 months thereafter. We estimate that more than 80% of all MSM currently diagnosed with HIV in Switzerland are followed in the cohort [8, 22].

The SHCS has been approved by the ethics committees of the participating institutions (Kantonale Ethikkommission Bern, Ethikkommission des Kantons St. Gallen, Comité départemental éthique des spécialités médicales et de médecine communautaire et de premier recours, Hôpitaux Cantonale de Genève, Kantonale Ethikkommission Zürich, Repubblica e Cantone Ticino—Comité Éthico Cantonale, Commission cantonale d'éthique de la recherche sur l'être humain, Canton de Vaud, Lausanne, Ethikkommission beider Basel for the SHCS and Kantonale Ethikkommission Zürich for the ZPHI). Written informed consent had been obtained from all participants.

Behavioral Matrix
This matrix contains trajectories of nsCAI over time for each patient. A binary variable determined the “nsCAI state” of each patient at time of interviewing (ie, engaging in nsCAI: yes or no). The succession of such states defined a patient’s nsCAI trajectory. The state was determined according to the patient’s answers to the questions: (i) Did you have sex with occasional partners in the last 6 months?; if yes, (ii) Did you have anal intercourse with these partners?; if yes, (iii) Did you use condoms all the time?. When changes in the nsCAI state were recorded, it was flipped at the midpoint between the 2 discrepant registries. This status only changed based on new information, that is, a patient who had been determined to engage in nsCAI was assumed to do so until something else could be determined based on a new interview. At least 1 report of nsCAI, a minimum of 2 nsCAI assessment records, and a follow-up time of 2 years or more were required for inclusion in these analyses.

Individual nsCAI trajectories were combined into a standardized matrix (behavioral matrix), which we constructed by transforming them using piecewise functions. The behavioral matrix recorded binary patients’ statuses over standardized 3-monthly updated intervals with a fixed time span (year 2000 to the first quarter of year 2018). Consequently, the behavioral matrix was sized (no. of persons × no. of time periods). The matrix only had valid entries for the periods where the patient in question was under follow-up and the outcome of the nsCAI assessment conclusive.

Inference of Behavioral Clusters
Based on the notion that similar trajectories of nsCAI may indicate concordant behavioral patterns, we inferred clusters of nsCAI trajectories (as recorded in the behavioural matrix) by means of agglomerative hierarchical clustering. We used a binary metric (ie, proportion of discordant bits, also known as Jaccard distance) for computing the distance matrix and the Ward method as agglomeration criterion [23].

To increase the likelihood of successful agglomeration in the status matrix, we considered information from mid-2001 (as sexual behavior questionnaires were first introduced in 2000). We utilized the R function hclust [24] to produce the clusters presented in this article.

Discriminatory Power of Sociodemographic Characteristics
We investigated whether a set of sociodemographic variables available from the cohort data could predict to which behavioral cluster a patient would belong. We simultaneously used boosted decision trees, k-nearest neighbors, and maximum likelihood methods to seek for associations between cluster membership and a set of sociodemographic variables including age, (calendar) time of registration, origin, and education level. We summarized the aggregated outcome of these analyses with receiver operating characteristic (ROC) curves. The toolkit for multivariate data analysis with ROOT (TMVA; https://root.cern.ch/ http://tmva.sourceforge.net/) served as the main tool for these analyses. In analogy with the “particle discovery” problem, we refer to background rejection and signal efficiency in this context as successful rejection of persons not belonging to a behavioral cluster and successful classification of persons belonging to a behavioral cluster, respectively.

Finally, we implemented a toy Monte Carlo “permutation test” for the significance (measured as a P value) of the classification based on sociodemographic variables. This test consisted of the iterative assigning of cluster labels to individuals at random, while conserving clusters’ size and computing the corresponding (simulated) ROC curves. We did this 3000 times for each cluster label and estimated the resulting P values as the number of iterations in which the simulated area under the...
ROC curve (AUC ROC) equaled or exceeded that of the original classification.

Computations were implemented in R (version 3.4.4), TMVA (version 4.2, running under ROOT version: 6.14), and python (version 3.7, libraries: pandas [25], numpy [26], scipy [27]). All software ran under Arch Linux x86_64 4.18.1.

RESULTS

Of 6025 HIV-diagnosed MSM ever active in the SHCS between 2000 and May 2018, 2766 reported nsCAI at least once. Of those, 2539 had at least 2 years of follow-up and 2 or more nsCAI data records. The analyses in this article were based on the latter group. Patients’ median age at enrollment in the cohort was 36 years (interquartile range [IQR]: 30–43), median follow-up over the study period was 10 years (IQR: 6–16 years), and 44% (1105/2539) reported (ever) using of recreational drugs. Table 1 shows characteristics of the patients included in the analyses.

Behavioral Matrix and Aggregated Trends

Figure 1A displays individual nsCAI trajectories. The median distance between trajectories defined by a binary metric was 0.8 (IQR: 0.4–1.0). The overall fraction of patients with nsCAI increased from 20% (95% confidence interval [CI]: 15–24%) in 2001 to 67% (95% CI: 64–70%) in 2018 (Figure 1B).

Distinct Behavioral Clusters

We assessed nsCAI trajectories for the top 4 hierarchie’s clusters (Figure 2A). In this case, this was equivalent to a cut at 1/3 of the full dendrogram height (Supplementary Figure S1). Figure 3 shows exemplary trajectories of nsCAI randomly retrieved from each of the resulting clusters.

We found dissimilar behaviors across clusters (colored in Figures 2A and 2B). All clusters displayed increasing nsCAI over the study period, and cluster 2 (blue) was the only one without a net increase in nsCAI between 2014 and 2018 (Figure 2B). Until about 2006, nsCAI trends for all clusters overlapped (Figure 2B). Cluster 4 (red) was the first to distinguish itself from the others, with consistently lower nsCAI until 2016 (Figure 2B). Median pairwise distances were 6.0 (IQR: 5.1–7.1), 4.9 (IQR: 3.2–5.6), 2.9 (IQR: 1.3–3.5), and 5.4 (IQR: 2.3–5.4) for clusters 1 to 4, respectively.

In the largest cluster (cluster 1, violet, 53% of total number of included patients) the fraction with nsCAI varied the least, going from 25% (95% CI: 19–31%) in 2001 to 43% (95% CI: 39–47%) in 2018 (Figure 2B). This test assessed the significance of the classification based on sociodemographic variables. $P$ values estimated through the permutation test ranged between .001 and .73 for cluster 2, and between .001 and .4 for cluster 4. This suggests no association between cluster membership and sociodemographic variables in these 2 clusters. In the same test for clusters 1 and 3, all simulated AUC were below those of the original classification. This constitutes

<table>
<thead>
<tr>
<th>Table 1. Characteristics of Patients Included in the Analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of MSM included</td>
</tr>
<tr>
<td>Age at registration in the SHCS (y, median [IQR])</td>
</tr>
<tr>
<td>Year of registration in the SHCS (y, median [IQR])</td>
</tr>
<tr>
<td>ART ever started (n [%])</td>
</tr>
<tr>
<td>Use of recreational drugs during follow-up* (n [%])</td>
</tr>
<tr>
<td>Origin</td>
</tr>
<tr>
<td>Education**</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: ART, antiretroviral therapy; IQR, interquartile range; MSM, men who have sex with men; SHCS, Swiss HIV Cohort Study.

*Includes injected and not injected heroin, cocaine, and others but excludes cannabis.

**Low: mandatory school or lower; High: university.
the strongest association we found between cluster membership and sociodemographic variables.

To identify potential prominent features, and in addition to the above described use of 3 machine learning algorithms, we directly examined the distributions of sociodemographic variables across clusters and their correlation, which are shown in Figure 6. Note that in Figure 6f, the distribution of year of registration in the cohort has separated peaks for clusters 1 and 3. Members of cluster 3 (green) had the latest average registration date across clusters, which suggests more recent HIV infections. By contrast, members of cluster 1 (violet) had the earliest average registration date across clusters. As depicted above, clusters 1 and 3 (violet and green) also yielded the strongest discriminatory power of sociodemographic variables. Neither visual inspection nor statistical tests indicated further prominent differences between clusters regarding variable distribution/correlation.

**DISCUSSION**

Behavioral clustering purely based on individual trajectories of nsCAI in MSM suggest that the continuous overall increase in nsCAI observed over the last 18 years was the consequence...
Figure 3. Exemplary nsCAI trajectories and cluster membership. Each horizontal line represents the trajectory of an individual. White: not under follow-up or invalid behavioural questionnaire record; green: no nsCAI; blue: nsCAI. Featured trajectories were selected at random. Hint: The estimated trajectories indicate that patient C1_P1 did not engage in nsCAI until 2011, from when he continued without interruptions. Patient C4_P2 did not engage in nsCAI until 2002, interrupted it in 2005, and engaged in nsCAI again in 2017. Abbreviation: nsCAI, condomless anal intercourse with nonsteady partners.

Figure 4. ROC curves. Variables tested for discriminatory power were education, origin, year of registration in the cohort, and age. ROC curve interpretation hint: larger areas between the ROC curve and the diagonal (gray) line indicate discriminatory power of the assessed variables. Abbreviations: BDT, boosted decision trees; kNN, k-nearest neighbors; likelihood, maximum likelihood; MVA, multivariate analysis; PCA, principal component analysis; ROC, receiver operating characteristic.
Figure 5. Outcomes of the toy Monte Carlo permutations tests comparing the area under the ROC (ROC AUC) displayed in Figure 4 (red) with that of 3000 runs with clusters labels assigned at random (blue). Abbreviations: AUC, area under the curve; BDT, boosted decision trees; kNN, k-nearest neighbors; likelihood, maximum likelihood, PCA, principal component analysis; ROC, receiver operating characteristic.

Figure 6. Configuration and correlations of sociodemographic variables across clusters. The diagonal (A, F, K, P) shows smoothed histograms, the rest of the panels shows correlations between these variables. Age: in years in 2018. Origin labels: 0 (other), 1 (white), 2 (black), 3 (Hispano-American), 4 (Asian), 5 (unknown). Education labels: 0 (Low), 1(Intermediate), 2 (High).
of collective, yet heterogeneous behaviors. These included drastic changes occurring over time periods that differed across clusters (Figure 2). The overall increase in this practice over the last 18 years is largely attributable to 2 behavioral clusters accounting for a third of the population (clusters 2 and 3). The largest behavioral cluster contained 53% of the study population and displayed the most stable behavior over time. Sociodemographic variables were predictive of cluster membership for 2 behavioral clusters containing 70% of the study population but not for the remaining 2 clusters, which displayed the most drastic changes over time.

To the best of our knowledge, this is the first published study to infer risk groups and to depict behavioral trends based purely on trajectories of sexual practices (nsCAI in this case). It did not assume a priori that persons’ characteristics such as age, origin, level of education, or year of HIV diagnosis explained their choices regarding sex. The availability of longitudinal, long-term records of nsCAI, which the SHCS has collected for almost 18 years, enabled this analysis.

The method outlined here is intuitive. Grounding the analyses on nsCAI alone allowed a compact presentation of the method and facilitated results interpretation. Although considering only 1 dynamic variable (nsCAI) may be seen as a limitation, this variable, often recorded in studies on sexual behavior, has shown to be a powerful predictor of other sexual behaviors and of STI transmission [4, 5, 7, 28, 29]. Adapting this method to include more variables is straightforward, and it is suitable for any setting with available longitudinal data on nsCAI or other quantities including clinical outcomes. We evaluated for discriminatory power variables available from the SHCS data, which are most commonly recorded in other longitudinal studies. However, we cannot exclude the existence of more predictive yet unmeasured sociodemographic characteristics. Moreover, this approach can be used within an explanatory mixed method research design [30, 31]: Hypotheses generated based on clusters could be explained by qualitative data that provide more comprehensive insights into behavioral change. Suitable methods alternative to our algorithmic approach (ie, hierarchical clustering) include model-based clustering such as longitudinal latent class analyses (LLCA). However, the outcome of hierarchical clustering offers an in-depth view of the risk structure of the population, and unlike LLCA, algorithmic clustering does not require model fitting, or hypothesis regarding data structure that could constrain the classification.

This study does not deal with the attribution of specific underlying mechanisms that bound patients within clusters together. A possible explanation for the behavioral patterns depicted in this article is that patients sharing behavioral clusters respond and adapt similarly to external information such as messages from the media, public campaigns aimed at reducing exposure to STI transmission, healthcare provider information, and scientific releases. For example, the acceleration in nsCAI in cluster 2 (Figure 2B, blue) coincided with the diffusion of the Swiss statement (part of a publication by Swiss researchers, which stated that people with HIV were not infectious if they were on effective antiretroviral therapy for at least 6 months and without any STI) [32]. This concept is closely related to the U(undetectable) = U(untransmissible) message [33, 34], which has been widely supported by subsequent studies [35]. The sharp rise of nsCAI in cluster 4 in 2017 may be associated with awareness resulting from the publication of landmark studies confirming the efficacy of preexposure prophylaxis (PrEP; from 2015 onward [36–38]) [39, 40] and with the rapid spread of chemsex (sexual activity under the influence of stimulant drugs such as methamphetamine, mephedrone, gamma-hydroxybutyrat/gamma-butyrolacton, or Ketamine) in Switzerland [41]. Importantly, decreasing condom use following the rollout of PrEP has been documented [40]. We believe the method outlined in this article could help identify triggers of behavioral change. Of note, the remarks in this paragraph are of a hypothetical nature, and proving or disproving them is beyond the scope of this article. A further study aimed at assessing these hypotheses by means of qualitative research is warranted.

Finally, STI propagate along sexual networks [42]. But sexual behavior may change unevenly within sexual networks if individuals sharing a sexual network do not share decision-making mechanisms and sexual behaviors [43]. We therefore think that the method depicted in this article is complementary to those concerning the characterization of contact structures (eg, inferred from transmission networks) [44–46].

In summary, we identified behavioral clusters based purely on the recognition of similarities and differences in longitudinal patterns of change in nsCAI. The method we proposed could help identify key target populations for behavioral interventions and meaningful risk groups for modeling of STIs. Both are key to achieving optimal allocation of resources to fight STI transmission. Available sociodemographic variables were found to be good predictors of behavioral clustering for the majority of the population but not for those men who displayed the most drastic changes in sexual behavior over time. A complete identification of such risk groups will require characterizing patients within clusters. For that purpose, and to understand drivers of changes in sexual behavior within clusters, further mixed methods studies combining quantitative and qualitative research are warranted.

**Supplementary Data**

Supplementary materials are available at Clinical Infectious Diseases online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

**Notes**

*Members of the Swiss HIV Cohort Study (SHCS).* Anagnostopoulos A, Battegay M, Bernasconi E, Boni J, Braun DL, Bucher HC, Calmy A, Cavassini M, Ciuflif A, Dollenmaier G, Egger M, Elzi L, Fehr J, Fellay J, Furrer H (Chairman of the Clinical and Laboratory Committee), Fux CA, Günthard HF (President of the SHCS), Haerry D (deputy of "Positive

Author contributions. L. S. V. and A. R. designed the study. L. S. V. and G. C. M. performed the analyses. L. S. V. and A. R. prepared the first draft manuscript, which was revised by all coauthors. G. W., D. L. B., J. E., K. E. A. D., E. B. P., H. F. G., and A. R. contributed to data acquisition. All authors contributed to the interpretation of analyses outcomes.

Acknowledgments. The authors thank all patients, physicians, and nurses associated with the SHCS.

Disclaimer. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the article.

Financial support. This work was performed within the framework of the SHCS, supported by the Swiss National Science Foundation (SNSF grants 177499 and 324730_179567) and by the SHCS grant 823. D. L. B. was supported by the University of Zurich’s Clinical Research Priority Program Viral Infections Diseases. R. D. K. was supported by the SNSF (grant 310030_155851).

Potential conflicts of interest. H. F. G. has received unrestricted research grants from Gilead Sciences, Roche; a grant from Systems X; grants from SHCS research foundation, National Institutes of Health, Yvonne Jacob Foundation, personal fees from Data Safety Monitoring Board, and consulting for Merck, Gilead, Teva, and Sandoz, outside the submitted work. A. R. reports support for advisory boards and/or travel grants from Janssen-Cilag, Merck Sharp & Dohme (MSD), Gilead Sciences, Abbvie, and Bristol-Myers Squibb, and an unrestricted research grant from Gilead Sciences. All remuneration went to his home institution and not to A. R. personally. D. L. B. reports support for advisory boards and/or travel grants from MSD, ViV, and Gilead Sciences. R. D. K. reports personal fees from Gilead Sciences. J. F. is a member of the Federal Commission for Sexual Health and reports support for advisory boards and/or travel grants from Janssen-Cilag, MSD, Gilead Sciences, Abbvie, Merck, ViV Healthcare and Bristol-Myers Squibb, and an unrestricted research grant from Gilead Sciences. All remuneration went to his home institution and not to J. F. personally. G. W. reports grants from Gilead Sciences and AbbVie. A. J. S. reports part-time employment in a governmental agency (Swiss Federal Office of Public Health). K. E. A. D. reports contributions to her home institution from Gilead Sciences and grants from Merck Sharp and Dohme. E. B. reports payments to his home institution from Gilead Sciences, MSD, ViV Healthcare, Pfizer, and Sandoz, and grants from Gilead Sciences, MSD, ViV, and Pfizer. All these relationships were outside this work. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that these relationships were outside this work. All authors have submitted the

References

23. Rumesburg C. Cluster analysis for researchers: Lulu. com. Lulu Press; 2004. Available at: https://scholar.google.ch/scholar?hl=de&as_sdt=0,5&q=Cluster+analysis&for_researchers=Rumesburg%202004&gs_sm%3AeE2vRPLWifoJ%3Ascholar.google.com%2F%2Foutput%3Dcite&scirp%3D%006b%3E3Dde


