DOI: 10.1111/hiv.12156

HIV Medicine (2014), **15**, 442–448

ORIGINAL RESEARCH

A comparison of computational models with and without genotyping for prediction of response to second-line HIV therapy

AD Revell,[1] MA Boyd,[2] D Wang,[1] S Emery,[2] B Gazzard,[3] P Reiss,[4,5] AI van Sighem,[5] JS Montaner,[6] HC Lane[7] and BA Larder[1]

[1]The HIV Resistance Response Database Initiative (RDI), London, UK, [2]The Kirby Institute for infection and immunity in society, University of New South Wales, Sydney, NSW, Australia, [3]Chelsea and Westminster Hospital, London, UK, [4]Department of Global Health, Academic Medical Centre of the University of Amsterdam, Amsterdam Institute for Global Health and Development, Amsterdam, The Netherlands, [5]Stichting HIV Monitoring, Amsterdam, The Netherlands, [6]BC Centre for Excellence in HIV and AIDS, Vancouver, BC, Canada and [7]National Institutes of Allergy and Infectious Diseases, Bethesda, MD, USA

Objectives

We compared the use of computational models developed with and without HIV genotype *vs.* genotyping itself to predict effective regimens for patients experiencing first-line virological failure.

Methods

Two sets of models predicted virological response for 99 three-drug regimens for patients on a failing regimen of two nucleoside/nucleotide reverse transcriptase inhibitors and one nonnucleoside reverse transcriptase inhibitor in the Second-Line study. One set used viral load, CD4 count, genotype, plus treatment history and time to follow-up to make its predictions; the second set did not include genotype. Genotypic sensitivity scores were derived and the ranking of the alternative regimens compared with those of the models. The accuracy of the models and that of genotyping as predictors of the virological responses to second-line regimens were compared.

Results

The rankings of alternative regimens by the two sets of models were significantly correlated in 60–69% of cases, and the rankings by the models that use a genotype and genotyping itself were significantly correlated in 60% of cases. The two sets of models identified alternative regimens that were predicted to be effective in 97% and 100% of cases, respectively. The area under the receiver-operating curve was 0.72 and 0.74 for the two sets of models, respectively, and significantly lower at 0.55 for genotyping.

Conclusions

The two sets of models performed comparably well and significantly outperformed genotyping as predictors of response. The models identified alternative regimens predicted to be effective in almost all cases. It is encouraging that models that do not require a genotype were able to predict responses to common second-line therapies in settings where genotyping is unavailable.

Keywords: antiretroviral, computational model, genotype, HIV, response prediction.

Accepted 24 February 2014

Correspondence: Dr Andrew D. Revell, 14 Union Square, London N1 7DH, UK. Tel/fax: +44 207 226 7314; e-mail: andrewrevell@hivrdi.org

Introduction

Combination antiretroviral therapy (cART) is capable of suppressing HIV replication and preventing disease progression for a number of years. However, in many cases, at some point, there will be a loss of viral control, often as a result of the emergence of HIV drug resistance in nonadherent patients or the development of drug-related toxicity, which will require a change to the drug regimen. In high-income countries, virological failure is generally detected quickly as a result of regular viral load monitoring. The selection of drugs for the next regimen can be made using genotypic antiretroviral resistance testing (GART) with rules-based interpretation indicating to which drugs the virus should still be sensitive [1]. Re-suppression of the virus is readily achieved in most cases and the selection of drugs only becomes challenging in deep salvage, after several lines of therapy have been exhausted, or with unusually complex cases.

The situation in low- and middle-income countries (LMIC) is different in that there are fewer drugs available from which to assemble an effective combination, GART is seldom available or affordable and, despite widespread advocacy for it to be made available, viral load testing is not yet in widespread routine use [2]. Therapy is often implemented in accordance with a simplified, public health approach made necessary as initiation and monitoring of cART in LMIC is often 'task-shifted' to nonphysician health care workers [3]. Third-line therapy, although included in World Health Organization (WHO) guidelines, is rarely available in publicly funded health care systems in LMIC.

Even for second-line cART there exists uncertainty about the optimum regimen. For example, the recent Second-Line study compared a regimen combining two drugs from classes that were new to the patient (raltegravir and ritonavir-boosted lopinavir) with a WHO-recommended regimen of ritonavir-boosted lopinavir plus two or three nucleoside or nucleotide reverse transcriptase inhibitors (NtRTIs) in patients failing first-line therapy comprising a nonnucleoside reverse transcriptase inhibitor (NNRTI) plus two NtRTIs [4]. The simplified, experimental dual therapy was found to be noninferior to the three- or four-drug standard of care regimen.

The HIV Resistance Response Database Initiative (RDI) has developed computational models to predict virological response to a change of ART following virological failure, as a tool to support optimum drug selection. Early models used viral loads, CD4 counts, treatment history and resistance mutations from GART, in addition to the drugs in the new regimen, to predict response. These models have regularly demonstrated 80% accuracy when tested with independent data sets and have been evaluated by expert HIV clinicians as being a useful clinical tool [5,6].

More recently the focus has been on modelling response without the need for GART, as a tool for settings where GART is unavailable [7]. The latest versions of these models have shown similar accuracy to modelling with GART in cross-study comparisons and have demonstrated significantly superior accuracy compared with the use of genotypic sensitivity scores derived from GART using rules-based interpretation [8].

In this retrospective study, we evaluated the performance of existing computational models that require a genotype (GT) and those that do not (NG) using an independent external data set, derived from the Second-Line study of patients experiencing confirmed virological failure of first-line antiretroviral regimens consisting of an NNRTI plus two NtRTIs in a range of settings, some with limited resources [4]. The study enrolled 558 participants with a median (interquartile range (IQR)) age of 38.5 (32.4–44.4) years; 55% of the participants were male; the median (IQR) weight was 62.7 (55.0–72.5) kg; and the median (IQR) estimated duration of HIV infection was 6 (3.6–8.7) years. At baseline, the median (IQR) CD4 T-cell count was 190 (95–293) cells/μL and the median (IQR) $\log_{10}$ plasma viral load was 4.3 (3.7–4.9) log HIV-1 RNA copies/mL. Forty-seven per cent of the cohort were at Centers for Disease Control and Prevention (CDC) stage C at baseline. The mode of HIV transmission was given as men who have sex with men (MSM) (12.9%), heterosexual (72.8%), injecting drug use (IDU) (1.1%), blood product (2.4%) and other (10.7%).

The participants were enrolled at 37 sites in Argentina, Australia, Chile, England, France, Hong Kong, India, Israel, Malaysia, Mexico, Peru, Nigeria, Singapore, South Africa and Thailand. The ethnicity of the participants was 7.6% Caucasian, 42.3% Asian, 36% African and 13.9% Hispanic. The great majority of participants enrolled in the trial had their first-line regimen monitored by either immunological (CD4) or clinical means. Very few participants received any form of virological monitoring while receiving their first-line regimen. The median time on first-line cART in the combined cohort was 219 weeks (minimum 24 weeks; maximum 953 weeks). Following prolonged exposure to first-line therapy with a lack of virological monitoring, the cohort manifested a substantial degree of resistance to both the NNRTI and NtRTI classes of ART, as a result of long-term incomplete suppression of HIV RNA [4].

The primary objective of this study was, for the first time, to compare the ability of the GT and NG models to predict virologically effective, alternative regimens for salvage of patients experiencing confirmed virological failure of first-line ART. Secondary objectives were to compare the consistency of both sets of computational

models with GART itself and to compare the accuracy of each set of models in predicting responses to the regimens initiated in the study with the accuracy of the other set of models and with that of GART.

Methods

The models used in this study were two sets of 10 random forest models, one set that includes the genotype [62 significant resistance mutations in reverse transcriptase (RT) and protease] in its input variable set and one that does not, as described in detail elsewhere [5,8,9]. The output of both sets of models was the probability of virological response (defined as a follow-up plasma viral load of < 50 copies/mL).

Both sets of models were trained without data from the Second-Line study and were validated using independent data sets. The results of the models' predictions and the actual virological responses observed in the validation sets were used to plot receiver-operator characteristic (ROC) curves, with the area under the curve (AUC) the key metric of the models' performance. The NG models achieved an AUC of 0.80 during independent validation and the GT models an AUC of 0.86.

Data were extracted from the Second-Line study database that met the criteria for a complete treatment change episode (TCE). The data required by the models for their predictions were as follows.

(1) A baseline viral load while on failing first-line therapy from a sample obtained no more than 8 weeks prior to initiation of the second-line regimen.
(2) A baseline CD4 count while on failing first-line therapy from a sample obtained no more than 12 weeks prior to initiation of the second-line regimen.
(3) A baseline genotype while on failing first-line therapy from a sample obtained no more than 12 weeks prior to initiation of the second-line regimen (for use with the GT models only, a list of 62 significant mutations in RT and protease).
(4) Treatment history: the names of the drugs in the failing first-line regimen.
(5) New treatment: the names of the drugs in the second-line regimen.
(6) Follow-up viral loads at 12 (8–16) weeks and 24 (20–28 weeks) after the initiation of second-line therapy.

In order to address the main objective of the study, the models were used to perform '*in silico*' analysis, i.e. used to predict the virological response to a range of alternative regimens. The input data (baseline viral load, CD4 count, treatment history and, in the case of the GT models, genotype) from the TCEs were entered into the two sets of models, which were used to make predictions of the probability of virological response for each of 99 three-drug regimens in use in clinical practice (identified as being used in at least 10 patients in the RDI database).

The regimens were ranked in order of estimated probability of response. The associations between the rankings produced by the GT and NG models were evaluated using linear regression analyses and the cases for which there was a significant ($P < 0.05$) correlation between the rankings from the GT and NG models identified.

The numbers of alternative three-drug regimens from the RDI list for which a virological response was predicted (the estimated probability of virological response was above the optimum operating point established during model validation) were obtained for each set of models. As the assumption of normality did not hold for these variables, a nonparametric signed-rank test was used for comparing the models.

To compare the consistency of the predictions of the computational models with GART, genotypic sensitivity scores (GSSs) were derived using the Stanford database mutation score system, HIVdbv6.2.0 (26 July 2013), using methodology described elsewhere [5]. Linear regression was again used to determine if the ranks of the 99 regimens obtained from the models (based on estimated probability of response) were correlated with those obtained from genotyping (ranks based on GSSs using the Stanford mutation score system).

The final objective was to assess and compare the predictive accuracies of the two sets of models with each other and with GART. This was achieved by comparing the models' predictions with the virological responses actually achieved in the study clinics. As the models were trained to predict virological response for changes to treatment involving a minimum of three drugs, this analysis was performed using data for those patients receiving ritonavir-boosted lopinavir plus two or three NtRTIs as second-line cART in the study.

The results were used to plot ROC curves, with the AUC the key metric of the models' performance. GSSs obtained using the Stanford resistance score system were also evaluated as predictors of the virological response observed in the clinic in the same way, for comparison with the models

Results

There were 471 TCEs extracted that met all the TCE criteria and were used for the study.

The rankings for the GT and NG models and GSSs were compared using regression analysis. There was a significant correlation between the rankings from the GT and NG

models in 60% of cases at 12 weeks and 69% of cases at 24 weeks of follow-up, as shown in Table 1.

The NG models identified one or more alternative regimens that were predicted to produce a virological response in 456 of the 471 cases at week 12 and 470 at week 24. The GT models predicted effective alternatives in all 471 cases at both time-points. The numbers of alternative regimens that were predicted to result in a virological response in each case (minimum, maximum, median and quartile) are listed in Table 2. There were no significant differences between the two sets of models.

The results of the analysis of the accuracy of the predictions of the two sets of models and of GSSs from GART are summarized in Table 3 and the ROC curve is presented in Fig. 1. At 12 weeks the AUC was 0.74, 0.72 and 0.55 for NG models, GT models and GSSs, respectively. At 24 weeks the figures were 0.66, 0.67 and 0.54, respectively.

Table 1 Comparison between the two sets of models and genotypic sensitivity scores (GSSs) in terms of ranking alternative three-drug regimens in order of probability of response

	Total *n*	*n* with significant correlation	%
NG *vs.* GT at 12 weeks	471	282	60
NG *vs.* GSSs at 12 weeks	471	211	45
GT *vs.* GSSs at 12 weeks	471	284	60
NG *vs.* GT at 24 weeks	471	326	69
NG *vs.* GSSs at 24 weeks	471	207	44
GT *vs.* GSSs at 24 weeks	471	284	60

GT, models that require a genotype; NG, models that do not require a genotype.

Table 2 The numbers of effective alternative three-drug regimens predicted for each case

Data set	*n*	Min	Q1	Median	Q3	Max	Stat*
NG (12 weeks)	456	1	38	53	73	99	
NG (24 weeks)	470	1	50	65	79	99	
GT (12 weeks)	471	1	42	71	86	99	ns
GT (24 weeks)	471	8	65	84	92	99	ns

GT, models that require a genotype; NG, models that do not require a genotype; ns, not significant.
*Statistical comparison between the NG and GT models.

Discussion

In this assessment of the ranking of alternative second-line regimens by estimated probability of virological response, using computational models incorporating (GT) and not incorporating (NG) genotypic resistance data at confirmed failure of a standard NNRTI + 2N(t)RTI regimen, we found that the GT and NG models were significantly correlated for the majority of cases (60% at 12 weeks and 69% at 24 weeks). It is reassuring that the two sets of models, developed 2 years apart and using substantially different information for their predictions, are aligned in their ranking of a standard list of regimens by probability of response for 60–70% of cases.

As might be expected, there was a greater degree of concordance between the predictions of the two sets of models and between the GT models and GSSs than between the NG models and GSSs.

The GT models identified more alternatives that were predicted to be effective than the NG models. In other words, the NG models were more 'conservative' in their predictions of response. Both sets of models identified more effective alternatives at week 24 than at week 12, suggesting that the models are predicting that some regimens would take some time to achieve a viral load below the target level.

Both sets of models were significantly more accurate predictors of virological response to second-line therapy comprising ritonavir-boosted lopinavir plus two or three NtRTIs than genotyping with Stanford rules-based interpretation. It is encouraging that the NG models were able to achieve this level of performance and this is consistent

Table 3 Performance of the models and genotypic sensitivity scores (GSSs) in predicting virological response to second-line therapy consisting of two or three nucleoside/nucleotide reverse transcriptase inhibitors plus ritonavir-boosted lopinavir

Model/data set	AUC	Sensitivity (%)	Specificity (%)	Overall accuracy (%)	Statistical comparison*
NG, 12 weeks	0.74	61	75	68	
NG, 24 weeks	0.66	63	67	64	
GT, 12 weeks	0.72	66	66	66	(1) ns
GT, 24 weeks	0.67	66	65	65	(1) ns
GSSs, 12 weeks	0.55	56	53	54	(2) $P < 0.05$
GSSs, 24 weeks	0.54	59	53	57	(2) $P < 0.05$

AUC, area under the curve; GT, models that require a genotype; NG, models that do not require a genotype; ns, not significant.
There were no statistically significant differences between the models at 12 and 24 weeks. The GSSs from the Stanford interpretation system were significantly less accurate as predictors of virological response at either 12 or 24 weeks than both sets of models ($P < 0.05$ for all comparisons).
*Statistical comparisons: (1) between the NG and GT models and (2) between the GSSs and each set of models.

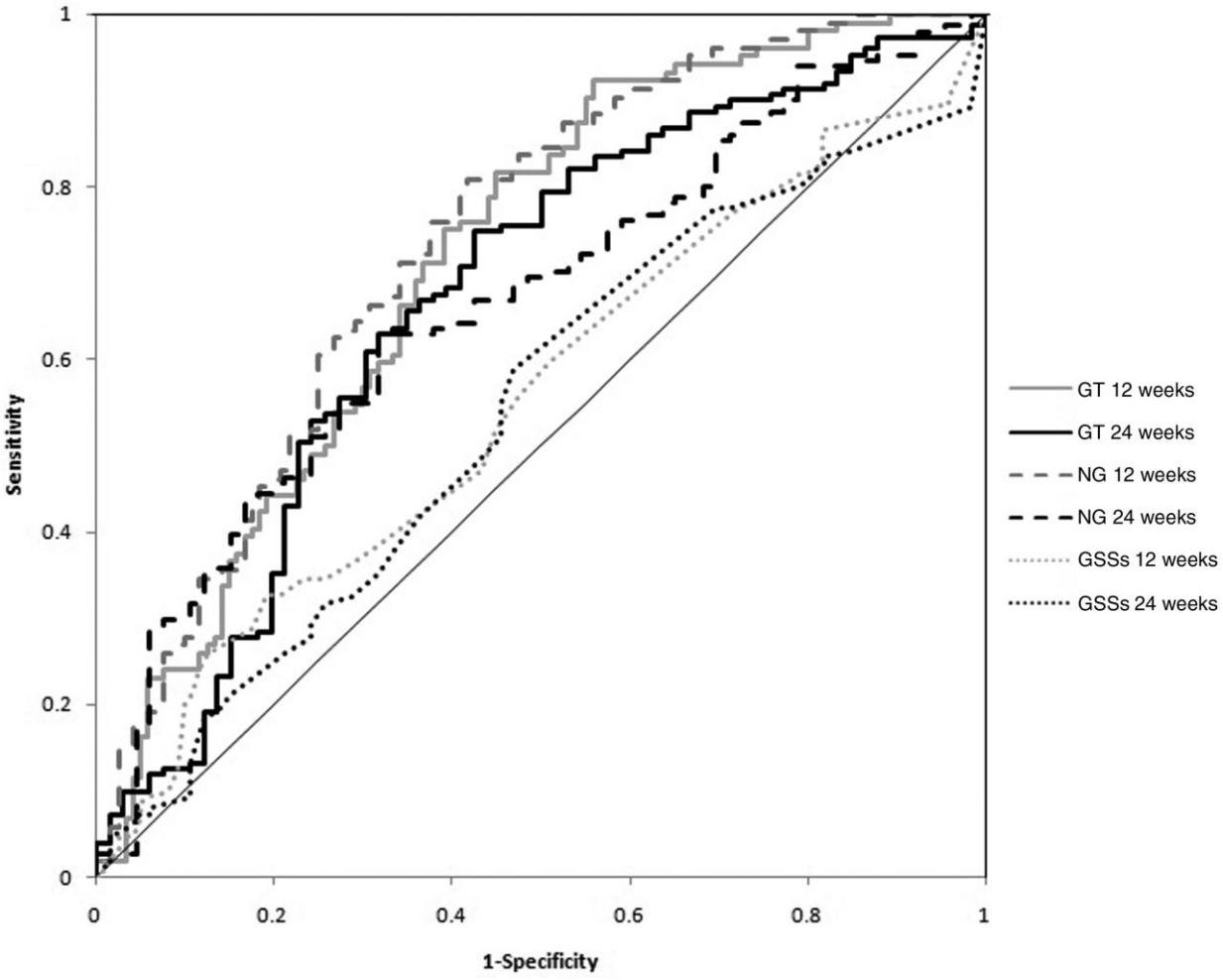

Fig. 1 Receiver-operator characteristic (ROC) curves for the models and genotypic sensitivity scores (GSSs) as predictors of the response to second-line therapy. GT, models that require a genotype; NG, models that do not require a genotype.

with the results of other studies [5,8]. It should be acknowledged that treatment decisions are not made on the basis of a genotype with rules-based interpretation alone; physicians include other information in their decisions, such as the genetic barrier associated with different inhibitors. These additional factors could of course also be taken into account when interpreting and applying the results of the models, which are intended to provide support to decision-making that is more predictive of outcomes than that provided by genotyping.

The performance of the models was somewhat inferior to that achieved in cross-validation during their development and their initial evaluation using large independent data sets from the same centres as the training data (AUC values of approximately 0.8). This is probably a consequence of the 'familiarity effect': models have consistently performed with greater accuracy for cases from familiar settings, from where the training data had been obtained, than for cases from unfamiliar settings, unrepresented in the training data, such as these Second-Line data.

The accuracy of the RDI models in this and numerous previous studies, and their consistent superiority as predictors of virological response to cART over genotyping with rules-based interpretation suggest that they may have significant utility to help guide salvage therapy. This is particularly relevant to LMIC where the range of drugs available is likely to be limited and genotyping unaffordable. The models described here are freely available online as part of the HIV Treatment Response Prediction System (HIV-TRePS) which also enables physicians to model therapy costs. In a recent retrospective study using data from an Indian cohort, the system identified alternative, locally available regimens with a higher probability of response and lower annual costs than those prescribed in

the clinic [10]. The debate about the optimum allocation of resources to drugs, diagnostics and systems in such settings is an important one that is set to continue [11]. While the cost-effectiveness of modelling would require a prospective controlled trial to establish definitively, it would appear that the models, which are free to use, at least have the potential to reduce both the failure rates and costs of ART, and so this approach should form an important part of such a debate.

The study had some shortcomings. An issue with the comparison of the *in silico* rankings by the two sets of models is that 471 different regressions were performed, increasing the likelihood of Type 1 error. At the 0.05 significance level used, 5% of the regressions performed (approximately 24) might be expected to have occurred by chance.

Conclusions

Computational models that do and do not require a genotype to predict virological response to ART performed comparably well in this analysis of data from the Second-Line study. Both sets of models significantly outperformed genotyping with rules-based interpretation as predictors of response, identified alternative regimens that were predicted to be effective in almost all cases and were similar in their rankings of alternative regimens by probability of response in around two-thirds of cases. It is particularly encouraging that models that do not require a genotype, developed specifically for use in LMIC, were able to predict responses to the WHO recommended second-line therapies. The results of these and other studies suggest that this approach could have a major role in maximizing the cost-effectiveness of ART in LMIC.

Acknowledgements

Cecilia Moore (The Kirby Institute for infection and immunity in society, University of New South Wales, Sydney, NSW, Australia) for the extraction of the Second Line data used in this study.

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 Rinkede 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); Adrian Streinu-Cercel (National Institute of Infectious Diseases, Prof. Dr. Matei Balş, Bucharest, Romania); Gerardo Alvarez-Uria (VFHCS, India).

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); 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 and Julie Metcalf (National Institutes of Health Clinic, Rockville, USA); Maria-Jesus Perez-Elias (Instituto Ramón y Cajal de Investigación Sanitaria, Madrid, Spain); Andrew Carr, Richard Norris and Karl Hesse (Immunology B 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), Carl Morrow and Robin Wood (Desmond Tutu HIV Centre, University of Cape Town, South Africa); Luminita Ene ('Dr. Victor Babes' Hospital for Infectious and Tropical Diseases, Bucharest, Romania); Gordana Dragovic (University of Belgrade, Belgrade, Serbia).

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).

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. 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.

Conflicts of interest: MAB has received grants from Abbott, BMS, BI, Gilead, Janssen and Merck.

References

1 Thompson MA, Aberg JA, Hoy JF *et al.* Antiretroviral treatment of adult HIV infection: 2012 recommendations of the international AIDS society-USA panel. *JAMA* 2012; **308**: 387–402.

2 Smith DM, Schooley RT. Running with scissors: using antiretroviral therapy without monitoring viral load. *Clin Infect Dis* 2008; **46**: 1598–1600.

3 World Health Organisation. *Consolidated Guidelines on the Use of Antiretroviral Drugs for Treating and Preventing HIV Infection: Recommendations for A Public Health Approach*. Geneva, WHO, 2013.
4 Second-Line Study Group. Ritonavir-boosted lopinavir plus nucleoside or nucleotide reverse transcriptase inhibitors versus ritonavir-boosted lopinavir plus raltegravir for treatment of HIV-1 infection in adults with virological failure of a standard first-line ART regimen (SECOND-LINE): a randomised, open-label, non-inferiority study. *Lancet* 2013; **381**: 2091–2099.
5 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**: 1855–1863.
6 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.
7 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**: 1406–1414.
8 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**: 1104–1110.
9 Revell AD, Wang D, DeWolf F *et al.* The development of new computational models for the HIV-TRePS online treatment selection tool. *XIX International AIDS Conference*. Washington DC, USA, July 2012 [Abstract TUPE091].
10 Revell AD, Alvarez-Uria G, Wang D *et al.* Potential impact of a free online HIV treatment response prediction system for reducing virological failures and drug costs after antiretroviral therapy failure in a resource-limited setting. *BioMed Res Int* 2013; doi 10.1155/2013/579741 [Epub ahead of print].
11 Katzenstein D. HIV RNA and Genotype in resource limited settings. Can we do better? *Clin Infect Dis* 2013; **58**: 110–112.