Journal of Antimicrobial Chemotherapy

An update to the HIV-TRePS system: the development of new computational models that do not require a genotype to predict HIV treatment outcomes

Andrew D. Revell[1]*, Dechao Wang[1], Robin Wood[2], Carl Morrow[2], Hugo Tempelman[3], Raph Hamers[4], Gerardo Alvarez-Uria[5], Adrian Streinu-Cercel[6], Luminita Ene[7], Annemarie Wensing[8], Peter Reiss[4,9], Ard I. van Sighem[9], Mark Nelson[10], Sean Emery[11], Julio S. G. Montaner[12], H. Clifford Lane[13] and Brendan A. Larder[1] on behalf of the RDI Study Group†

[1]HIV Resistance Response Database Initiative (RDI), London, UK; [2]Desmond Tutu HIV Centre, University of Cape Town, Cape Town, South Africa; [3]Ndlovu Care Group, Elandsdoorn, South Africa; [4]Department of Global Health, Academic Medical Centre of the University of Amsterdam, Amsterdam Institute for Global Health and Development, Amsterdam, The Netherlands; [5]Rural Development Trust (RDT) Hospital, Bathalapalli, AP, India; [6]National Institute of Infectious Diseases Prof. Dr. Matei Balş, Bucharest, Romania; [7]Dr. Victor Babes' Hospital for Infectious and Tropical Diseases, Bucharest, Romania; [8]Department of Virology, Medical Microbiology, University Medical Centre, Utrecht, The Netherlands; [9]Stichting HIV Monitoring, Amsterdam, The Netherlands; [10]Chelsea and Westminster Hospital, London, UK; [11]Kirby Institute, University of New South Wales, Sydney, Australia; [12]BC Centre for Excellence in HIV/AIDS, Vancouver, Canada; [13]National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA

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

Objectives: The optimal individualized selection of antiretroviral drugs in resource-limited settings is challenging because of the limited availability of drugs and genotyping. Here we describe the development of the latest computational models to predict the response to combination antiretroviral therapy without a genotype, for potential use in such settings.

Methods: Random forest models were trained to predict the probability of a virological response to therapy (<50 copies HIV RNA/mL) following virological failure using the following data from 22 567 treatment-change episodes including 1090 from southern Africa: baseline viral load and CD4 cell count, treatment history, drugs in the new regimen, time to follow-up and follow-up viral load. The models were assessed during cross-validation and with an independent global test set of 1000 cases including 100 from southern Africa. The models' accuracy [area under the receiver-operating characteristic curve (AUC)] was evaluated and compared with genotyping using rules-based interpretation systems for those cases with genotypes available.

Results: The models achieved AUCs of 0.79–0.84 (mean 0.82) during cross-validation, 0.80 with the global test set and 0.78 with the southern African subset. The AUCs were significantly lower (0.56–0.57) for genotyping.

Conclusions: The models predicted virological response to HIV therapy without a genotype as accurately as previous models that included a genotype. They were accurate for cases from southern Africa and significantly more accurate than genotyping. These models will be accessible via the online treatment support tool HIV-TRePS and have the potential to help optimize antiretroviral therapy in resource-limited settings where genotyping is not generally available.

Keywords: antiretroviral therapy, resource-limited settings, genotyping

Introduction

Optimizing antiretroviral therapy over the long term can be very challenging in resource-limited settings where the newest drugs are not available and genotyping to help select drugs to overcome or avoid resistance is not generally affordable. The HIV Resistance Response Database Initiative (RDI) has developed computational models to assist in the selection of the most effective combinations of drugs from those available. The models are able to predict accurately the virological response to combination antiretroviral therapy

with or without genotypic information, in the latter case basing their predictions on viral loads, CD4 counts, treatment history and time to follow-up.[1]

This approach has been assessed in prospective clinical pilot studies involving highly experienced HIV physicians in well-resourced settings and has been found to be a useful clinical tool.[2] The RDI models are made freely available as a treatment support tool, the HIV Treatment Response Prediction System (HIV-TRePS), via www.hivrdi.org.[3]

In a retrospective study, another group also reported the successful development of predictive models that performed as well as the predictions made by HIV physicians and virologists, albeit without the benefit of a full treatment history.[4]

Our models have been trained using data collected from multiple sources around the world, with well-resourced countries being the main contributors to date. The RDI now has data from ~110000 patients and our models have become more accurate as more data have become available with which to train them. The previous set of models that do not require a genotype were trained with data from well-resourced countries only and achieved a mean area under the receiver-operating characteristic (ROC) curve of 0.77, both during cross-validation and with an independent test set from the same settings, but a value of 0.60 (range 0.58–0.65) with cases from clinics in southern Africa.[1]

Here we report on the development of new models that do not require a genotype using the largest training set to date, including for the first time over a thousand cases from southern Africa, their evaluation with global and southern African independent test sets and a comparison with genotyping as a predictor of the virological response to antiretroviral therapy following virological failure.

It is important to develop new models to power the online treatment support tool on a regular basis to keep the system up to date with developments in clinical practice, particularly the development of new drugs and combinations. It is therefore anticipated that this update will be the first of a series that detail the development and evaluation of successive, improved generations of the HIV-TRePS system.

Methods

Clinical data

The package of data collected when antiretroviral therapy is changed, for whatever reason, is termed a 'treatment-change episode' (TCE).[5] TCEs from cases of virological failure were extracted from the RDI database when all the following data were available: the on-treatment baseline plasma viral load (with the sample taken ≤8 weeks prior to the treatment change); the on-treatment baseline CD4 cell count (≤12 weeks prior to treatment change); the baseline regimen (the drugs the patient was taking prior to the change); the antiretroviral treatment history; the drugs in the new regimen; a follow-up plasma viral load determination taken between 4 and 52 weeks following the introduction of the new regimen; and the time to that follow-up viral load (in order that the models could be trained to predict responses at different times).

The TCEs were censored using rules established in previous studies and published elsewhere.[3] For example, TCEs involving drugs not adequately represented in the database (<500 cases) were excluded; in this case, this applied to tipranavir, maraviroc and rilpivirine.

Computational model development

The qualifying TCEs were used to train a committee of 10 random forest (RF) models to predict the probability of the follow-up viral load being <50 copies/mL, using methodology described in detail elsewhere.[3,6] The following input variables were used, with changes from the previously published models highlighted: the baseline viral load ($\log_{10}$ copies HIV RNA/mL); the baseline CD4 count (cells/mm^3); the treatment history, comprising 20 binary variables coding for any experience of zidovudine, didanosine, stavudine, abacavir, lamivudine, emtricitabine, tenofovir DF, efavirenz, nevirapine, etravirine, indinavir, nelfinavir, saquinavir, amprenavir, fosamprenavir, lopinavir, atazanavir, darunavir, enfuvirtide and raltegravir [the previous models used just five treatment history variables: zidovudine, lamivudine/emtricitabine, any non-nucleoside reverse transcriptase inhibitor (NNRTI), any protease inhibitor (PI) and enfuvirtide]; the antiretroviral drugs in the new regimen, comprising 19 variables covering zidovudine, didanosine, stavudine, abacavir, lamivudine, emtricitabine, tenofovir DF, efavirenz, nevirapine, etravirine, indinavir, nelfinavir, saquinavir, (fos)amprenavir, lopinavir, atazanavir, darunavir, enfuvirtide and raltegravir (raltegravir was not covered by the previous models); and the time from the treatment change to the follow-up viral load (number of days). The output variable was the follow-up viral load coded as a binary variable: ≤1.7 log or 50 copies/mL=1 (response) and >1.7 log or 50 copies/mL=0 (failure). The models were trained to produce an estimate of the probability of the follow-up viral load being ≤50 copies/mL as the output. The previous models used 400 copies/mL as the output.

The performance of the models as predictors of virological response was evaluated by using the models' estimates of the probability of a response and the responses observed in the clinics (as a binary response variable: response versus failure) to plot ROC curves and assessing the areas under the ROC curves (AUCs). In addition, the optimum operating point (OOP; the cut-off point above which the models' prediction is classified as a prediction of response) for the models was derived during cross-validation and used to obtain the overall accuracy, sensitivity and specificity of the system. Finally, in order to evaluate further the potential clinical utility of the models, we calculated the positive and negative predictive value of different cut-offs for the probability of a response given by the models when the expected response rate to antiretroviral therapy is 40%, 60% and 80%, which cover the most commonly reported response rates after virological failure in developed and developing countries.[7,8] This is an approach that has not been used to evaluate the models in previous studies.

Data partition

The qualifying TCEs were partitioned using methods described elsewhere.[3,6] The following datasets resulted: a training set of 22567 TCEs and an independent test set of 1000 TCEs from 1000 patients. For the first time, the training data included a substantial number of cases from resource-limited settings, including 1090 from southern Africa. The 'global' test set was constructed to include a subset of 100 TCEs from 100 patients in southern Africa. Overall data from 38 sources, involving patients in more than 40 countries, were involved in the study. The data were considerably more numerous and heterogeneous in terms of their origin than was the case in previous studies.

Validation

The committee of 10 RF models was developed using a 10× cross-validation scheme.[3,6] The RF models were then validated externally by providing them with the baseline data from the global and southern African test set TCEs, obtaining the predictions of virological response from the 10 models, averaging them and comparing these averaged predictions with the responses observed in the clinic. The differences in performance of the models with the different test sets were tested for statistical significance using DeLong's test.[9]

Comparison with genotyping

Genotypic sensitivity scores (GSSs) were obtained for cases in the global independent test set with genotypes available from within 12 weeks before the start of the new regimen, using three rules-based interpretation systems in common clinical use as an aid to treatment selection: ANRS v2012.09, Rega v8.0.2 and Stanford HIVdb v6.2.0, all obtained on 29 May 2013, using methodology described elsewhere (the GSS for each drug in the new regimen was obtained and added to give a total score for the regimen).[3] These GSSs were then used as predictors of virological response and their performance was compared with that of the models.

In silico analysis to identify effective alternatives to failed regimens in southern Africa

In order to assess the potential of the models to help avoid treatment failure in a resource-limited setting, where models that do not require a genotype may be of most use, they were used to identify antiretroviral regimens that were predicted to be effective for the cases in southern Africa. Of particular interest were those cases where the regimen selected in the clinic failed to re-suppress the virus. Baseline data were used by the models to make predictions of response for alternative three-drug regimens in common use comprising only those drugs that were in use in those settings at the time (defined as drugs that were represented in the data from that particular clinic up until the date that the new regimen was started). For all these tests, the OOP that was derived during model development was used, as a test of how generalizable the system is.

Results

Characteristics of the datasets

The baseline, treatment and response characteristics of the datasets are summarized in Table 1. The 100 test cases from southern Africa had higher baseline viral loads (a median of 4.32 $\log_{10}$ copies/mL) and lower CD4 counts (a median of 163 cells/mm^3) than the training and test data as a whole (median baseline viral loads of 3.83 and 3.97, and CD4 counts of 268 and 260, respectively). They had less previous drug exposure (a median of three versus four drugs), a greater previous exposure to NNRTIs (94% versus 62%–63%) and a lower exposure to PIs (11% versus 69% and 63%, respectively). More of the southern African cases were changing onto two nucleoside(-tide) reverse transcriptase inhibitors [N(t)RTIs]+PI (70% versus 32%–34%), consistent with most being first-time treatment failures following the treatment protocols in the region. There were fewer virological failures among the southern African cases (48% versus 66% and 64%, respectively), reflecting their earlier stage of treatment.

Results of the modelling

Cross-validation

The performance characteristics from the ROC curves of the 10 individual models during cross-validation and independent testing are summarized in Table 2. The 10 models achieved AUC

Table 1. Characteristics of the TCEs in the training and test sets

	Training set	Global test set	Southern Africa test set
TCEs	22567	1000	100
Patients	7568	1000	100
male, *n*	5055	661	36
female, *n*	1584	218	56
not known, *n*	929	121	8
age (years), median	40	39	35
Baseline data			
baseline VL ($\log_{10}$ copies/mL), median (IQR)	3.83 (2.83–4.84)	3.97 (2.98–4.97)	4.32 (3.62–5.01)
baseline CD4 (cells/mm^3), median (IQR)	268 (128–407)	260 (123–387)	163 (65–362)
Treatment history			
no. of previous drugs, median (IQR)	4 (3–7)	4 (3–6)	3 (3–4)
N(t)RTI experience, *n* (%)	22478 (100)	998 (100)	100 (100)
NNRTI experience, *n* (%)	13998 (62)	634 (63)	94 (94)
PI experience, *n* (%)	15532 (69)	630 (63)	11 (11)
New regimens, *n* (%)			
2 N(t)RTI+PI	7114 (32)	335 (34)	70 (70)
2 N(t)RTI+NNRTI	4801 (21)	228 (23)	22 (22)
3 N(t)RTIs+PI	1587 (7)	74 (7)	2 (2)
3 N(t)RTIs	1472 (7)	59 (6)	3 (3)
3 N(t)RTIs+NNRTI	1031 (5)	40 (4)	
2 N(t)RTIs	783 (3)	23 (2)	2 (2)
2 N(t)RTIs+NNRTI+PI	604 (3)	21 (2)	
4 N(t)RTIs	574 (3)	20 (2)	
N(t)RTI+NNRTI+PI	400 (2)	29 (3)	
N(t)RTI+PI	372 (2)	10 (1)	
other	3829 (17)	239 (24)	1 (1)

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

Table 2. Results of the modelling

Model	AUC	sensitivity (%)	specificity (%)	OA (%)	OOP
Cross-validation (10×, $n = 22485$)					
1	0.84	67	83	78	0.42
2	0.79	71	73	73	0.36
3	0.80	64	78	74	0.40
4	0.83	66	82	77	0.41
5	0.83	72	79	77	0.40
6	0.81	60	82	75	0.45
7	0.81	64	82	76	0.43
8	0.84	69	83	78	0.42
9	0.83	63	86	78	0.48
10	0.82	61	84	76	0.45
Mean	0.82	66	81	76	0.422
95% CI	0.78–0.85	58–74	74–88	73–80	0.36–0.49
Testing with a global test set ($n = 1000$) using an OOP of 0.42					
CAP	0.80	66	79	74	
95% CI	0.77–0.82	61–71	76–82	71–77	
Testing with the Southern African cases ($n = 100$) using an OOP of 0.42					
CAP	0.78	81	60	71	
95% CI	0.69–0.87	67–90	45–74	61–80	

OA, overall accuracy; CAP, committee average prediction.

values during cross-validation ranging from 0.79 to 0.84, with a mean of 0.82. The overall accuracy ranged from 73% to 78% (mean 76%), the sensitivity ranged from 60% to 72% (mean 66%) and the specificity ranged from 73% to 86% (mean 81%). The OOP ranged from 0.36 to 0.48 with a mean of 0.42.

Testing with the independent set of 1000 TCEs

The committee of 10 models achieved an AUC of 0.80 based on their predictions of response for the 1000 test TCEs. The overall accuracy was 74%, the sensitivity 66% and the specificity 79%. The ROC curve for the committee is presented in Figure 1.

The results of calculating the positive and negative predictive values of different cut-offs for the probability of response estimated by the models, when the expected response rate to antiretroviral therapy was 40%, 60% and 80%, are presented in Figure 2. The positive predictive value for the virological response was good for regimens with a high probability of response estimated by the models. Even when the expected response rate was 40%, the positive predictive value of the models for antiretroviral regimens with a probability of response >50% was nearly 70%. The models were also able to predict failures when the probability of response was low (<10%), but their predictive power was less satisfactory when the expected response rate was 80%.

Testing the models with the independent test set of 100 TCEs from southern Africa

The committee of 10 models achieved an AUC of 0.78. The overall accuracy was 71%, the sensitivity 81% and the specificity 60%. The ROC curve for the committee is also presented in Figure 1.
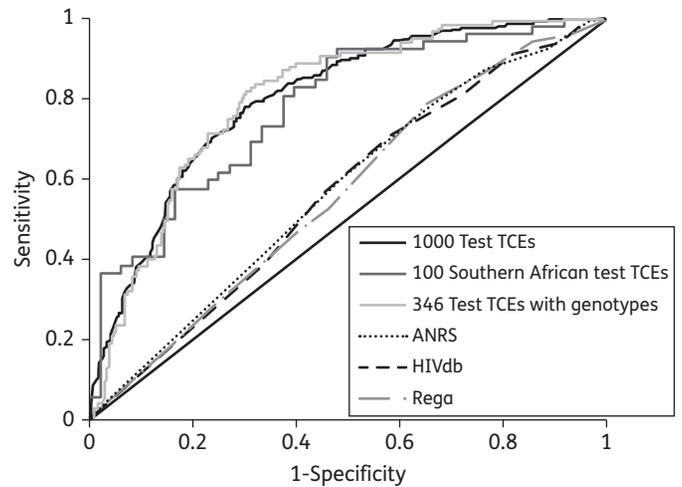

Figure 1. ROC curves for the committee of RF models tested with a global test set ($n = 1000$), the southern African cases ($n = 100$), the test cases with genotypes available ($n = 346$) and for GSS using three common interpretation systems ($n = 346$).
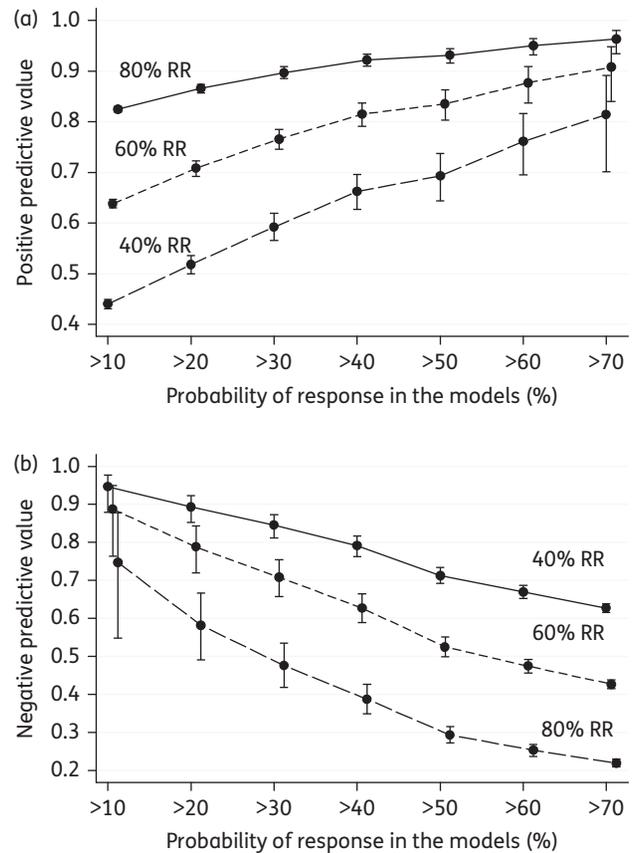

Figure 2. Positive (a) and negative (b) predictive value of several cut-off points for the probability of response given by the models when the response rate (RR) to antiretroviral therapy is 40%, 60% and 80%.

Table 3. Comparison of model predictions versus GSS for the 346 test TCEs with genotypes

Prediction system	AUC	Sensitivity (%)	Specificity (%)	Overall accuracy (%)	P (GSS versus models)
Total ANRS score	0.5714	51	58	55	<0.0001
Total HIVdb score	0.5657	53	57	56	<0.0001
Total Rega score	0.5628	52	54	53	<0.0001
Models	0.8009	65	80	75	

Table 4. *In silico* modelling to identify potentially effective alternative regimens for the southern African cases

	All cases (100)	Failures (48)
Percentage of cases for which alternative three-drug regimens were predicted to be effective	76	65
Median number of alternatives	14.5	14
Percentage of cases for which alternative three-drug regimens were predicted to be more effective than the regimen selected	85	96
Median number of alternatives	7	9

Comparing the predictive accuracy of the models versus genotyping

Of the 1000 TCEs in the global test set, genotypes were available for 346. The AUC values for the GSSs obtained using the three genotype interpretations systems were 0.57 (ANRS), 0.56 (Rega) and 0.57 (Stanford HIVdb) (Table 3). All were significantly less accurate predictors of virological response than the models (P<0.0001).

In silico analysis

Considering all 100 cases from southern Africa, the models were able to identify one or more three-drug regimens that were predicted to be effective (the estimated probability of the follow-up viral load being above the OOP derived during cross-validation), comprising only those drugs available at that time in that centre in 76 cases (Table 4). In these cases, the median number of alternative regimens that were predicted to be effective was 14.5. The models identified alternatives with a higher estimated likelihood of response than the regimen actually used in the clinic, but not necessarily above the OOP, in 85 of 100 cases.

Forty-eight of the 100 cases from southern Africa experienced virological failure following the introduction of a new regimen in the clinic. The models were able to identify one or more locally available three-drug regimens that were predicted to be effective in 31 (65%) of these cases. The median number of alternative 'effective' regimens identified was 14. The models identified alternatives with a higher estimated likelihood of response than the regimen actually used in the clinic in 46 (96%) of the failures.

Discussion

These latest computational models, which do not require a genotype for their predictions, predicted virological response to a change in antiretroviral therapy following virological failure with a level of accuracy that is comparable to that of previous RDI models that used a genotype in their predictions and were significantly more accurate than genotyping with rules-based interpretation. The overall accuracy of the models was comparable when tested with cases from well-resourced and resource-limited settings (southern Africa). However, the specificity of predictions of the models, using the OOP derived during cross-validation with data predominately related to well-resourced settings, was reduced and sensitivity was increased for cases from southern Africa compared with a global test set.

Overall, the models exhibited higher specificity than sensitivity using the OOP derived during cross-validation, meaning that they were less likely to produce false-positive results (classifying regimens as effective when they were not) than false negatives (classifying regimens as failures when they were effective). This 'conservative' performance bias is probably desirable from a clinical perspective. In addition, the analysis of positive and negative predictive values suggests that the models are more robust in predicting responses than failures, which is interesting in terms of utility in clinical practice. The main aim of the HIV-TRePS system is to offer clinicians a tool for selecting an effective antiretroviral regimen for patients who experience virological failure. These results indicate that regimens with a high probability of response estimated by the models are likely to work, even in heavily pretreated patients, where the expected response rate to salvage therapy is low.

The models were able to identify alternative three-drug regimens, comprising locally available drugs that were predicted to produce a virological response for a substantial proportion of the cases from southern Africa, including the virological failures. In almost all cases of failure, the models were able to identify regimens with a higher predicted probability of success than the regimen that had failed. However, the reduced level of specificity for the southern African cases means that caution should be exercised in the interpretation of these data and of the 'fail/respond' classification using the standard OOP for cases from this setting. It raises the question of whether a different, setting-specific OOP might be used to increase the specificity of the system for patients from this setting.

It is important to note that one of the input variables for these models was the plasma viral load, which previous studies have shown to be important for the predictive accuracy of the models.[10,11] Although viral load monitoring is not routine in most

resource-limited settings, it is now recommended in the latest WHO guidelines as the preferred approach to monitoring the success of antiretroviral therapy and diagnosing treatment failure.[12] As technological advances enable lower test costs and simpler machines that require less infrastructure, maintenance and technical expertise become available, so the use of the viral load is likely to increase in clinical practice.[13]

The study has some limitations. First, it was retrospective and, as such, no firm claims can be made for the clinical benefit that use of the system as a treatment support tool could provide. Another possible shortcoming, inherent in such studies and discussed in previous publications, is that the cases used are, by definition, those with complete data around a change of therapy and therefore they may not be truly representative of the general patient population. Nevertheless, the size of the training dataset, the variety of sources and settings from which the data were collected and the range of clinical practice represented are positive factors in considering the potential robustness and generalizability of the models' performance.

Conclusions

This study builds on the findings from the previous one published in this journal, which was the first to describe computational models that predict virological response to antiretroviral therapy for patients in resource-limited settings without a genotype, with an accuracy comparable to that of genotyping with rules-based interpretation. As predicted in that publication, these latest models, trained with a very large dataset including, for the first time, cases from resource-limited settings, showed improved accuracy for cases from southern Africa, significantly outperformed genotyping and were comparably accurate for cases from resource-limited and well-resourced settings.

A full validation of this approach as a clinical tool would require a prospective, controlled clinical trial. Nevertheless, the results suggest this approach has the potential to reduce virological failure and improve patient outcomes in resource-limited settings. It can provide clinicians with a practical tool to support optimized treatment decision-making in the absence of resistance tests and where expertise may be lacking in the context of a public health approach to antiretroviral roll-out and management.

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, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government.

References

1 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.

2 Larder BA, Revell AD, Mican J *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.

3 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.

4 Zazzi M, Kaiser R, Sönnerborg A *et al*. Prediction of response to antiretroviral therapy by human experts and by the EuResist data-driven expert system (the EVE study). *HIV Med* 2010; **12**: 211–8.

5 Larder BA, 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.

6 Wang D, Larder BA, Revell AD *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.

7 Yazdanpanah Y. Multidrug resistance: a clinical approach. *Curr Opin HIV AIDS* 2009; **4**: 499–506.

8 Ajose O, Mookerjee S, Mills EJ *et al*. Treatment outcomes of patients on second-line antiretroviral therapy in resource-limited settings: a systematic review and meta-analysis. *AIDS* 2012; **26**: 929–38.

9 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44**: 837–45.

10 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.

11 Prosperi MCF, Rosen-Zvi M, Altman A *et al*. Antiretroviral therapy optimisation without genotype resistance testing: a perspective on treatment history based models. *PLoS One* 2010; **5**: e13753.

12 WHO. *Consolidated Guidelines on the Use of Antiretroviral Drugs For Treating and Preventing HIV Infection: Recommendations For a Public Health Approach*. Geneva: WHO, 2013.

13 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.