MEETING TARGETS AND MAINTAINING EPIDEMIC CONTROL (EPIC) PROJECT

COOPERATIVE AGREEMENT NO.
7200AA19CA00002

A guide to client risk segmentation to optimize the impact of HIV programming

Suggested citation: Cassell MM. A guide to client risk segmentation to optimize the impact of HIV programming. Durham (NC): FHI 360; 2021.

ACKNOWLEDGMENTS

This guide was developed and drafted by Michael Cassell. Inputs and comments from FHI 360 and EpiC staff are gratefully acknowledged, with particular thanks to Pai-Lien Chen, Christa Fischer Walker, Kayla Stankevitz, Chris Akolo, Meghan DiCarlo, and Navindra Persaud. Without the generous engagement and experience sharing of the EpiC Indonesia, Thailand, and Vietnam teams, this work would not have been possible. This guide was edited by Stevie Daniels.

This resource was made possible by the generous support of the American people through the United States Agency for International Development (USAID) and the U.S. President's Emergency Plan for AIDS Relief (PEPFAR). The contents are the responsibility of the EpiC project and do not necessarily reflect the views of USAID, PEPFAR, or the United States Government.

EpiC is a global cooperative agreement (7200AA19CA00002) led by FHI 360 with core partners Right to Care, Palladium International, Population Services International (PSI), and Gobee Group.

Contents

Preface

This guide aims to orient HIV program implementers to some basic principles and practices they can apply to improve individual client and overall program outcomes through enhanced use of routinely collected program data to inform client-centered differentiation of HIV services.

Programs have traditionally relied on epidemiological studies to identify the characteristics of priority clients, but these studies are sometimes challenged to supply the timely, context-relevant information needed to support continuous HIV service adaptation and improvement based on client experiences and needs. Conversely, as coverage of lifesaving HIV prevention and treatment services becomes more inclusive, client experiences contain increasingly valuable clues about how best to close outstanding access gaps. Program data are not typically considered to be representative of unreached populations, but the potential to apply program data to distinguish continuously between individuals who progress across the HIV service continuum, and those who do not, may increase as coverage expands.

Opportunities to conduct more granular analyses to direct and drive program improvement have also expanded with improvements over time in the quality and detail of routinely available program data. Unfortunately, these expanding opportunities can sometimes themselves serve as barriers to data use. Among other challenges, concerns about analytic complexity, available time, relative benefits, and opportunity costs can all prevent HIV service providers from turning to the data they have in hand as a compass to direct action.

To help overcome some of these barriers, this guide is an illustrative road map for analysis and application of routine program data that offers providers low-threshold entry points, while illuminating opportunities to build upon these foundations to enhance the complexity and utility of analyses over time. As the primary objective of the described analysis strategies is the evolution and differentiation of services to be more responsive to clients, this guide places major emphasis on data *use*. Indeed, a primary objective of the client risk segmentation approach is the enhanced integration of data analysis and use into a continuous quality improvement cycle that benefits clients.

Some of the strategies described herein can be initiated by just one committed individual, but all stand to benefit from engagement of a broader, multidisciplinary team. As the potential benefits of invoking more complex analyses are revealed, it may become necessary to identify statistical experts who can support their rigorous pursuit. Similarly, clients and community members will have crucial insights to share on how programs can or should be adapted to address issues identified as part of more detailed analyses of program data.

In summary, client risk segmentation is intended to reflect an inclusive process for continuously asking and answering two critical questions about HIV service delivery:

1. *Who is being left behind?*
2. *What can we do better to serve these individuals?*

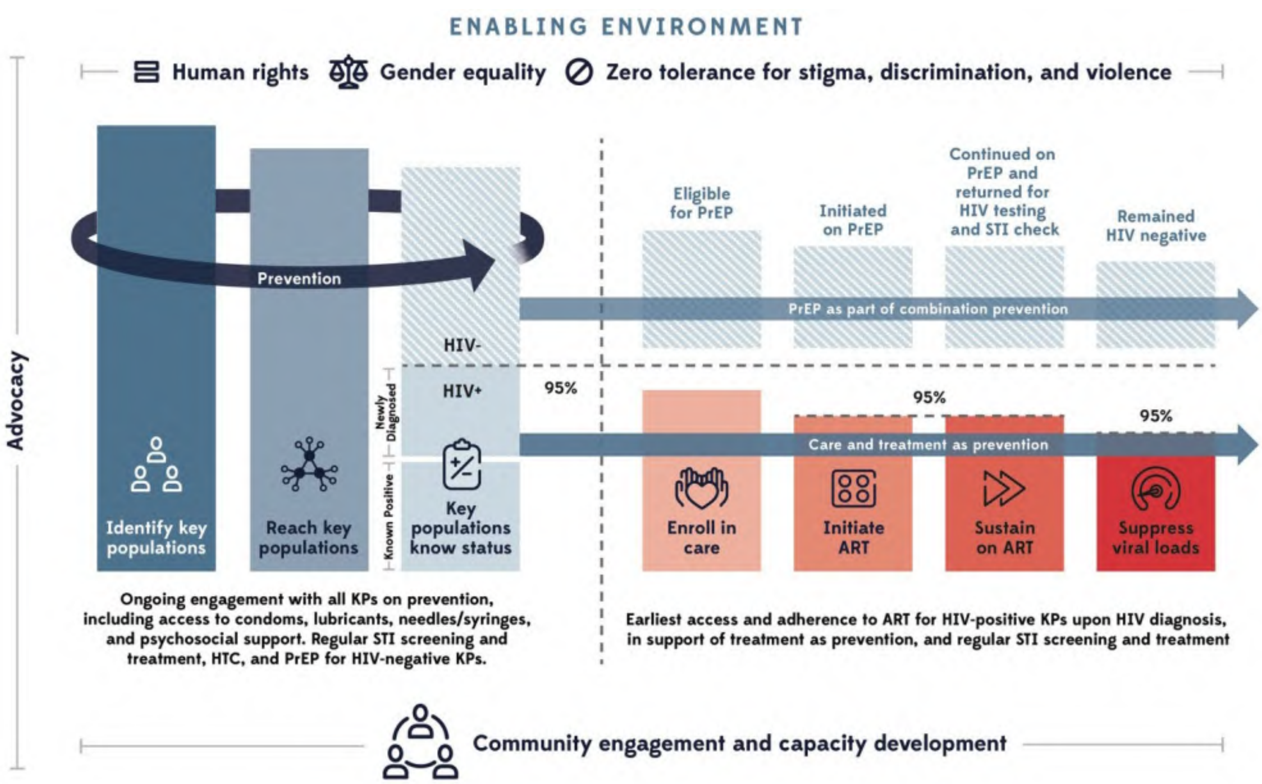
Ideally, this guide will serve as a useful entry point for much longer and more detailed conversations about these topics, which are informed by current program data.

Background

More than four decades into the global response to HIV, advances in science and innovations in practice have together placed HIV epidemic control within reach. Antiretroviral (ARV) medications have been proven to prevent HIV-related illness and death — as well as ongoing transmission of the virus. HIV pre-exposure prophylaxis (PrEP), correct and consistent condom use, and several other effective resources exist to prevent HIV infection.

An HIV cascade framework, like the one depicted in Figure 1, is often used to track progress toward public-health impact by documenting coverage of proven prevention and treatment services among individuals facing the greatest HIV infection risks. By responding to the differentiated preferences and needs of un- and underserved-individuals, programs can help to raise the bars across the cascade and plausibly achieve HIV epidemic control and an end to the

Figure 1. The EpiC HIV Cascade Framework



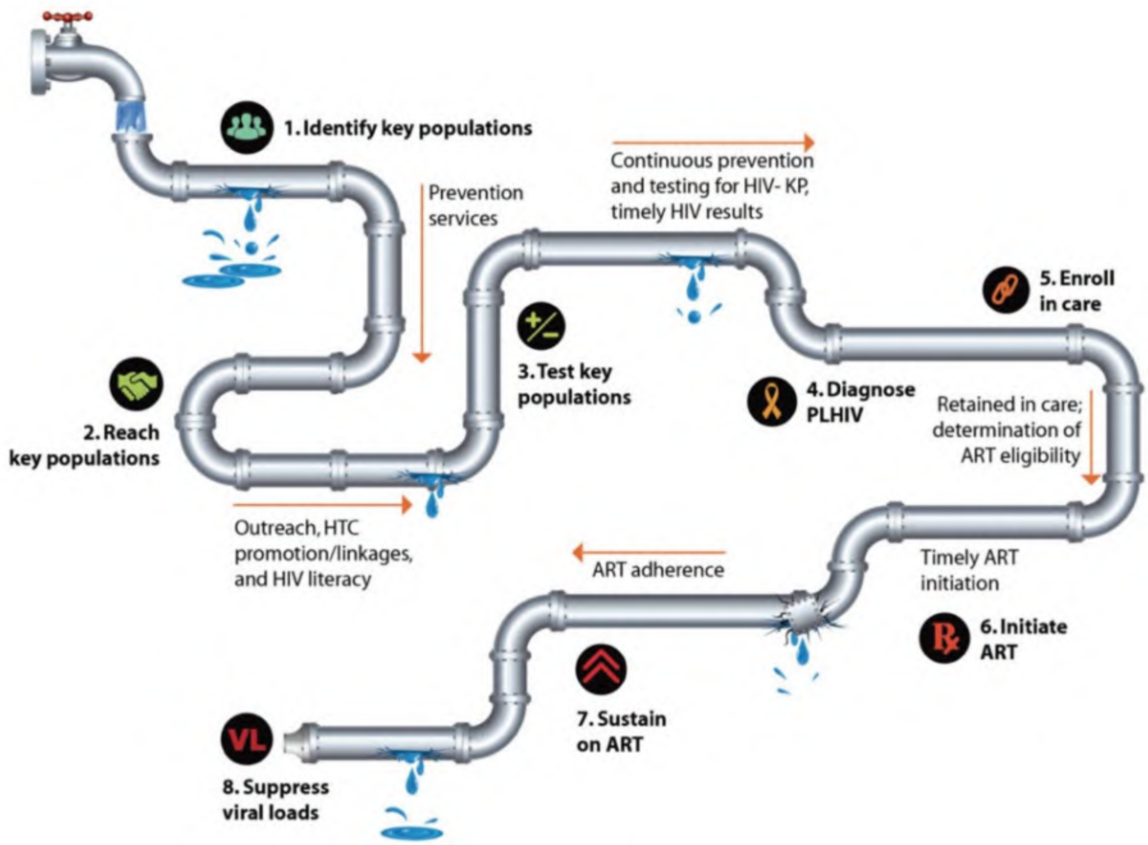
suffering associated with AIDS.

A review of current cascade progress reveals that almost 38 million people globally are living with HIV, 84 percent of whom now know their HIV status.[1] However, only about 73 percent of

[1] UNAIDS. Global commitments, local action: after 40 years of AIDS, charting a course to end the pandemic. Geneva: UNAIDS, 2021.

people living with HIV (PLHIV) are receiving lifesaving HIV treatment, and about 1.5 million people are still newly infected each year. To achieve global targets set for 2025, we need to support treatment access for an additional 6.5 million people, prevent an additional 1.1 million people from becoming HIV infected annually, and prevent almost 440,000 additional AIDS-related deaths each year.[1] Despite substantial global progress, the cascade reality can resemble a system of leaky pipes, as depicted in Figure 2.[2] At critical junctures from initial identification of individuals with elevated risks all the way to sustained service access and improved health outcomes, certain individuals are unfortunately falling out of care or are never engaged. For many individuals, engagement in HIV cascade services is nonlinear, reflecting a dynamic cycle of barriers and facilitators to engagement that are constantly evolving in the face of personal, social, and structural influences.[3]

Figure 2. Leaks in Progress across the Cascade



[2] Adapted from: Kilmarx PH, Mutasa-Apollo T. Patching a leaky pipe: the cascade of HIV care. Curr Opin HIV AIDS. 2013 Jan;8(1):59-64.

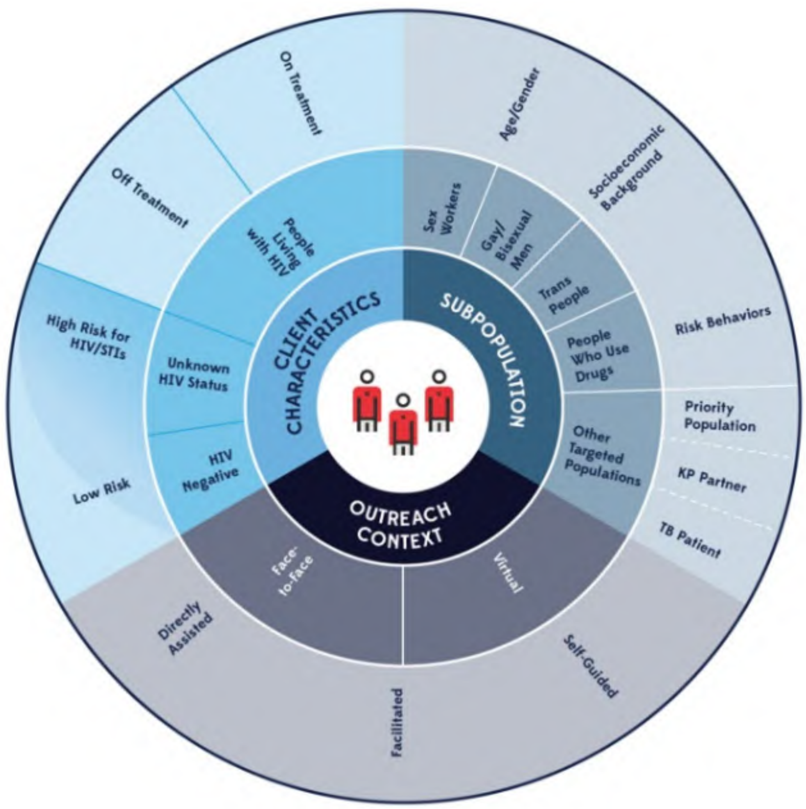
[3] Ehrenkranz P, Rosen S, Boulle A, Eaton JW, Ford N, Fox MP, et al. (2021) The revolving door of HIV care: revising the service delivery cascade to achieve the UNAIDS 95-95-95 goals. PLoS Med. 2021;18(5):e1003651. doi: 10.1371/journal.pmed.1003651.

To close these gaps and achieve and sustain HIV epidemic control, it is increasingly clear that efforts must be client-centered and focused on the specific experiences and needs of those who are currently being left behind.

The EpiC differentiated services delivery (DSD) model (Figure 3) aims to segment populations to optimize the allocation of time and resources according to individual preferences and needs.[4] Put simply, evidence — in the form of persistent gaps in the cascade — shows there are no "one size fits all solutions," and we need to differentiate service and support options based on the specific characteristics of clients facing greater risks. To do this, a far deeper appreciation of who these people are is needed.

As information systems have evolved to track and improve individual client and overall HIV cascade outcomes safely and accurately, so too have program opportunities and responsibilities to analyze routine program data to identity population segments and clients facing elevated risks.

Figure 3. EpiC Differentiated Services Delivery (DSD) Model



For example, programs can develop tailored and preferred service solutions by identifying the differentiating characteristics of clients who are more likely to receive positive results from HIV testing, not be linked to antiretroviral therapy (ART) following HIV diagnosis, not sustain access to treatment or achieve viral suppression, or not initiate or continue PrEP, and others like them.

[4] Grimsrud A, Bygrave H, Doherty M, et al. Reimagining HIV service delivery: the role of differentiated care from prevention to suppression. J Int AIDS Soc. 2016;19(1):21484.

Client risk segmentation: Making the link between data and practice

Client risk segmentation (client segmentation) helps programs focus effort and differentiate services by taking a granular look at the characteristics that distinguish *individuals* who meet certain HIV cascade criteria — such as being newly diagnosed, initiating HIV treatment, or achieving HIV viral suppression — from those that do not. By summarizing these differences, programs can optimize the focus and impact of their outreach, targeted testing, and **case management** efforts by devoting more time and resources to individuals with greater needs, and fewer to those who do not want or need additional support.

Typically, client segmentation will identify differences in terms of age, sex, risk behavior, education, media use, location, experiences of violence, and other characteristics that can provide important insights to guide programming. With this deeper understanding in hand, some examples of the way programs may then apply segmentation include:

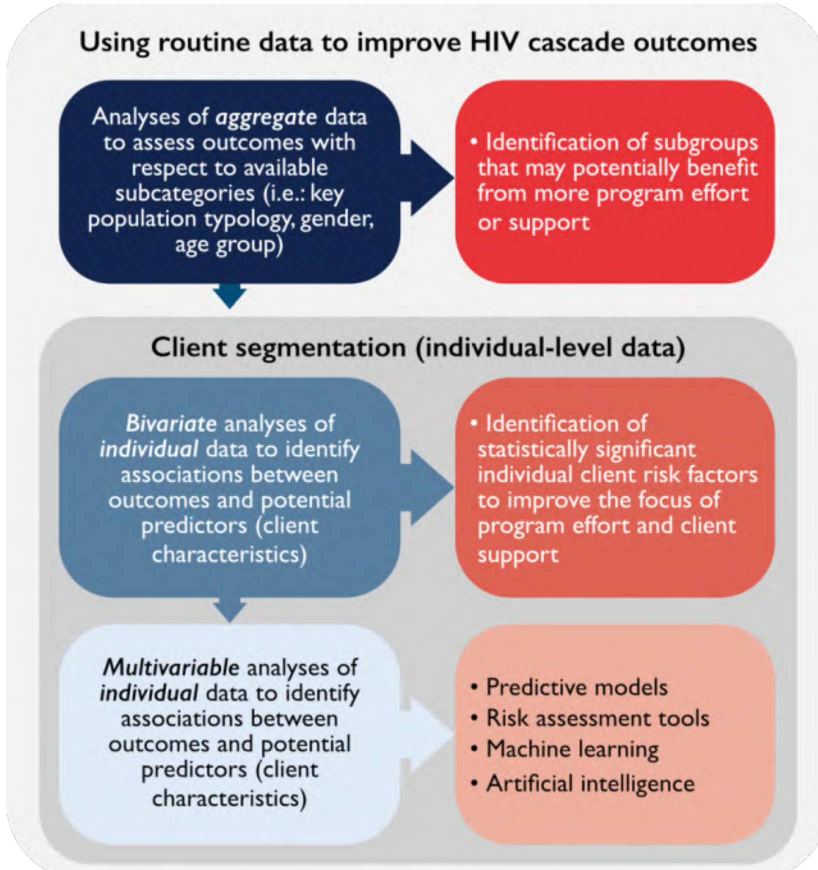
- Improving the targeting of testing, prevention, and treatment linkage and continuation services
- Engaging peer mobilizers, navigators, and supporters with characteristics that are similar to those of individuals facing the greatest risks
- Optimizing provider resources including risk assessment tools, case management and client support protocols and operating procedures, and job aids to meet the needs of high-risk populations
- Designing client-facing educational and promotional materials to meet the needs of those at highest risk for treatment interruption and viral load failure

HOW CLIENT SEGMENTATION DIFFERS FROM OTHER APPLICATIONS OF ROUTINE PROGRAM DATA TO SUPPORT CONTINUOUS QUALITY IMPROVEMENT

Most programs already conduct rigorous reviews of program data and achievements to identify opportunities for program improvement. Indeed, under the U.S. President's Emergency Plan for AIDS Relief (PEPFAR), supported partners are required to disaggregate achievements by key and priority population typology, gender, and age group, and are expected to investigate and address varying levels of achievements by site.

PEPFAR has established a "minimum standard" expectation that all programs employ confidential, client-specific unique identifying codes (UICs) to improve tracking and support of individual clients across the HIV cascade. The expanding availability of individual-level UIC data in routine data systems unlocks the potential of programs to conduct much more granular and potentially impactful analyses to identify characteristics of clients with the greatest needs. As depicted in Figure 4, the move from aggregate-level analysis to individual-level analysis is what primarily distinguishes client segmentation from other efforts to use routine data to drive program improvement. Programs can conduct *bivariate* analyses to identify sociodemographic, behavioral, and other client characteristics that are associated with a significantly higher likelihood of a specific HIV cascade outcome of interest — such as receiving a positive-HIV test result or experiencing interruption in treatment (IIT). With this information in hand, programs can adapt their programming and prioritize support for current and prospective clients with these characteristics.

Figure 4. Distinguishing Client Segmentation from Aggregate Analyses of Routine Program Data



Identification of a set of client characteristics having bivariate associations with an outcome of interest enables programs to develop and validate multivariable models that predict the likelihood of that outcome occurring based on multiple client-level variables. *Multivariable* analyses are not an essential component of client segmentation, but enable more sophisticated applications of risk assessment tools, machine learning, and artificial intelligence. The use of multivariable models allows teams to segment client populations into different categories based on the likelihood of clients experiencing an outcome and the observable characteristics they have in common.

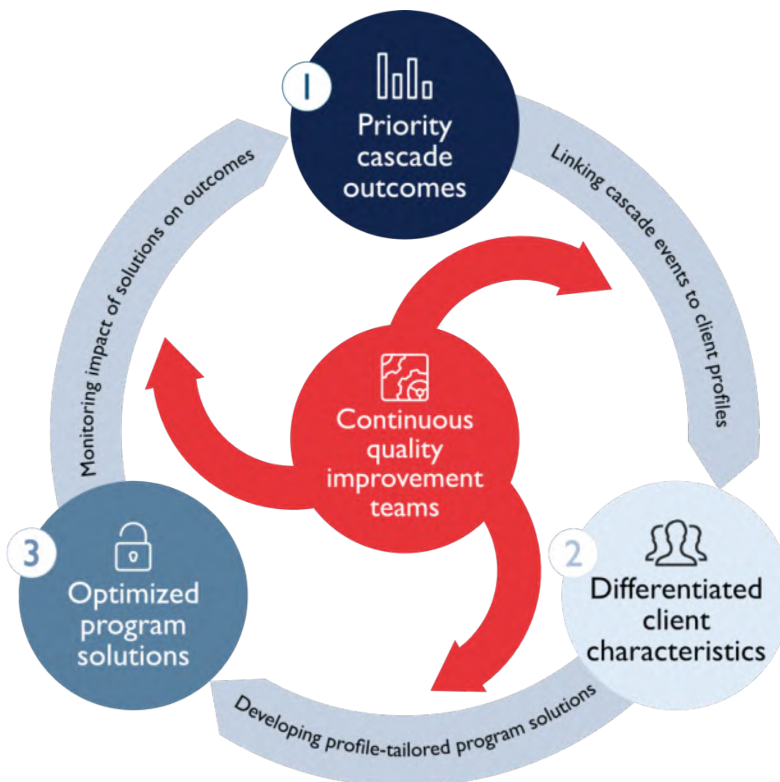
Machine learning is defined as a set of methods for getting computers to recognize patterns in data and use these patterns to make predictions. Computer programs (models) that can access and learn from data are developed. **Artificial intelligence** uses computers for making decisions or recommendations in an automated way. Automated decisions might be directly implemented or suggested to a human decision-maker. Artificial intelligence can be thought of as "smart automation," and machine learning as computer-generated "data-driven predictions." Machine learning is a subset of artificial intelligence and enables data-driven predictions to inform decisions.[5] As such, machine learning may be applied to support, extend, and sustain client segmentation, but is not a required or essential component.

Client segmentation steps

The primary objectives of client segmentation are to improve individual client and overall HIV cascade outcomes. For the approach to have an impact, the results must be *applied* by programmers to refine practice so that services are better catered to client preferences and needs.

Client segmentation focuses not only on data analysis, but also on implementation of a continuous quality improvement cycle in which analyses are linked to program actions, and the potential impacts of these actions are evaluated. Throughout this cycle, client segmentation generates critical insights in three core areas (Figure 5):

Figure 5. Implementing Client Segmentation in a Continuous Quality Improvement Cycle



1. *Identification of priority outcomes* — gaps or leaks in the cascade as evidenced by program performance issues.

[5] Managing Machine Learning Projects In International Development: A Practical Guide. USAID, Vital Wave, and DAI. https://www.usaid.gov/sites/default/files/documents/Vital_Wave_USAID-AIML-FieldGuide_FINAL_VERSION_1.pdf

2. *Identification of priority clients* — individuals with characteristics that are significantly associated with outcomes of interest, such as those who are more likely to have a positive HIV test result, or are more likely to experience IIT.
3. *Identification of priority solutions* — in particular, modifications to services or support to make these more responsive to the experiences, preferences, needs, and social and contextual circumstances of priority clients. Solutions should ideally consider issues of service access, convenience, availability, relevance, and "friendliness" toward clients, among other factors.

The introduction and implementation of client-centered solutions based on client segmentation can close access and overall program performance gaps while enabling programs to adapt to emerging needs. Indeed, service adaptations made as a result may have broader systemic benefits by virtue of freeing up staff and resources to devote less time to lower priority tasks and more time to newly identified actions. As historical issues are remedied and new priorities are identified, client segmentation can continuously provide fresh insight to fuel improvement. Client segmentation involves eight basic steps implemented by multidisciplinary teams. Figure 6 illustrates the organization of these steps in relation to the continuous improvement cycle, and Table 1 describes the steps in more detail.

Figure 6. Eight Client Segmentation Steps

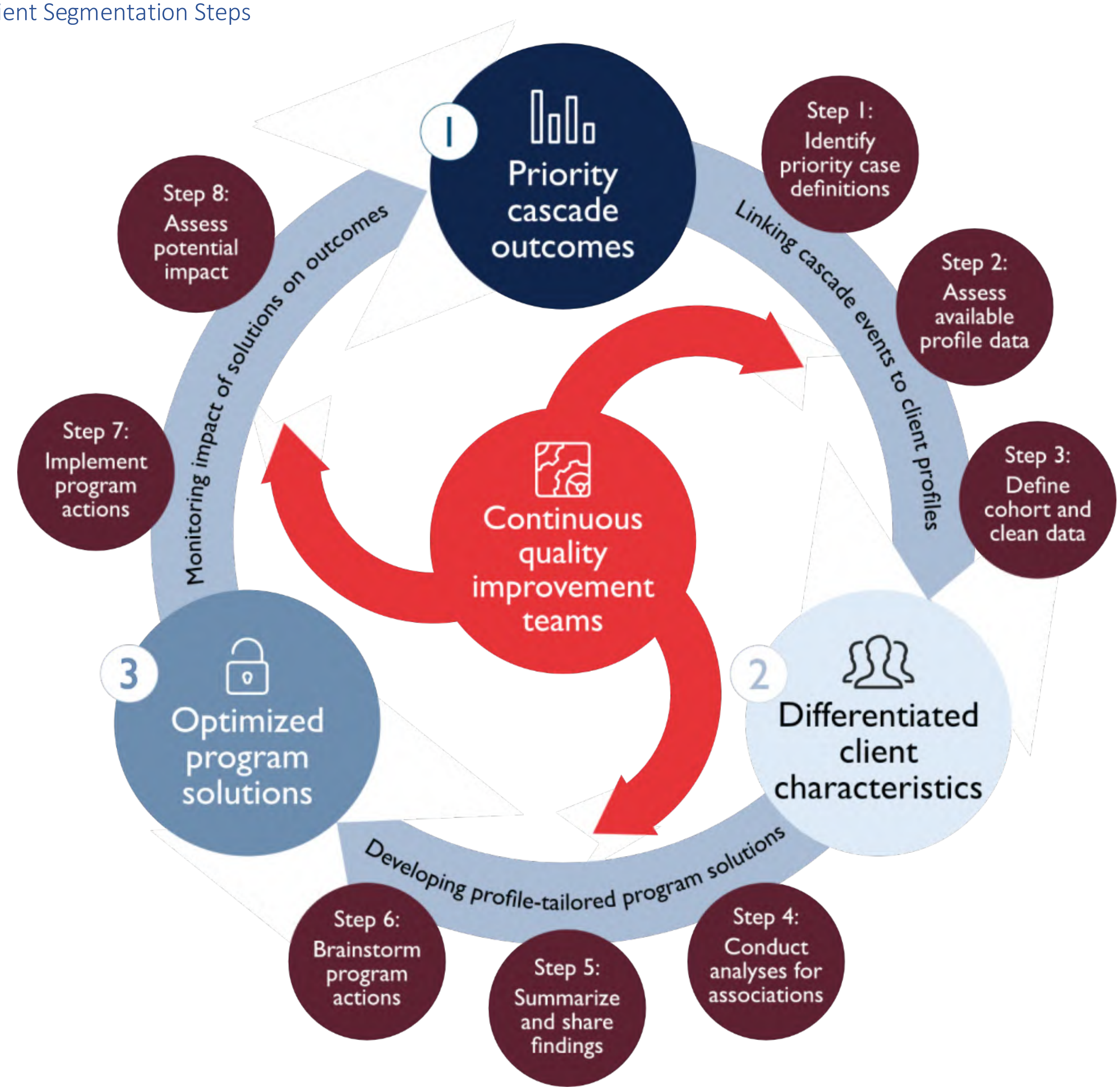


Table 1. Eight Illustrative Steps to Implement Client Segmentation

Step	Detail
1. Identify priority outcome(s)	Review current program data to identify performance challenges and priority outcomes for improvement across the HIV cascade. Possible examples include improving case-finding rates; reducing seroconversion rates; improving linkages to treatment initiation; reducing rates of IIT, morbidity, or death; and improving linkages to viral load testing or rates of viral load suppression.
2. Identify what individual client characteristics can be linked to the priority outcome(s)	Review sociodemographic, risk, and program exposure data stored in program databases. Determine which client characteristics can be linked by UIC to the outcome of interest for individual clients. Work with implementers to filter out characteristics that have limited or no potential to sharpen the focus of program effort or client support. The remaining list will serve as potential "predictors" in the client-segmentation analysis.
3. Define client cohort and generate a clean dataset	Define the client cohort as a specific group of individuals contributing outcome and possible predictor data based on their program experiences within a specific time period. For example, if a program wants to identify the characteristics of HIV testing clients who are more likely to receive a positive test result, it might define its cohort as all individuals who received an HIV test and their result within a specific period of implementation.
4. Develop and implement an analysis plan	Develop an analysis plan that will support quantification of the associations between possible predictors and the outcome(s) of interest. This may involve chi-square tests to identify significant bivariate associations, calculation of bivariate odds ratios to characterize the magnitude of significant associations, and the use of logistic regression for multivariable analyses.
5. Summarize and share findings	Identify and implement activities to summarize and share results of the analyses with implementers, community members, and policymakers. Consider dissemination of the findings through a variety of channels to maximize exposure such as workshops, webinars, interactive dashboards, and/or short briefs.
6. Brainstorm program actions	Work in close partnership with implementers, community members, and policymakers to develop a set of specific, prioritized, and timebound program improvement actions in response to the summarized analysis results. Teams may employ crowdsourcing or incentivized competitions to foster inclusive and active participation in the generation of client-segmentation-informed program pivots or solutions. Consider the implications of client segmentation for: ▪ Revisions to standard operating procedures and job aids ▪ Revisions to educational and promotional materials

Step	Detail
	▪ Targeting HIV testing and other services ▪ Differentiated client support and management ▪ Geographic and population prioritization ▪ Resource allocation In developing program actions, consider available evidence from the literature and from other implementation experience of good and promising practices that may be relevant and responsive to the client segmentation analysis findings. Develop strategies to monitor implementation and to assess potential impact on client or program-level outcomes (see step 8).
7. Implement program actions	Support the implementation of prioritized program improvement actions, ensuring that protocols, training, staffing, policies, and systems are in place to facilitate maximum coverage and potential impact. It is vital for programs to commit time and resources for these potential solutions. Because client segmentation focuses on client-centered prioritization, it is possible that implementation of the identified solutions will entail devoting less energy to lower priority historical tasks, freeing up resources to advance the newly identified actions.
8. Closely monitor and assess the impact of action plans	Document the process applied to conduct client segmentation, share results, and generate potential solutions; as well as the specific actions taken. Look for opportunities to monitor trends in performance that may change with the introduction of novel client-segmentation-inspired service approaches or adaptations. Assess changes over time in the profiles of clients who face the greatest risks, as these changes should occur as a result of the new actions.

In selecting one or more priority outcomes for client segmentation analyses, it may be useful to think in terms of cascade events that the program tracks routinely, and the ways in which these events can be used to identify high-priority clients.

Annex I depicts ways in which routine indicator and other data can be used to segment the client population. For example, if a program is challenged to engage previously undiagnosed PLHIV as part of its HIV testing services (HTS) efforts, taking a closer look at how the HTS clients who receive positive results differ from those who receive negative results may prove useful.

The same approach can be applied to a wide range of outcomes across the prevention-to-treatment cascade. To gain insights to improve treatment continuation performance, programs can assess the ways in clients who experience IIT differ from those who do not, or the ways in which clients with detectable viral loads differ from those who are virally suppressed. To improve engagement in PrEP or index testing, programs can identify differences in the characteristics of those who decline these services versus those who accept them.

A closer look at client-segmentation analyses

Client-segmentation is the analysis of relationships between outcomes of interest and client characteristics at the *individual* level.

Identifying the ways that clients experiencing specific outcomes differ from those who do not requires that programs have individual-level client characteristic data on members of both groups, and that both groups are included in the cohort for analysis. The approach to analysis of client-segmentation data is similar to what is sometimes referred to as a “risk factors” or “predictors” analysis and draws upon traditional epidemiological case-control analysis strategies.

To use these strategies with routine program data, careful consideration must be devoted to understanding the quality and components of these data. As noted in the client segmentation “steps,” it is essential to identify what opportunities exist to link outcomes of interest to specific client characteristics at the individual level before diving into any detailed analyses.

One way to accomplish this is to “map” outcomes of interest to available client characteristics as illustrated in the example table in Annex II. In this example, individual client characteristics that can be linked by UIC to outcomes of interest are checked. This “map” can be used to prioritize analyses of potential “predictors” of each outcome of interest based on consideration of the relevance of these predictors to opportunities for program improvement.

After identifying one or more outcomes and associated potential predictors, the program must define the client cohort for analysis. A cohort should be representative of the *client* population — taking care not to omit specific individuals or groups of clients in a manner that may skew the analysis outcomes. The cohort should be sufficiently large to facilitate the

Garbage In, Garbage Out

The results of any analysis can only be as good as the quality of the analyzed data. The client segmentation analyses described in this section follow a stepwise process (see Annex III) that entails assessments of data quality and only advances analyses to additional levels of complexity to the extent that the quality of the available data allow. Key initial considerations include the volume and completeness of observations in the program data. Generally, to observe simple associations between an outcome and a client characteristic, a dataset should include at least 50 individuals in each of the four possible categories of experiencing vs. not experiencing the outcome of interest, and exhibiting vs. not exhibiting the characteristic of interest. The larger the dataset, the more power a team will have to identify significant relationships between an outcome of interest and one or more client characteristics. Care must also be taken to minimize and account for missing information. Because the reasons data may be missing can be associated with the outcome of interest, analyses of indicators with many missing values can produce misleading results. At a minimum, these limitations should be considered in conducting analyses, and should be specified in sharing findings.

detection of significant differences in the characteristics of clients who experienced the outcome(s) of interest and those who did not. Ultimately, the cohort size needed to detect significant differences will vary according to the frequencies of different client characteristics and outcomes, as well as the strengths of the associations between these; so consultation with a data analysis expert is desirable in cohort selection. If a team has a concern that the client observations within a specific period may not be enough to detect differences, one thing to consider is expanding the period of observation to include more clients.

Based on the cohort definition, a clean dataset can be generated that includes unique rows for all individuals in the cohort, and columns for the client-specific values of the outcome of interest and the selected predictor variables. For this step, a strategy may be needed to deduplicate or merge client records.

For example, if a program is looking at HTS results among individuals tested in the past year, clients who tested more than once during this period may have multiple records in the database. The team will need to employ a decision rule about which record to use with respect to the outcome of interest (typically the most recent); which client characteristic data to use (also typically the most recent); and how to capture and integrate relevant information from other records (such as creating a column to reflect how many prior tests the individual has had during the period of interest prior to the most recent).

After cleaning the data, teams will develop a multi-level, stepwise analysis plan. As depicted in Annex III, these levels can generally be organized into:

1. *Univariate analyses* – generating descriptive statistics such as counts, missing data, range, median, mean, and standard deviation for each outcome and potential predictor.
2. *Bivariate analyses* – employing statistical tests to determine the association between unique potential predictors and the outcome(s) of interest.
3. *Multivariable analyses* – employing statistical tests to determine the simultaneous relative relationships between more than one potential predictor and the outcome of interest.

Should the univariate analyses highlight data quality and completeness issues that would call into question the value or interpretation of additional inquiry, teams should aim to address or resolve the underlying data issues prior to proceeding to conduct bivariate analyses. Similarly, teams should consider the utility and limitations of the findings from bivariate analyses prior to proceeding to conduct multivariable analyses.

UNIVARIATE ANALYSES

The primary function of the univariate analyses is to characterize the dataset, assessing frequencies of different outcome values and other client characteristics. Descriptive statistics on the outcome and each potential predictor variable quantify the number of clients falling into specific categories with respect to these variables, and identify the proportion of clients with missing values with respect to any given variable. A substantial number of missing values for any given variable can complicate or constrain the potential to conduct meaningful bivariate analyses.

To illustrate this point, consider a scenario in which a team is exploring associations between HIV testing outcomes and client characteristics among 1,000 HTS clients reached within a quarter: running descriptive statistics on the outcome of interest — HIV test results — reveals that 100 clients received a positive test result, 400 had a negative result, and 500 had tests but no recorded results. Because the results of 500 individuals are missing, it is impossible to assess bivariate associations between their other characteristics and their testing outcome. One potential solution is to exclude the individuals with missing data from the bivariate analyses. But, in this example, excluding half of the HTS client population would raise substantial questions about whether the associations observed for the half for which the team has outcome data would still hold true for the half with missing data. For example, if outcome data are missing because individuals with reactive HIV screening results experience delays in the reporting of confirmatory test results, dropping the missing clients can result in misleading interpretation of the data.

BIVARIATE ANALYSES

Once a determination has been made that the frequency of missing data — and particularly of missing outcome data — is unlikely to impact the relevance and interpretation of statistical associations between client characteristics and the outcome of interest, teams can proceed with bivariate analyses. As depicted in Annex III, a range of options is available for bivariate analysis, and these should be prioritized in consultation with data analysis specialists according to available analytic capacities, analytic platforms, and ease of interpretation. These include:

1. Cross-tabulations
2. Bivariate logistic regression with identified reference groups
3. Calculation of simplified odds ratios for binary predictors

Cross-tabulations

The cross-tabulations typically involve the use of chi-square tests for categorical variables to identify statistical associations between values of a potential predictor and the outcome of interest. Evidence of a statistical association is reflected in the test generating a p-value that is at or below a certain threshold. Put simply, a p-value reflects the likelihood of seeing the observed relationship between the outcome and the potential predictor if in fact no association exists. Typically, a p-value of <0.05 is considered as evidence of statistical significance. The interpretation is that there would be a less than five percent chance of experiencing the observed relationship between the outcome of interest and the potential predictor if we initially assume that no relationship exists.

For predictor values that are statistically associated with the outcome of interest, cross-tabulations also allow for the calculation of odds ratios, illustrating how many times more or less likely individuals with a specific predictor value are to experience the outcome of interest, *as compared to other individuals with an identified reference value for that predictor*.

For example, let's imagine that a team was interested in exploring the relationship between client district of residence and the likelihood of a positive HIV test result. If clients could come from one of five districts, the program would select one of these districts as a reference group and report the odds of a positive test result for clients residing in other districts relative to those residing in the reference district of residence.

For ordinal variables — categorical variables that have mathematical meaning — such as age-group categorizations, similar cross-tabulations can be accomplished using the Mantel-Haenszel test.

Bivariate logistic regression

All the analytic results generated from cross-tabulations — including p-values and odds ratios — can also be derived through bivariate logistic regression. While this approach is somewhat more complex than cross-tabulation in terms of underlying calculations, most statistical software can help teams generate logistic regression output with a few clicks. Put simply, logistic regression models the probability of an outcome occurring — or not — based on the values of one or more predictor variables. Bivariate logistic regression predicts the probability of an outcome based on the values for just one possible predictor such as district of residence in the example above.

With logistic regression, odds ratios can be calculated for values of both categorical and ordinal predictor variables with respect to a specified reference group. So, a bivariate analysis of testing outcomes by age groups could generate, for example, predictions of how much greater or lesser the likelihood of a positive test result is for members of the 20- to 24-year-old age group or the 25- to 29-year-old age group as compared to members of the 15- to 19-year-old age group.

Two noteworthy advantages of opting to pursue bivariate analyses via logistic regression versus cross-tabulation: First, the use of logistic regression for the bivariate analyses opens opportunities for a seamless transition into multivariable analysis, for which logistic regression is the preferred analysis strategy. Second, logistic regression generates a model that calculates the probability of a specific outcome based on predictor variable values. Having the capacity to generate such prediction calculations based on observed historical data affords opportunities to optimize risk screening tools and to pursue machine learning and artificial intelligence applications of client segmentation.

Simplified odds ratios

Before diving into more detail on approaches and applications of multivariable client-segmentation analyses, recalling the primary objectives of client segmentation is key. Teams are encouraged to reflect carefully on their bivariate findings and whether these may be sufficient to meet their quality improvement needs.

The intent of client segmentation is not to infer causal relationships, but rather to provide insights into the characteristics or preferences of clients facing elevated risks. If, for example, a program found that clients who said their favorite color was blue were twice as likely to drop out of care than clients with a different favorite color, we would not attribute their risk to the fact that they like blue but the program might well choose to incorporate a question about clients' favorite color in the next version of risk screening tools.

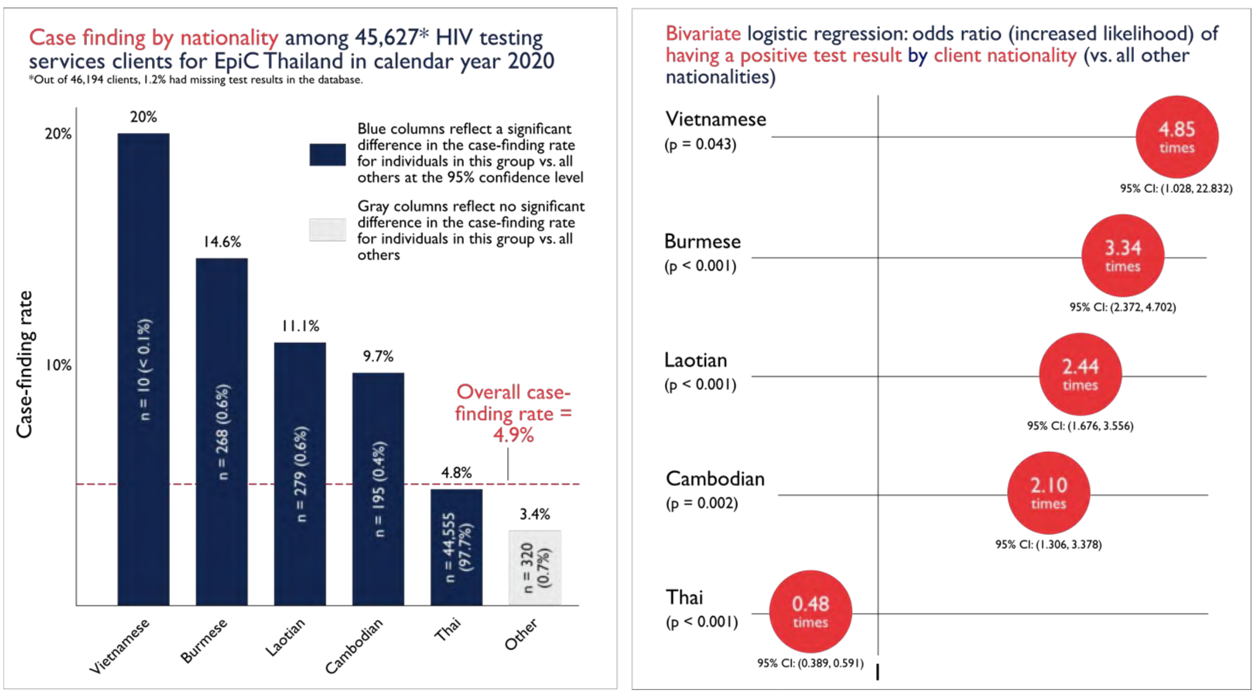
Because the key to unlocking the potential impact of client segmentation is providing simple and unambiguous insights that inspire programmatic action, making the findings of client-segmentation analyses as clear as possible is a primary priority. Bivariate analyses that compare probabilities for multiple values of a predictor variable against one selected reference category can in some instances make interpretation elusive. To simplify interpretation, it is possible for the bivariate analyses to recode variable values as new "on or off" *dummy variables* to compare the probability of an outcome for clients with a characteristic versus all other clients without that characteristic.

As an example, imagine that a program interested in improving case-finding rates is applying client segmentation to assess how the probability of a positive HIV test result varies according to an HTS client's identification with a specific risk group. Let's imagine that the overall case-finding rate across all HTS clients is five percent, and that clients could fall into four different risk-group categories, two of which have average case-finding rates significantly below the overall rate, and two of which have rates significantly above the overall rate.

If the program chooses the group with the lowest average case-finding rate as the "reference," the odds-ratio calculations may find that individuals in all the other groups will have a significantly higher likelihood of receiving a positive test result — even though individuals in at least one of these groups may be *less* likely to receive positive test results than the overall cohort of clients.

To help normalize the odds-ratio comparisons and overcome this challenge, programs can alternatively conduct separate analyses to calculate odds ratios reflecting the likelihood of a positive test result for all individuals identifying with group one, versus all other clients, and can repeat these analyses for members of the other groups. In this manner, the calculated odds ratios reflect how the likelihood of the outcome of interest is greater or lesser for individuals with a characteristic as compared to all other clients without that characteristic. Calculated this way, focusing HTS on individuals having characteristics associated with significantly greater odds of a positive test result should help improve overall case-finding rates. An example is depicted in Figure 7, which illustrates findings from a bivariate analysis of potential associations between client nationality and likelihood of receiving a positive HIV test result among all HIV testing clients in the EpiC Thailand program in calendar year 2020. Clients of Vietnamese, Burmese, Laotian, and Cambodian nationalities all had a significantly higher likelihood of receiving a positive HIV test result when compared to clients of all other nationalities.

Figure 7. Bivariate Analysis of HIV Test Results by HIV Testing Services Client Nationality, EpiC Thailand, 2020



MULTIVARIABLE ANALYSES

Should teams identify more than one predictor characteristic that is statistically associated with the outcome of interest through the bivariate analyses, they can pursue multivariable analyses to assess the extent to which each of these characteristics *together* are associated with the outcome of interest. A straightforward way to accomplish this is to employ multivariable logistic regression in a manner similar to the application of bivariate logistic regression described above.

A multivariable logistic regression model will predict the likelihood of an outcome of interest based on the simultaneous values for more than one predictor variable. Put simply, if a team found through bivariate analyses that levels of reported condom use, drug use, and gender each had p-values reflecting statistically significant associations with the likelihood of client receiving a positive test result, then a multivariable model can be used to predict the likelihood of a positive test result based on a client's characteristics with respect to all three of these predictors at the same time.

In this manner, the pursuit of multivariable analyses can supply teams with a predictive model, which in turn can be applied to risk screening tools and broader machine learning and artificial intelligence applications to differentiate and improve client support based on program data. In addition, multivariable analyses will allow teams to tease out the relative associations between more than one client characteristic and the outcome of interest, given that some of these characteristics may overlap, resulting in their having similar associations with the outcome of interest.

For example, imagine that the team referenced above constructs a multivariable logistic regression model based on the predictor variables it identified. The model will generate p-values and odds ratios for values of each of the predictors similar to the output of the bivariate analyses. However, in conducting the multivariable analyses, the team finds that the odds ratios have changed by virtue of having been *adjusted* to account for the influences of the other variables in the model and that the p-values associated with reported condom use are no longer significant.

Perplexed by this, the team investigates further and finds that reported condom use and reported drug use were highly correlated; individuals who reported drug use also did not report consistent condom use. As a result, condom use did not contribute significantly to the predictive power of the model and, indeed, the team could simplify its predictive model to exclude the condom use variable.

This simplification is a subtle but noteworthy potential benefit of multivariable analyses. Based on the results of the bivariate analysis alone, the team could focus on expanding HTS promotion to both individuals who report drug use and those who report inconsistent condom use. In doing so, it would be quite reasonable for the team to expect to see an increase in their HTS case-finding rates over time. The multivariable analysis does not negate the observed bivariate associations, but does suggest that the team might be able to achieve similar results more efficiently simply by focusing on clients who report drug use — as drug use and inconsistent condom use were highly correlated and were having a similar predictive influence on the likelihood of a client receiving a positive test result.

Another possible benefit of multivariable analyses is to control for the potential influences of specific variables, to see if other variables still have their own independent association with the outcome of interest. For example, let's imagine that our initial bivariate analyses have revealed a strong correlation between engagement of clients in urban vs. rural geographies and the likelihood of a client receiving a positive test result. In addition, our initial analyses have revealed an association between reported alcohol use and the likelihood of a positive test result. Now, we are wondering if alcohol use is independently associated with testing outcome or just tends to be more prevalent in urban geographies that have a higher background prevalence of undiagnosed HIV infection. To explore this, we can include both geography and reported alcohol use in a multivariable model as potential predictors of testing outcome. If alcohol use remains a significant predictor of testing outcome in a model that includes geography, then we can conclude that the association between alcohol use and the likelihood of a positive testing outcome appears to be independent of geography.

Model validation

Once a multivariable model has been generated from observations in a specified cohort, it can be cross-validated for relevance to a broader set of historical or new observations. Cross-validation affords teams opportunities to quantify the predictive power of the model as new data becomes available over time, and to revisit or adjust the model as program efforts and other factors influence the relationships between client characteristics and the likelihood of their experiencing specific outcomes. Indeed, while we will not cover the details of this approach in this overview, one application of machine learning technologies is to train a computer to adapt and improve a predictive model over time thorough routine model cross-validation on an expanding set of observations.

For teams that would like to perform a simple cross-validation independent of a machine-learning approach, the steps are as follows:

- Identify a set of observations as a data validation dataset. This validation dataset should be at least 20% to 30% of the size of the original model dataset. Observations can be randomly set aside from the original cohort, or the most recently available observations can be used for the validation dataset.
- Build a multivariable predictive model from the observations in the cohort. If the validation dataset has been generated by extracting 20% to 30% of records from the original cohort, then the predictive model should be built upon analyses of the 70% to 80% of remaining observations.
- Generate predictive probabilities from the multivariable model for all the different possible values of predictor variables in the model.
- Identify practical prediction "cut-off" values that can be used to compare the consistency of model predictions with actual outcomes from the validation dataset. This will entail establishing a model predictive probability threshold above which an outcome is considered to have occurred, and below which it is considered to have not occurred.

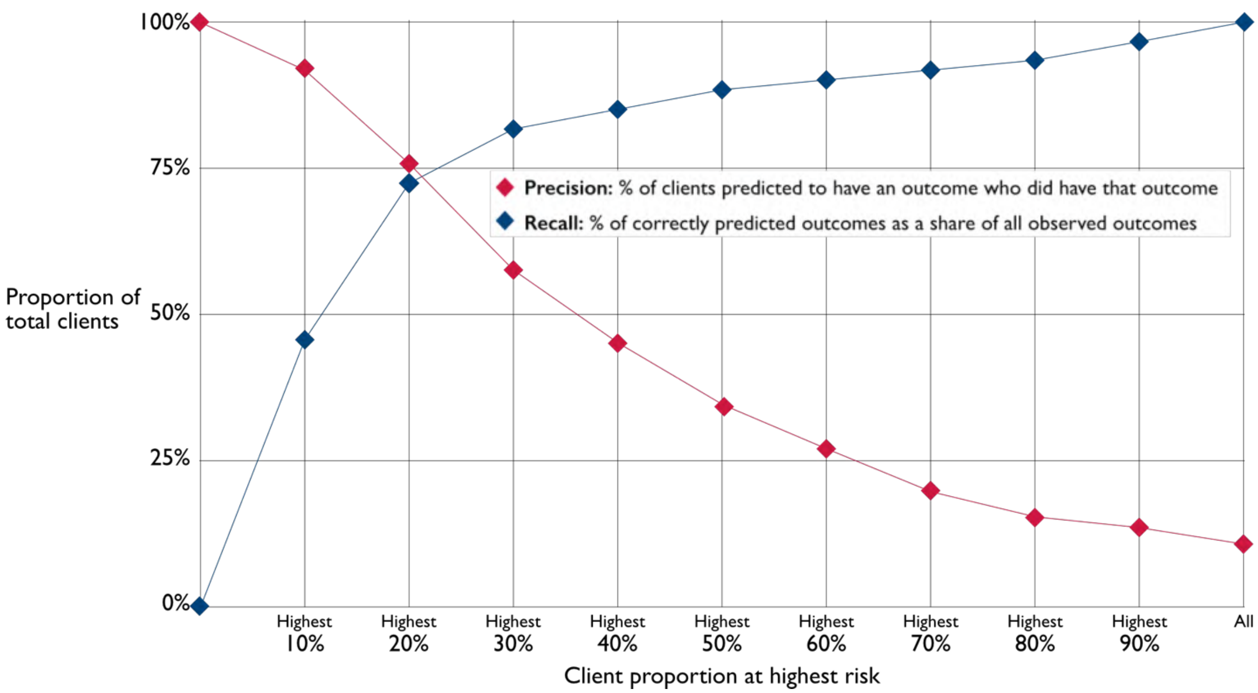
- Classify predicted model outcomes based on the identified cut-off values.
- Evaluate the accuracy of the model predictions by comparing the predicted outcome for clients with certain characteristics with the observed outcome for clients with those characteristics from the validation dataset. The model accuracy can be quantified by calculating positive predictive values and negative predictive values:
  - *Positive-predictive value:* The proportion of positive model predictions that are in fact positive in observed cases.
  - *Negative-predictive value:* The proportion of negative model predictions that are in fact negative in observed cases.

An alternative approach to assessing model validity in which a relatively small overall proportion (i.e., 5% to 20%) of the client population experience the outcome of interest is to plot the relationship between model precision and recall. Precision is defined as the percent of clients predicted to experience an outcome who did experience that outcome. Recall is the percent of correctly predicted outcomes as a proportion of all individuals who experienced that outcome.

Imagine that a team has developed a model that predicts the likelihood of HIV testing clients receiving positive results, and anticipates that no more than 20% of the total testing client population will have positive test results. The team could then take a closer look at the 20% of the client population that the model predicts is most likely to receive positive test results and calculate both precision and recall for this group. Do most of these clients that the model predicts are at highest risk receive positive test results (precision)? Do these 20% of clients represent a large proportion of all the clients who received positive test results (recall)?

By plotting the relationship between precision and recall for specific segments of the population by predicted level of risk — looking for example at the 5% predicted to have the greatest risk, then the top 10%, 15%, 20%, 25%, 30%, and so on — a team can gain useful insights relevant to model validity (Figure 8). For example, if the team found that the precision and recall lines cross at the 75% mark for the 20% of clients predicted to face the highest risks of a positive test result, this would suggest that 75% of these "highest-risk" clients did in fact receive positive test results, and that 75% of all positive test results occurred among this 20% of clients.

Figure 8. Illustrative Model Performance Validation by Plotting Precision vs. Recall



Achieving and evaluating client segmentation impact

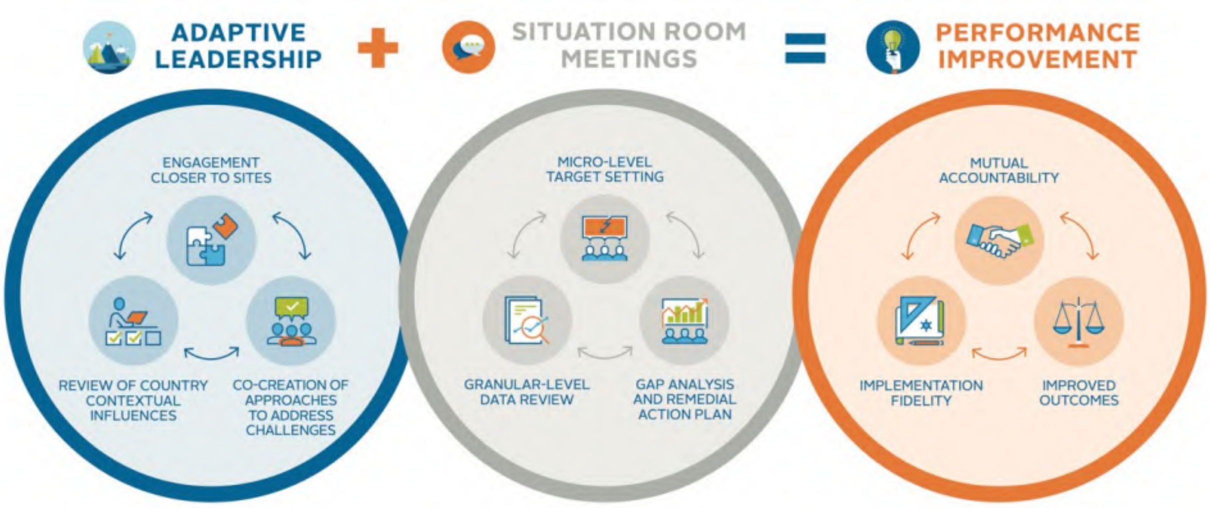
The opportunities to apply client segmentation to analyses of routine program data are considerable, but as depicted in the eight client-segmentation "steps," these efforts only get us halfway around the envisioned cycle of continuously applying client segmentation to improve client and cascade outcomes.

To maximize the potential impact of client segmentation, teams should ensure they have resources and plans to invest in program improvement. Prior to undertaking any analyses, teams need:

1. Systematic processes and procedures to review data collaboratively and generate potential programmatic solutions and remedies
2. Clear strategies to integrate client-segmentation findings into these processes

Client-segmentation can both inform and benefit from participatory initiatives to promote performance improvement, like FHI 360's total quality leadership and accountability approach depicted in Figure 9.

Figure 9. FHI 360's Total Quality Leadership and Accountability Framework



Whether teams are implementing a more structured approach as depicted here, or other performance improvement efforts, it may be useful to identify from the outset expectations about how client segmentation can be integrated with and add value to existing efforts to apply data to inform action.

One specific opportunity to consider is whether the results of client-segmentation analyses can be applied to refine and improve the risk assessment tools teams are currently using to be more responsive to localized client data. Figure 10 depicts a risk screening tool from the EpiC Long Term Adherence Guide[6] that matches potential client risks to programmatic solutions. Client segmentation may help to identify which of these — or potentially other factors — are most closely associated with client risk in a specific country or risk population context. Accordingly, it may also help programs to prioritize among these and other potential solutions for clients, keeping in mind that the risk-screening concept is intended to help focus additional support but not remove client options for service access according to their preferences and felt needs.

[6] Levitt D, Lillie T. Long-term HIV treatment adherence for key populations: program considerations. Durham (NC): FHI 360; 2020. Available from: https://www.fhi360.org/sites/default/files/media/documents/epic-long-term-hiv-adherence-guide.pdf.

Figure 10. A Risk Assessment Tool from EpiC Long-Term HIV Treatment Adherence Guide

PROVIDER SUPPORTIVE INTERVENTIONS	
KEY BARRIERS	**SERVICE/TOOLS /SUPPORT**
☐ Migrant labor	☐ Regular phone contact/SMS check-ins. Frequency:___________ *Confirm when client plans to travel, record below.* ☐ SMS/phone appointment reminders. Frequency:_____________ ☐ Referral to multi-month ART dispensation (6 months or more) ☐ Plan for ensuring access to ART while away ☐ Identify additional contacts in case of travel / migrant labor
☐ History of violence ☐ Concerned about violence with disclosure	☐ Violence prevention/response counseling ☐ Counseling on disclosure/partner notification ☐ Referral to legal aid
☐ Substance use	☐ Substance use counseling ☐ Identification of a treatment buddy/other social support ☐ Referral to drug rehabilitation
KEY POPULATION SPECIFIC ☐ Likely to change venue/location for sex work ☐ History of arrest/imprisonment/police harassment ☐ KP status not known by partners/family ☐ Young/adolescent KP ☐ Compound risk factors (e.g., transgender woman sex worker, homeless, medication interactions with hormone therapy, etc.)	☐ Psychosocial counseling by provider trained in KP service provision ☐ KP-specific violence prevention/response counseling and support ☐ Referral to legal aid ☐ Counseling on family disclosure/partner notification and referral ☐ Referral to social support services (for adolescents, homeless, etc.) ☐ Referral to specialized clinical care for transgender women
☐ Poverty/unable to miss work ☐ Transportation ☐ Insurance status	☐ Identification of a treatment buddy/other social support ☐ Referral to community adherence group/dispensation ☐ Referral to social services (e.g., national insurance scheme)
☐ Fear of disclosure	☐ Counseling on disclosure/partner notification
☐ Depression/mental health	☐ Psychosocial counseling ☐ Referral to professional mental health support
☐ Knowledge (understanding ART and HIV) ☐ Myths/beliefs	☐ Education on HIV/ART ☐ Written instructions ☐ Weekly pill box for medications (provide/recommend)
☐ Lack of social support ☐ Stigma	☐ Psychosocial counseling ☐ Counseling on disclosure/partner notification ☐ Referral to community adherence/peer support group ☐ Referral to legal aid
☐ 0 Treatment fatigue/discomfort/side effects	☐ Counseling on management of side effects ☐ Referral to clinician ☐ Navigation support (accompaniment to appointment with clinician)
☐ 0 Non-national	☐ Language Interpretation/referral to appropriate services
☐ Other:	☐ Other:
Priority: ☐ Higher risk of missed appointments ☐ Lower risk of missed appointments	Comments:
Referrals:	Follow-up date: DD/MM/YY Estimated time of day: ________ AM / PM (circle one)
Provider Name (Print):	Provider signature:

Teams may also consider summarizing the results of client segmentation to generate profiles or “archetypes” of clients who may face greater risks to help providers focus additional support and help stimulate thinking about and co-creation of potential service solutions to better serve clients with these risks. That said, teams pursuing this approach should also take care to ensure that in doing so they do not further amplify stigma or discrimination among or toward specific categories of clients.

ASSESSING THE POTENTIAL IMPACT OF CLIENT SEGMENTATION

The purpose of client-segmentation is to generate impactful action, evident in the form of improved client and overall HIV cascade outcomes. Improvements should be observed with respect to the specific performance issues for the specified outcomes of interest (i.e., IIT, lack of viral load uptake or suppression, progression to advanced HIV disease, poor case finding, limited engagement in index testing).

In addition, teams should document the process applied to conduct client-segmentation across the eight steps, in particular summarizing how results were shared, potential solutions generated, and what changes were made to act upon the results. These actions are the true "outputs" of client segmentation, without which even the most detailed and potentially illuminating client segmentation will not have impact.

Finally, teams should monitor trends in performance that may change with the introduction of novel client-segmentation-inspired service approaches or adaptations. With epidemic shifts, the profiles of clients facing the greatest risks will also change, and thus client-segmentation will need to be repeated and updated. The cross-validation approaches described above may be useful in this regard.

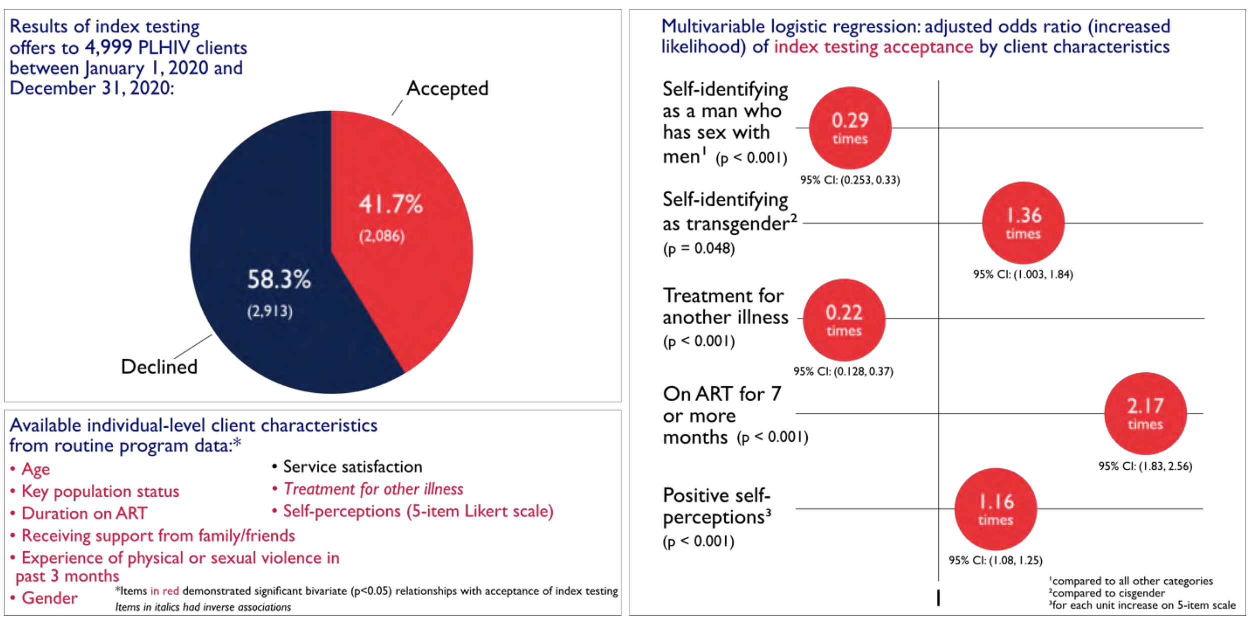
Case study: EpiC Indonesia

With the transition of support from the LINKAGES project to the new EpiC project in Indonesia, the EpiC team sought to take advantage of the rich health information systems and connections forged with LINKAGES support to conduct more granular analyses of routine program data to improve individual client and overall HIV cascade outcomes. In the first half of FY21, the team applied client segmentation to generate insights relevant to three priority program areas based on historical challenges.

The first of these was index testing. As facilitating voluntary, safe, and ethical participation in index testing at scale remains a challenge in Indonesia and globally, the team reviewed existing program data to assess the differentiating characteristics of: (1) index testing clients who accepted participation in index testing (vs. those who declined) and (2) those acceptors who referred HIV-positive contacts (vs. no contacts or HIV-negative contacts).

The program supported offers of index testing to 4,999 people living with HIV (PLHIV) who were clients in calendar year 2020 (Figure 11). Of these, almost 42% accepted index testing, and about 58% declined. The available client-level characteristics that could be linked to the decision to accept or decline included age, key or priority population status, length of time on antiretroviral treatment (ART), whether reported receiving support from family or friends, reported experiences of physical or sexual violence in the past three months, reported service satisfaction, gender, current treatment for other illnesses, and positive vs. negative self-perceptions. In bivariate (chi-square) analyses, each variable had a significant association (at the 95% confidence interval) with the likelihood of accepting or declining participation in index testing, except client reports of service satisfaction. Of note, only 18 of 4,999 clients (0.4%) reported not being satisfied with the services they experienced.

Figure 11. Identifying Characteristics of PLHIV Clients More Likely to Accept Voluntary Index Testing Offers



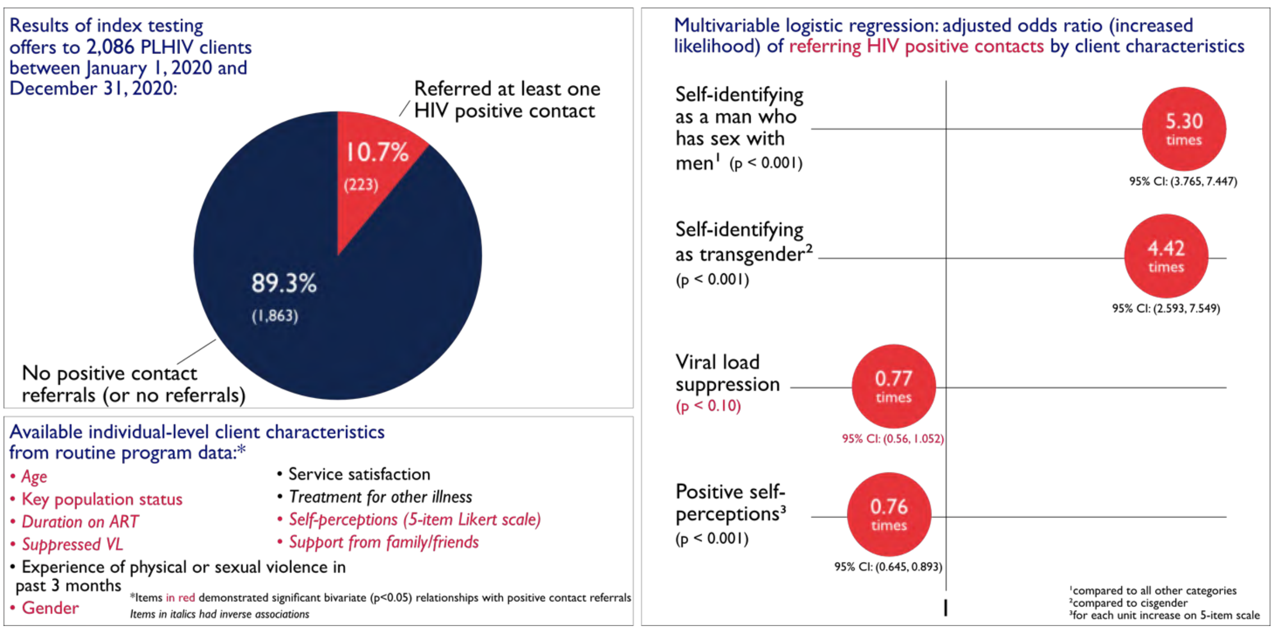
The team constructed a multivariable logistic regression model to characterize the relationship between each of the variables with significant bivariate associations and the likelihood of accepting or declining index testing while accounting for the potential influence of all other variables. This predictive model was iteratively reduced to remove variables that in the presence of the influence of other items no longer had a significant association with the likelihood of index testing acceptance. In the final multivariable predictive model, clients who self-identified as men who have sex with men (MSM)[7] were substantially less likely to accept index testing than clients of all other populations while controlling for other variables. Those who were on ART for seven or more months were more than two times as likely to accept index testing than those who had

[7] In this context, this means that male clients reported to HIV program staff that they had sex with at least one other man in the past year. This does not necessarily mean that these clients identify as gay. Transgender clients reported to program staff that their gender identity differs from their sex at birth.

been on ART for a shorter duration. The “receiving support” and “experienced physical or sexual violence” variables lost their significance in the multivariate model, as other variables in the model better explained variability in the likelihood of acceptance. Only 12 individuals in the dataset reported experiences of violence, and in the bivariate analysis these individuals were remarkably much more likely to accept index testing (10 of 12 accepted).

Figure 12 depicts an index-testing subanalysis the team conducted looking at the characteristics of index testing acceptors who were more likely to refer positive contacts (vs. negative or no contacts). Among the 2,086 clients who accepted index testing, almost 11% referred at least one positive contact. Again, variables with bivariate associations are in red at bottom left.

Figure 12. Identifying Characteristics of PLHIV Clients More Likely to Refer Contacts Who Test Positive

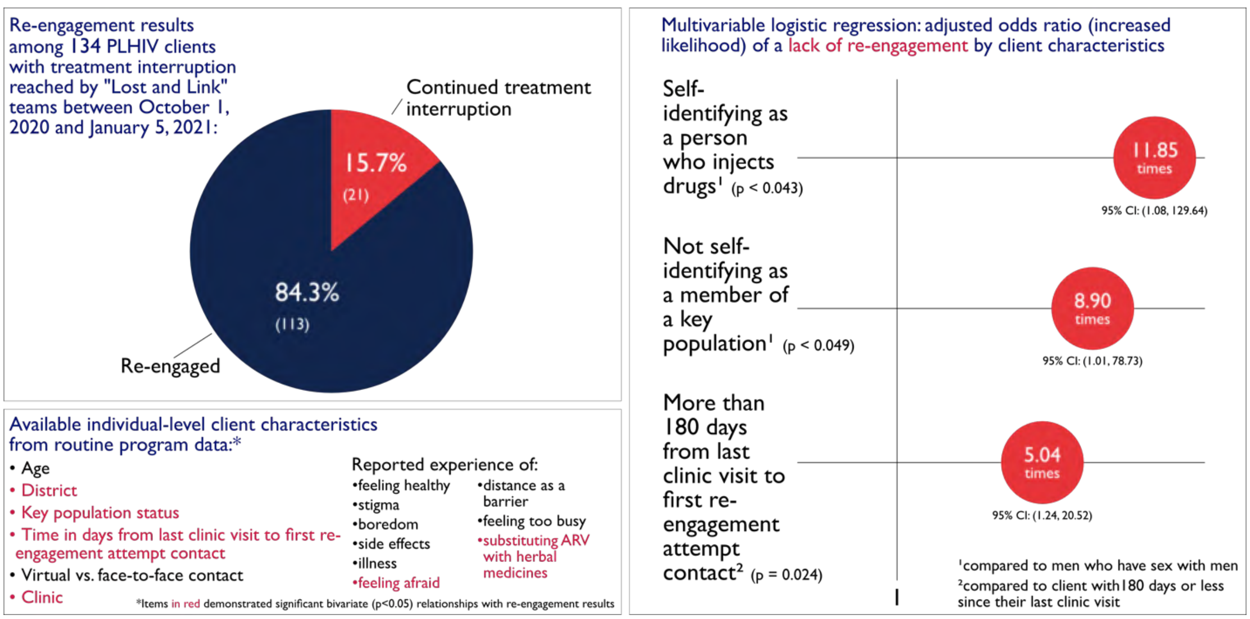


In the multivariable model, MSM were more than five times more likely than other clients to refer positive contacts in sharp contrast to the previous analysis illustrating they were also less likely to accept participation in index testing. Those with more positive self-perceptions were less likely than others to refer positive contacts. At the less significant 90% confidence level, those who were virally suppressed were less likely to refer positive contacts.

From these analyses, the team is working with community partners to develop strategies to tailor index testing to address the differentiated preferences of MSM, and to ensure “no missed opportunities” to promote participation in index testing among MSM, transgender individuals, and individuals who have been on treatment for six months or less or have not achieved viral suppression. The team has also supported the establishment of community and client feedback initiatives that are now being scaled up with support from the Global Fund to Fight AIDS, Tuberculosis, and Malaria to help HIV programming become more broadly and routinely responsive to clients.

The team also focused a client-segmentation lens to optimize efforts to re-engage PLHIV clients who have experienced IIT through the program's "Lost and Link" initiative. The analysis depicted in Figure 13 looks at characteristics of PLHIV who had experienced treatment interruption and were reached by the team's re-engagement initiative between October 1, 2020, and January 5, 2021, but were less likely to be successfully re-engaged after being reached.

Figure 13. Identifying Characteristics of "Lost and Link" PLHIV Clients Less Likely to be Successfully Re-engaged in Treatment



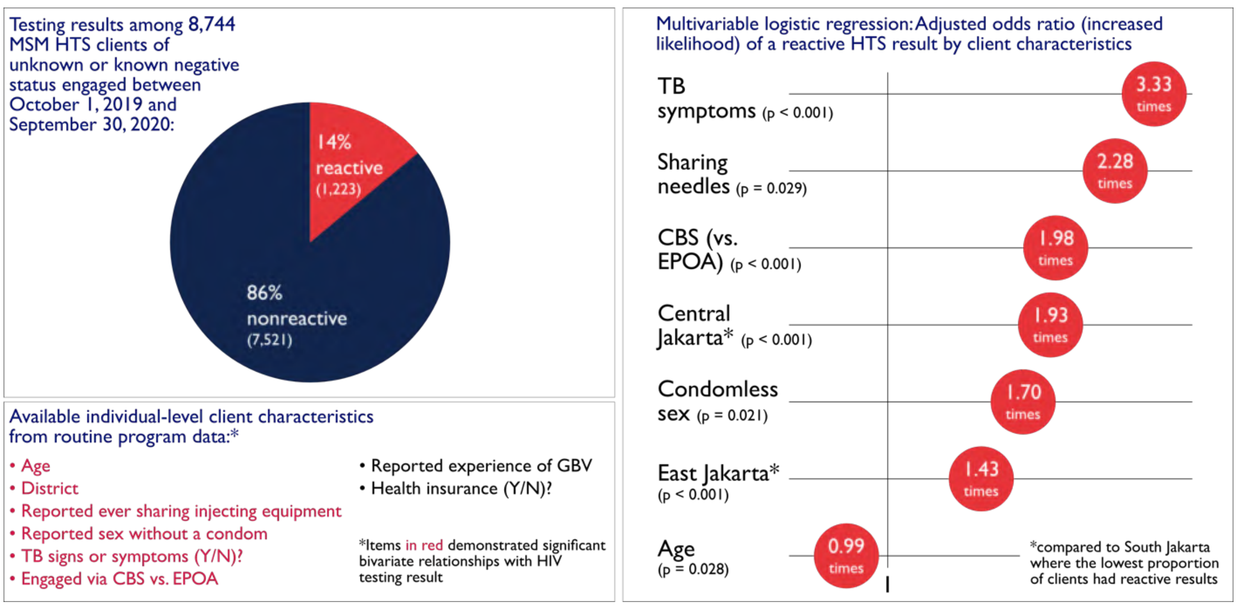
Injecting drug use and delays in the time between the last clinic visit and the re-engagement contact were associated with a greater likelihood of a lack of re-engagement. The community-based organization (CBO) partners were predominantly affiliated with the MSM community, which may have challenged their efforts to support re-engagement of other KP members as well as PLHIV who did not self-identify as KP. As a result of this analysis, the program is focusing on the activation of more inclusive and broadly representative case management and re-engagement support, as well as swift action to support clients with missed appointments and re-engage clients experiencing treatment interruption.

To improve the focus and impact of HTS, the team also applied client-segmentation to identify the characteristics of MSM clients more likely to receive positive HIV test results.

Among 8,744 MSM clients in FY20 reporting unknown or known negative status prior to testing, 14% received reactive results (Figure 12). Available individual client characteristics that could be linked to testing outcome in the dataset were: age, district of service delivery, self-reported ever sharing injecting equipment, self-reported sex without a condom, having tuberculosis (TB) signs or symptoms, method of client engagement,[8] self-reported ever experience of gender-based violence, and access to health insurance. In bivariate (chi-square) analyses, all but gender-based violence and access to health insurance had a significant association (at the 95% confidence interval) with the likelihood of receiving a reactive HIV test result.

In the multivariable logistic regression model, individuals reached in Central or East Jakarta, those who reported condomless sex or sharing needles, and those with TB symptoms were significantly more likely to receive a reactive test result (Figure 14).

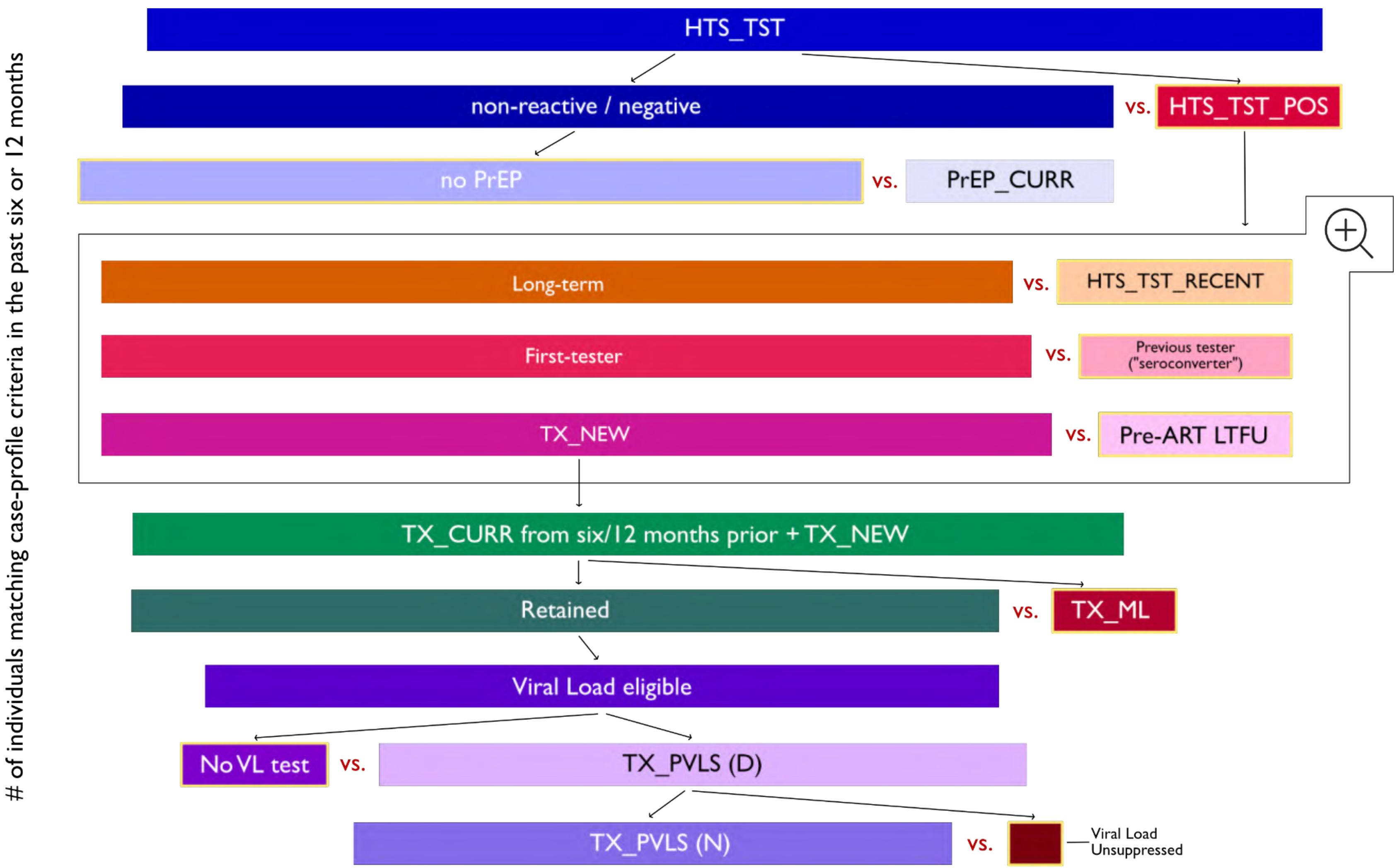
Figure 14. Identifying Characteristics of MSM HTS Clients More Likely to Have Reactive HIV Test Results



Historically, the program has focused on the need to ensure HIV testing and links to treatment as appropriate for all potential TB cases, but had not explicitly prioritized TB symptoms as a risk factor for HIV among KPs. The team is now working to strengthen TB screening and links to TB services, as well as the use of TB symptom data, to help focus HIV case-finding efforts among KPs. The team is also exploring options to introduce enhanced risk assessment tools and to explore other potential risk factors — such as reported engagement in chemsex — to differentiate and focus efforts.

[8] The program engages prospective HTS clients through a variety of online and off-line strategies. These include traditional outreach efforts undertaken by trained and salaried community based supporters (CBSs), as well as informal incentivized network referrals through untrained peer mobilizers (PMs) as part of the enhanced peer outreach approach (EPOA). More information about EPOA is available at: https://www.fhi360.org/sites/default/files/media/documents/resource-linkages-enhanced-peer-outreach-implementation.pdf.

Annex I: Illustrative "case" comparisons based on HIV cascade events



Annex II: Illustrative "mapping" of available individual-level outcome data, and differentiating sociodemographic, risk, and other client characteristics

	Potential Differentiators of Risk	Outcome of interest and source						
		HTS_POS	HIV Recency	Not linked to PrEP	PrEP LTFU	ART LTFU	VL Untested	VL Unsupp
Sociodemographic	Sex	√	√					
	Age (year)	√	√					
	KP group	√	√					
	Province for care	√	√					
	District for care							
	Commune (residence)							
	Occupation							
	Education							
Risk factors	Inconsistent condom use	√						
	Experiences of violence							
	Non-injecting drug use	√						
	Injecting drug use	√						
	Index partner							
	Multiple partners							
	Sexually transmitted infections	√						
HIV testing-related project data	Test modality	√	√					
	HIV tested before	√	√					
	Reasons for testing							
	Recency result		√					
	HIV pos before							
	Reasons not enrolling ART							
	Time between confirmation and initiation							
ART-related project data	Missed ART clinic appointment							
	Regimen (first/second-line)							
	On MMD							
	Viral load coverage							
	Viral load suppression							
	Received viral load testing							
PrEP-related project data	Missed PrEP appointment							
	ED or non-ED PrEP							
	PrEP risk group							

Annex III: Illustrative "road map" for conducting client-segmentation analyses