Original article

Investigation of expert rule bases, logistic regression, and non-linear machine learning techniques for predicting response to antiretroviral treatment

Mattia CF Prosperi[1,2], Andre Altmann[3], Michal Rosen-Zvi[4], Ehud Aharoni[4], Gabor Borgulya[5], Fulop Bazso[5], Anders Sönnerborg[6], Eugen Schülter[7], Daniel Struck[8], Giovanni Ulivi[1], Anne-Mieke Vandamme[9], Jurgen Vercauteren[9] and Maurizio Zazzi[10] on behalf of the EuResist and Virolab study groups*

[1]Computer Science and Automation Department, Roma Tre University, Rome, Italy
[2]Informa, Rome, Italy
[3]Max Planck Institute for Informatics, Saarbrücken, Germany
[4]IBM Haifa Research Lab, Haifa, Israel
[5]KFKI Research Institute for Particle and Nuclear Physics of the Hungarian Academy of Sciences, Budapest, Hungary
[6]Karolinska Institute, Stockholm, Sweden
[7]University of Cologne, Cologne, Germany
[8]Centre de Recherche Public-Santé, Luxembourg, Luxembourg
[9]Rega Institute, Katholieke Universiteit Leuven, Leuven, Belgium
[10]University of Siena, Siena, Italy

*Corresponding author: E-mail: ahnven@yahoo.it

Background: The extreme flexibility of the HIV type-1 (HIV-1) genome makes it challenging to build the ideal antiretroviral treatment regimen. Interpretation of HIV-1 genotypic drug resistance is evolving from rule-based systems guided by expert opinion to data-driven engines developed through machine learning methods.

Methods: The aim of the study was to investigate linear and non-linear statistical learning models for classifying short-term virological outcome of antiretroviral treatment. To optimize the model, different feature selection methods were considered. Robust extra-sample error estimation and different loss functions were used to assess model performance. The results were compared with widely used rule-based genotypic interpretation systems (Stanford HIVdb, Rega and ANRS).

Results: A set of 3,143 treatment change episodes were extracted from the EuResist database. The dataset included patient demographics, treatment history and viral genotypes. A logistic regression model using high order interaction variables performed better than rule-based genotypic interpretation systems (accuracy 75.63% versus 71.74–73.89%, area under the receiver operating characteristic curve [AUC] 0.76 versus 0.68–0.70) and was equivalent to a random forest model (accuracy 76.16%, AUC 0.77). However, when rule-based genotypic interpretation systems were coupled with additional patient attributes, and the combination was provided as input to the logistic regression model, the performance increased significantly, becoming comparable to the fully data-driven methods.

Conclusions: Patient-derived supplementary features significantly improved the accuracy of the prediction of response to treatment, both with rule-based and data-driven interpretation systems. Fully data-driven models derived from large-scale data sources show promise as antiretroviral treatment decision support tools.

Introduction

Treatment of HIV type-1 (HIV-1) infection has rapidly evolved, from palliative use of single drugs to very effective strategies, employing combinations of multiple drugs inhibiting different steps of the viral life cycle. Associating ≥3 antiretroviral drugs (combination antiretroviral therapy [cART]) has led to significant decreases in HIV-1-related morbidity and mortality by reducing viral replication to undetectable levels [1,2].

Notwithstanding this success, virus eradication is still not achieved and antiretroviral treatment must be

administered continuously [3]. A complex interplay among tolerability, adherence issues and pharmacokinetics can cause suboptimal drug exposure and incomplete viral suppression, ultimately resulting in selection of HIV-1 variants with reduced susceptibility to some or most antiretroviral compounds [4,5]. The viral mutants can display different degrees of decreased susceptibility to the ongoing treatment regimen and cross-resistance to other related compounds, leading to virological rebound in subsequent treatments and accelerated disease progression. As long as drug resistance accumulates, it is crucial to choose appropriate drugs and build a salvage therapy achieving substantial and durable control of viral replication. The number of HIV-1 mutations considered to be associated with drug resistance, as indicated by the reference International AIDS Society (IAS) list [5], has increased with the expanding arsenal of available drugs. As a consequence, predicting drug susceptibility from HIV-1 genotype has become increasingly challenging over time [6].

Earlier attempts aimed at adjusting treatment were mainly based on estimates of *in vitro* resistance to individual drugs. Predicting the phenotype to single drugs from the viral genotype has been accomplished with increasing accuracy. Multiple linear regression, decision trees (DT) and support vector machine models trained on a sufficient number of genotype-phenotype pairs can explain up to 80% of phenotypic variance [7]. Genotype interpretation systems have later attempted to correlate resistance mutations with virological treatment outcome, thus translating HIV-1 genetic data into clinically relevant information [8]. Many of these systems have been developed and some have evolved into widely used online treatment decision support tools [9]. However, periodical updates are required to incorporate new knowledge as well as include novel compounds into the system. At present, these algorithms are derived from different types of data sources correlating HIV-1 genotype with *in vitro* phenotype, virological outcome and treatment history. Such data are critically revised by expert panels and processed through a number of methods [10]. Because of variations in scoring functions and evaluation procedures, discordances indeed exist among methods [11–13]. More recently, machine learning approaches have been investigated with different technologies, as artificial neural networks modelling [14], fuzzy-rule-based systems [15,16], probabilistic models focusing on viral evolutionary pathways through mutagenetic trees (MT) [7,17] and *in vivo* viral fitness landscape estimation with Bayesian network learning [17,18]. An MT model is currently implemented in the THEO web service.

Most of the commonly used algorithms predict susceptibility to individual drugs rather than to a combination regimen, except for a set of guidelines provided by Rega and for the data-driven approaches of ANN and MT. However, the widespread use of HIV-1 genotyping as an integral part of routine monitoring of infected patients is providing a unique opportunity for the compilation of large observational databases, in which the baseline HIV-1 genotype and cART can be correlated with virological and immunological follow-up data, together with an extended collection of patient and virus information. Additional data can be thus incorporated into the model to help improve the accuracy of the prediction of the response to therapy.

The EU-funded EuResist project (IST-2004-027173) is regularly gathering and updating clinical and virological data from an increasing number of HIV-1-infected patients in Italy, Germany, Luxembourg and Sweden. The data are being used to train multiple response models to antiretroviral therapy and develop a novel data-driven metamodel, to be implemented as a treatment decision support tool freely available on the web [19]. In this paper, a large training dataset derived from the EuResist database was used to explore several machine learning approaches and compare their performance with respect to those of the state of the art genotypic interpretation systems.

Methods

Dataset

The November 2007 release of the EuResist database was used to extract instances in the form of treatment change episodes (TCEs), a term coined by the HIV Resistance Response Database Initiative (RDI) [7] and later tentatively standardised by the Forum for Collaborative HIV Research [20]. The EuResist Standard Datum (ESD) definition is basically compliant with the indications given by the Forum for the short-term response, but refined to include additional variables. Each record in the minimal ESD form included a new treatment regimen, a baseline HIV-1 genotype (protease and reverse transcriptase, coded as mutations present in the IAS list) and a follow-up viral load measured at 4–12 weeks of unmodified therapy. If available, the HIV-1 RNA load obtained at 0–12 weeks before treatment start was collected, provided that no other therapies were started and stopped during this time window. When multiple baseline or follow-up data were available, the values closest to the therapy start date and to the 8-week follow-up time point were used, respectively. The treatment was labelled as successful when the follow-up viral load was undetectable or ≥2 log lower than the baseline value. Because of the inclusion of many viral load data obtained with first-generation assays, the 500-copy threshold was applied for the definition of undetectable viral load. To evaluate the performance of the models at a later follow-up time point,

a medium-term outcome was also collected as the HIV-1 RNA measurement closest to week 24 (within 20–28 weeks) of treatment.

No restrictions on therapies were considered, that is, suboptimal treatment regimens made of <3 drugs were included. Indeed, it was previously demonstrated that such TCEs improve data-driven models [14]. The following compounds approved by the Food and Drug Administration and by the European Medicines Agency were considered: the nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs) lamivudine, abacavir, zidovudine, stavudine, zalcitabine, didanosine, emtricitabine (FTC) and tenofovir disoproxil fumarate; the non-NRTIs (NNRTIs) efavirenz and nevirapine; and the protease inhibitors (PIs) amprenavir/fosamprenavir, atazanavir, indinavir, lopinavir, nelfinavir, full-dose ritonavir, boosting dose ritonavir and saquinavir. Darunavir, tipranavir and etravirine were excluded from the analysis because of an insufficient number of ESD instances. In addition, treatments including drugs without available genotypic prediction (fusion or entry inhibitors and integrase inhibitors) were excluded to avoid any potential confounding effect.

In addition to the above mentioned variables, the complete ESD included the following attributes: viral subtype as determined by a BLAST search [21] on an updated reference sequence set [22], similarity to consensus B calculated by local alignment [23,24], presence or absence of all the individual reverse transcriptase and protease mutations not included in the IAS list but present with frequency ≥3%, patient age, gender, ethnicity (Asian, African, Caucasian, Hispanic or other), route of HIV-1 infection (drug user, heterosexual, homosexual/bisexual, blood products or mother-to-child transmission), previous >12-month exposure to individual drug classes, previous >12-month exposure to individual drugs, number of previous therapy lines (any change in the combination therapy for any reason), number of drugs included in the cART, baseline $CD4^+$ T-cell count and baseline $CD4^+$ T-cell percentage (same time window as for baseline viral load). Patients with incomplete treatment history were excluded. Mutations were encoded as multi-dimensional vectors in [0,1] and were extracted from nucleotide sequences using Smith–Waterman–Gotoh local alignment [23,24] with post-alignment detection of ambiguities and repositioning of insertions or deletions in the correct coding frame. In general, missing values in numeric and nominal attributes were replaced by means and modes of the corresponding features in the data, respectively, on the basis of our previous work [19] showing the benefit derived from this approach for the multivariable logistic regression (LR) in the same setting.

To further analyse interactions among variables, the following two- and three-way higher-order interactions were considered: drug × drug, drug × drug × drug, drug × mutation, drug × previous drug class exposure, drug × previous drug class exposure and mutation × mutation. The resulting total input space size thus included >10,000 features, but was reduced to 1,466 variables by filtering out features with frequency <3%.

Statistical learning models

Among the wide range of machine learning algorithms currently available as off-the-shelf packages, the following were considered as base learners: LR [25], DT [26], random forests (RF) [27] and rule bases (RB) [28,29]. These technologies are complementary and can reveal different aspects of the modelled scenarios [30]. LR, DT and RB are understandable, as they provide a set of coefficients and significance values, a decision graph and a set of rules in natural language, respectively. RF show powerful performances and limited over-training, but are often criticized for their black-boxed nature. However, a measure of the relevance of each variable can be calculated from any RF model, contributing to interpretation of the model. LR models exploit a linear combination of variables, whereas DT, RB and RF can produce non-linear functions. LR can exploit non-linear functions by encoding higher-order variable interactions.

Feature selection is the problem of finding a (minimal) subset of attributes that leads to maximal performances, given a loss function, a learning model (or an ensemble of them) and a procedure to test a statistical hypothesis [31]. Feature selection techniques can be grouped into three methodologies: filter, wrapper and embedded [32]. In this study, the first and last methods were applied to the machine learning models. More precisely, filter methods based on univariable analysis (χ^2 and rank-sum tests) or correlation-based feature selection [33] for LR, embedded selection based on information gain for DT [26], Akaike's information criterion (AIC) [34] and ridge shrinkage for LR [25], repeated incremental pruning to produce error reduction [28] and ripple-down rule learner [29] for RB.

Accuracy (the proportion of correct classifications) and area under the receiver operating characteristic curve (AUC) [35] were adopted to evaluate the model goodness of fit. The AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one, whereas a receiver operating characteristic (ROC) plot provides an easy to understand evaluation of true-positives (sensitivity) versus false-positives (1-specificity) tradeoffs. Robust extra-sample error estimation was obtained by 10-fold cross-validation (CV) [30]. Multiple independent runs of 10-fold CV were executed, obtaining a normal distribution for loss functions. In this way, two different models can be evaluated and compared (each against the other or each against a null

hypothesis) using a Student's *t* statistic. However, the usage of a naive Student's *t*-test can be biased because of the sample overlap. Thus, the adjusted Student's *t*-test proposed by Nadeau and Bengio [36] for multiple CV runs was applied. Specifically, 90% of the data was used to execute multiple 10-fold CV, whereas 10% was kept out as an independent test set. In addition, learning models were also tested using data from different clinics (Italian versus non-Italian centres). A performance comparison on the test set was done by bootstrap analysis [30].

Reference rules-based algorithms

The models were compared with the following genotype interpretation algorithms: Stanford HIVdb (version 4.3.6), Rega (version 7.1.1) and ANRS (version 16). The algorithms were evaluated on the whole ESD dataset, except for zalcitabine-containing records, as these interpretation systems no longer consider zalcitabine. Drug scores were 0 for resistant, 0.5 for intermediate-resistant, and 1 for susceptible rating in ANRS. For HIVdb, continuous scores were calculated using the scoring matrices of the website. Calculation of Rega scores was the same for ANRS, with the following changes as indicated by the algorithm developers: NNRTI were scored 0.25 for intermediate and ritonavir-boosted PI were scored 0.75 and 1.5 for intermediate and susceptible, respectively. The arithmetic sum of the scores gained for the individual drugs included in the regimen was taken as an overall indicator for efficacy of the specific cART, an approach usually referred to as the genotypic susceptibility score (GSS). The Rega algorithm also suggests different target scores depending on the patient's treatment history, but in this analysis the thresholds were optimized as a part of the modelling process. The GSS of interpretation algorithms were analysed as predictors of the virological response to treatment using univariable or multivariable analysis. Raw GSS and adjustment attributes (the same set as defined above for the machine learning models) were fed to LR, whereas parameter optimization and model comparison were made via multiple 10-fold CV and adjusted Student's *t*-test.

Statistical software

All the analyses were carried out using R open source software for statistical computing [37] and the Weka data mining suite [38].

Results

Descriptive statistics

A total of 3,143 ESD instances were obtained from the EuResist database: 2,831 records were kept as a training set and 312 as a test set. The 8-week virological outcome was successful in 68.21% of cases. The prevalence of successful therapies was kept the same in the training and the test set. Missing information was 4% for baseline HIV-1 RNA load, 9% for baseline CD4$^+$ T-cell counts, and 18% for route of HIV-1 infection. Table 1 summarizes the baseline characteristics of the dataset and Figure 1 shows the frequency of use of individual drugs and drug combinations in the TCEs.

Table 1. Summary of the EuResist training dataset (n=2,831)

Parameter	Value
Mean patient age, years (±SD)	42 (13)
Cases from male patients, %	70
Cases from drug users, %	27
Cases from homosexual men, %	32
Cases from heterosexual patients, %	38
Cases from Caucasian patients, %	72
Cases from African patients, %	22
Cases with previous exposure to NRTI, %	74
Cases with previous exposure to NNRTI, %	41
Cases with previous exposure to PI, %	58
Median log baseline HIV-1 RNA load, copies/ml (IQR)	4.4 (3.8–5.0)
Median baseline CD4$^+$ T-cell percentage (IQR)	16.5 (11–20)
Median baseline CD4$^+$ T-cell count, cells/mm^3 (IQR)	255 (137-397)
Median previous treatment lines (IQR)	3 (1–6)
Median drugs included in the cART (IQR)	3 (1–4)
Subtype B sequences, %	83
Subtype C sequences, %	3
Subtype 02_AG sequences, %	2.5
Subtype F1 sequences, %	2.3
Median baseline IAS NRTI mutations (IQR)	1 (0–3)
Median baseline IAS NNRTI mutations (IQR)	0 (0–1)
Median baseline IAS PI mutations (IQR)	3 (2–5)

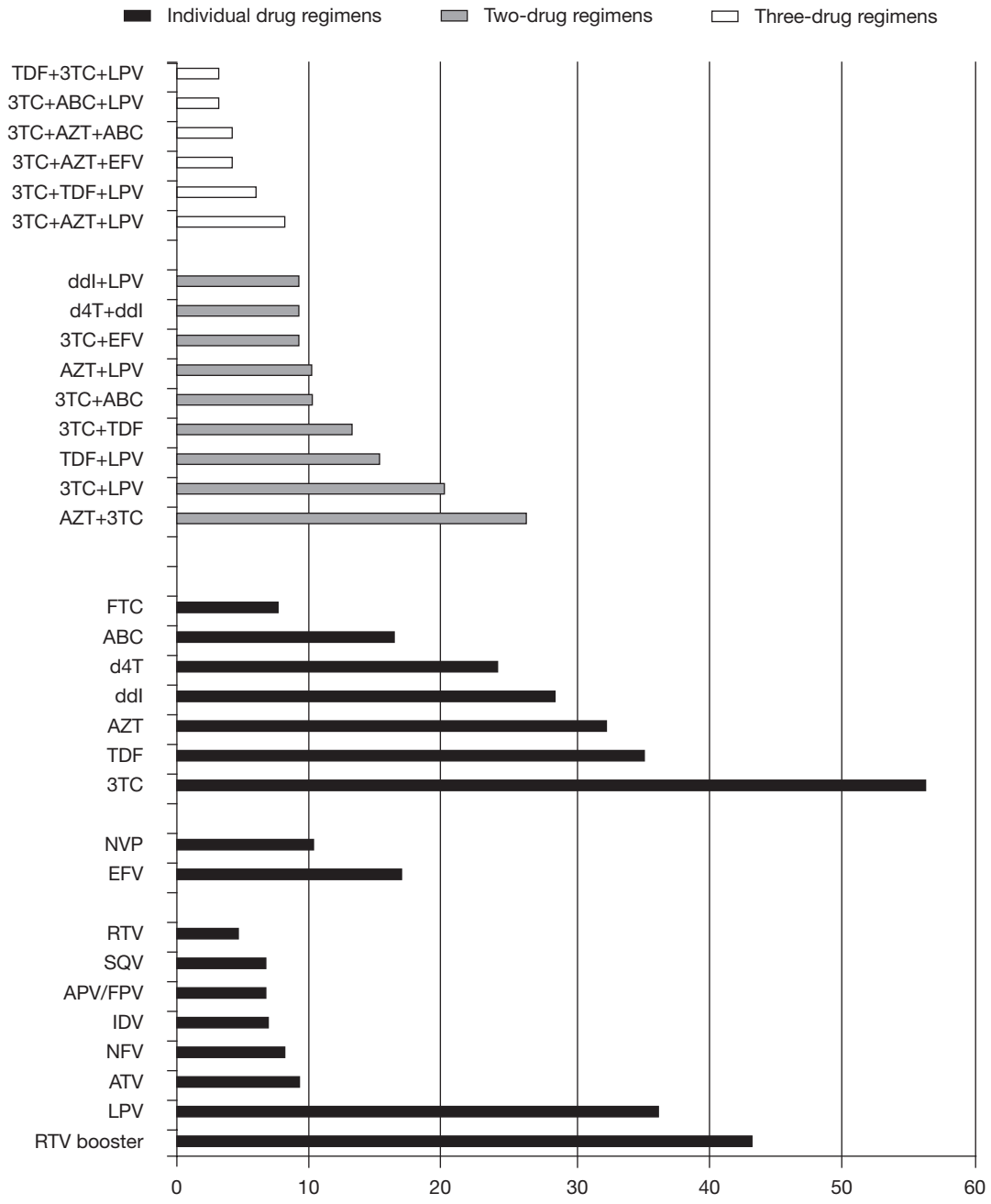
cART, combination antiretroviral therapy; HIV-1, HIV type-1; IAS, International AIDS Society; IQR, interquartile range; NNRTI, non-nucleoside/nucleotide reverse transcriptase inhibitor; NRTI, nucleoside/nucleotide reverse transcriptase inhibitor; PI, protease inhibitor.

Model assessment

To evaluate the contribution of the different input variables to the performance of the model, a set of different input spaces were tested using LR; specifically, a null reference model with no features (predicting simply the majority class; Model 1); only cART (Model 2); cART and IAS mutations (Model 3); cART, IAS mutations, baseline HIV-1 RNA load and CD4$^+$ T-cell counts (Model 4); complete ESD without higher-order interactions (Model 5); and complete ESD including higher-order interactions (Model 6).

Because the null model depends solely on the prevalence of the majority class, its accuracy was 68.21% (±SD 0.03) with an AUC of 0.5. Table 2 shows the results for the LR models. The simplest model, Model 2, improved significantly with respect to the null model only in terms of AUC. Instead, Model 3 improved both

Figure 1. Percentage of use of individual drugs and most frequent two- or three-drug cART regimens in the 2,831 EuResist Standard Datum instances of the training dataset



ABC, abacavir; APV/FPV, amprenavir/fosamprenavir; ATV, atazanavir; AZT, zidovudine; ddI, didanosine; d4T, stavudine; EFV, efavirenz; FTC, emtricitabine; IDV, indinavir; LPV, lopinavir; NVP, nevirapine; RTV, ritonavir; SQV, saquinavir; TDF, tenofovir disoproxil fumarate; 3TC, lamivudine.

accuracy and AUC. Model 4 significantly increased AUC only. The usage of complete ESD, that is, Model 5, significantly improved both accuracy and AUC with respect to Models 1 to 4 and was the highest performance bound achieved by a linear combination of variables. The usage of higher-order interactions, that is, Model 6, significantly increased the AUC (accuracy 75.63% ±2.32 and AUC 0.76 ±0.03 in multiple CV). Accuracy and AUC on the test set were comparable with the multiple CV average values obtained in the training set for all models.

When testing non-linear techniques, the complete ESD was used, excluding higher-order interactions (because the models are expected to account for non-

linear interactions by definition). The optimal number of trees in RF was selected using a parameter optimization on a bootstrap sample of the data, in the discrete space {10, 30, 60, 100}, maximizing accuracy. Table 3 summarizes the validation results. RF reached an accuracy of 76.16% ±2.22 and an AUC of 0.77 ±0.03 in multiple CV, equivalent to the best LR model (that is, Model 6) without requiring the cumbersome higher-order interaction coding. Finally, rule finding algorithms and DT achieved a significantly greater accuracy with respect to the null model and LR Model 2, but the AUC was always lower than for RF and for LR Models 3 to 6.

When analysing test set error distributions using 100 bootstrap samples, we found that RF and LR Model 6 were superior both in accuracy and AUC with respect to the LR Model 5 (Students *t*-test $P<0.0001$). RF and LR Model 6 were not significantly different (Student's *t*-test $P=0.5$).

Comparison of ANRS, Rega and Stanford HIVdb interpretation algorithms

The accuracy and AUC were 71.74% ±1.72 and 0.68 ±0.03 for Stanford HIVdb, 73.89% ±1.81 and 0.69 ±0.03 for Rega, 72.96% ±1.78 and 0.69 ±0.03 for ANRS, respectively. All the scores were significantly associated with the outcome response in univariable LR (all $P<0.0001$). The only statistically significant difference among the systems was a greater accuracy of Rega with respect to Stanford HIVdb ($P<0.05$). Of note, when applying a combined LR model on the three scores together (that is, using all three algorithms), results improved to 73.90% ±2.01 accuracy and 0.70 ±0.03 AUC. The AUC under the combined model was significantly higher than any single score estimate ($P<0.05$). However, the LR Model 6 was superior to all algorithms in accuracy or AUC. Figure 2 plots the true-positive rate (sensitivity) against the false-positive rate (1-specificity) for the interpretation algorithms and LR Model 6 with higher-order interactions showing ROC curves for CV predictions.

In order to resemble the setting of the complete ESD, the ANRS, Rega or Stanford HIVdb scores were then included in an LR model adjusted for age, gender, mode of HIV-1 transmission, ethnicity, subtype, consensus B similarity, previous usage of NRTIs, NNRTIs or PIs, number of drugs in cART, number of previous treatment lines, presence of boosted PI, baseline viral load, baseline CD4$^+$ T-cell count and

Table 2. Multiple cross-validation and test set evaluation results for different logistic regression models

Input space	Feature selection	Multiple 10-fold CV[a]		Test set evaluation	
		Accuracy, % (±SD)	AUC (±SD)	Accuracy, %	AUC
cART	AIC	68.49 (1.79)[b]	0.643 (0.035)[b]	68.27	0.619
IAS mutations + cART	AIC	74.08 (2.22)	0.725 (0.037)[b]	74.36	0.713
IAS mutations + cART + baseline markers	AIC	74.66 (2.18)	0.741 (0.033)[b]	75.64	0.721
Full EuResist Standard Datum	CFS + AIC	75.43 (2.30)	0.746 (0.034)[b]	75.64	0.742
Full EuResist Standard Datum + higher order interactions	Univariable filter + AIC	75.63 (2.32)	0.761 (0.032)	75.32	0.773

[a]From 30 independent runs. [b]Significantly worse (*P*-value <0.05) than the best model under adjusted Student's *t*-test. AIC, Akaike's information criterion; AUC, area under the receiver operating characteristic curve; cART, combination antiretroviral therapy; CFS, correlation-based feature selection; CV, cross-validation; IAS, International AIDS Society.

Table 3. Multiple cross-validation and test set evaluation results for different non-linear classifiers

Input space	Method	Feature selection	Multiple 10-fold CV[a]		Test set evaluation	
			Accuracy, % (±SD)	AUC (±SD)	Accuracy, %	AUC
Full EuResist Standard Datum	Decision tree	Recursive partitioning	73.24 (2.14)[b]	0.695 (0.040)[b]	73.40	0.638
	Random forest	60 Random trees with 10 features	76.16 (2.22)	0.767 (0.035)	75.64	0.778
	Rules	RIPPER	73.98 (2.33)	0.664 (0.032)[b]	70.19	0.646
	Rules	RIDOR	72.80 (1.65)[b]	0.597 (0.029)[b]	69.87	0.552

[a]From 30 independent runs. [b]Significantly worse (*P*-value <0.05) than the best model under adjusted Student's *t*-test. AUC, area under the receiver operating characteristic curve; CV, cross-validation; RIDOR, ripple-down rule learner; RIPPER, repeated incremental pruning to produce error reduction.

baseline CD4+ T-cell percentage. Using stepwise AIC selection, AUC and accuracy increased and indeed became comparable to the LR Model 6. The Stanford HIVdb algorithm achieved 76.19% ±1.75 accuracy and 0.76 ±0.03 AUC, Rega achieved 75.57% ±1.83 accuracy and 0.75 ±0.03 AUC and ANRS achieved 75.70% ±1.67 accuracy and 0.75 ±0.03 AUC. All the scores remained significantly associated with outcome ($P<0.0001$). No statistical difference was found when comparing the adjusted models with each other. Finally, upon applying a combined LR model on the three scores together plus the adjustments, the overall performances reached 75.93% ±1.73 accuracy and 0.76 ±0.03 AUC, without significant improvement with respect to any of the single algorithms.

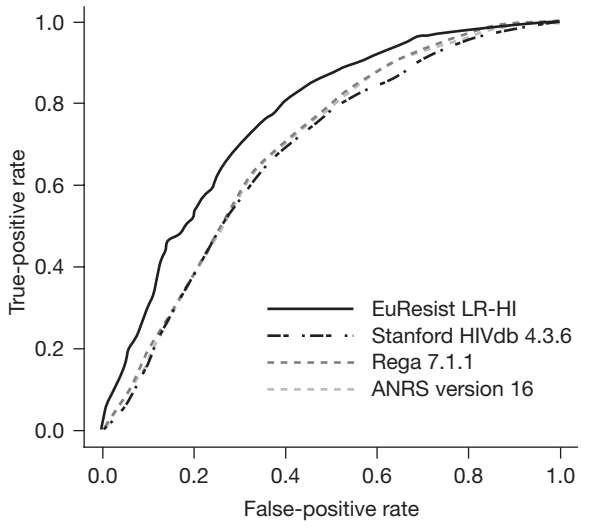
Additional performance evaluation using data from different clinics

The data-driven LR Model 6 was tested comparatively with two different EuResist datasets, comprising data coming from Italian versus non-Italian clinics (n=1,445 and 1,698 respectively). Non-Italian TCEs came from Germany (60%), Sweden (27%) and Luxembourg (13%). A model was trained on the first set, also executing 10-fold CV, and then the same model was tested on the second dataset. The 10-fold CV yielded an accuracy of 75.84% and AUC of 0.75, whereas the test set evaluation achieved 73.73% accuracy and 0.73 AUC. Thus, training on <50% of the data available yielded a robust model predicting new instances from different geographical sources with a performance decrease of only 2% with respect to both measures.

Prediction of medium-term response

A query on the EuResist database for a 24-week virological outcome returned 2,778 ESD instances. Of these, 2,579 instances were used for training and 199 instances for testing; the latter originated from the test set defined for the short-term analysis. In total, 1,837 out of 2,579 (71.2%) were also comprehensive of the 8-week response, whereas for 742 (28.8%) instances only information of the 24-week outcome was available. The cross-tabulation between short- and medium-term outcomes indicated 406 (22.1%) failures at both follow-up time points, 1,134 (61.7%) successes at both follow-up time points, 155 (8.4%) failures at 8 weeks and successes at 24 weeks, and 142 (7.7%) successes at 8 weeks and failures at 24 weeks. The short-term outcome was significantly associated to the medium-term outcome (odds ratio 20.9; $P<0.0001$ under univariable LR). The proportion of successful treatments was comparable in the complete medium-term dataset (69.9%), the short-term and medium-term outcome containing subset (69.7%) and the medium-term-only subset (68.8%).

Figure 2. ROC curves for EuResist logistic regression model with higher-order interactions and Rega, Stanford and ANRS algorithms[a]



Data points were obtained from cross-validated predictions. [a]Overall scores from univariable models. LR-HI, logistic regression model with higher-order interactions; ROC, receiver operating characteristic.

The LR Model 6 was trained on the 24-week ESD and evaluated for medium-term prediction of virological outcome. A 10-fold CV performed on the complete set of 2,778 instances yielded an accuracy of 79.3% and AUC of 0.834. In a second analysis, the performance of the short-term prediction LR Model 6 was tested with respect to the medium-term outcome. A 10-fold CV training was carried out on the instances for which the short-term outcome was available (that is, train plus test, n=2,036) and the short-term prediction was tested separately against the medium-term label and against the set of medium-term-only instances. The prediction of the medium-term label in the short and medium-term paired dataset yielded 80.3% accuracy and 0.82 AUC. The prediction for the medium-term-only test set yielded 78.4% accuracy and 0.78 AUC.

Discussion

In this study, patient and HIV-1 data were modelled to derive a therapy optimization model for the prediction of short-term virological outcome. LR models were compared against rule-based genotypic interpretation systems and against several non-linear techniques. Different variable space encodings were investigated: from a basic input space made of genotype and cART information, we explored the addition of supplementary patient data and higher-order

variable interactions. Extra-sample error estimation was assessed via multiple CV, and model comparison was carried out using adjusted Student's *t*-test. It was shown that baseline markers (HIV-1 RNA load and $CD4^+$ T-cell counts), additional mutations, patient demographics and information from past treatments did improve performances over the basic (genotype and cART) model. LR with higher-order interactions and random forest modelling provided the best performances in terms of accuracy and AUC, supporting the hypothesis that non-linear relationships among variables play a significant role. Conversely, automated rule finding algorithms and single DT models did not prove to be a satisfactory model framework in terms of performance, especially for AUC.

Among the most widely used interpretation algorithms, Stanford HIVdb, ANRS and Rega performed similarly, yet they were inferior to the data-driven LR with higher-order interactions or RF models trained on the EuResist database. However, when the scores derived from these interpretation algorithms were coupled with baseline markers, patient demographics and treatment history covariates in a LR model, the prediction performances were comparable to the fully data-driven models. This indicates that the benefit derived from the usage of popular rule-based systems in clinical practice can be further increased by considering additional and usually available attributes. Accordingly, several studies have shown the prognostic value of baseline parameters other than the HIV-1 genotype [39–44].

Data-driven approaches, especially non-linear models, might account for dynamics that are not exploited by current rule-based systems (for example, drug interactions). However, variable interpretation becomes more challenging, although methods and measures for assessing variable importance for non-linear techniques are available [45]. In rule-based systems, over time, the genotypic scores have been tuned on several specific scenarios (for example, studies focusing on single-drug clinical outcomes [46]); hence they might be capable of including contributions of rare mutations to drug resistance or susceptibility. The contribution of rare mutations might not be significant from a statistical perspective, but it can be validated via tailored experiments (for example, executing *in vitro* cultures on viral clones with selected mutations). Such mutations can be added easily to rule-based interpretation systems, whereas they usually would not be selected as significant features via data-driven modelling, because of their infrequent appearance.

Further improvements for the prediction systems could probably come from incorporating supplementary information related to the host, such as human leukocyte antigen pattern or other genetic traits suggested to play a role in disease progression and treatment response [47,48]. Unfortunately, such data will hardly be available for a large number of patients in the near future. Thus, the current performance might be close to an upper bound, although meta-models (that is, combinations of different machine learners or algorithms) could be worth testing.

It must be noted that the state of the art interpretation algorithms have not been designed for predicting response at any specific follow-up time point, whereas the LR or RF engines described here were specifically trained on 8-week response instances; however ongoing analyses indicate that the LR engine trained on the 8-week outcome predicts 8-week and 24-week outcomes with comparable efficiency. The follow-up time for definition of success or failure can be set differently depending on the aim of the prediction system. Baseline genotype shows the most prevalent viral species present at the beginning of therapy. The initial effect of therapy is basically directed against that virus population. However, in most, if not all, pretreated patients, other viral species are archived at undetectable levels and can re-emerge at later time points. These hidden species can have a major effect on medium-term response to treatment, but usually cannot be computed as an input variable for training a predictive model (although treatment history or past genotypes might be a surrogate source of information). In addition, adherence-related confounding factors are also time-dependent. Based on these considerations, the Forum for HIV Research has indeed proposed an 8-week and a 24-week TCE definition. When developing the system described in our work, a short-term follow-up was preferred to avoid time-related confounding factors, with a plan to extend the system towards 24-week and 48-week prediction once a reasonably good performance was achieved with the short-term 8-week model.

Treatment success is also not consistently defined across different studies and is subject to changes over time as along as antiretroviral therapy options evolve. In principle, the choice of a binary encoding based on undetectable viral load is in agreement with the current view that any detectable viraemia is indicative of treatment failure. Because of the limited follow-up time in our short-term model, a decrease >2 log in viral load was also labelled as a success, allowing us to consider cases where a high baseline HIV-1 RNA load would decrease to undetectable levels at later time points. By contrast, the ANN system developed by the RDI was shown to perform well at predicting the viral load change at 4–40 weeks following the treatment switch [14]. Although our model and the RDI method cannot be compared, both systems are data-driven and interestingly showed the benefit of additional patient-derived features. At present, a drawback of the data sources available for training the models is the inclusion of viral load measurements obtained with laboratory assays with different sensitivity thresholds.

For example, we found that our models trained using the truncation at 500 copies/ml predicted success at <50 copies/ml HIV-1 RNA significantly worse than at <500 copies/ml HIV-1 RNA (see Additional file). For the future we foresee setting the VL response threshold at 50 HIV-1 RNA copies/ml, applying additional statistical methods [46] to account for truncated data, obtained from first-generation assays. In addition, expanding the data sources over time would probably be advised for considering only viral load measurements generated by ultra-sensitive tests.

Another potential limitation of this study is the use of a dataset not reflecting the current antiretroviral treatment options but rather an historical archive of therapies. Regimens including the latest generation boosted PI, such as tipranavir and darunavir, were in fact not available in a sufficiently large number to train the models. Similarly, drugs belonging to the newly licensed classes, such as raltegravir and maraviroc, were not included. Shortage of data for novel items is an inherent weakness of all the data-driven systems that require a complex and time-consuming data collection process. Nevertheless, the EuResist project has gained growing attention from related research projects and companies and this can contribute to speeding up the future uptake of data for novel drugs. For example, data from some clinics of the Virolab consortium have recently been uploaded. Large comparative analysis of multiple rules-based algorithms has also suffered from a lack of recent drug data. Yet, they have undoubtedly contributed to the advancement of knowledge on HIV-1 resistance. By contrast, algorithms derived from small datasets have been typically subject to continuous revisions, sometimes confusing HIV practitioners. This temporal gap between the availability of drugs in the market and their respective TCE-like data are expected to be filled with continuous feeding of large multicentric databases by international initiatives. The models must clearly be retrained with every major update of the dataset.

The LR model with higher-order interactions presented here is one of three prediction engines implemented by the EuResist consortium; the other two models are based on Bayesian networks and MT. The combination of the three models has recently been proven to be successful in further improving prediction performances [19,49]. The system has been operative as a public web service since August 2008 [50]. Specifically, the system runs using an input set in the form of a ESD instance (with the baseline genotype as mandatory information and, optionally, baseline HIV-1 RNA, $CD4^+$ T-cell counts, mode of HIV-1 transmission and treatment history) and will give a rank of suitable cARTs with a corresponding probability of short-term virological success and confidence intervals.

Acknowledgements

MCFP was involved in manuscript writing, experimental design, machine learning model execution and model validation. AA provided model validation and additional performance tests. MR-Z, GU and GB contributed to theoretical aspects of machine learning procedures and validation, and manuscript revision. EA was involved in data extraction and formatting after study design. FB provided support for feature selection and variable importance measures. AS and MZ contributed to study design and biological/medical expertise, Swedish and Italian HIV cohort data management and data transfer. ES was involved in data cleansing and quality checks, German HIV cohort data management and data transfer. DS was involved in Luxembourg HIV cohort data management and data transfer. A-MV and JV provided supervision for rule-based algorithms, execution of Rega rules and Virolab HIV data cohort transfer.

Francesca Incardona provided coordination of EuResist study group and public relations with related EU-funded studies and Yardena Peres provided EuResist integrated data base design, maintenance and update.

This work was supported in part by the EuResist GEIE (former IST-2004-027173), the FWO (G.0611.09) and the Virolab (EU IST STREP 027446) projects, and the IAP P6/41.

Disclosure statement

The authors declare no competing interests.

Additional file

The additional file 'Supplementary material' can be accessed at www.intmedpress.com

References

1. Mocroft A, Vella S, Benfield TL, *et al*. Changing patterns of mortality across Europe in patients infected with HIV-1. *Lancet* 1998; **352:**1725–1730
2. Richman DD. HIV chemotherapy. *Nature* 2001; **410:**995–1001.
3. Geeraert L, Kraus G, Pomerantz RJ. Hide-and-seek: the challenge of viral persistence in HIV-1 infection. *Annu Rev Med* 2008; **59:**487–501.
4. Zaccarelli M, Tozzi V, Lorenzini P, *et al*. Multiple drug class-wide resistance associated with poorer survival after treatment failure in a cohort of HIV-infected patients. *AIDS* 2005; **19:**1081–1089.
5. Johnson VA, Brun-Vézinet F, Clotet B, *et al*. Update of the drug resistance mutations in HIV-1: 2007. *Top HIV Med* 2007; **15:**119–125.
6. Vercauteren J, Vandamme AM. Algorithms for the interpretation of HIV-1 genotypic drug resistance information. *Antiviral Res* 2006; **71:**335–342.
7. Beerenwinkel N, Daumer M, Oette M, *et al*. Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res* 2003; **31:**3850–3855.

8. Van Laethem K, Vandamme AM. Interpreting resistance data for HIV-1 therapy management - know the limitations. *AIDS Rev* 2006; **8:**37–43.
9. Liu TF, Shafer RW. Web resources for HIV type 1 genotypic-resistance test interpretation. *Clin Infect Dis* 2006; **42:**1608–1618.
10. Brun-Vézinet F, Costagliola D, Khaled MA, *et al*. Clinically validated genotype analysis: guiding principles and statistical concerns. *Antivir Ther* 2004; **9:**465–478.
11. Stürmer M, Doerr HW, Staszewski S, Preiser W. Comparison of nine resistance interpretation systems for HIV-1 genotyping. *Antivir Ther* 2003; **8:**239–244.
12. Ravela J, Betts BJ, Brun-Vézinet F, *et al*. HIV-1 protease and reverse transcriptase mutation patterns responsible for discordances between genotypic drug resistance interpretation algorithms. *J Acquir Immune Defic Syndr* 2003; **33:**8–14.
13. Snoeck J, Kantor R, Vandamme AM, *et al*. Discordances between interpretation algorithms for genotypic resistance to protease and reverse transcriptase inhibitors of human immunodeficiency virus are subtype dependent. *Antimicrob Agents Chemother* 2006; **50:**694–701.
14. Larder B, Wang D, Revell A, *et al*. The development of artificial neural networks to predict virological response to combination HIV therapy. *Antivir Ther* 2007; **12:**15–24.
15. De Luca A, Ulivi G, Vendittelli M, *et al*. Construction, training and clinical validation of an interpretation system for genotypic HIV-1 drug resistance based on fuzzy rules revised by virological outcomes. *Antivir Ther* 2004; **9:**583–593.
16. Prosperi M, Ulivi G. Evolutionary fuzzy modelling for drug resistant HIV-1 treatment optimisation. In Abraham A, Grosan C, Pedrycz W (Editors). *Engineering evolutionary intelligent systems: studies in computational intelligence*. Heidelberg: Springer Berlin 2008; pp. 251–287.
17. Altmann A, Beerenwinkel N, Sing T, *et al*. Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. *Antivir Ther* 2007; **12:**169–178.
18. Deforche K, Camacho R, Van Laethem K, *et al*. Estimation of an *in vivo* fitness landscape experienced by HIV-1 under drug selective pressure useful for prediction of drug resistance evolution during treatment. *Bioinformatics* 2008; **24:**34–41.
19. Rosen-Zvi M, Altmann A, Prosperi M, *et al*. Selecting anti-HIV therapies based on a variety of genomic and clinical factors. *Bioinformatics* 2008; **24:**i399–i406.
20. Cozzi-Lepri A. Initiatives for developing and comparing genotype interpretation systems: external validation of existing rule-based interpretation systems for abacavir against virological response. *HIV Med* 2008; **9:**27–40.
21. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990; **215:**403–410.
22. Leitner T, Korber B, Daniels M, Calef C, Foley B. *HIV-1 subtype and circulating recombinant form (CRF) reference sequences, 2005*. Los Alamos National Laboratory, Los Alamos, NM, USA. (Updated 9 September 2007. Accessed 8 May 2009.) Available from http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/RefSeqs2005/RefSeqs05.html
23. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981; **147:**195–197.
24. Gotoh O. An improved algorithm for matching biological sequences. *J Mol Biol* 1982; **162:**705–708.
25. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. *Appl Stat* 1992; **41:**191–201.
26. Breiman L, Friedman JH, Olshen RA, *et al*. *Classification and regression trees*. Belmont, CA: Wadsworth International Group 1984.
27. Breiman L. Random Forests. *Machine Learning* 2001; **45:**5–32.
28. Cohen WW. Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning*. San Francisco: Morgan Kaufmann 1995; pp. 115–123.
29. Gaines BR, Compton P. Induction of ripple-down rules applied to modeling large databases. *Journal of Intelligent Information Systems* 1995; **5:**211–228.
30. Hastie T, Tibshirani H, Friedman J. *The elements of statistical learning*. New York: Springer 2001.
31. Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V. Feature selection for SVMs. *Adv Neural Inf Process Syst* 2000; **13:**668–674.
32. Kohavi R. Wrappers for feature subset selection. *Artificial Intelligence* 1997; **97:**273–324
33. Hall MA. *Correlation-based feature selection for machine learning* [dissertation]. Hamilton, New Zealand: Waikato University 1998.
34. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr* 1974; **19:**716–723.
35. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* 2006; **27:**861–874.
36. Nadeau C, Bengio Y. Inference for the generalisation error. *Machine Learning* 2003; **52:**239–281.
37. R Development Core Team. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing 2008.
38. Witten IH, Frank E. *Data mining: practical machine learning tools and techniques*. 2nd Ed. San Francisco: Morgan Kaufmann 2005.
39. Mellors JW, Rinaldo CR, Jr, Gupta P, *et al*. Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science* 1996; **272:**1167–1170.
40. Kelley CF, Barbour JD, Hecht FM. The relation between symptoms, viral load, and viral load set point in primary HIV infection. *J Acquir Immune Defic Syndr* 2007; **45:**445–448.
41. De Milito A, Titanji K, Zazzi M. Surrogate markers as a guide to evaluate response to antiretroviral therapy. *Curr Med Chem* 2003; **10:**349–365.
42. Dragsted UB, Mocroft A, Vella S, *et al*. Predictors of immunological failure after initial response to highly active antiretroviral therapy in HIV-1-infected adults: a EuroSIDA study. *J Infect Dis* 2004; **190:**148–155.
43. Hoen B, Cooper DA, Lampe FC, *et al*. Predictors of virological outcome and safety in primary HIV type 1-infected patients initiating quadruple antiretroviral therapy: QUEST GW PROB3005. *Clin Infect Dis* 2007; **45:**381–390.
44. Hulgan T, Shepherd BE, Raffanti SP, *et al*. Absolute count and percentage of $CD4^+$ lymphocytes are independent predictors of disease progression in HIV-infected persons initiating highly active antiretroviral therapy. *J Infect Dis* 2007; **195:**425–431.
45. Rogers J, Gunn S. Identifying feature relevance using a random forest. In Saunders C, Grobelnik M, Gunn S, Shawe-Taylor J (Editors). *Lecture notes in computer science: theoretical computer science and general issues*. Vol. 3940. *Subspace, latent structure and feature selection*. Heidelberg: Springer 2006.
46. Breen R. Regression models: censored, sample selected or truncated data. In Dickens G (Editor). *Quantitative applications in the social sciences*. Vol. 111. Thousand Oaks, CA: SAGE Publications, Inc. 1996.
47. Fellay J, Shianna KV, Ge D, *et al*. A whole-genome association study of major determinants for host control of HIV-1. *Science* 2007; **317:**944–947.
48. Haas DW. Human genetic variability and HIV treatment response. *Curr HIV/AIDS Rep* 2006; **3:**53–58.
49. Altmann A, Rosen-Zvi M, Prosperi M, *et al*. Comparison of classifier fusion methods for predicting response to anti HIV-1 therapy. *PLoS One* 2008; **3:**e3470.
50. The EuResist Network. *EuResist prediction engine*. (Accessed 8 May 2009.) Available from http://engine.euresist.org

Accepted for publication 5 January 2009