J Antimicrob Chemother 2018; **73**: 2186–2196
doi:10.1093/jac/dky179 Advance Access publication 8 June 2018

Journal of Antimicrobial Chemotherapy

2018 update to the HIV-TRePS system: the development of new computational models to predict HIV treatment outcomes, with or without a genotype, with enhanced usability for low-income settings

Andrew D. Revell[1]*, Dechao Wang[1], Maria-Jesus Perez-Elias[2], Robin Wood[3], Dolphina Cogill[3], Hugo Tempelman[4], Raph L. Hamers[5], Peter Reiss[5,6], Ard I. van Sighem[6], Catherine A. Rehm[7], Anton Pozniak[8], Julio S. G. Montaner[9], H. Clifford Lane[7] and Brendan A. Larder[1] on behalf of the RDI Data and Study Group†

[1]The HIV Resistance Response Database Initiative (RDI), London, UK; [2]Instituto Ramón y Cajal de Investigación Sanitaria, Madrid, Spain; [3]Desmond Tutu HIV Centre, University of Cape Town, Cape Town, South Africa; [4]Ndlovu Care Group, Elandsdoorn, South Africa; [5]Departments of Internal Medicine and Global Health, Academic Medical Centre of the University of Amsterdam, Amsterdam Institute for Global Health and Development, Amsterdam, The Netherlands; [6]Stichting HIV Monitoring, Amsterdam, The Netherlands; [7]National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA; [8]Chelsea and Westminster Hospital, London, UK; [9]BC Centre for Excellence in HIV/AIDS, Vancouver, Canada

*Corresponding author. Tel: +44 207 226 7314; Fax: +44 207 226 7314; E-mail: andrewrevell@hivrdi.org
†Members are listed in the Acknowledgements section.

Received 14 December 2017; returned 24 March 2018; revised 10 April 2018; accepted 17 April 2018

Objectives: Optimizing antiretroviral drug combination on an individual basis can be challenging, particularly in settings with limited access to drugs and genotypic resistance testing. Here we describe our latest computational models to predict treatment responses, with or without a genotype, and compare their predictive accuracy with that of genotyping.

Methods: Random forest models were trained to predict the probability of virological response to a new therapy introduced following virological failure using up to 50 000 treatment change episodes (TCEs) without a genotype and 18 000 TCEs including genotypes. Independent data sets were used to evaluate the models. This study tested the effects on model accuracy of relaxing the baseline data timing windows, the use of a new filter to exclude probable non-adherent cases and the addition of maraviroc, tipranavir and elvitegravir to the system.

Results: The no-genotype models achieved area under the receiver operator characteristic curve (AUC) values of 0.82 and 0.81 using the standard and relaxed baseline data windows, respectively. The genotype models achieved AUC values of 0.86 with the new non-adherence filter and 0.84 without. Both sets of models were significantly more accurate than genotyping with rules-based interpretation, which achieved AUC values of only 0.55–0.63, and were marginally more accurate than previous models. The models were able to identify alternative regimens that were predicted to be effective for the vast majority of cases in which the new regimen prescribed in the clinic failed.

Conclusions: These latest global models predict treatment responses accurately even without a genotype and have the potential to help optimize therapy, particularly in resource-limited settings.

Introduction

The development of approximately 30 HIV drugs acting at six different points in the virus life cycle and the expansion of access to therapy around the world is a great success story.[1] The current United Nations Programme on HIV/AIDs (UNAIDS) target for 2020 is for 90% of infected people to be diagnosed and 90% of them to be on therapy with 90% of those treated having suppressed virus ('90–90–90'). The last target is critical in order not only to prevent disease progression, morbidity and mortality but to decrease the spread of the virus.[2,3] A major threat to this is the development of HIV drug resistance, often linked to poor adherence and interruptions to drug supplies in some settings.

A recent report from the WHO and others showed that the prevalence of HIV drug resistance among patients in public health ART programmes has increased from 11% to 29% since the global expansion of ART to low- and middle-income countries (LMICs) began in 2001.[1]

When treatment fails, the combination of antiretroviral agents should be changed in order to resuppress the virus. In most

well-resourced countries the selection of a new combination is individualized by expert physicians using information including the patient's treatment history and the results of a genotypic resistance test.[4–6] However, resistance testing is relatively expensive and only moderately predictive of response to treatment.[7]

The challenge of individually optimized drug selection in LMICs is even greater as resistance tests are typically unavailable or unaffordable and drug options are limited.[8] In the absence of routine viral load monitoring, therapy failure is often detected late and regimen switch decisions are based on standard protocols rather than individualized. The result can be suboptimal regimen selection, failure to achieve viral resuppression and further resistance selection, which may limit future therapeutic options and can be transmitted to others.[9]

The HIV Resistance Response Database Initiative (RDI) has collected biological, clinical and treatment outcome data for more than 200 000 HIV-1 patients around the world over a period of 16 years. From these data, we have used machine learning to develop models to predict HIV-1 treatment outcomes and to identify optimal, individualized therapies.[10–15] We have developed models that use information from genotypic resistance tests in their predictions and others that do not. The most recent models, developed using large data sets from around the world then tested with independent test sets predicted virological response with an overall accuracy (OA) of around 80% with a genotype and 74% without.[14,15]

The models are used to power an online treatment decision support tool, the HIV Treatment Response Prediction System (HIV-TRePS). To keep this system as current as possible in terms of the inclusion of new drugs and reflection of current clinical practice it is essential that new models are regularly developed using the latest data.

Here we report the development of two new sets of random forest (RF) models that estimate the probability of combinations of antiretroviral drugs reducing the plasma viral load to undetectable (<50 copies HIV RNA/mL):

1. Models that do not require a genotype for their predictions (no-genotype or NG models), trained using a large global data set and intended for use in LMICs without access to genotyping. We compared standard models (NG1) with experimental models developed using new highly permissive data inclusion criteria (NG2) to increase utility in LMICs where clinic visits can be infrequent.
2. Global models that use a viral genotype in their predictions (global genotype or G models). These were developed using data screened for likely non-adherence using an experimental filter. Cases of discordance between virological failure observed in the clinic and predictions of response by both our current models and genotyping with rules-based interpretation were excluded (G2) and the resultant models compared with 'standard' models (G1).
3. For the first time there were sufficient data without genotypes for the NG models to be trained to predict outcomes for three drugs not previously covered: tipranavir, maraviroc and elvitegravir.

The accuracy of all the models was ascertained and they were evaluated as potential tools to support optimized, individualized treatment decision-making in the RDI's HIV-TRePS system. This paper represents the latest update alluded to in our previous publications of modelling.[13–15]

Methods

Clinical data

Treatment change episodes (TCEs) were collected from cases in which ART was changed following virological failure.[10] TCEs for development of the NG models had all the following data available: on-treatment baseline plasma viral load (obtained ≤8 weeks prior to treatment change for the standard models, NG1, and ≤12 weeks for the experimental models, NG2); on-treatment baseline CD4 cell count (≤12 weeks prior to treatment change for NG1 and ≤16 weeks for NG2); the drugs in use prior to the change; ART history; drugs in the new regimen; follow-up plasma viral load obtained 4–52 weeks following introduction of the new regimen and time to that follow-up. A similar extraction was subsequently performed for TCEs that also had an on-treatment genotype (protease and reverse transcriptase sequence ≤12 weeks prior to treatment change) for the development of the genotype models.

The TCEs were censored using rules established in previous studies and published in detail elsewhere.[16]

Data partition for models without genotypes (NG)

The qualifying TCEs were partitioned using methods described elsewhere.[11,16] The partition scheme is illustrated in Figure 1. For the NG2 models, with the expanded baseline data windows, a training set of 50 270 TCEs and an independent test set of 3000 TCEs, one each from 3000 patients not represented in the training set, were obtained. For the NG1 models, the standard baseline data windows were applied to the NG2 training and test sets resulting in training and test sets reduced to 43 239 and 2500 TCEs.

Data partition for models with genotypes (G)

The data that included baseline genotypes were partitioned into a master pool of 18 242 training TCEs and 1000 test TCEs. The data were screened for possible non-adherence using our standard filter (which excludes cases with a baseline viral load of ≤3.0 $\log_{10}$ HIV RNA and an increase in viral load of ≥2.0 following the introduction of a new regimen selected with a recent genotype available). This resulted in a training set of 18 188 and a test set of 997. The master pool of TCEs was then screened using a trial filter for possible non-adherence. Cases in which the new regimen used in the clinic failed to achieve virological response were passed through HIV-TRePS. Those predicted to respond (follow-up viral load ≤50 copies) were extracted and genotypic sensitivity scores (GSSs) were obtained using the Stanford HIVDB, REGA and ANRS interpretation systems. Cases with a GSS of ≥2 (two or more active drugs) in all three systems were then excluded. This removed ∼5% of the TCEs, resulting in a training set of 17 378 and test set of 940.

Computational model development

The two NG training sets of TCEs were used to train two committees of five RF models, to estimate the probability of the follow-up viral load being less than 50 copies/mL, using methodology described elsewhere.[11,13] The following 47 input variables were used (new variables underlined): baseline viral load ($\log_{10}$ copies HIV RNA/mL); baseline CD4 count (cells/mm^3); treatment history—22 binary variables coding for experience of zidovudine, didanosine, stavudine, abacavir, lamivudine, emtricitabine, tenofovir disoproxil fumarate, efavirenz, nevirapine, etravirine, indinavir, nelfinavir, saquinavir, amprenavir, fosamprenavir, lopinavir, atazanavir, darunavir, enfuvirtide, raltegravir, tipranavir and maraviroc; antiretroviral drugs in the new regimen—23 variables as above with the addition of elvitegravir; and time to follow-up (days). The output variable was the follow-up viral load as

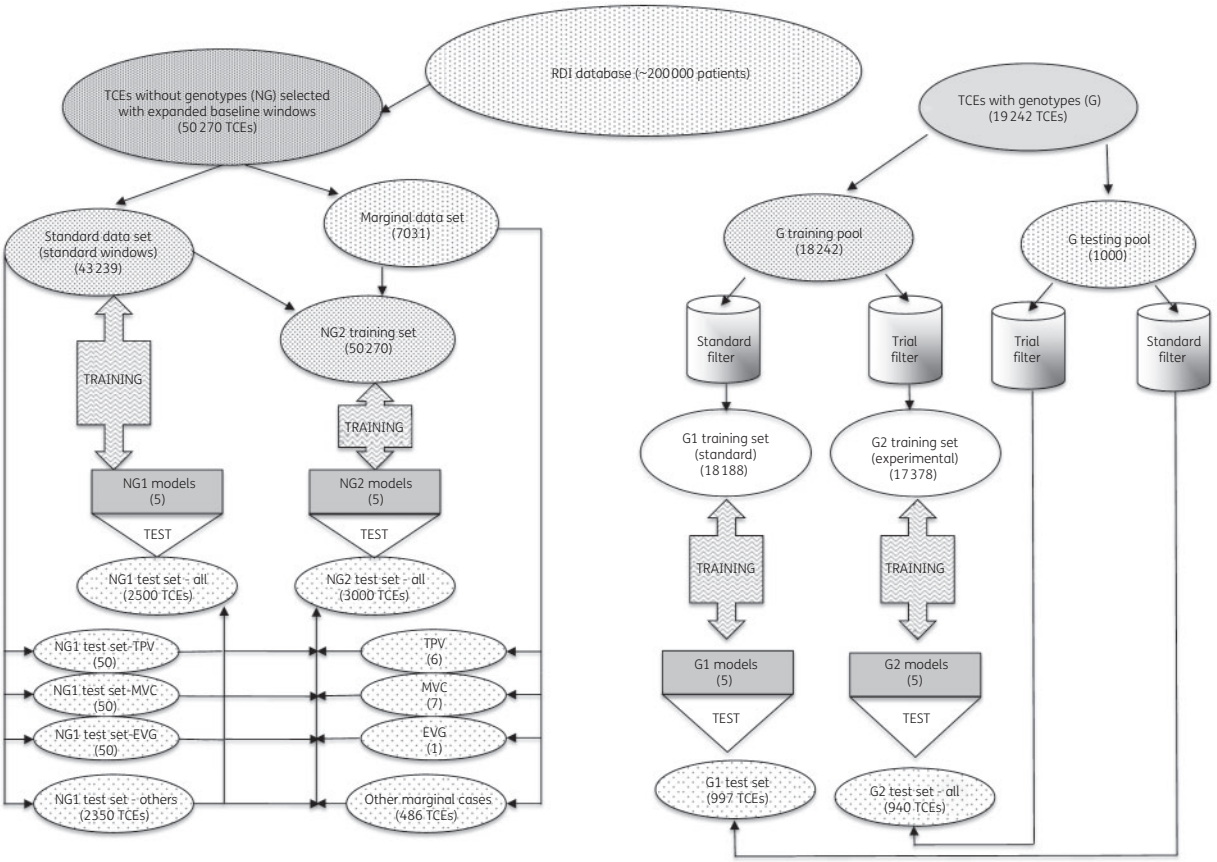

Figure 1. Partition scheme. EVG, elvitegravir; MVC, maraviroc; TPV, tipranavir.

a binary variable: ≤1.7 log or 50 copies/mL = 1 (response) and >1.7 log or 50 copies/mL = 0 (failure).

The genotype models used 105 input variables including the above but without raltegravir as a historical drug and without maraviroc or tipranavir in the new regimen because of insufficient data with genotypes. History of maraviroc use was a new variable. In addition, the following 62 mutations, detected in the baseline genotype were used: HIV reverse transcriptase mutations ($n = 33$): M41L, E44D, A62V, K65R, D67N, 69 insert, T69D/N, K70R, L74V, V75I, F77L, V90I, A98G, L100I, L101I/E/P, K103N, V106A/M, V106I, V108I, Y115F, F116Y, V118I, 138A/G/K, Q151M, V179D/F/T, Y181C/I/V, M184V/I, Y188C/L/H, G190S/A, L210W, T215F/Y, K219Q/E and P236L; and protease mutations ($n = 29$): L10F/I/R/V, V11I, K20M/R, L24I, D30N, V32I, L33F, M36I, M36L/V, M46I/L, I47V, G48V, I50V, I50L, F53L, I54 (any change), 58E, L63P, A71(any change), G73(any change), T74P, L76V, V77I, V82A/F/S, V82T, I84V/A/C, N88D/S, L89V and L90M. The mutations were selected on the basis of the IAS-USA mutation list as well as previous modelling studies.[17]

Validation and independent testing

Each of the four committees of five RF models was developed using a 5× cross-validation scheme.[11,16] For each partition the model's estimates of the probability of response for the validation cases was compared with the actual response observed in the clinic and the best-performing model selected for the final committee. For each of the five final models, the optimum operating point (OOP) was identified (the cut-off for the probability of response being classed as response versus failure that gave the best performance overall).

The performance of the models as predictors of response was then evaluated using the independent test cases. The models' estimates of the probability of response and the responses observed in the clinics for these cases were used to plot receiver operator characteristic (ROC) curves and assess the area under the ROC curve (AUC). In addition, the average OOP, derived during cross-validation, was used to obtain the OA (the percentage of outcomes that were correctly predicted), the sensitivity and the specificity of the models.

Comparison of the accuracy of the models versus rules-based interpretation of the genotype

GSSs were obtained for test cases with genotypes that could be fully interpreted by three rules-based genotype interpretation systems in common use: ANRS, REGA and Stanford HIVDB. The three systems were accessed online on 7 September 2017 and the GSSs calculated by adding the score for each drug in the regimen, with full sensitivity scored as 1, partial as 0.5 and no response as 0. These scores were then used as predictors of response and the performance compared with that of the models.[16]

In silico analysis to evaluate the potential of the models to help avoid treatment failure

In order to evaluate further the potential clinical utility of the models, we assessed their ability to identify alternative, practical regimens that were predicted to be effective (probability of virological response above the OOP), or more likely to be effective than the regimens introduced in the clinic. Lists of regimens in regular clinical use were identified from the RDI database. The baseline data for all test TCEs in which the new regimen comprised three or more drugs were entered into the models and predictions obtained for the regimens on the drug lists that had no more drugs than

the regimen used in the clinic. Since the NG models are used primarily in LMICs, we wanted to avoid modelling regimens that are unavailable in such settings, as this could overestimate the system's utility. The analysis was therefore repeated using test cases from sub-Saharan countries only and modelling alternative regimens comprising only those drugs that were in use in those countries at the time of data collection.

Results

Characteristics of the data sets

The baseline, treatment and response characteristics of the data sets are summarized in Tables 1 and 2. The training sets for NG models comprised 43 239 TCEs using the standard baseline data windows and 50 270 using the expanded baseline data windows. The data sets have very comparable baseline data with a median plasma viral load of ∼3.8 $\log_{10}$ copies/mL and CD4 count of ∼260 cells/mm^3. The median number of previous drugs used in the patients' treatment was 4–5, with almost all exposed to nucleoside or nucleotide reverse transcriptase inhibitors [N(t)RTIs], around two-thirds having been exposed to NNRTIs and two-thirds to protease inhibitors (PIs). There was a broad range of new regimens represented in the data, the most common being two NRTIs and one PI (32%–36%), followed by two NRTIs and an NNRTI (18%–23%).

The characteristics of the TCEs with genotypes are summarized in Table 2. The training sets comprised 18 188 TCEs using the standard non-adherence filter and 17 378 using the new filter. The sets have very comparable baseline data with a median plasma viral load of ∼4.3 $\log_{10}$ copies/mL (about half a log higher than the NG data) and median CD4 count of ∼230 cells/mm^3, slightly lower than the NG data. The treatment history was similar to that of the NG data, as was the range of new regimens other than somewhat fewer patients changing to NNRTI-based regimens (∼10% versus 20%).

Results of the modelling without a genotype

The performance characteristics from the ROC curves of the models during cross-validation and independent testing are summarized in Table 3. The NG1 models achieved AUC values during cross-validation of 0.83 to 0.84, with a mean of 0.84. The OA was 77% to 78% (mean = 77%), the sensitivity was 71% for all five models and the specificity ranged from 80% to 81% (mean = 80%). The OOP was 0.42–0.43.

The NG2 models achieved very similar results with AUC values during cross-validation of 0.83 to 0.84 and a mean of 0.83. OA ranged from 76% to 77% (mean = 76%), sensitivity again was 71% and specificity ranged from 79% to 80% (mean = 80%).

Independent testing

The NG1 models achieved an AUC of 0.82 in independent testing (Figure 2). OA was 75%, sensitivity 72% and specificity 77%. The NG2 models achieved an AUC value of 0.81, OA of 75%, sensitivity was 73% and specificity 76%. The performance of the two sets of models during independent testing was not significantly different ($P = 0.84$).

When NG2 models were tested using only the 500 test cases with baseline data that fell outside of the standard windows, the AUC was 0.79, OA 73%, sensitivity and specificity each 73%. There were no significant differences between the performance of NG2 models with those cases with baseline data inside versus outside the standard windows ($P = 0.29$).

When the two sets of models were tested only with those cases involving each of the three new drugs (50 cases in each subset) the AUC values ranged from 0.75 to 0.89.

Comparing the predictive accuracy of the models versus genotyping

Of the 3000 TCEs in the global NG test set, 634 had genotypes available that were suitable for full interpretation by the three interpretation systems. The ROC curves are presented in Figure 2, alongside those for the models. The AUC values for the GSS were 0.56 (ANRS), 0.58 (Stanford HIVDB) and 0.55 (REGA) (Table 4). All were significantly less accurate predictors of virological response than the NG models, both sets of which achieved an AUC of 0.81 for these cases ($P < 0.0001$).

Results of the modelling with a genotype

The performance characteristics from the ROC curves of these models during cross-validation and independent testing are summarized in Table 5. The five G1 models achieved AUC values during cross-validation ranging from 0.84 to 0.86, with a mean of 0.86. OA was 78% to 79% (mean = 79%), sensitivity 71% to 73% (mean = 73%) and specificity 78% to 82% (mean = 81%). The OOP was 0.42.

The five G2 models (using data with the new non-adherence filter) achieved AUC values during cross-validation ranging from 0.86 to 0.89, with a mean of 0.88. OA was 79% to 82% (mean = 80%), sensitivity 77% to 81% (mean = 79%) and specificity 79% to 82% (mean = 81%). The OOP was again 0.42.

Independent testing

When tested with the independent G1 test cases using the OOP developed in cross-validation, the G1 models achieved an AUC of 0.84 (Figure 3). The OA was 76%, sensitivity 72% and specificity 80%. G2 models achieved an AUC value during testing with the G2 test set of 0.86, OA of 79%, sensitivity of 74% and specificity of 83%. Again, the G2 models' performance was slightly better than G1 but there was no statistically significant difference between the two sets of models ($P = 0.25$).

When G1 models were tested using the G2 test set, performance improved slightly (AUC from 0.84 to 0.85 and OA from 76% to 77%) but was not as good as with the G2 models. Conversely, the performance of the G2 models worsened when tested with the G1 test set (AUC reduced from 0.86 to 0.83 and OA from 79% to 76%), but remained comparable with the performance of the G1 models.

Comparing the predictive accuracy of the G2 genotype models versus genotyping

GSSs were generated for 856 test TCEs in which the drugs in the new regimens were fully covered by the interpretation systems. The genotype systems achieved AUC values of 0.60–0.63, compared with 0.86 using the G2 models and 0.84 for G1 (Figure 3).

Table 1. Demographic characteristics of the TCEs without genotypes (NG)

	NG1 training set	NG1 test set	NG2 training set	NG2 test set
TCEs, *n*	43 239	2500	50 270	3000
Patients, *n*	13 970	2500	15 850	3000
Age (years), median	42	41	42	41
Gender, *n*				
male	29 290	1647	33 751	1947
female	8347	544	10 157	704
not known	5602	309	6362	349
Geographical sources of TCEs, *n*				
Argentina	112	11	177	27
Australia	481	33	505	34
Brazil	3	1	5	1
Canada	3554	197	4214	238
Germany	4679	244	5077	266
India	330	37	469	57
Italy	1418	86	1632	97
Japan	116	6	133	8
Mexico	308	28	415	34
Netherlands	6173	368	7298	464
Romania	434	51	603	68
Serbia	0	0	1	0
South Africa	3032	303	4190	451
Spain	4727	226	5856	258
UK	9530	451	10 735	507
USA	1876	77	2116	95
sub-Saharan Africa (country unknown)	52	9	66	10
unknown (from multinational cohorts/trials)	6414	372	6758	385
Baseline data, median (IQR)				
baseline VL (log_{10} copies/mL)	3.86 (2.75–4.73)	3.81 (2.65–4.69)	3.83 (2.75–4.7)	3.78 (2.66–4.66)
baseline CD4 (cells/mm^3)	260 (134–417)	256 (130–405)	261 (139–420)	260 (135–417)
Treatment history				
number of previous drugs, median (IQR)	5 (3–8)	4 (3–6)	5 (3–7)	4 (3–6)
N(t)RTI experience, *n* (%)	43 119 (99.7)	2496 (99.8)	50 142 (99.7)	2995 (99.8)
NNRTI experience, *n* (%)	28 653 (66.3)	1675 (67.0)	33 398 (66.4)	2026 (67.5)
PI experience, *n* (%)	30 030 (69.5)	1551 (62.0)	34 461 (68.6)	1776 (59.2)
number of previous regimens, median (IQR)	4 (3–9)	4 (2–7)	4 (3–9)	3 (2–7)
New regimens, *n* (%)				
2 N(t)RTIs + 1 PI	13 915 (32.2)	861 (34.4)	16 649 (33.1)	1065 (35.5)
2 N(t)RTIs + 1 NNRTI	7960 (18.4)	550 (22)	9446 (18.8)	690 (23)
3 N(t)RTIs + 1 PI	2963 (6.9)	167 (6.7)	3354 (6.7)	189 (6.3)
3 N(t)RTIs	1921 (4.4)	94 (3.8)	2179 (4.3)	111 (3.7)
3 N(t)RTIs + 1 NNRTI	1493 (3.5)	56 (2.2)	1661 (3.3)	66 (2.2)
2 N(t)RTIs	1211 (2.8)	63 (2.5)	1542 (3.1)	79 (2.6)
2 N(t)RTIs + 1 NNRTI + 1 PI	1305 (3)	63 (2.5)	1510 (3)	74 (2.5)
4 N(t)RTIs	750 (1.7)	37 (1.5)	860 (1.7)	43 (1.4)
1 N(t)RTI + 1 NNRTI + 1 PI	883 (2)	43 (1.7)	1019 (2)	49 (1.6)
1 N(t)RTI + 1 PI	725 (1.7)	42 (1.7)	860 (1.7)	49 (1.6)
other	10 113 (23.4)	524 (21)	11 190 (22.3)	585 (19.5)

VL, viral load; N(t)RTI, nucleoside or nucleotide reverse transcriptase inhibitor.

The OA values were 57%–59% for genotyping compared with 80% for the G2 models and 77% for G1. All three genotype interpretation systems were significantly worse at predicting responses than both sets of models ($P < 0.00001$).

In silico analysis

The NG models were able to identify alternative regimens that were predicted to be effective for 97% (NG1) to 98% (NG2) of cases (Table 6). They were able to identify alternative regimens

Table 2. Demographic characteristics of the TCEs including a genotype (G)

	G1 training set	G1 test set	G2 training set	G2 test set
TCEs, *n*	18 188	997	17 378	940
Patients, *n*	6844	997	6700	940
Gender, *n*				
male	11 364	601	10 887	565
female	2700	150	2544	141
unknown	4124	246	3947	234
Geographical sources of TCEs, *n*				
Australia	327	24	307	22
Canada	1773	104	1616	99
Germany	1729	84	1617	74
India	82	6	70	5
International	1015	51	1009	51
Italy	869	69	852	66
Japan	115	7	113	7
Netherlands	1365	74	1297	68
Romania	34	1	34	1
South Africa	176	14	142	12
Spain	2095	116	2017	105
sub-Saharan Africa	44	5	40	4
UK	2767	118	2653	114
USA	544	41	521	37
unknown	5253	283	5090	275
total	18 188	997	17 378	940
Baseline data, median (IQR)				
baseline VL (log_{10} copies/mL)	4.23 (3.47–4.89)	4.3 (3.51–4.9)	4.27 (3.5–4.9)	4.3 (3.58–4.9)
baseline CD4 (cells/mm^3)	229 (112–380)	240 (113–382)	226 (110–377)	234 (112–380)
Treatment history				
number of previous drugs, median (IQR)	5 (3–8)	5 (3–7)	5 (3–8)	5 (3–7)
N(t)RTI experience, *n* (%)	18 124 (99.7)	990 (99.3)	17 317 (99.6)	933 (99.3)
NNRTI experience, *n* (%)	12 019 (66.1)	641 (64.3)	11 483 (66)	602 (64)
PI experience, *n* (%)	13 015 (71.6)	717 (71.9)	12 536 (72.1)	684 (72.8)
number of previous regimens, median (IQR)	3 (2–7)	3 (2–6)	3 (2–7)	3 (2–6)
New regimens, *n* (%)				
2 N(t)RTIs + 1 PI	5887 (32.4)	310 (31.09)	5463 (31.4)	278 (29.6)
2 N(t)RTIs + 1 NNRTI	1686 (9.3)	100 (10.03)	1597 (9.2)	97 (10.3)
3 N(t)RTIs + 1 PI	1855 (10.2)	105 (10.53)	1745 (10.0)	99 (10.5)
3 N(t)RTIs	792 (4.4)	32 (3.21)	781 (4.5)	32 (3.4)
3 N(t)RTIs + 1 NNRTI	651 (3.6)	41 (4.11)	634 (3.6)	40 (4.3)
2 N(t)RTIs	490 (2.7)	29 (2.91)	488 (2.8)	28 (3.0)
2 N(t)RTIs + 1 NNRTI + 1 PI	763 (4.2)	50 (5.02)	735 (4.2)	47 (5.0)
1 PI + 1 integrase inhibitor	0 (0)	0 (0)	0 (0)	0 (0)
4 N(t)RTIs	424 (2.3)	17 (1.71)	413 (2.4)	17 (1.8)
1 N(t)RTI + 1 NNRTI + 1 PI	409 (2.2)	38 (3.81)	392 (2.3)	37 (3.9)
1 N(t)RTI + 1 PI	345 (1.9)	14 (1.40)	344 (2.0)	12 (1.3)
other	4886 (26.9)	261 (26.18)	4786 (27.5)	253 (26.9)

VL, viral load; N(t)RTI, nucleoside or nucleotide reverse transcriptase inhibitor.

comprising only those drugs in common use with a higher probability of response (but not necessarily above the response classification threshold of the models) for all cases. For the cases in which the new regimen failed in the clinic, the models were able to identify alternatives that were predicted to be effective in 96% and with a higher probability of response in 100%.

Using only locally available drugs (lamivudine, abacavir, zidovudine, didanosine, efavirenz, emtricitabine, lopinavir, nevirapine,

Table 3. Results of the modelling without a genotype (NG)

	NG1 models (standard data windows)					NG2 models (expanded data windows)				
	AUC	sensitivity (%)	specificity (%)	OA (%)	OOP	AUC	sensitivity (%)	specificity (%)	OA (%)	OOP
Cross-validation during model development										
model										
1	0.84	71	81	78	0.43	0.84	71	80	77	0.42
2	0.84	71	80	77	0.42	0.83	71	79	76	0.42
3	0.83	71	80	77	0.42	0.83	71	79	76	0.41
4	0.83	71	80	77	0.42	0.83	71	79	76	0.42
5	0.83	71	80	77	0.43	0.83	71	79	76	0.43
mean	0.84	71	80	77	0.42	0.83	71	80	76	0.42
min	0.83	71	80	77	0.42	0.83	71	79	76	0.41
max	0.84	71	81	78	0.43	0.84	71	80	77	0.43
Independent testing										
main test set (NG1, $n = 2500$; NG2, $n = 3000$)	0.82	72	77	75	0.42	0.81	73	76	75	0.42
marginal cases[a] ($n = 500$)						0.79	73	73	73	0.42
new drug subsets										
EVG ($n = 50$)	0.81	80	79	80	0.61	0.75	74	79	76	0.63
MVC ($n = 50$)	0.89	88	89	88	0.52	0.83	74	74	74	0.52
TPV ($n = 50$)	0.84	71	73	73	0.35	0.85	70	82	80	0.43

EVG, elvitegravir; MVC, maraviroc; TPV, tipranavir.
[a]Cases with baseline data outside the standard windows used for M1.

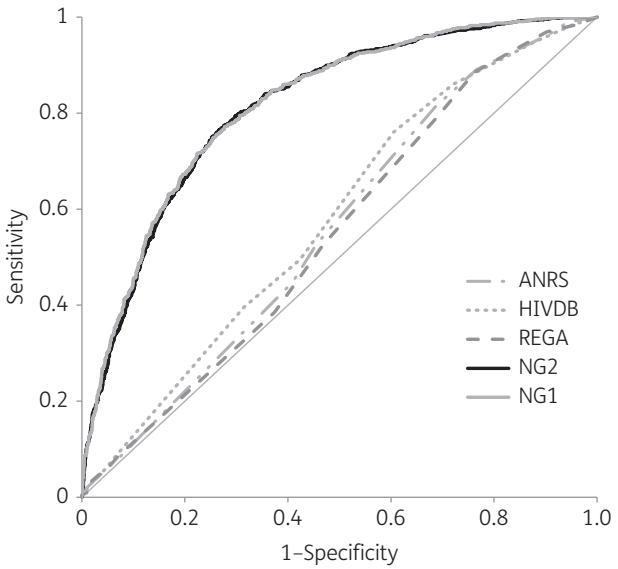

Figure 2. ROC curves for the NG models versus genotyping (GSS).

raltegravir and tenofovir disoproxil fumarate) for 450 cases from sub-Saharan Africa, the NG models were able to identify alternative regimens that were predicted to be effective for 95% of cases and 92%–93% for cases in which the new regimen failed in the clinic.

The genotype models identified alternative regimens that were predicted to give a response for 93% of the test cases and regimens with a higher probability of response for 99.9% (Table 6). For patients who experienced virological failure in the clinic, the models identified alternatives that were predicted to give a response for 90% and with a higher probability of response than the regimen in the clinic for all 100%.

Table 4. Comparison of model predictions versus GSS for test TCEs with genotypes

Prediction system	AUC	Sensitivity (%)	Specificity (%)	OA (%)	*P* (GSS versus either models)
NG1 models	0.82	71	77	75	
NG2 models	0.81	68	78	74	
Total ANRS score	0.56	53	55	54	<0.0001
Total HIVDB score	0.58	40	68	56	<0.0001
Total REGA score	0.55	53	53	53	<0.0001
G1 models	0.84	73	80	77	
G2 models	0.86	76	82	80	
Total ANRS score	0.61	53	61	58	<0.0001
Total HIVDB score	0.63	54	63	59	<0.0001
Total REGA score	0.60	57	57	57	<0.0001

Discussion

These latest computational models, developed using our largest databases, are the most accurate predictors of response to combination ART to date. They include, for the first time, tipranavir, maraviroc and elvitegravir.

Both sets of models achieved AUC values over 0.80 in cross-validation and independent testing. The results replicated and

Table 5. Results of modelling with a genotype (G)

	G1 models (standard non-adherence filter)					G2 models (experimental filter)				
	AUC	sensitivity (%)	specificity (%)	OA (%)	OOP	AUC	sensitivity (%)	specificity (%)	OA (%)	OOP
Cross-validation during model development										
model										
1	0.86	71	78	78	0.41	0.86	78	79	79	0.43
2	0.86	73	82	79	0.43	0.89	81	82	82	0.42
3	0.86	73	82	79	0.43	0.89	81	81	81	0.42
4	0.84	73	80	78	0.42	0.87	77	80	79	0.41
5	0.86	73	82	79	0.43	0.88	79	81	80	0.41
mean	0.86	73	81	79	0.42	0.88	79	81	80	0.42
min	0.84	71	78	78	0.41	0.86	77	79	79	0.41
max	0.86	73	82	79	0.43	0.89	81	82	82	0.43
Independent testing										
test set										
G1 ($n = 997$)	0.84	72	80	76	0.42	0.83	75	76	76	0.42
G2 ($n = 940$)	0.85	72	82	77	0.42	0.86	74	83	79	0.42

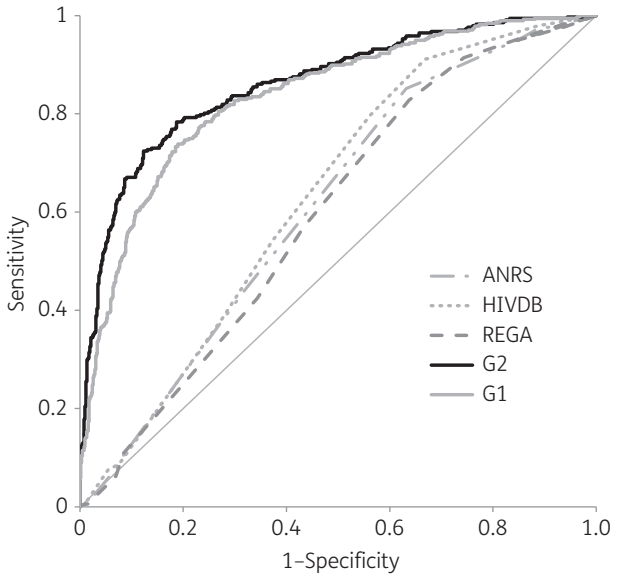

Figure 3. ROC curves for the G models versus genotyping (GSS).

reinforced previous findings that our models are substantially more accurate predictors of virological response to combination therapy than viral genotyping with rules-based interpretation.[14,15] It is encouraging that this superiority was maintained despite eliminating a number of cases in which the GSS and the response observed in the clinic were discordant, in an attempt to exclude non-adherent patients.

The expanded 'windows' for baseline data make these models more practical for use in LMICs where visits for laboratory monitoring are relatively infrequent. Moreover, the broad range of settings represented in the study data suggests that these findings and the potential utility of these models are highly generalizable. Nevertheless, in some LMICs viral loads or CD4 counts may lay outside even these extended windows, if they are available at all, preventing use of the system. Given the size of the RDI database, research looking at models that can accept missing values is warranted.

The use of a new more stringent filter for presumed non-adherence removed a greater proportion of available TCEs than the standard filter and led to a small numerical increase in performance for the genotype models.

The NG models presented here can predict outcomes to 23 different drugs, including some relatively recently approved inhibitors that are not routinely available in LMICs. Users of the HIV-TRePS system are able to exclude any drugs that are not locally available from the modelling. The *in silico* results for cases from LMICs, using a highly restricted list of locally available drugs, demonstrated the potential of the models to improve virological response rates nevertheless, underlining their applicability for LMICs.

A key input variable for these models was the plasma viral load, which studies have shown to be important for the accuracy of the models.[18] Although viral load monitoring is not routine in LMICs, it is now recommended in WHO guidelines for monitoring ART response.[19] Accurate statistics on viral load monitoring in LMICs are scarce. However, a recent study of its scale-up in sub-Saharan Africa showed the percentage of patients with viral loads ranged from 3% (Côte d'Ivoire and Tanzania) to 96% in Namibia. Of 11 million patients on ART in the region, 5 million were estimated to have access to viral load monitoring.[20] As technological advances enable lower costs and point-of-care testing, the use of viral load is likely to increase.[21,22]

The study has some limitations. Firstly, it was retrospective and no firm claims can be made for the clinical benefit of using the system as a treatment support tool. This would require large, prospective clinical trials, for example comparing outcomes for patients with a change to their treatment managed under standard of care (SoC) versus SoC plus the HIV-TRePS report.

The models described here were trained to estimate the probability of response to therapy using a definition of response of <50 copies HIV RNA/mL. Although recent data strongly suggest that low-level viraemia predicts virological failure, differences

Table 6. Results of *in silico* analysis for the G and NG models

Test set	Measure	NG1/G1	NG2/G2
NG models			
all cases with ≥3 drugs in their new regimen	alternatives predicted to be effective (%)	97	98
2315 cases for NG1 and 2783 for NG2[a]	alternatives with higher probability of response than regimen used in clinic (%)	100	100
subset of the above that failed the new regimen introduced in the clinic	alternatives predicted to be effective (%)	96	96
1433 cases for NG1 and 1716 for NG2[a]	alternatives with higher probability of response than regimen used in clinic (%)	100	100
G models			
all cases with ≥3 drugs in their new regimen	alternatives predicted to be effective (%)	93	93
892 cases for G1 and 841 for G2[a]	alternatives with higher probability of response than regimen used in clinic (%)	99.9	99.9
subset of the above that failed the new regimen introduced in the clinic	alternatives predicted to be effective (%)	90	90
564 cases for G1 and 513 for G2[a]	alternatives with higher probability of response than regimen used in clinic (%)	100	100

[a]Only those cases with ≥3 drugs in the new regimen were used in these analyses.

persist in the definition of virological response used in the clinic.[23] The US AIDS Clinical Trials Group (ACTG) define virological failure as a confirmed viral load >200 copies/mL.[6] The WHO, however, defines virological failure as persistent plasma RNA levels ≥1000 copies/mL after 3 months with adherence support.[19] Other groups are using 400 copies/mL.[24]

The models may predict that a certain combination of drugs is likely to fail (viral load >50 copies/mL) with no indication of the probability of the viral load being below a different cut-off, e.g. 1000 copies/mL. Studies are now ongoing to develop models that predict absolute viral load value over time following a treatment change.

Conclusions

Computational models developed using large, heterogeneous data sets with relatively permissive rules governing the timing of baseline data can be highly accurate predictors of virological response to combination ART, even without a genotype. Such models are of enhanced utility in settings with infrequent laboratory monitoring.

Attempts to remove any possible contamination of training with data from non-adherent patients continue. A new filter for presumed non-adherence removed around 5% of training data and led to a very small increase in accuracy.

The models were able to predict responses to tipranavir, maraviroc and elvitegravir for the first time and with accuracy comparable with that of other antiretrovirals, again expanding the utility of the system.

These latest models are better predictors of response to therapy than genotyping with rules-based interpretation, even when those models do not use a genotype for their predictions. Since use of these models is free of charge, this suggests that scarce funds in LMICs would be better spent on antiretroviral drugs and viral load testing than on genotyping. This would enable a greater range of treatments to be offered, treatment failure to be detected earlier and optimal, individualized treatment change decisions made using the models.

Full validation of this approach as a clinical tool would require a prospective, controlled clinical trial. Nevertheless, the results suggest that these models have the potential to reduce virological failure and improve patient outcomes in all parts of the world, with particular utility in LMICs. The use by clinicians of this tool to support optimized treatment decision-making in the absence of resistance tests could also combat the development of drug resistance and its contribution to treatment failure, disease progression and onward viral transmission.

The global models described in this paper are freely available to use online through the HIV-TRePS system at http://www.hivrdi.org/treps.

Acknowledgements

Members of the RDI Data and Study Group

The RDI wishes to thank all the following individuals and institutions for providing the data used in training and testing its models:

Cohorts

Peter Reiss and Ard van Sighem (ATHENA, The Netherlands); Julio Montaner and Richard Harrigan (BC Center for Excellence in HIV & AIDS, Canada); Tobias Rinke de Wit, Raph Hamers and Kim Sigaloff (PASER-M cohort, The Netherlands); Brian Agan, Vincent Marconi and Scott Wegner (US Department of Defense); Wataru Sugiura (National Institute of Health, Japan); Maurizio Zazzi (MASTER, Italy); Rolf Kaiser and Eugen Schuelter (Arevir Cohort, Köln, Germany); Adrian Streinu-Cercel (National Institute of Infectious Diseases "Prof. Dr. Matei Balş", Bucharest, Romania); Gerardo Alvarez-Uria (VFHCS, India); Maria-Jesus Perez-Elias, (CORIS, Spain); Tulio de Oliveira (SATuRN, South Africa).

Clinics

Jose Gatell and Elisa Lazzari (University Hospital, Barcelona, Spain); Brian Gazzard, Mark Nelson, Anton Pozniak and Sundhiya Mandalia (Chelsea and Westminster Hospital, London, UK); Colette Smith (Royal Free Hospital, London, UK); Lidia Ruiz and Bonaventura Clotet (Fundacion Irsi Caixa, Badelona, Spain); Schlomo Staszewski (Hospital of the Johann Wolfgang Goethe-University, Frankfurt, Germany); Carlo Torti (University of Brescia, Brescia, Italy); Cliff Lane, Julie Metcalf and Catherine A. Rehm (National Institutes of Health Clinic, Rockville, USA); Maria-Jesus Perez-

Elias (Instituto Ramón y Cajal de Investigación Sanitaria, Madrid, Spain); Stefano Vella and Gabrielle Dettorre (Sapienza University, Rome, Italy); Andrew Carr, Richard Norris and Karl Hesse (Immunology and Ambulatory Care Service, St. Vincent's Hospital, Sydney, NSW, Australia); Emanuel Vlahakis (Taylor's Square Private Clinic, Darlinghurst, NSW, Australia); Hugo Tempelman and Roos Barth (Ndlovu Care Group, Elandsdoorn, South Africa), Robin Wood, Carl Morrow and Dolphina Cogill (Desmond Tutu HIV Centre, University of Cape Town, South Africa); Chris Hoffmann (Aurum Institute, Johannesburg, South Africa and Johns Hopkins University, Boston, USA); Luminita Ene ("Dr Victor Babes" Hospital for Infectious and Tropical Diseases, Bucharest, Romania); Gordana Dragovic (University of Belgrade, Belgrade, Serbia); Ricardo Diaz and Cecilia Sucupira (Federal University of Sao Paulo, Sao Paulo, Brazil); Omar Sued and Carina Cesar (Fundación Huésped, Buenos Aires, Argentina); Juan Sierra Madero (Instituto Nacional de Ciencias Medicas y Nutricion Salvador Zubiran, Mexico City, Mexico); Pachamuthu Balavskrishnan and Shanmugam Saravanan (YRG Care, Chennai, India).

Clinical trials

Sean Emery and David Cooper (CREST); Carlo Torti (GenPherex); John Baxter (GART, MDR); Laura Monno and Carlo Torti (PhenGen); Jose Gatell and Bonventura Clotet (HAVANA); Gaston Picchio and Marie-Pierre deBethune (DUET 1 & 2 and POWER 3); Maria-Jesus Perez-Elias (RealVirfen); Sean Emery, Paul Khabo and Lotty Ledwaba (PHIDISA).

Funding

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. This research was supported by the National Institute of Allergy and Infectious Diseases.

Transparency declarations

None to declare.

Disclaimer

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, and the mention of trade names, commercial products or organizations does not imply endorsement by the US Government.

References

1 WHO. *HIV Drug Resistance Report 2017*. www.who.int/hiv/pub/drugresistance/hivdr-report-2017/en/.

2 Montaner JS, Lima VD, Harrigan PR *et al*. Expansion of HAART coverage is associated with sustained decreases in HIV/AIDS morbidity, mortality and HIV transmission: the "HIV Treatment as Prevention" experience in a Canadian setting. *PLoS One* 2014; **9**: e87872.

3 Nkambule R, Nuwagaba-Biribonwoha H, Mnisi Z *et al*. Substantial progress in confronting the HIV epidemic in Swaziland: first evidence of national impact. In: *Abstracts of the Ninth IAS Conference on HIV Science, Paris, France, 2017*. Abstract MOAX0204LB. International AIDS Society, Geneva, Switzerland.

4 Günthard HF, Aberg JA, Eron JJ *et al*. Antiretroviral treatment of adult HIV infection: 2014 recommendations of the International Antiviral Society–USA Panel. *JAMA* 2014; **312**: 410–25.

5 Williams I, Churchill D, Anderson J *et al*. British HIV Association guidelines for the treatment of HIV-1-positive adults with antiretroviral therapy 2012. *HIV Med* 2014; **15** Suppl 1: 1–85.

6 Panel on Antiretroviral Guidelines for Adults and Adolescents. *Guidelines for the Use of Antiretroviral Agents in HIV-1-Infected Adults and Adolescents*. Department of Health and Human Services. http://aidsinfo.nih.gov/contentfiles/lvguidelines/AdultandAdolescentGL.pdf.

7 Degruttola V, Dix L, D'Aquila R *et al*. The relation between baseline HIV drug resistance and response to antiretroviral therapy: re-analysis of retrospective and prospective studies using a standardized data analysis plan. *Antivir Ther* 2000; **5**: 41–8.

8 Conradie F, Wilson D, Basson A *et al*. The 2012 southern African ARV drug resistance testing guidelines. *South Afr J HIV Med* 2012; **13**: 162–7.

9 Gupta RK, Hill A, Sawyer AW *et al*. Virological monitoring and resistance to first-line highly active antiretroviral therapy in adults infected with HIV-1 treated under WHO guidelines: a systematic review and meta-analysis. *Lancet Infect Dis* 2009; **9**: 409–17.

10 Larder B, Wang D, Revell A *et al*. The development of artificial neural networks to predict virological response to combination HIV therapy. *Antivir Ther* 2007; **12**: 15–24.

11 Wang D, Larder B, Revell A *et al*. A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy. *Artif Intell Med* 2009; **47**: 63–74.

12 Larder BA, Revell A, Mican JM *et al*. Clinical evaluation of the potential utility of computational modeling as an HIV treatment selection tool by physicians with considerable HIV experience. *AIDS Patient Care STDS* 2011; **25**: 29–36.

13 Revell AD, Wang D, Wood R *et al*. Computational models can predict response to HIV therapy without a genotype and may reduce treatment failure in different resource-limited settings. *J Antimicrob Chemother* 2013; **68**: 1406–14.

14 Revell AD, Wang D, Wood R *et al*. An update to the HIV-TRePS system: the development of new computational models that do not require a genotype to predict HIV treatment outcomes. *J Antimicrob Chemother* 2014; **69**: 1104–10.

15 Revell AD, Wang D, Wood R *et al*. An update to the HIV-TRePS system: the development and evaluation of new global and local computational models to predict HIV treatment outcomes, with or without a genotype. *J Antimicrob Chemother* 2016; **71**: 2928–37.

16 Revell AD, Wang D, Boyd MA *et al*. The development of an expert system to predict virological response to HIV therapy as part of an online treatment support tool. *AIDS* 2011; **25**: 1855–63.

17 Wensing AM, Calvez V, Huldrych F *et al*. 2015 update of the drug resistance mutations in HIV-1. *Top Antivir Med* 2015; **23**: 132–41.

18 Revell AD, Wang D, Harrigan R *et al*. Modelling response to HIV therapy without a genotype: an argument for viral load monitoring in resource-limited settings. *J Antimicrob Chemother* 2010; **65**: 605–7.

19 WHO. *Consolidated Guidelines on the Use of Antiretroviral Drugs for Treating and Preventing HIV Infection. Recommendations for a Public Health Approach, Second Edition*. Geneva, Switzerland: WHO, 2016. http://www.who.int/hiv/pub/arv/arv-2016/en/.

20 Lecher S, Ellenberger D, Kim AA *et al*. Scale up of HIV viral load monitoring—seven sub-Saharan African countries. *MMWR Morb Mortal Wkly Rep* 2015; **64**: 1281–304.

21 Stevens WS, Scott LE, Crowe SM. Quantifying HIV for monitoring antiretroviral therapy in resource-poor settings. *J Infect Dis* 2010; **201** Suppl 1: S16–26.

22 Roberts T, Cohn J, Bonner K *et al*. Scale-up of routine viral load testing in resource-poor settings: current and future implementation challenges. *Clin Infect Dis* 2016; **62**: 1043–8.

23 Hermans LE, Moorhouse M, Carmona S *et al*. Effect of HIV-1 low-level viraemia during antiretroviral therapy on treatment outcomes in WHO-guided South African treatment programmes: a multicentre cohort study. *Lancet Infect Dis* 2018; **18**: 188–97.

24 Hassan AS, Nabwera HM, Mwaringa SM *et al*. HIV-1 virological failure and acquired drug resistance among first-line antiretroviral experienced adults at a rural HIV clinic in coastal Kenya: a cross-sectional study. *AIDS Res Ther* 2014; **11**: 9.