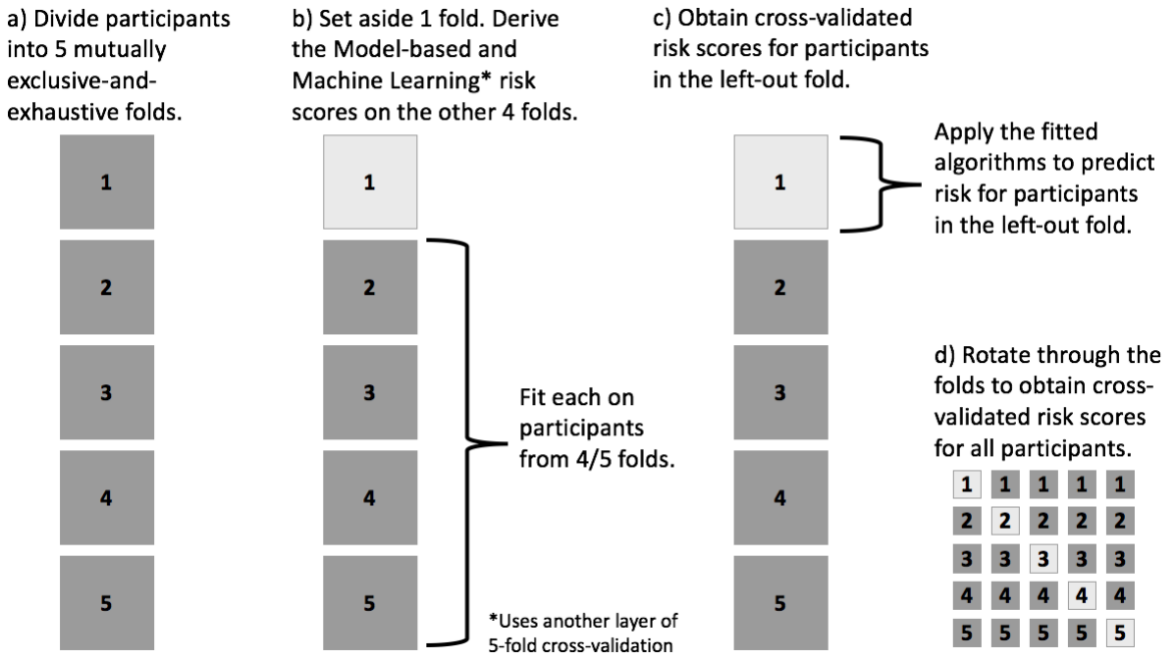
SUPPLEMENTARY MATERIALS

Description of Super Learner algorithm: Super Learner is machine learning algorithm that combines predictions from a pre-specified set of candidate learners together[1] and is thus considered an "ensemble" or "stacking" method[2,3]. More detailed overviews of the algorithm can be found in [4–6], with applications described in [7–10], among others. Briefly, to implement Super Learner, the user specifies (1) a "library" of candidate prediction algorithms, (2) a measure of performance (i.e. a loss function); and (3) a cross-validation scheme. In this application, our candidate algorithms were LASSO (*glmnet)*, generalized additive models (*gam*) after screening based on LASSO, stepwise logistic after screening based on univariate correlations with the outcome, and logistic regression with main terms for the factors in the Risk Group score (See Supplementary Table 1). We used the negative log-likelihood loss function as our measure of performance. We employed five-fold cross-validation, in which individuals were divided into five mutually exclusive-and-exhaustive groups (called "folds"; Supplementary Figure 1), with all repeat tests from a given individual included in the same fold. Using data on 4/5 folds, we trained each of the candidate algorithms, and then applied these algorithms to predict HIV seroconversion in the remaining, left-out fold. Specifically, for individuals in the remaining "validation" fold, we estimated performance by evaluating the negative log-likelihood. We then rotated through the folds, such that each individual served once in the validation fold. After this process, one option would be to average the five cross-validated estimates of performance for each algorithm in turn and select the candidate algorithm with the best performance (i.e. lowest average cross-validated loss estimate). Instead, Super Learner goes one step further and combines the prediction algorithms. By regressing the observed outcomes on the cross-validated predictions from each candidate and then normalizing the coefficients, we can obtain the best weighted, convex combination of predictions from all candidates. We refer the reader to [16,17] for worked examples. We also note that in this application two-layers of cross-validation were employed: one for construction of the machine learning risk score, and a second for evaluation of performance, as detailed in Supplementary Figure 1.

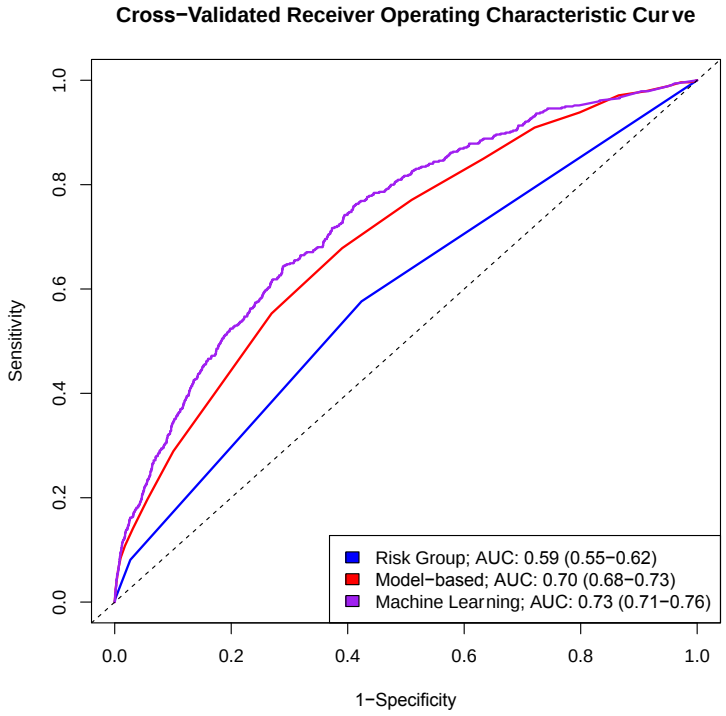
Supplementary Table 1: Complete list of risk factors and their descriptions, including handling of missing data. Model-based and Machine Learning risk scores considered all variables as candidate predictors; variables included in the Risk Group strategy denoted with *.

	INDICATORS FOR
DEMOGRAPHICS	
Age	Age groups (in years): 25-34, 35-44, 45-54, 55+; reference as 15-24
Sex	Male sex; reference as female sex
Marital status*	Single, married, widowed*, divorced/separated; reference as missing
Polygamy	In a polygamous marriage, not in a polygamous marriage; reference as unknown
Relation	Familial relation to head-of-household; reference as otherwise
Education	Completed primary school, completed secondary school or more; reference as less than primary or unknown
Occupation strata	Formal sector (teacher, student, government worker, military worker, health worker, or factory worker); High-risk informal sector (fishmonger, fisherman, bar owner, bar worker, transport, or tourism); Low-risk informal sector (farmer, shopkeeper, market vendor, hotel worker, homemaker, household worker, construction worker, or mining); Jobless; reference as other (including unknown)
Student	Student; reference as otherwise
Transportation*	Transportation (truck, taxi, motorcycle, bike, boat) drivers; reference as otherwise
Fishing*	Fisherman or fishmonger; reference as otherwise
Bar worker*	Bar worker or bar owner; reference as otherwise
Hotel worker	Hotel or restaurant worker; reference as otherwise
Shopkeeper	Shopkeeper or market vendor; reference as otherwise
Alcohol*	Any alcohol use*, no alcohol use; reference as unknown
Region	Uganda-West; Kenya; reference as Uganda-East
MOBILITY	
Inmigrant	Moved into the community after baseline; reference as otherwise
Baseline stable resident	Living <6 months outside community in the year preceding baseline; reference as otherwise
Mobile resident	Living 1+ month outside the community; reference as otherwise
Shifted residence	Shifted residence in past year; did not shift residence in past year; reference as unknown
Nights away	Nights spent away from home in past month grouped as 0-few, less than half of the month, more than half of the month, most nights, every night; reference as unknown
HEALTH	
Health fair attendance	Attended the health fair; reference as otherwise
Contraceptive use	Using contraceptives; not using contraceptives; reference as unknown
Pregnant	Pregnant; not pregnant; reference as male or unknown
Live birth	At least 1 live birth in the past year; 0 live births in the past year; reference as male or unknown
Male circumcision	Traditional male circumcision; medical male circumcision; not circumcised; reference as unknown or female
SPOUSES	
Unknown status	Spouse has known HIV status; spouse has unknown HIV status; reference as otherwise
Serodiscordant*	Spouse is HIV-infected; spouse is HIV-uninfected; reference as otherwise
Serodiscordant and male	Spouse is HIV-infected and male; reference as otherwise
Serodiscordant and circumcision	Spouse is HIV-infected and male partner is not circumcised; reference as otherwise
Serodiscordant and polygamous	Spouse is HIV-infected and the marriage is polygamous; reference as otherwise
Serodiscordant and unsuppressed	Spouse is HIV-infected with HIV RNA level >500 copies/mL; reference as otherwise
HOUSEHOLD FACTORS	
Wealth	Quintiles based on a principle components analysis of household wealth survey and calculated at the level of the household; reference as unknown
HIV-unknown adult	At least 1 adult whose HIV status is unknown in the household; no adults with unknown HIV status in the household; reference as unknown
HIV-infected adult	At least 1 HIV-infected adult in the household; no HIV-infected adults in the household; reference as unknown
HIV-infected adult of the opposite sex	At least 1 adult of the opposite sex and HIV-infected in the household; no adults of the opposite sex and HIV-infected in the household; reference as unknown
INTERACTIONS	
Young woman*	Woman aged 15-24 years; reference as otherwise
Female bar worker	Female bar worker or bar owner; reference as otherwise
Wealthy male	Male in highest socioeconomic strata; reference as otherwise
Young pregnancy	Woman aged 15-24 years and reporting current pregnancy; reference as otherwise
Young mother	Woman aged 15-24 years and reporting at least 1 live birth in the past year; reference as otherwise

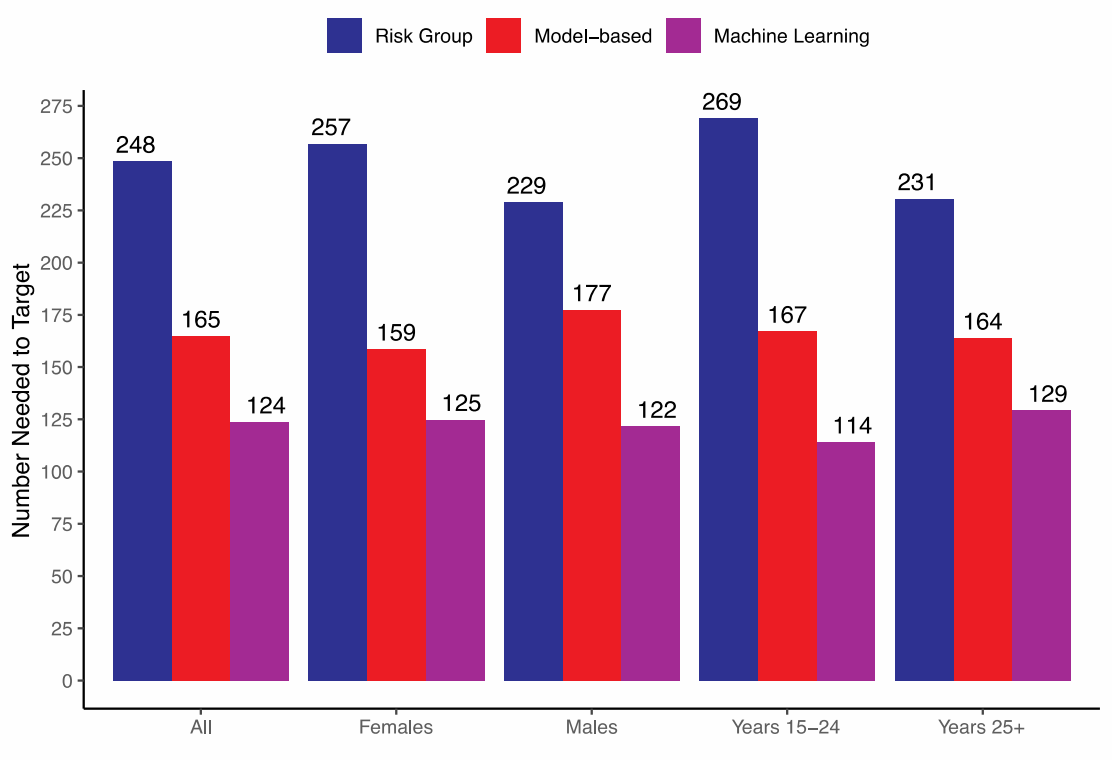
Supplementary Figure 1: Schematic of the sample-splitting procedure used to generate cross-validated risk scores for all participants.



Supplementary Figure 2: Cross-validated receiver operating characteristic curves for three HIV risk scoring algorithms: Risk Group (score as sum of known risk factors), Model-based, and Machine Learning. Area under the curve (AUC), reflecting an approach's ability to score a person who acquired HIV as higher risk than a person who did not, is given in the legend with 95% confidence intervals in parentheses. Influence curve-based variance estimates were used to construct 95% confidence intervals and evaluate whether Machine Learning improved the AUC[15]. Dashed black indicates a random guess.



Supplementary Figure 3: Cross-validated estimates of the number needed to target (NNT, equal to 1/positive predictive value) to achieve 50% sensitivity for correct classification of seroconversions.



Supplementary Table 2: By region, cross-validated efficiency of each candidate targeting strategy, defined as the proportion of the population that would have been classified as high-risk and targeted for intensified prevention (rate of positive predictions) to achieve at least 50% sensitivity within each region, and corresponding region-specific number needed to target (NNT, equal to 1/positive predictive value).

		Rate of Positive Predictions (%)			**Number Needed to Target (NNT)**		
		Risk Group	Model-based	Machine Learning	Risk Group	Model-based	Machine Learning
Uganda-West	All	43%	34%	20%	253	183	133
	Females	49%	36%	21%	250	177	138
	Males	35%	31%	19%	258	195	125
	Years 15-24	59%	28%	19%	222	147	104
	Years 25+	35%	36%	21%	287	203	152
Uganda-East	All	44%	44%	26%	505	432	318
	Females	57%	51%	30%	588	440	351
	Males	27%	35%	20%	372	419	272
	Years 15-24	57%	36%	22%	807	1085	642
	Years 25+	34%	49%	28%	353	331	252
Kenya	All	40%	31%	24%	145	129	103
	Females	57%	37%	28%	155	138	107
	Males	17%	23%	17%	112	115	94
	Years 15-24	51%	25%	20%	156	133	110
	Years 25+	33%	35%	26%	135	128	99

Supplementary Table 3: By region, cross-validated sensitivity that would have been achieved by targeting 45% of the region-specific population, and corresponding region-specific number needed to target (NNT, equal to 1/positive predictive value).

		Sensitivity (%)			Number Needed to Target (NNT)		
		Risk Group	Model-based	Machine learning	Risk Group	Model-based	Machine learning
Uganda-West	All	56%	60%	78%	253	183	189
	Females	63%	65%	80%	250	177	200
	Males	45%	52%	74%	258	195	173
	Years 15-24	78%	55%	77%	222	147	164
	Years 25+	42%	62%	78%	287	203	204
Uganda-East	All	54%	64%	68%	505	432	418
	Females	52%	63%	64%	588	440	456
	Males	58%	66%	74%	372	419	360
	Years 15-24	90%	43%	62%	807	1085	805
	Years 25+	45%	69%	69%	353	331	332
Kenya	All	60%	69%	74%	145	139	132
	Females	71%	72%	81%	155	140	132
	Males	41%	64%	62%	112	138	133
	Years 15-24	78%	64%	74%	156	140	131
	Years 25+	50%	72%	74%	135	139	133

Supplementary Table 4: Sensitivity achieved by each strategy when targeting a fixed proportion of the young adult (15-24 years) overall.

Limiting the rate of positive predictions	Risk Group*	Model-based	Machine Learning
10%	13%	22%	38%
20%	13%	34%	59%
30%	13%	51%	67%
40%	13%	66%	81%
50%	13%	75%	85%
60%	79%	83%	91%

*A strategy to target any young adult with at least two known risk factors (score≥2) would have achieved 13% sensitivity, while a strategy to target any young adult with at least one known risk factors (score≥1) would have achieved 79% sensitivity.

References:

1. van der Laan MJ, Polley EC, Hubbard AE. Super Learner. Statistical Applications in Genetics and Molecular Biology **2007**; 6:Article 25.
2. Wolpert DH. Stacked Generalization. Neural Networks **1992**; 5:241–259.
3. Breiman L. Stacked Regressions. Mach Learn **1996**; 24:49–64.
4. Polley EC, Rose S, Laan MJ van der. Super Learner. In: van der Laan MJ, Rose S, eds. Targeted Learning: Causal Inference for Observational and Experimental Data. New York Dordrecht Heidelberg London: Springer, 2011.
5. Petersen M, Balzer L. Introduction to Causal Inference. 2014. Available at: www.ucbbiostat.com.
6. Naimi AI, Balzer LB. Stacked Generalization: An Introduction to Super Learning. European Journal of Epidemiology **2018**; 33:459–464.
7. Sinisi SE, Polley EC, Petersen ML, Laan MJ van der. Super learning: an application to the prediction of HIV-1 drug resistance. Stat Appl Genet Mol **2007**; 6:Epub.
8. Petersen ML, LeDell E, Schwab J, et al. Super learner analysis of electronic adherence data improves viral prediction and may provide strategies for selective HIV RNA monitoring. J Acquir Immune Defic Syndr **2015**; 69:109–118.
9. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med **2015**; 3:42–52.
10. Baćak V, Kennedy EH. Principled Machine Learning Using the Super Learner: An Application to Predicting Prison Violence. Sociological Methods & Research **2019**; 48:698–721.
11. Hastie TJ. Generalized additive models. In: Chambers JM, Hastie TJ, eds. Statistical models in S. Boca Raton: Chapman & Hall, 1992.
12. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B **1996**; 58:267–288.
13. Friedman JH, Hastie TJ, Tibshirani RJ. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software **2010**; 33:1–22.
14. Hastie T. gam: Generalized Additive Models. 2018. Available at: http://CRAN.R-project.org/package=gam.
15. LeDell E, Petersen M, van der Laan M. cvAUC: Cross-Validated Area Under the ROC Curve Confidence Intervals. Available at: https://CRAN.R-project.org/package=cvAUC. Accessed 15 July 2019.