Prediction of response to antiretroviral therapy by human experts and by the EuResist data-driven expert system (the EVE study)

M Zazzi,1 R Kaiser,2 A Söönerborg,3 D Struck,4 A Altmann,5 M Prosperi,6 M Rosen-Zvi,7 A Petroczi,8 Y Peres,9 E Schüller,2 CA Boucher,10 F Brun-Vezinet,11 PR Harrigan,12 L Morris,13 M Obermeier,14 C-F Perno,15 P Phanuphak,16 D Pillay,17 RW Shafer,18 A-M Vandamme,19 K van Laethem,19 AMJ Wensing,20 T Lengauer5 and F Incardona21

1Department of Molecular Biology, University of Siena, Siena, Italy, 2Institute of Virology, University of Cologne, Cologne, Germany, 3Department of Infectious Diseases, Karolinska Institutet, Stockholm, Sweden, 4CRP-Santé, Laboratory of Retrovirology, Luxembourg, Luxembourg, 5Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Saarbrücken, Germany, 6Clinic of Infectious Diseases, Catholic University of Rome, Rome, Italy, 7Machine Learning and Data Mining group, IBM Research Labs, Haifa, Israel, 8School of Life Sciences, Kingston University, Kingston upon Thames, UK, 9Knowledge Management Group, IBM Research Labs, Haifa, Israel, 10Department of Virology, Erasmus Medical Center, Erasmus University, Rotterdam, The Netherlands, 11Laboratoire de Virologie, Hôpital Bichat Claude Bernard, Paris, France, 12BC Centre for Excellence in HIV/AIDS, University of British Columbia, Vancouver, Canada, 13AIDS Virus Research Unit, National Institute for Communicable Diseases, Johannesburg, South Africa, 14Department of Virology, Ludwig-Maximilians-University Munich, Munich, Germany, 15Department of Experimental Medicine, University of Rome Tor Vergata, Rome, Italy, 16HIV-NAT/Thai Red Cross AIDS Research Centre, Bangkok, Thailand, 17Department of Infection and Immunity, University College London, London, UK, 18Division of Infectious Diseases, Department of Medicine, Stanford University, Stanford, CA, USA, 19Rega Institute for Medical Research, Katholieke Universiteit Leuven, Leuven, Belgium, 20Department of Medical Microbiology, University Medical Center Utrecht, Utrecht, The Netherlands and 21R&D, Informa S.R.L., Rome, Italy

Objectives
The EuResist expert system is a novel data-driven online system for computing the probability of 8-week success for any given pair of HIV-1 genotype and combination antiretroviral therapy regimen plus optional patient information. The objective of this study was to compare the EuResist system vs. human experts (EVE) for the ability to predict response to treatment.

Methods
The EuResist system was compared with 10 HIV-1 drug resistance experts for the ability to predict 8-week response to 25 treatment cases derived from the EuResist database validation data set. All current and past patient data were made available to simulate clinical practice. The experts were asked to provide a qualitative and quantitative estimate of the probability of treatment success.

Results
There were 15 treatment successes and 10 treatment failures. In the classification task, the number of mislabelled cases was six for EuResist and 6–13 for the human experts [mean ± standard deviation (SD) 9.1 ± 1.9]. The accuracy of EuResist was higher than the average for the experts (0.76 vs. 0.64, respectively). The quantitative estimates computed by EuResist were significantly correlated (Pearson $r = 0.695$, $P < 0.0001$) with the mean quantitative estimates provided by the experts. However, the agreement among experts was only moderate (for the classification task, inter-rater $κ = 0.355$; for the quantitative estimation, mean ± SD coefficient of variation = 55.9 ± 22.4%).

Conclusions
With this limited data set, the EuResist engine performed comparably to or better than human experts. The system warrants further investigation as a treatment-decision support tool in clinical practice.

Keywords: antiretroviral therapy, drug resistance, genotype, HIV type 1, prediction systems

Accepted 8 June 2010

Correspondence: Dr Maurizio Zazzi, Section of Microbiology, Department of Molecular Biology, University of Siena, Policlinico S. Maria alle Scotte, Viale Bracci 16, I-53100 Siena, Italy. Tel: +39 0577 233863; fax: +39 0577 233870; e-mail: zazzi@unisi.it

DOI: 10.1111/j.1468-1293.2010.00871.x
Introduction

Monitoring the development and evolution of antiretroviral drug resistance is an integral part of the clinical management of HIV type 1 (HIV-1)-infected patients [1]. Although novel classes of anti-HIV-1 compounds have been made available recently, most of the treatment regimens are still based on combinations of nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs) and protease inhibitors (PIs). These drugs have been used for many years and there is extensive information on the correlation between mutations in the HIV-1 pol gene and changes in susceptibility to the individual NRTIs, NNRTIs and PIs [2]. This knowledge has been translated into expert-based algorithms whereby a specific pattern of HIV-1 pol mutations can be interpreted as conferring complete, intermediate or no resistance to each of the available drugs [3]. Such systems are regularly updated by expert panels periodically reviewing the latest in vitro and in vivo antiretroviral resistance data and accordingly adjusting the algorithm rules. Indeed, the most widely used rule-based algorithms have been shown to be helpful in predicting response to treatment in patients harbouring drug-resistant virus [4]. However, given the complexity of HIV-1 drug resistance, the inferred drug susceptibilities derived by different systems may diverge [5–7]. Moreover, HIV-1 drug resistance experts agree that selection of a treatment regimen must also be based on additional factors including patient clinical status and commitment to therapy, previous exposure to antiretroviral drugs, and past HIV-1 genotype information. In fact, interpretation of HIV-1 genotype by one or more experts in the field can improve virological treatment outcome with respect to simple indication of the susceptibility to individual drugs shown in a resistance test report [8–10]. Thus, HIV-1 genotyping complemented by expert advice is considered the best procedure to take into account HIV-1 drug resistance when building an antiretroviral regimen.

More recently, data-driven drug susceptibility prediction systems have started to be explored through different statistical learning methods. Large genotype-to-phenotype and/or genotype-to-virological response correlation data sets are required to train such systems. The derived model is typically applied to predict drug activity against a given HIV-1 genotype. For instance, the proprietary VircoType system was trained on tens of thousands of genotype-phenotype pairs and can reliably estimate in vitro resistance to individual drugs for any specific set of mutations based on multiple linear regression [11]. Clinical cut-off values derived from statistical learning are applied to estimate the in vivo activity of each drug against the virus [12]. Using a large genotype-to-virological response training data set, researchers of the Resistance Response Database Initiative (RDI) group have developed an artificial neural network method to predict the change in viral load caused by a given therapy in the presence of a specific HIV-1 mutant [13]. The same group has also shown that the model can use additional data such as the patient CD4 cell count and summary indicators of previous treatment exposure to increase the accuracy of the prediction [13]. Finally, the EuResist consortium has developed a novel system based on a combination of three statistical learning models to predict the probability of short-term treatment success based on HIV-1 genotype and, when available, supplementary patient data [14]. In contrast to the VircoType and all rule-based algorithms, the RDI system and the EuResist engine are intended to predict the virological success of a combination regimen, rather than the activity of the individual drugs, thus providing more clinically oriented guidance for building an antiretroviral therapy regimen.

The aim of this study was to compare the performance of the EuResist system with that of human experts predicting short-term virological outcomes in a set of 25 past treatment cases with complete clinical and virological information.

Methods

The EuResist engine (http://engine.euresist.org/) has been trained and validated on around 3000 treatment change episodes (TCEs) extracted from the EuResist integrated database (EIDB), a collection of HIV-1 resistance data from four European nationwide study cohorts (Germany, Italy, Luxembourg and Sweden). Briefly, a TCE was defined as a treatment switch with baseline genotype and viral load obtained at maximum 12 weeks before the therapy change and a follow-up viral load measured after 8 (4–12) weeks of the same uninterrupted treatment. Success was defined as a decrease of baseline viral load by at least 2 log10 HIV-1 RNA copies/mL or suppression of viral load to undetectable levels. The prediction system combines three independent models into a classification of the treatment as a success or failure at 8 weeks [14]. A number of different ensemble methods were explored with the aim of finding the optimal way to combine the different models [15]. The EuResist system output is the mean of the three probability values returned by the three individual engines and varies between 0 and 1; a value of >0.5 indicates success and a value of ≤0.5 indicates failure. Based on validation studies carried out on historical cases, the combined engine predicts the correct outcome in 76% of cases [14] and is more stable than any single engine [15]. The system accepts as input the HIV-1 genotype (mandatory) as a list of
mutations or as a whole sequence and any of the following information when available: patient age and sex, route of infection, baseline viral load and CD4 cell count, the number of previous treatment lines, and binary indicators of previous use of the individual NRTIs, NNRTIs and PIs. These optional input variables have been shown to increase the accuracy of at least one of the three individual engines.

For the EuResist system vs. expert (EVE) comparison, 12 top-level international HIV-1 drug resistance experts were invited to take part in the study, and enrolment was closed when the first 10 declared their availability. Experts were recruited among scientists with highly documented activity in the field based on long-standing and relevant visibility as authors of peer-reviewed articles and presentations at HIV-specific international conferences. All of the EuResist data come from patients treated in Europe. Six of the experts contacted were chosen from Europe and, in order to determine whether working in a different region with possibly different drug prescription attitudes could have an impact on predicting treatment outcome for European patient cases, six experts from non-European countries were invited to participate. The 10 experts composing the final panel are listed as coauthors of the study (C.A.B, F.B.-V, P.R.H., L.M., M.O., C.F.P., P.P., D.P, R.W.S. and A.-M.V.). A total of 25 TCEs were randomly extracted from a subset of the EIDB validation data set (i.e. the cases were excluded from training the EuResist system) for which the treatment regimen consisted of exactly three drugs (a ritonavir-boosted PI being considered a single drug), the baseline viral load was at least 10 000 copies/mL and the baseline genotype included at least one major resistance mutation according to the contemporary International AIDS Society (IAS) definition [2]. The TCEs were provided via an online interactive questionnaire that could be partially filled in and saved for later completion. Each of the experts received a private username and password that could be used to view and fill in the questionnaire anonymously. Only European or non-European origin was retained by the system; the identities of the individual experts could not be determined. Upon completing and closing the questionnaire, the expert was given a result page where she/he could see her/his own choices together with the actual outcomes and the EuResist predictions.

For each TCE, the web interface showed the baseline HIV-1 viral load and genotype (both as a list of mutations with respect to consensus B and as a fasta file), the treatment applied and summary indicators of patient history including the number of previous treatment lines, the months of cumulative previous use of each antiretroviral drug, and past genotypes. In addition, the full history was available as a Microsoft Excel file reporting all available CD4 cell counts, viral load measurements and treatment changes over time. Of note, there was no available information about patient adherence to treatment, although treatment records originally labelled with poor adherence had been removed when building the EIDB.

Experts were instructed to categorically label each of the 25 treatments as a 'success' or a 'failure'; and provide a quantitative estimate for this prediction expressed as probability of success in the range 0–100%, with values higher than 50% indicating success. This estimate was requested so that the evaluation data could be used to make a quantitative comparison between the expert opinion and the EuResist system output. In addition, experts were asked if they had used any of the following expert systems while completing the evaluation: Stanford HIVdb (http://hivdb.stanford.edu/pages/alg/hivdb.html), Agence Nationale de Recherche sur le SIDA (ANRS) rules (www.hivfrenchresistance.org/table.html), Rega rules (www.regakuleuven.be/cev/index.php?id=30), the IAS reference mutation list (http://iasusa.org/resistance_mutations/index.html), geno2pheno (www.geno2pheno.org) and HIV-Grade (www.hiv-grade.de/cms/grade/homepage.html).

The agreement among experts was evaluated by computing the multirater free-marginal kappa statistics for the qualitative prediction [16] and the coefficient of variation for the quantitative prediction. The trade-off between specificity and sensitivity for labelling a treatment as successful was evaluated by receiver operating characteristics (ROC) analysis [17], where the area under the ROC curve (AUC) was used as an indicator of the performance of a binary classifier (success/failure), with AUC values up to 1. The agreement between human experts and the expert system for the quantitative prediction was evaluated using Pearson correlation coefficients. The absence of systematic error was checked on a Bland–Altman plot with the limit of agreement set as mean ± 1.96 SD.

Results

Data set and expert panel

The 25 TCEs randomly chosen from the EIDB included 16 PI-based and four NNRTI-based treatments all coupled with two NRTIs. The remaining therapies included four cases of concurrent use of one PI and one NNRTI with one NRTI and a single treatment of four NRTIs. The year of therapy spanned 2001–2006 with the single exception of the four-NNRTI treatment, which was administered in 1998. Of the 20 therapies including a PI, 17 had a boosted PI, two had unboosted atazanavir and one had nelfinavir.

Table 1 shows the baseline characteristics of the 25 patients included in the case file. In addition to the baseline TCE-defining information (viral load, genotype and therapy), a median [interquartile range (IQR)] of 15 (8–25) viral
Number of available previous genotypes 1 (0–3)
Number of available previous CD4 cell counts 14 (10–30)
Number of available previous viral load measurements 15 (8–25)
Number of NNRTI mutations at baseline 1 (0–2)
Number of NRTI mutations at baseline 3 (3–4)
Number of previous treatment lines 5 (3–6)
Baseline CD4 count (cells/μL) 299 (134–412)
Number of previous treatment lines 5 (3–6)
Number of NNRTI mutations at baseline 3 (3–4)
Number of NRTI mutations at baseline 2 (0–3)
Number of available previous viral load measurements 15 (8–25)
Number of available previous CD4 cell counts 14 (10–30)
Number of available previous genotypes 1 (0–3)

IQR, interquartile range.

load measurements, 14 (10–30) CD4 cell counts and 1 (0–3) genotype were available from past patient histories.

Six and four of the 10 experts participating in the study were from European and non-European countries, respectively. Eight of the experts declared the use of one to four rule-based expert systems while two declared the use of none.

Prediction of treatment response

Figure 1 shows the predictions made by the 10 experts and by the EuResist engine for each of the individual TCEs. Overall, 15 of the 25 TCEs met the criteria for definition of virological success. The EuResist engine mislabelled six cases; three successes and three failures (accuracy 0.76). The mean ± SD number of incorrect calls made by the human experts was 9.1 ± 1.9 (mean ± SD accuracy 0.64 ± 0.07), with only one expert making the same number of errors as EuResist and all the others making more (range 8–13). Overall, there were apparently more failures mislabelled as successes than the opposite (mean ± SD 5.3 ± 2.7 vs. 3.8 ± 1.6, respectively) but the difference was not significant and reflected the uneven distribution of failures and successes in the data set (Table 2). Also, European and non-European experts did not differ in their performance (mean ± SD number of wrong calls 9.8 ± 1.7 vs. 8.0 ± 1.6, respectively), nor did they show different use of the expert systems. There was no correlation between the number of expert systems consulted and the number of errors made.

When ROC analysis was applied to determine the sensitivity and specificity of prediction of treatment success, EuResist was found to be not significantly better than the mean prediction computed by the human experts, nor was it better than any of the individual experts (Fig. 2). The only significant difference in performance was between the best and worst experts, as measured by the area under the ROC curve (P = 0.011).

The agreement among the experts in terms of binary classification of success and failure was only fair, as revealed by the relatively low kappa multirater agreement value (0.355). There were only five (20%) cases where all the experts made the same prediction. In all of these, the outcome was as predicted and the EuResist system prediction agreed with the opinion of the experts.

The mean ± SD coefficient of variation for the quantitative prediction made by the experts for the individual TCEs was also relatively high (55.9 ± 22.4%). However, the significant correlation between the quantitative prediction generated by EuResist and the average quantitative prediction provided by the experts showed a strong positive relationship (Pearson r = 0.695, P < 0.0001), with considerable inter-individual variation.

According to the Bland–Altman plot (Fig. 3), the difference between the quantitative predictions given by the experts and by the EuResist engine is independent of the mean of the two values, indicating that there was no systematic error related to the magnitude of the predicted probability.

Analysis of cases mislabelled by the EuResist engine and by most human experts

A closer look at the individual TCEs revealed four cases where the EuResist engine as well as eight or nine of the human experts made incorrect calls. Two patients had an unexpectedly successful response to treatment. TCE case number 12843 had an HIV-1 genotype showing NRTI resistance, the key PI mutations G48V, V82A and L90M and thymidine analogue mutation (TAM) pattern 1 with a T215C revertant variant. The patient was treated with stavudine, abacavir and lopinavir/ritonavir and had a partial response, with a reduction in HIV-1 RNA load from 72 300 to 314 copies/mL, representing a 2.36 log10 copies/mL reduction, which met the definition of success. TCE case number 14503 referred to a patient treated with stavudine, efavirenz and lopinavir/ritonavir who had a very low CD4 count nadir (8 cells/μL) and a high baseline viral load (794 328 copies/mL). The HIV-1 genotype included the PI mutations G48V, V82C and I84V, the NNRTI mutation Y181I and the TAM mutations M184V and NNRTI (Y181C) mutations were present, the baseline protease was wild type. However, there was a past genotype record showing I84V. The viral load did not decrease at all. Case 43708 referred to a patient treated with...
three-class therapy consisting of boosted atazanavir in combination with zidovudine and efavirenz. Baseline and one past HIV-1 genotypes were identical, showing major NRTI mutations (K65R, L74V, Y115F and M184V) and minor or uncommon NNRTI mutations (V90I and G190Q) but a wild-type protease. The viral load decreased by only 1.48 log10 copies/mL at the planned 8-week observation, thus meeting the definition of failure. However, a more pronounced decrease by 3.07 log10 copies/mL was recorded at an earlier time-point, indicating transient success.

**Discussion**

Although the correlation between HIV-1 genotype and drug susceptibility in vitro has been one of the foundations of the incorporation of HIV-1 drug resistance testing into
clinical practice, genotype interpretation systems have gradually evolved into more clinically oriented tools designed to predict response to treatment in vivo. Accordingly, currently available rule-based systems have been partly derived from statistical learning based on virological response data. Next-generation, fully data-driven engines, including the RDI system [13] and EuResist [14], have been developed to predict response to a combination of drugs rather than to the individual drugs, thus moving a step further towards clinical needs. The EuResist model computes an input that consists of HIV-1 genotype and therapy information, complemented by several optional patient and virus features, and returns the probability of short-term success for any given combination treatment.

We were able to make a retrospective comparison of the performance of the EuResist engine with 10 HIV drug resistance experts’ opinions on a set of 25 cases derived from patients harbouring drug-resistant virus. The number of cases was deliberately limited so that it would take a reasonable amount of time for the participants to complete the study. As a cautionary note, it must be taken into account that the cases were selected from the EIDB rather than from an external source, although these cases have never been used during the development of the EuResist model. Moreover, the EIDB, including data from more than 100 different clinics in four countries, is likely to represent great diversification in drug prescription attitudes and patient populations.

**Table 2** Ability to call failure and success and overall accuracy of the binary prediction for the 10 human experts and for the EuResist expert system

<table>
<thead>
<tr>
<th>Rater</th>
<th>Ability to call failure (specificity) (%)</th>
<th>Ability to call success (sensitivity) (%)</th>
<th>Overall accuracy</th>
<th>AUC (standard error)</th>
<th>P-value (comparing AUCs against EuResist)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert 1 (E)</td>
<td>80.00</td>
<td>53.30</td>
<td>0.64</td>
<td>0.727 (0.102)</td>
<td>0.542</td>
</tr>
<tr>
<td>Expert 2 (E)</td>
<td>70.00</td>
<td>53.30</td>
<td>0.60</td>
<td>0.733 (0.101)</td>
<td>0.669</td>
</tr>
<tr>
<td>Expert 3 (NE)</td>
<td>50.00</td>
<td>66.70</td>
<td>0.60</td>
<td>0.693 (0.106)</td>
<td>0.347</td>
</tr>
<tr>
<td>Expert 4 (E)</td>
<td>70.00</td>
<td>66.70</td>
<td>0.68</td>
<td>0.743 (0.099)</td>
<td>0.703</td>
</tr>
<tr>
<td>Expert 5 (E)</td>
<td>50.00</td>
<td>73.30</td>
<td>0.64</td>
<td>0.747 (0.098)</td>
<td>0.696</td>
</tr>
<tr>
<td>Expert 6 (NE)</td>
<td>30.00</td>
<td>93.30</td>
<td>0.68</td>
<td>0.683 (0.108)</td>
<td>0.406</td>
</tr>
<tr>
<td>Expert 7 (E)</td>
<td>60.00</td>
<td>60.00</td>
<td>0.60</td>
<td>0.663 (0.110)</td>
<td>0.368</td>
</tr>
<tr>
<td>Expert 8 (NE)</td>
<td>70.00</td>
<td>80.00</td>
<td>0.76</td>
<td>0.853 (0.076)</td>
<td>0.433</td>
</tr>
<tr>
<td>Expert 9 (E)</td>
<td>80.00</td>
<td>26.70</td>
<td>0.48</td>
<td>0.653 (0.111)</td>
<td>0.182</td>
</tr>
<tr>
<td>Expert 10 (NE)</td>
<td>60.00</td>
<td>73.30</td>
<td>0.68</td>
<td>0.783 (0.092)</td>
<td>0.971</td>
</tr>
<tr>
<td>Average expert</td>
<td>62.00</td>
<td>64.70</td>
<td>0.64</td>
<td>0.777 (0.091)</td>
<td>0.917</td>
</tr>
<tr>
<td>EuResist</td>
<td>70.00</td>
<td>80.00</td>
<td>0.76</td>
<td>0.787 (0.091)</td>
<td>NA</td>
</tr>
</tbody>
</table>

AUC, area under the receiver operating characteristics curve; E, European; NE, non-European; NA, not applicable.

**Fig. 2** Receiver operating characteristics (ROC) curves for the EuResist engine and for the best, worst and average expert probability of success as predictors of response to treatment. The inset shows the area under the curve (AUC) values with the associated standard error (SE) for each rater.

**Fig. 3** Bland–Altman analysis of the quantitative estimate of treatment success provided by the experts (mean value) and computed by the EuResist engine. Each data point represents one of the 25 individual cases. The solid line is the mean difference between the experts’ and the EuResist prediction, and the dotted lines are the limits of agreement [±1.96 standard deviation (SD)].
Overall, the EuResist engine performed at least as well as the human experts. The lowest number of incorrect calls in the binary classification of success and failure was in fact made by EuResist and by only one of the experts. To mimic clinical practice, the experts had access to the entire available patient history, including all CD4 cell counts and viral load measurements, past treatments and HIV-1 genotypes. It should be noted that the current version of EuResist does not include past viraemia levels and only simple surrogate markers of previous drug exposure, less detailed than those made available to the experts, are taken into account. Thus, the experts could consider some extra information over and above that considered by the expert system. However, it could be argued that the experts did not have any familiarity with the patients and the design thus failed to reproduce the real scenario where doctor-patient interaction plays a key role, particularly in assessing patient commitment to therapy. A prospective study comparing standard of care supplemented or not by the EuResist system is required to evaluate appropriately the potential role of the engine in clinical practice. By design, this study did not allow assessment of whether (and by how much) taking into account the patient and virus data not included in the minimal TCE definition increased the accuracy of the prediction. However, such additional information has been consistently found to increase accuracy in several recent studies using rule-based or data-driven systems [13,18,19].

The correlation between the average quantitative prediction made by the experts and the quantitative prediction computed by EuResist was statistically significant. However, the agreement among the individual experts was rather low, both in the binary classification and in the quantitative score. This highlights the complexity of choosing an antiretroviral treatment in patients harbouring drug-resistant virus which results in frequent discordances in experts’ opinions. Consistent with this complexity, it should also be emphasized that the best result achieved in this study still labelled incorrectly as much as one-quarter of the treatment cases. Unknown adherence issues and the possibility that hidden drug-resistant minority species impaired response to treatment are among the most likely, although not verified, reasons for prediction errors.

The inclusion of some currently obsolete therapies (e.g. use of nelfinavir or stavudine in five cases) and the lack of novel antiretroviral drug classes in the test data set may have been a limitation of the study. However, most of the therapies were not outdated and in addition are clearly relevant for most of the low- to middle-income areas where antiretroviral coverage has recently expanded. The free web service provided by the EuResist network may be particularly effective in these settings. Several high-genetic-barrier drugs such as darunavi, tipranavir and etravirine could not be considered for training the EuResist engine because of a shortage of data and thus could not be included in the study data set. The updated version of the EuResist engine recently made available online (version 2.0) can now also compute the response to these three drugs. It remains to be established how the expert system would perform with respect to human experts for these high-genetic-barrier drugs. This is clearly relevant because predicting the activity of such drugs is crucial in the current antiretroviral therapy situation, at least in Western countries. Also, drugs belonging to novel classes such as integrase inhibitors and coreceptor antagonists cannot be included in the computations because of the scarcity of available treatment cases and/or a lack of virus genotype information.

The TCE definition itself had its own limitations. First, a short follow-up time was employed because EuResist was trained to predict response at 8 weeks. Short-term response is directly related to antiviral activity on the majority virus population and is usually less complicated by confounding factors, such as adherence or toxicity, than long-term response. However, with the availability of novel well-tolerated long-lasting therapies, the goal shifts to prediction of longer-term response. While the aim of the study was to predict the 8-week response because the EuResist engine had been trained on that follow-up time, post hoc intention-to-treat analysis at 24 weeks (not shown) confirmed an accuracy of 0.78 for EuResist compared with an average accuracy of 0.71 for the human experts. The next update of the EuResist engine is also planned to focus on the 24-week response. Secondly, the definition of virological success was based on a single follow-up viral load measurement. In some cases, treatment success was reached at a later time-point under the same therapy (data not shown), making definition of the case as a failure questionable [15].

Despite the limitations, this study suggests that a data-driven and clinically oriented expert system can predict the response to antiretroviral therapy as accurately as, and in most cases better than, HIV drug resistance experts. The engine is not intended to replace the HIV specialist but rather to be an advisory tool. Updates and upgrades are required to exploit the full potential of this and other data-driven expert systems. Treatment response data from patients treated with the novel drugs are critically needed to enable new regimens to be included in the engine set. Integrating new drugs into the system has required more than 1 year because of the need to collect a sufficient amount of training data and retrain and validate the system. Clearly, early access to drug resistance data derived from Phase III clinical trials, once the drugs have been licensed, is a critical step for reducing this delay. Also, the TCE collection must include instances from patients
infected with all the different HIV-1 clades to weight a possible impact of HIV-1 natural variability on treatment. An expanded, publicly available TCE repository could be the best way of providing a common source for training and testing treatment decision support tools. It is hoped that the scientific community and regulatory bodies will endorse such an initiative to further improve clinical management of HIV-1 drug resistance.

Acknowledgments

This work was presented at the Eighth European HIV Drug Resistance Workshop, Sorrento, Italy, 17–19 March 2009. The EuResist Project was funded by the European Community under FP6 (IST-2004-027173). The EuResist Network has been supported by grants from Abbott and Pfizer and is part of the European Community’s Seventh Framework Programme (FP7/2007–2013) under the project ‘Collaborative HIV and Anti-HIV Drug Resistance Network (CHAIN)’ (grant agreement number 223131).

References