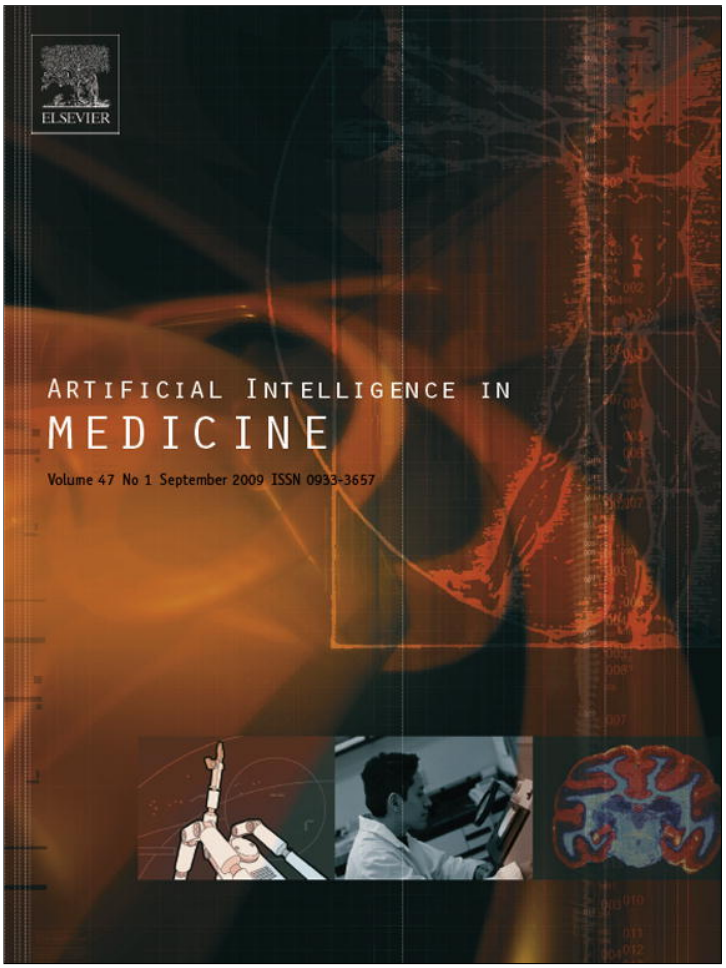

Artificial Intelligence in Medicine (2009) **47**, 63—74

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy

Dechao Wang [a], Brendan Larder [a], Andrew Revell [a,*], Julio Montaner [b], Richard Harrigan [b], Frank De Wolf [c], Joep Lange [d], Scott Wegner [e], Lidia Ruiz [f], María Jésus Pérez-Elías [g], Sean Emery [h], Jose Gatell [i], Antonella D'Arminio Monforte [j], Carlo Torti [k], Maurizio Zazzi [l], Clifford Lane [m]

[a] *The HIV Resistance Response Database Initiative (RDI), 14 Union Square, London, UK*
[b] *BC Centre for Excellence in HIV/AIDS, Vancouver, Canada*
[c] *Netherlands HIV Monitoring Foundation, Amsterdam, The Netherlands*
[d] *Academic Medical Centre of the University of Amsterdam, Amsterdam, The Netherlands*
[e] *Uniformed Services University of the Health Sciences, Bethesda, MD, USA*
[f] *Fundació irsiCaixa, Badalona, Spain*
[g] *Ramón y Cajal Hospital, Madrid, Spain*
[h] *National Centre in HIV Epidemiology and Clinical Research, Sydney, Australia*
[i] *Hospital Clinic of Barcelona, Barcelona, Spain*
[j] *University of Milan (on behalf of ICONA), Milan, Italy*
[k] *Institute for Infectious and Tropical Diseases, University of Brescia (on behalf of the Italian MASTER Cohort), Brescia, Italy*
[l] *University of Siena (on behalf of the Italian ARCA database), Siena, Italy*
[m] *National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA*

Received 21 August 2008; received in revised form 16 April 2009; accepted 10 May 2009

KEYWORDS
HIV;
Artificial neural networks;
Support vector machines

Summary

Objective: HIV treatment failure is commonly associated with drug resistance and the selection of a new regimen is often guided by genotypic resistance testing. The interpretation of complex genotypic data poses a major challenge. We have developed artificial neural network (ANN) models that predict virological response to therapy from HIV genotype and other clinical information. Here we compare the accuracy of ANN with

* Corresponding author. Tel.: +44 020 7226 7314; fax: +44 020 7226 7314.
E-mail address: andrewrevell@hivrdi.org (A. Revell).

0933-3657/$ — see front matter
doi:10.1016/j.artmed.2009.05.002

Random forests;
Treatment decision support techniques;
Antiretroviral treatment;
Antiviral drug resistance

alternative modelling methodologies, random forests (RF) and support vector machines (SVM).

Methods: Data from 1204 treatment change episodes (TCEs) were identified from the HIV Resistance Response Database Initiative (RDI) database and partitioned at random into a training set of 1154 and a test set of 50. The training set was then partitioned using an *L*-cross (*L* = 10 in this study) validation scheme for training individual computational models. Seventy six input variables were used for training the models: 55 baseline genotype mutations; the 14 potential drugs in the new treatment regimen; four treatment history variables; baseline viral load; CD4 count and time to follow-up viral load. The output variable was follow-up viral load. Performance was evaluated in terms of the correlations and absolute differences between the individual models' predictions and the actual ΔVL values.

Results: The correlations (r^2) between predicted and actual ΔVL varied from 0.318 to 0.546 for ANN, 0.590 to 0.751 for RF and 0.300 to 0.720 for SVM. The mean absolute differences varied from 0.677 to 0.903 for ANN, 0.494 to 0.644 for RF and 0.500 to 0.790 for SVM. ANN models were significantly inferior to RF and SVM models.

The predictions of the ANN, RF and SVM committees all correlated highly significantly with the actual ΔVL of the independent test TCEs, producing r^2 values of 0.689, 0.707 and 0.620, respectively. The mean absolute differences were 0.543, 0.600 and 0.607 $\log_{10}$ copies/ml for ANN, RF and SVM, respectively. There were no statistically significant differences between the three committees.

Combining the committees' outputs improved correlations between predicted and actual virological responses. The combination of all three committees gave a correlation of $r^2 = 0.728$. The mean absolute differences followed a similar pattern.

Conclusions: RF and SVM models can produce predictions of virological response to HIV treatment that are comparable in accuracy to a committee of ANN models. Combining the predictions of different models improves their accuracy somewhat.

This approach has potential as a future clinical tool and a combination of ANN and RF models is being taken forward for clinical evaluation.

1. Introduction

Despite the approval of more than 20 antiretroviral drugs, HIV treatment failure due to drug resistance still occurs. HIV genotyping is recommended by a range of HIV treatment guidelines and is commonly employed to help the selection of a new regimen to re-establish viral suppression [1—3]. However, the complexity of resistance patterns and the expanding range of therapeutic options available have made the interpretation of genotype results in order to optimise virological treatment response extremely challenging [1]. A number of interpretation systems have been developed that relate HIV genotype to single antiretroviral drug susceptibility using different 'rules' or algorithms [for example, 4—7] and relational databases have been used to predict resistance to specific drugs by matching a test genotype with archived genotypic and phenotypic data [8,9]. There is no recognised standard interpretation system and different systems can produce different results from the same genotype [10—13].

Several groups have explored the use of bioinformatics to address the challenges of genotype interpretation and response prediction [14 for a review]. For example, artificial neural networks (ANN) [15], decision trees [16], support vector machines (SVM) [9] or phenotype matching in relational databases [17] have all been used to predict phenotype from genotype. Other groups have gone further to relate the predicted phenotype of individual drugs to virological response. However, the relationship between phenotype and response to combination therapy is not well characterized and attempting to infer response from genotype via the intermediate step of predicted phenotype has serious limitations [18]. Most of the groups that have attempted this have related predicted phenotype to a categorical prediction of response, with cut-offs in predicted fold-changes in phenotypic sensitivity linked to clinical response [e.g. 19]. However, in terms of potential clinical utility, a strong case can be made for predicting response to combination therapy (rather than individual drugs) as a continuous variable [20], *directly* from genotype. Given the complexity of the drug and genotype permutations the main obstacle facing this approach is the size of the dataset required [21].

The HIV Resistance Response Database Initiative (RDI) is a not-for-profit organization set up to

establish a large clinical database and develop bioinformatic techniques to define the relationships between HIV resistance and virological response to treatment. It is hoped that this approach might potentially overcome some of the limitations of current interpretation systems [22]. The development of the database is an international collaboration and data from more than 50,000 HIV patients have already been provided by a variety of private and public research groups.

The ultimate aim is to develop computational models that are able to predict treatment response accurately from genotype and other clinically relevant information, which will then be made freely accessible as an aid to treatment selection.

We recently demonstrated that ANN models trained with datasets from multiple clinical sources can be accurate predictors of virological response to combination therapy [23]. Here we tested the accuracy of two alternative computation modelling methods, namely random forests (RF) and SVM, and compare their performance individually and in combination with that of ANN models, using the same dataset.

The principle of RF is to grow many decision trees in parallel. For a given sample, votes are carried out over all the trees in the forest. The individual trees are built using different sets of samples from the original training dataset. In each node of a tree, the splitting feature is selected from a randomly chosen sample of features. In RF modelling, the training datasets of the individual trees are built by bootstrap replication, leaving about one-third of the samples out of the bootstrap sample, which are used for validation. The injection of randomness makes RF highly resistant to over-fitting [24,25]. A disadvantage of RF is that the model is complex and cannot be visualised like a single tree [25].

The principle of SVM is to map the data into a high-dimensional feature space and then perform linear regression. SVM searches for a global solution and does not control model complexity by keeping the number of input variables small [26,27]. It is considered more resistant to 'over-fitting' based on the training dataset and, therefore, potentially more generalisable to new data [28]. The drawback of SVM is its high algorithmic complexity [29].

2. Materials and methods

2.1. Data

The basic package of information that is used for the training of the RDI's models is the treatment change episode (TCE) as illustrated in Fig. 1 [30]. This comprises key information required by the models from a patient who has had a new treatment started, in order to develop a prediction of virological response. It includes baseline genotype, viral load, CD4+ T-lymphocyte (CD4) count and other information as well as the follow-up viral load value: the response variable that the models are being trained to predict.

Based on the results of previous studies, the RDI applied the following criteria to the TCEs in order to optimise the training and performance of the models:

1. A baseline genotype must be available from a plasma sample taken no more than 12 weeks prior to treatment change date
2. Baseline viral load no more than 8 weeks prior to treatment change date
3. Baseline CD4 count no more than 12 weeks prior to treatment change date
4. Details of at least one previous treatment available
5. Follow-up viral load available from between 4 and 48 weeks after treatment change date.

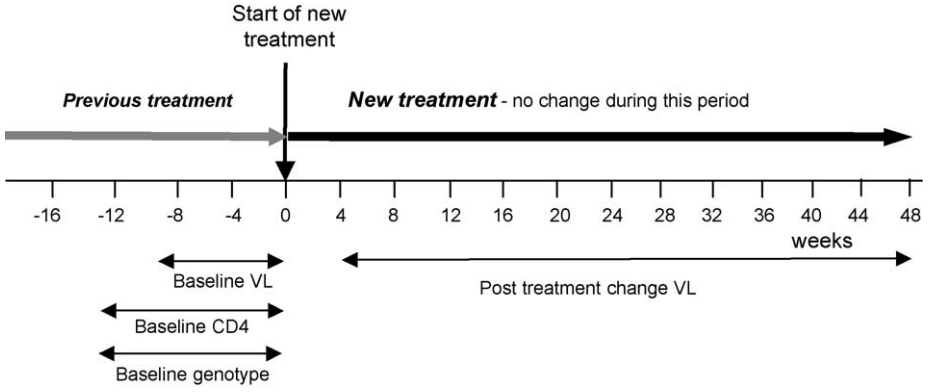

Figure 1 Treatment change episode (TCE).

Results of previous studies have shown that including multiple TCEs from the same patient-treatment change, using follow-up viral loads at different times enhances the accuracy of prediction of the models. However, as a precaution against over-training, the number of TCEs permitted from a single treatment change, using different follow-up viral loads was restricted to a maximum of three, each more than 28 days apart. The RDI database has, on average, approximately two TCEs per treatment change under this rule and this was reflected in data used in this study.

2.2. Input variables

Seventy six input variables were used in the training of all the models:

a. 55 baseline genotypic mutations (coded as binary variables where 1 = the mutation was present and 0 = it was not). The mutations were selected on the basis of frequency in the RDI database, their relative impact on treatment responses in previous multivariate analyses (data on file) and the HIV resistance literature (HIV reverse transcriptase codons: M41L, E44D, A62V, K65R, D67N, 69 insert, T69D/N, K70R, L74V, V75I, F77L, A98G, L100I, L101I/E, K103N, V106A, V108I, Y115F, F116Y, V118I, Q151M, V179D, Y181C, M184V, Y188C/L/H, G190S/A, L210W, T215Y, T215F, K219Q/E, P236L; HIV protease mutations: L10F/I/R/V, K20M/R, L24I, D30N, V32I, L33F, M36I, M46I/L, I47V, G48V, I50V, I50L, F53L, I54V/L, L63P, A71V/T, G73S/A, V77I, V82A/F/S, V82T, I84V, I84A, N88S, L90M)
b. Drugs in the new combination regimen (coded as binary variables where 1 = the drug was used in the regimen and 0 = it was not). All 14 drugs appearing in the training and test datasets were covered (zidovudine, didanosine, stavudine, abacavir, lamivudine, tenofovir, efavirenz, nevirapine, indinavir, nelfinavir, ritonavir as a protease inhibitor booster, saquinavir, amprenavir, lopinavir)
c. Four treatment history variables (binary variables coding for any historical exposure to zidovudine, lamivudine, any non-nucleoside reverse transcriptase inhibitor (NNRTI) and any protease inhibitor (PI)
d. Baseline viral load (log copies HIV RNA/ml)
e. Baseline CD4 count (cells/ml)
f. Time to follow-up viral load (number of days),

all as described previously [4].

The output variable was follow-up viral load (log copies RNA/ml) [4].

2.3. Development of computational models

The development of computational models consists of the following phases (Fig. 2): (a) partitioning the data, (b) training individual models using ANN, RF, and SVM machine learning methods, (c) forming model committees for each of ANN, RF and SVM methods, and (d) combining the outputs of the committees.

2.3.1. Data partitioning

The data used in this study comprised 1204 TCEs taken from the RDI database. Because of the large number of input variables and limited TCEs available at the time, the size of the training dataset was maximised while retaining an independent test set sufficient for statistical testing. Therefore 50 test TCEs and 1154 training TCEs were selected, at random except for the constraint that TCEs from the same patient could not appear in both the training and test datasets. The 1154 TCEs in the training set

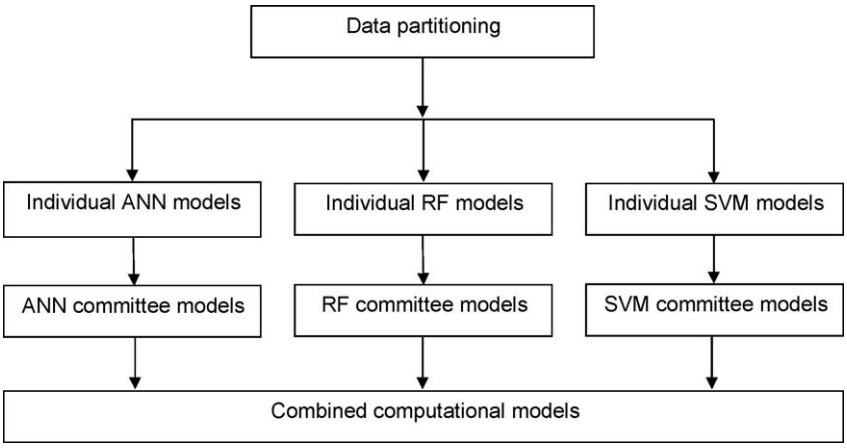

Figure 2 The structure of the computational models.

were partitioned using an L-cross ($L = 10$ in this study) validation scheme for training individual computational models.

2.3.2. ANN models

ANN models were trained using sub-training datasets, which were assumed to be independently drawn from the joint distribution of (X,Y) and comprised $N(m + 1)$ patterns (x_k,y_k), $k = 1, 2,\ldots,N$, Y is the observed follow-up viral load. The predicted follow-up viral load was estimated using (1).

$$o(x;w) = g(\sum_{j=1}^{H} w_{0j} h_j) \tag{1}$$

$$h_j = \sigma(\sum_{i=1}^{m} w_{ji} x_i + w_j), \quad j = 1,2,\ldots,H \tag{2}$$

$$\varepsilon = \frac{1}{2}\sum_{k=1}^{N}[o(x_k;w) - y_k]^2 \tag{3}$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{4}$$

$$\bar{c}(x) = \frac{1}{L}\sum_{l=1}^{L} o_l(x;w) \tag{5}$$

where x represents m observed input variables, w_{0j} was the output weight from a hidden node j to an output node and g was a linear output function. The value of a hidden layer node h_j, $j = 1, 2,\ldots,H$ (H was the number of hidden nodes) was given by (2), w_{ji} was input weight from an input node i to a hidden node j, w_j was threshold weight from an input node that had a constant value 1 to a hidden node j, x_i was the value at the input node i, representing either a mutation (0/1) or a drug (0/1), or a treatment history variable (0/1), or the baseline viral load (a real number), or the baseline CD4 count (an integer), or the time to follow-up (a real number), and σ was a sigmoid function and was defined by Eq. (4). The weights between nodes were obtained using a back-propagation algorithm to minimise the error function defined in Eq. (3) by iteratively adjusting the weights between the interconnections [15,23]. The predicted follow-up viral load from the committee of ANN model was estimated by Eq. (5).

2.3.3. RF models

A random forest model is a group of tree predictors $f(x;\theta_t)$, $t = 1, 2,\ldots,T$, where x represents m observed input variables with associated random vector X and θ_t are independent and identically distributed random vectors. The training dataset was assumed to be independently drawn from the joint distribution of (X, Y) and comprised $K(m + 1)$ patterns (x_i, y_i), $i = 1, 2,\ldots,N$ (N is sample size for each tree), which was a sub set of $N(m + 1)$ patterns used in the ANN modelling. The random forest prediction was calculated by Eq. (6).

$$f(x) = \frac{1}{T}\sum_{t=1}^{T} f(x;\theta_t) \tag{6}$$

According to the law of large numbers, $E_{X,Y}(Y - f(X))^2 \to E_{X,Y}(Y - E_\theta f(X;\theta))^2$ if $T \to \infty$. The training procedure of RF models included the following steps: Firstly, a bootstrap sample was drawn from the whole training dataset. Secondly, a tree was built for each bootstrap sample at each node, the best split among a randomly selected subset of input variables was chosen. The tree building was stopped when the tree was grown to the maximum size (the number of cases in a node is below a threshold of 5). Thirdly, these steps were repeated to generate a sufficiently large number (a variable from 200 to 500) of trees. The RF model was trained using the random forest package in R [31]. The predicted follow-up viral load from the committee of RF models was estimated by Eq. (7).

$$\bar{r}(x) = \frac{1}{L}\sum_{l=1}^{L} f_l(x) \tag{7}$$

2.3.4. SVM models

The same $N(m + 1)$ patterns (x_k, y_k) as used in the ANN modelling were utilised to train individual SVM models. For an ε-insensitive loss function, the SVM model parameters were determined by Eq. (8)

$$\begin{aligned}\max_{\alpha,\beta} W(\alpha,\beta) = \max_{\alpha,\beta}\Big(&\sum_{i=1}^{N}\alpha_i(y_i - \varepsilon) - \beta_i(y_i + \varepsilon) \\ &- \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \beta_i)(\alpha_j - \beta_j)K \\ &\times (x_i, x_j)\Big)\end{aligned} \tag{8}$$

with constraints, defined by Eq. (9),

$$\begin{aligned}&0 \le \alpha_i,\ \beta_i \le C, \quad i = 1,2,\ldots,N \\ &\sum_{i=1}^{N}(\alpha_i - \beta_i) = 0\end{aligned} \tag{9}$$

where the kernel function is defined by Eq. (10)

$$K(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2} \tag{10}$$

The follow-up viral load predicted by the SVM model is given by Eq. (11)

$$s(x) = \sum_{p=1}^{V}(\alpha_p^* - \beta_p^*)K(x_p, x) \tag{11}$$

where V is the number of support vectors. The SVM models were trained using LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). The predicted follow-up viral load from the committee of SVM models was estimated by Eq. (12).

$$\bar{s}(x) = \frac{1}{L}\sum_{l=1}^{L} s_l(x) \quad (12)$$

2.3.5. Testing the models

The performance of the models was assessed by providing them with the input data from the independent test dataset and then comparing the predictions of the models with the actual change in viral load (ΔVL) from those test TCEs. The comparison was made in terms of the correlation between actual and predicted ΔVL (r^2 generated from Pearson product—moment correlations) and the mean absolute difference between the actual and the predicted ΔVL. The performance of the ANN, RF and SVM committees was assessed using the committee average prediction for each TCE [23].

2.3.6. Combining model outputs

In addition to comparing the accuracy of the predictions of the three machine learning methods, their outputs were combined in three different ways: (a) the mean of the predictions from ANN, RF and SVM committee models, $\mu = \sum_{i=1}^{3} \mu_i/3$, where, μ_i, i = 1,2,3, represent the predicted viral load from the ANN, RF, and SVM committee models, respectively, (b) the weighted mean by the inverse variance of the differences between the predicted and actual ΔVL values in the test dataset, $\mu = \sum_{i=1}^{3} w_i\mu_i / \sum_{i=1}^{3} w_i$, where $w_i = 1/_i$, v_i is the variance and (c) the weighted mean by the r^2 values between the predicted and the actual viral loads, $\mu = \sum_{i=1}^{3} w_i\mu_i / \sum_{i=1}^{3} w_i$, where $w_i = R_i^2$.

2.4. Estimate of confidence intervals

The output of the computational models is the predicted follow-up viral load for a patient with a specific set of baseline variables, including HIV genotype. In order to aid decision-making in selecting a new therapeutic regimen for the patient, it is important to provide uncertainty estimates (for example, confidence intervals) associated with the predictions of virological response. Suppose that $t = t(x)$ is the true viral load we want to approximate and that $D = \{(x_k, y_k), k = 1, 2, \ldots, n\}$ is an independent test dataset, where $y = t + \varepsilon$, ε is the noise with $E(\varepsilon) = 0$. The mean squared error of the model predictions o is defined by $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - o_i)^2$. The expectation of MSE can be expressed as

$$\begin{aligned} E(\text{MSE}) &= \frac{1}{n}\sum_{i=1}^{n} E(y_i - o_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(E(y_i - t_i)^2 + E(t_i - o_i)^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}(E(\varepsilon^2) + E(t_i - E(o_i))^2 + (E(o_i) - o_i)^2) \end{aligned}$$

The first term in the right side of the equation is the variance of noise, the second term is the model bias squared, and the third term is the variance of the model predictions. The variance of noise is independent of the computational models. It can be estimated by training an auxiliary computational model on the residuals of the committee predictions [32]. However, this is computationally time-consuming because it involves training many models to ensure a reliable estimate. The model bias vanishes completely if the computational models are perfectly trained. Most re-sampling based techniques neglect the contribution of the model bias to the total error variance [32,33]. The model variance can be estimated by a committee of computational models. In this paper we constructed confidence intervals for the predicted viral loads based on the model variance only.

Table 1 Performance of individual ANN, RF and SVM models with independent test data.

	Correlations (r^2)			Mean absolute difference scores		
	ANN	RF	SVM	ANN	RF	SVM
Model 1	0.380	0.59	0.660	0.859	0.644	0.500
Model 2	0.360	0.625	0.520	0.736	0.640	0.604
Model 3	0.331	0.628	0.370	0.732	0.623	0.732
Model 4	0.352	0.749	0.620	0.735	0.494	0.576
Model 5	0.435	0.654	0.430	0.733	0.639	0.728
Model 6	0.358	0.716	0.300	0.691	0.583	0.790
Model 7	0.546	0.717	0.490	0.730	0.595	0.625
Model 8	0.318	0.666	0.720	0.677	0.632	0.507
Model 9	0.421	0.647	0.630	0.903	0.625	0.553
Model 10	0.436	0.751	0.480	0.772	0.578	0.624
Mean	0.394	0.674	0.522	0.757	0.605	0.624
SD	0.068	0.056	0.135	0.071	0.046	0.098

3. Results

The correlations and absolute differences between the individual models' predictions and the actual ΔVL values are summarised in Table 1. These results reveal marked differences between the different methods. The r^2 of the individual ANN models varied from 0.318 to 0.546, with a mean (SD) of 0.394 (0.068) and a coefficient of variation of 18%. The r^2 of the individual RF models varied from 0.590 to 0.751, with a mean (SD) of 0.674 (0.056) and a coefficient of variation of 8%. The r^2 of the individual SVM models varied from 0.300 to 0.720, with a mean (SD) of 0.522 (0.135) and a coefficient of variation of 26%. Individual SVM and ANN models varied more in the accuracy of their predictions than did the RF models.

The mean absolute differences between the actual ΔVL values and the predictions by individual ANN models varied from 0.677 to 0.903, with a mean (SD) of 0.757 (0.071) and a coefficient of variation of 9%. For the RF models this value varied from 0.494 to 0.644, with a mean (SD) of 0.605 (0.046) and a coefficient of variation of 8%. For the SVM models, this value varied from 0.500 to 0.790, with a mean (SD) of 0.624 (0.098) and a coefficient of variation of 16%. Again on average, the RF models gave the smallest variations. However, the variations between the ANN models and the RF models were not markedly different. In summary, the performance of the individual ANN models was significantly inferior to that of individual RF and SVM models in terms of the correlations ($p < 0.0001$ and $p < 0.05$, respectively) and absolute differences ($p < 0.0001$ and $p < 0.01$) between predicted and actual viral load values.

The predictions of the ANN, RF and SVM committees all correlated highly significantly with the actual ΔVL of the independent test TCEs, producing r^2 values of 0.689, 0.707 and 0.620, respectively ($p < 0.00001$). The scatter plots of predicted versus actual ΔVL values are presented in Fig. 3. The mean absolute differences between the models' predictions and the actual ΔVL values were 0.543, 0.600 and 0.607 $\log_{10}$ copies/ml for ANN, RF and SVM, respectively. The performance measures for the model committees are summarised in Table 2. The

Table 2 Summary of results for ANN, RF and SVM committees.

	Independent test data	
	r^2	Mean absolute difference
ANN	0.689	0.543
RF	0.707	0.600
SVM	0.620	0.607

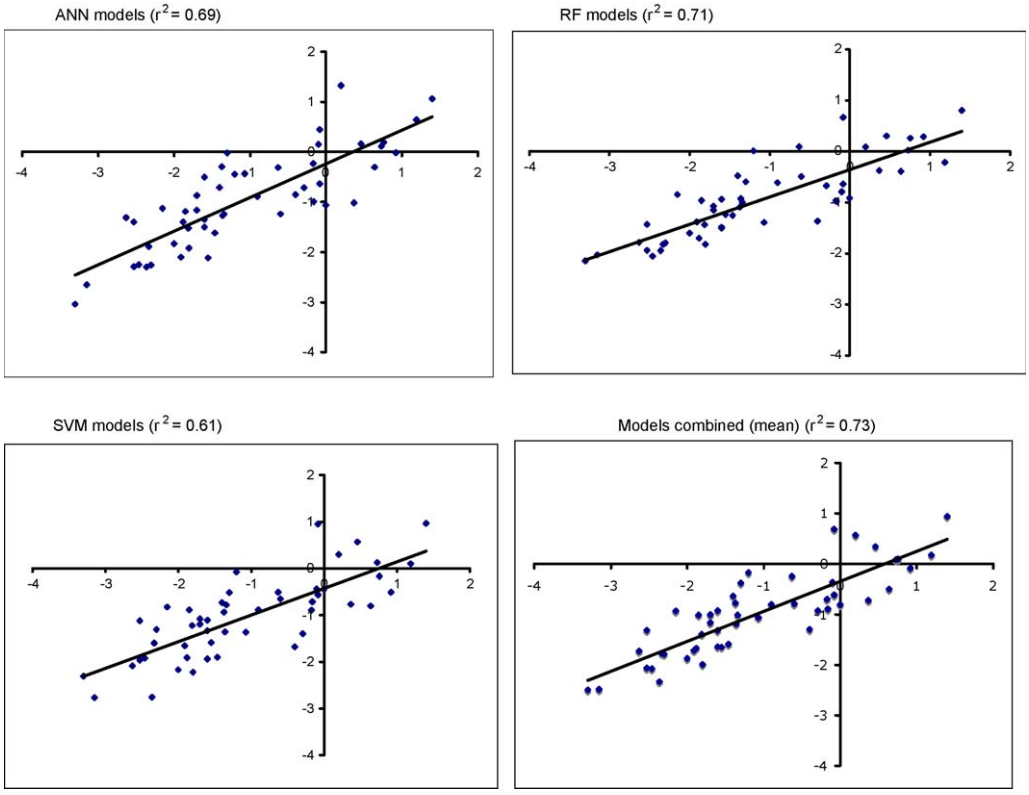

Figure 3 Actual versus predicted viral load change for an independent test dataset.

Table 3 Summary of results for combined models.

	Equally weighted		r^2 weighted		IV weighted	
	r^2	Mean absolute difference	r^2	Mean absolute difference	r^2	Mean absolute difference
ANN and RF	0.747	0.543	0.747	0.542	0.746	0.543
ANN and SVM	0.701	0.551	0.703	0.549	0.704	0.548
RF and SVM	0.686	0.579	0.690	0.580	0.690	0.580
ANN, RF and SVM	0.728	0.547	0.731	0.546	0.730	0.547

performance of the ANN committee in terms of the r^2 is comparable to that of the RF committee and numerically superior to that of the SVM committee. There were no statistically significant differences between the three modelling methods (ANOVA, $F = 0.42$, d.f. = 2, $p = 0.74$). However, in contrast to ANN, the performance of the individual SVM and particularly RF models was not markedly different to that of the committees.

The results of combining computational models using differently weighted methods are summarised in Table 3. For the equally weighted method, combining the committees' outcomes produced different correlations between predicted and actual virological responses. The r^2 varied from 0.686 for a combination of RF and SVM models up to 0.747 for a combination of RF and ANN models. The combination of predictions from all three committee models gave a correlation of $r^2 = 0.728$. The mean absolute differences between predicted and actual virological responses followed a similar pattern as the correlations, ranging from 0.543 log copies for a combination of RF and ANN models up to 0.579 log copies for a combination of RF and SVM models. The results obtained using r^2 weighted and the inverse variance weighted methods were not markedly different to those obtained using the equally weighted method as shown in Table 3.

There were no statistically significant differences between the predictions from each of the four combinations of models' outputs or between the combined outputs and the individual outputs in terms of absolute differences between predicted and actual virological responses ($F = 0.32$, $p = 0.93$). Overall, the combined predictions of the RF and ANN models achieved the highest correlation with actual responses. Comparison of the absolute difference between predicted and actual responses for this combination of models approached statistically superiority over those of the SVM model, which produced the least accurate predictions ($p = 0.086$).

Table 4 lists the actual follow-up viral loads, the predicted follow-up viral loads and 95% confidence intervals relating to the combined predictions of the ANN and RF committees for the independent test dataset. In 44 out of 50 patients (88%), the predicted 95% confidence intervals included the actual follow-up viral loads. For three of the remaining six patients, the differences between the actual follow-up viral loads and the upper or lower limit of the predicted 95% confidence intervals were less than or equal to the 0.1 $\log_{10}$ copies/ml. For the other three patients, the differences between the actual follow-up viral loads and the upper or lower limit of the predicted 95% confidence intervals were less than or equal to the 0.2 $\log_{10}$ copies/ml. The results obtained from a combination of ANN, RF, and SVM committee models are shown in Table 5. The predicted 95% confidence intervals were similar to those obtained by a combination of ANN and RF committee models, albeit slightly wider. This is in accordance with the performance of these combined models as shown in Table 3.

In order to test if the models performed differently for patients with a short versus a long time to follow-up, the 50 test TCEs were divided into two sets of 25: those with the shortest time to follow-up (range = 4—13 weeks, mean = 8) and those with the longest the longest (range = 13—47 weeks, mean = 26). The performance of the models was found to be comparable for both groups. The correlations between the combined predictions and the actual ΔVL yielded an r^2 value (mean absolute difference score) of 0.69 (0.54) and 0.76 (0.56), respectively for the TCEs with shorter and longer follow-up times.

4. Discussion

In terms of the main measure of the correlation between predicted and actual virological response, individual ANN models performed significantly worse in their predictions of virological response to HIV therapy than RF and SVM models and their predictions were significantly more variable that were those of RF models.

The use of a model committee substantially improved the accuracy of the ANN predictions.

Table 4 The actual and predicted follow-up viral loads by a combination of ANN and RF models and the 95% CI.

Patient	Actual viral load	Predicted viral load	95% CI		Patient	Actual viral load	Predicted viral load	95% CI	
			Low VL	Upper VL				Low VL	Upper VL
1	2.6	2.1	1.7	4.7	26	5.5	4.4	2.5	5.4
2	5.4	4.7	3	5.8	27	1.9	2.4	1.8	2.8
3	1.7	2.5	1.7	3.1	28	1.7	2	1.7	2.5
4	2.7	3.5	1.7	5.2	29	2	2.5	1.7	4.8
5	2.7	2.6	1.8	3.1	30	4.5	3.9	2.6	5.5
6	1.7	2.4	1.7	2.8	31	2.6	2.4	1.7	3.4
7	1.7	2	1.7	2.2	32	5	4.6	2.4	5.8
8	4.5	3.6	2.5	5.6	33	4.9	4	2.3	5.8
9	1.7	2.7	1.8	3.1	34	2.9	3.1	1.8	4.6
10	2.7	2.9	1.8	5	35	2.3	2.7	1.8	4.8
11	4.7	4	2.7	5.5	36	5.9	5.1	3.9	5.7
12	1.9	2.7	1.8	4.1	37	1.7	2.5	1.7	3.8
13	1.7	2.8	1.9	5	38	3.2	3.7	2.3	5.8
14	2.5	3.4	1.7	5.8	39	1.9	1.9	1.7	2.2
15	1.9	2.6	1.7	4.6	40	1.7	2	1.7	3.1
16	2.8	3.5	1.9	5.8	41	2.8	3	1.8	4.2
17	2.3	2.8	1.7	4.8	42	3.2	2.9	1.9	3.5
18	3.5	3.1	1.8	5.2	43	2.1	2	1.7	2.4
19	1.7	2.9	1.7	4.8	44	4.7	4.5	2.7	5.8
20	2.5	2.8	1.7	4.9	45	3.8	4	1.9	5.7
21	5.7	4.6	2.4	5.8	46	5.2	4.7	3.3	5.7
22	5.9	4.9	2.9	5.8	47	1.7	2.9	1.9	5.1
23	2.6	2.9	2.1	3.2	48	2.6	2.8	2.4	3.8
24	2.6	2.8	2	3	49	1.7	1.8	1.7	2
25	2.7	3.7	1.7	5.6	50	1.7	2	1.7	2.3

For example, the r^2 of the ANN committee was 0.689 while the average of the r^2 of the individual ANN models was just 0.394. The use of a committee also improved the predictions of RF and SVM somewhat. In fact, this is not surprising as the correlations between the actual and predicted values made by averaging the predictions of individual models of each TCE is always greater than or equal to the average of the correlations between the actual viral load values and the predictions for individual models. However the improvement seen with committees of RF and SVM was not as great as for ANN, with the committees failing to out-perform the best individual models. As a result, the ANN committee was comparable in accuracy to the RF committee (although not as accurate as four of the individual RF models) and numerically but not significantly superior to the SVM committee.

The smallest coefficient of variation was achieved in the case of the RF models, indicating that the individual RF models were quite stable in predicting virological response. This may be because the individual RF models were composed of many different tree models. These tree models were built using different sets of samples from the original training dataset. In each node of a tree, the splitting feature was selected from a randomly chosen sample of features. In RF modelling, the training datasets of the individual trees were built by bootstrap replication, leaving about one-third of the samples out of the bootstrap sample. It is the injection of randomness that is likely to have made the individual RF models highly stable. This is in accordance with the studies reported by Liaw and Breiman [24,25]. In contrast, the individual SVM models gave the largest coefficient of variation. The correlation between predicted and actual viral loads was more than 10% worse for the SVM committee model than for the ANN and RF committee models, suggesting that SVM may have a poor generalisability. This is inconsistent with the study reported by Furey [28], in which SVM was used to classify cancer tissue samples using microarray expression data. The discrepancy may be due to the fact that in the Furey study a hold-one-out test scheme was used to test the generalisability of SVM classifiers, while we used an actual independent test dataset. Nevertheless, it was somewhat surprising that the SVM models did not perform better than they did.

Combining the predictions from models trained using different machine learning methods for each test case generally improved the correlations between predicted and actual virological responses

Table 5 The actual and predicted follow-up viral loads by a combination of ANN + RF + SVM models the 95% CI.

Patient	Actual viral load	Predicted viral load	95% CI		Patient	Actual viral load	Predicted viral load	95% CI	
			Low VL	Upper VL				Low VL	Upper VL
1	2.6	1.9	1.5	4.7	26	5.5	4.3	2.5	5.4
2	5.4	4.7	3.2	5.7	27	1.9	2.5	1.8	3.1
3	1.7	2.5	1.7	3.1	28	1.7	2.1	1.7	2.5
4	2.7	3.4	1.9	5.2	29	2	2.4	1.5	4.8
5	2.7	2.6	1.8	3.1	30	4.5	4	2.6	5.5
6	1.7	2.5	1.7	3	31	2.6	2.5	1.7	3.4
7	1.7	1.8	1.3	2.2	32	5	4.3	2.4	5.7
8	4.5	3.7	2.5	5.2	33	4.9	3.9	2.3	5.6
9	1.7	2.7	1.8	3.1	34	2.9	3	1.8	4.6
10	2.7	2.9	1.8	5	35	2.3	2.8	1.8	4.8
11	4.7	3.8	2.7	5.5	36	5.9	4.9	3.3	5.8
12	1.9	2.7	1.8	3.8	37	1.7	2.6	1.7	3.8
13	1.7	2.6	1.8	4.7	38	3.2	3.6	2.3	5.8
14	2.5	3.2	1.7	5.8	39	1.9	1.8	1.2	2.2
15	1.9	2.6	1.7	4.6	40	1.7	2.1	1.7	3.1
16	2.8	3.4	1.9	5.8	41	2.8	3.1	1.8	3.7
17	2.3	2.8	1.7	4.8	42	3.2	3	1.9	3.5
18	3.5	3	1.8	5.1	43	2.1	1.9	1.2	2.4
19	1.7	2.9	1.7	4.8	44	4.7	4.4	2.7	5.8
20	2.5	2.7	1.7	4.9	45	3.8	3.8	1.9	5.7
21	5.7	4.5	2.4	5.8	46	5.2	4.8	3.3	5.7
22	5.9	5.1	2.9	5.8	47	1.7	3	1.9	3.3
23	2.6	3.1	2.1	3.9	48	2.6	2.9	2.2	3.8
24	2.6	2.7	2	3	49	1.7	1.7	1.2	2
25	2.7	3.6	1.7	5.6	50	1.7	1.7	1.1	2.3

to treatment, although this was not statistically significant. The performance of the combined models seemed not to be affected by the combination methods, suggesting that a simple averaging method could be used to combine different computational models. The accuracy of the models' predictions also appears to be independent of time to follow-up up to one year. The combination of models that achieved the highest correlation and smallest mean absolute difference score was a combination of RF and ANN. This combination of modelling methods has been taken forward into clinical testing as a treatment decision support tool.

The mean absolute difference between the predictions of the best models and the actual virological responses, at approximately 0.5 log HIV RNA/ml, was comparable to the limit of reproducibility of the viral load assays in current use. A corollary is that approximately half of the predictions (44% for the RF and ANN combined outputs) were out by more than 0.5 log. While this is a concern, reducing this is likely to require improvements in the assays themselves. In this, as in previous studies (data on file) the models were somewhat conservative (making more under-estimates of virological response than over estimates). In addition, substantial over-estimates of response may well have been due to poor therapy-adherence, which is clearly beyond the scope of such a system.

The clinical utility of models such as these is arguably more dependent on ranking different regimens accurately in order of virological response, than on the absolute accuracy of their predictions, which is why our primary measure of accuracy is the correlation between predicted and actual response.

The results of confidence interval estimates revealed that the 95% confidence intervals estimated by the RF committee model were usually narrower compared to those estimated by ANN or SVM committee models. This was probably due to the fact that a RF model consists of a large number of trees which are trained using different sets of samples randomly selected with replacement. Due to the nature of sampling, these tree models might be highly dependent. This implies that different RF models may be correlated to some extend and may suggest that 95% confidence intervals constructed using the RF committee model alone may be problematic. However, our results have demonstrated that using a combination of ANN, RF, and/or SVM models provides reliable estimates of 95% confidence intervals for the predicted viral loads.

It is possible that construction of confidence intervals by taking into account all the uncertainties,

including model bias and variance of noise, would give a more accurate estimate. Nevertheless, the results from this study have shown that in about 90% of cases in the independent test dataset the 95% confidence intervals cover their actual viral loads, and that in the remaining cases the actual viral loads fall in windows by enlarging their 95% confidence intervals by 0.2 $\log_{10}$ copies/ml. This suggests that the variance of noise and model bias appears to have had, at most, only a small impact on confidence interval estimates for the predicted viral loads using a combination of computational models.

A potential limitation to this study was the size of the dataset available and, as a result, the comparatively small size of the independent test set. We have since collected additional data from more recent clinical practice. A search of these data enabled us to construct an additional test set of 50 TCEs. When the ANN, RF and SVM committees were tested using these more recent data, performance was not significantly different from that achieved with the original, contemporaneous test data, although the r^2 values were reduced to 0.543, 0.564 and 0.475 compared to 0.689, 0.707 and 0.620 with the original test datasets. As with the original test data, there were no significant differences between the different models' performance but the trend for SVM models to perform less well than the other methods was replicated.

Future studies are being planned and undertaken with larger test sets as we now have larger numbers of TCEs available for modelling. In addition, as well as developing new models to predict absolute virological response to all the available antiretroviral drugs, we are developing models to estimate the probability of the viral load falling below the limit of detection of assays in widespread use (currently 50 copies/ml).

In conclusion, RF and SVM model committees are able to predict virological response to HIV therapy with accuracy that does not differ significantly from that of ANN. Individual RF proved the most accurate and consistent of the individual models and ANN and RF committees provided the best combined performance. The results of this study provide the basis for ongoing larger studies and support to the RDI's drive to develop a treatment decision tool using a combination of computational models.

Acknowledgements

This research has been funded with Federal Funds from the National Cancer Institute, National Institutes of Health, under contract No. NO1-CO-12400.

The authors would also like to acknowledge the following institutions and research groups for the provision of data to the RDI: National Institute of Allergy and Infectious Diseases, USA; BC Centre for Excellence in HIV/AIDS, Vancouver, BC Canada; USA Military HIV Research Program; ICONA; The Italian ARCA database; Hospital Clinic of Barcelona, Spain; Fundació IrsiCaixa, Badalona, Spain; Northwestern University Division of Infectious Diseases, Chicago, USA; National Centre in HIV Epidemiology and Clinical Research, Sydney, Australia; Gilead Sciences, Foster City, USA; Ramón y Cajal Hospital, Madrid, Spain; University of Brescia, Italy; Community Programs for Clinical Research on AIDS (CPCRA), USA; Athena database, HIV Monitoring Foundation, Amsterdam, Netherlands; Hôpital Timone, Marseilles, France; Royal Free Hospital, London, UK; AIDS Research Center, National Institute of Infectious Diseases, Tokyo, Japan; Hospital of the Johann Wolfgang Goethe-University, Frankfurt, Germany; Chelsea and Westminster Hospital, London, UK.

References

[1] Hirsch MS, Günthard HF, Schapiro JM, Brun-Vézinet F, Clotet B, Hammer SM, et al. Antiretroviral drug resistance testing in adult HIV-1 infection: 2008 recommendations of an International AIDS Society-USA panel. Clin Infect Dis 2008;47(2): 266—85.

[2] Department of Health and Human Services Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. Washington, DC: Department of Health and Human Services; May 4, 2006.

[3] Vandamme AM, Sönnerborg A, Ait-Khaled M, Albert J, Asjo B, Bacheler L, et al. Updated European recommendations for the clinical use of HIV drug resistance testing. Antivir Ther 2004;6:829—48.

[4] Ormaasen V, Sandvik L, Asjo B, Holberg-Petersen M, Gaarder P, Bruun J. An algorithm-based genotypic resistance score is associated with clinical outcome in HIV-1-infected adults on antiretroviral therapy. HIV Med 2004;5:400—6.

[5] Wang K, Jenwitheesuk E, Samudrala R, Mittler JE. Simple linear model provides highly accurate genotypic predictions of HIV-1 drug resistance. Antivir Ther 2004;9:343—52.

[6] Sturmer M, Doerr HW, Preiser W. Variety of interpretation systems for human immunodeficiency virus type 1 genotyping: confirmatory information or additional confusion? Curr Drug Targets Infect Disord 2003;3:373—82.

[7] Shapiro JM, De Luca A, Harrigan PR, Hellmann N, McCreedy B, Pillay D, et al. Resistance assay interpretation systems vary widely in method and approach. Antivir Ther 2001;6(Suppl. 1):131.

[8] Gallego O, Martin-Carbonero L, Aguero J, de Mendoza C, Corral A, Soriano V. Correlation between rules-based interpretation and virtual phenotype interpretation of HIV-1 genotypes for predicting drug resistance in HIV-infected individuals. J Virol Methods 2004;121:115—8.

[9] Beerenwinkel N, Däumer M, Oette M, Korn K, Hoffmann D, Kaiser R, et al. Geno2pheno: estimating phenotypic drug

resistance from HIV-1 genotypes. Nucleic Acids Res 2003;31: 3850—5.
[10] Shafer RW, Gonzales MJ, Brun-Vezinet F. Online comparison of HIV-1 drug resistance algorithms identifies rates and causes of discordant interpretations. Antivir Ther 2001; 6:101.
[11] Torti C, Quiros-Roldan E, Keulen W, Scudeller L, Lo Caputo S, Boucher C, et al. Comparison between rules-based human immunodeficiency virus type 1 genotype interpretations and real or virtual phenotype: concordance analysis and correlation with clinical outcome in heavily treated patients. J Infect Dis 2003;188:194—201.
[12] Sturmer M, Doerr HW, Staszewski S, Preiser W. Comparison of nine resistance interpretation systems for HIV-1 genotyping. Antivir Ther 2003;8:239—44.
[13] De Luca A, Cingolani A, Di Giambenedetto S, Trotta M, Baldini F, Rizzo MG, et al. Variable prediction of antiretroviral treatment outcome by different systems for interpreting genotypic human immunodeficiency virus type 1 drug resistance. J Infect Dis 2003;15:1934—43.
[14] Beerenwinkel N, Sing T, Lengauer T, Rahnenführer J, Roomp K, Savenkov I, et al. Computational models for the design of effective therapies against drug resistant HIV strains. Bioinformatics 2005;21(21):3943—50.
[15] Wang D, Larder B. Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks. J Infect Dis 2003;188:653—60.
[16] Beerenwinkel N, Schmidt B, Walter H, Kaiser R, Lengauer T, Hoffmann D, et al. Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. Proc Natl Acad Sci USA 2002;99(12): 8271—6.
[17] Larder BA, Kemp SD, Hertogs K. Quantitative prediction of HIV-1 phenotypic drug resistance from genotypes: the virtual phenotype (VirtualPhenotype). Antivir Ther 2000;5(Suppl. 3):49.
[18] Brun-Vezinet F, Costagliola D, Ait Khaled M, Calvez V, Clavel F, Clotet B, et al. Clinically validated genotype analysis: guiding principles and statistical concerns. Antivir Ther 2004;9(4):465—78.
[19] Bachelor LT, Winters B, Nauwelaers D, Rinehart A, McGregor M, Harrigan R, et al. Estimation of phenotypic clinical cutoffs for VirtualPhenotype through meta analyses of clinical trial and cohort data. Antivir Ther 2004;9:S154.
[20] Swanstrom R, Bosch RJ, Katzenstein D, Cheng H, Jiang H, Hellmann N, et al. Weighted phenotypic susceptibility scores are predictive of the HIV-1 RNA response in protease inhibitor-experienced HIV-1-infected subjects. J Infect Dis 2004;190:886—93.
[21] DiRienzo G, DeGruttola V. Collaborative HIV resistance-response database: sample size for detection of relationships between HIV-1 genotype and HIV-1 RNA response using a non-parametric approach. Antivir Ther 2002;7:S71.
[22] Larder BA, DeGruttola V, Hammer S, Harrigan R, Wegner S, Winslow D, et al. The international HIV resistance response database initiative: a new global collaborative approach to relating viral genotype treatment to clinical outcome. Antivir Ther 2002;7:S84.
[23] Larder B, Wang D, Revell A, Montaner J, Harrigan R, DeWolf F, et al. The development of artificial neural networks to predict virological response to combination HIV therapy. Antivir Ther 2007;12:15—24.
[24] Liaw A, Wiener M. Classification and regression by random forest. R News 2002;2—3:18—22.
[25] Breiman L. Random forests. Mach Learn 2001;45:5—32.
[26] Onkamo P, Toivonen H. A survey of data mining methods for linkage disequilibrium mapping. Hum Genom 2006;2: 336—40.
[27] Listgarten J, Damaraju S, Poulin B, Cook L, Dufour J, Driga A, et al. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. Clin Cancer Res 2004;10:2725—37.
[28] Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000;16(10):906—14.
[29] Bartlett P, Shawe-Taylor J. Generalization performance of support vector machines and other pattern classifiers. In: Schoelkopf B, Burges CJC, Smola AJ, editors. Advances in Kernel methods—support vector learning. Boston: MIT Press; 1999. p. 43—54.
[30] Larder BA, Wang D, Revell A, Lane C. Neural network model identified potentially effective drug combinations for patients failing salvage therapy. In: The 2nd IAS conference on HIV pathogenesis and treatment, Paris, France, 13—16 July; 2003 [Poster LB39].
[31] Breiman and Cutler's random forests for classification and regression. http://cran.r-project.org/web/packages/randomForest/index.html [accessed 30 November 2007].
[32] Shrestha DL, Solomatine DP. Machine learning approaches for estimation of prediction interval for the model output. Neural Networks 2006;19:225—35.
[33] Heske T. Bias/variance decompositions for likelihood-based estimators. Neural Comput 1998;10:1425—33.