Relative contributions of baseline patient characteristics and the choice of statistical methods to the variability of genotypic resistance scores: the example of didanosine

Lambert Assoumou1,2*, Allal Houssaïni1,2, Dominique Costagliola1–3, Philippe Flandre1,2,4 on behalf of the Standardization and Clinical Relevance of HIV Drug Resistance Testing Project from the Forum for Collaborative HIV Research

1INSERM, U943, Paris, F-75013 France; 2UPMC Univ Paris 06, UMR S 943, Paris, F-75013 France; 3Department of Infectious Diseases, Assistance Publique Hôpitaux de Paris (AP-HP) Pitié-Salpêtrière Hospital, Paris, France; 4Laboratory of Virology, Assistance Publique Hôpitaux de Paris (AP-HP) Pitié-Salpêtrière Hospital, Paris, France

*Corresponding author. INSERM U943, 56 Bd V Auriol, BP 335, 75625 Paris Cedex 13, France. Tel: +33(0)1-42-16-42-80; Fax: +33(0)1-42-16-42-61; E-mail: klassoumou@ccde.chups.jussieu.fr

Received 4 November 2009; returned 5 December 2009; revised 13 January 2010; accepted 16 January 2010

Background: Our aim was to investigate the respective role of statistical methodology and patients’ baseline characteristics in the selection of mutations included in genotypic resistance scores.

Methods: As an example, the FORUM database on didanosine including 1453 patients was used. We split this population into four samples based on countries of enrolment (France n=474, Italy n=440, USA/Canada n=219, others n=320). We used both a continuous outcome measure (the viral load reduction at week 8) and a binary outcome measure (viral load decline at week 8, 0.6 log10 or /210.6 log10) and both parametric and non-parametric methods for each outcome.

Results: Overall, 61 distinct mutations were selected by at least one method in at least one data set. The variability due to baseline characteristics varies from 79% to 88%, i.e. for a given method applied to the four data sets. 80% of the mutations are selected only once. The variability induced by the methodology varies from 49% to 56%, i.e. for a given data set /2450% of the mutations are selected by at least two methods.

Conclusions: Baseline patient characteristics contribute more than the choice of statistical method to the variability of the mutations to be included in the genotypic resistance scores.

Introduction

Selection of drug-resistant HIV-1 variants is one of the main factors limiting the efficacy of antiretroviral therapy. Several rule-based systems have been developed to interpret HIV-1 resistance genotyping results for a given drug. Most use different approaches and methodologies. For example, the HIV RT and Protease Sequence Database (HIVdb) algorithm assigns a drug penalty score for each resistance mutation. These penalty scores are derived from the literature linking mutations and antiretroviral drugs, including correlations between the genotype and the treatment history, the genotype and the phenotype, and the genotype and clinical outcome. The scores are updated as new studies are published and are occasionally modified on the basis of user feedback. The French Agence National de Recherches sur le SIDA (ANRS) HIV-1 genotypic drug resistance interpretation system is mostly based on correlations between drug resistance mutations and virological outcome. Other statistical methods such as neural networks, support vector machine, random forests and LARS/LASSO regression have been used for building genotypic resistance interpretation systems.

Recent studies have shown substantial disagreement among these algorithms for the interpretation of genotypic resistance mutations, and variable ability to predict the virological response when accounting for other important predictors of response. This is not surprising, as these interpretation rules were developed by different groups, often in different countries, sometimes in different clinical settings, and using distinct virological outcomes and statistical methods. It is important to determine how much of the variability in the mutation selected and the weight of each mutation arises from baseline characteristics of the patients and how much arises from the choice of statistical methodology.

To address this issue, we used, as an illustrative example, the Forum for Collaborative HIV Research didanosine database.
The aim of this study was to investigate the respective contributions of the statistical methodology and the baseline population characteristics to the variability of the selected mutations to be included in the genotypic resistance interpretation systems, using didanosine as an example.

Patients and methods

Study population

We used pooled data from the following 13 sources (clinical trials and clinic-based cohorts): the Aquitaine Cohort, France; the Antiretroviral Resistance Cohort Analysis (ARCA) database, Italy; the British Columbia Cohort, Canada; the EuroSIDA cohort, Europe; the Italian Cohort Naïve Antiretrovirals (ICONA), Italy; the Institut d’Investigacions Biomediques August Pi i Sunyer (IDIBAPS) Barcelona Hospital Clinic cohort, Spain; the Ramon y Cajal Madrid Hospital clinical database, Spain; the Jaguar trial, Bristol-Myers Squibb (BMS), France; the ANRS Narval 88 trial, France; the Swiss HIV Cohort Study, Switzerland; the AIDS Clinical Trial Group (ACTG), USA; the Catholic University Sacro Cuore (UCSC) Cohort, Italy; and the UK HIV Resistance Response Database Initiative (RDI), UK and UK HIV Cohort Study (CHIC) database, UK. Patients were included if they met the following criteria:

1. virological failure (clinician’s judgement) before beginning a regimen including didanosine (either used for the first time or recycled)
2. a genotypic resistance test performed during the failing regimen (<12 weeks before the start of the new regimen)
3. viral load >500 copies/mL during the failing regimen (<12 weeks before the start of the new regimen); this was considered the baseline viral load
4. at least one viral load measured 4–12 weeks (8 week viral load) or 16–32 weeks (24 week viral load) after starting the new (didanosine-containing) regimen
5. no changes in therapy between the time of the baseline viral load or resistance test and the start of the new regimen, nor between the time of the start of the new regimen and the time of the analyses (4–12 weeks for the 8 week analyses)
6. no evidence of inadequate adherence to the didanosine-containing regimen.

More details on the data extraction procedure and the items collected can be found elsewhere.

HIV-1 RNA extraction and sequencing

HIV-1 RNA was isolated by means of commonly used assays, and full sequence analyses of HIV-1 protease and reverse transcriptase reading frames were performed according to study protocols, using the genotype that was the closest to the date of initiation for the didanosine-containing regimen. Reference HIV strains (pNL4.3, HXB2 and consensus B) varied from study to study, but data were standardized so that coding for resistance was not ambiguous.

Statistical analyses

Four distinct samples were defined according to the country of enrolment [France n=474, Italy n=440, USA/Canada n=219 and others (Spain/UK/Switzerland) n=320]. Two outcome measures were considered: a continuous outcome defined as the viral load reduction at week 8 and a binary outcome defined as a fall in viral load of \( \geq 0.6 \log_{10} \) at week 8 (yes/no), given the period when the various studies were performed (between May 1993 and May 2004). For each country-based data set and for each type of outcome, a parametric method and a non-parametric method were applied. The variability due to baseline characteristics was estimated by computing the proportion of mutations retained by each method when applied to the four geographic data sets. The variability due to the choice of statistical methods was estimated by computing the proportion of mutations retained when the four methods were applied to each data set.

We looked at all possible mutations at each position among positions 20–245 of the reverse transcriptase gene.

Method 1. We tested the association between each possible mutation at each position among positions 20–245 of the reverse transcriptase gene and the continuous outcome measure by using the Mann–Whitney statistic. Mutations with \( P \) values \(<0.20\) were retained. We used the Jonckheere–Terpstra removal procedure to select the final set of mutations most strongly associated with the viral load response and used for the genotypic score. From the initial set of \( K \) mutations retained from the first analysis all mutations are removed one by one to investigate all combinations of \( K \)–1 mutations. The combination of \( K \)–1 mutations providing the lowest \( P \) value with the Jonckheere–Terpstra test was retained. In the second step, mutations were again removed one by one to compare the different combinations of \( K \)–2 mutations; the combination providing the lowest \( P \) value was again retained, and so on. The procedure stopped when removing a mutation did not provide a lower \( P \) value than the previous \( P \) value. This procedure has been used successfully. The Jonckheere–Terpstra test, for ordered alternatives, has been specially designed for the context of a continuous response. Indeed, when comparing groups of patients with distinct numbers of resistance mutations one would expect that patients with no resistance mutations would have a better virological response than patients with one mutation, who in turn would have a better response than patients with two mutations, and so on. Mutations associated with a poorer virological response were scored as \((-1\) and those associated with a better virological response were scored as \((+1\). We looked at all possible mutations at each position among positions 20–245 of the reverse transcriptase gene.

Method 2. Linear regression models, accounting for the censoring of viral load measurements due to assay lower limits, of the week 8 reduction in viral load from baseline were adjusted for baseline viral load, the exact number of weeks between baseline and the week 8 viral load measurements, and the ANRS genotypic sensitivity score for all drugs received apart from didanosine in the prescribed combination. The score of a drug was 1, 0.5 and 0 when rated susceptible, intermediate and resistant, respectively, according to the ANRS interpretation system.

We looked at all possible mutations at each position among positions 20–245 of the reverse transcriptase gene.

The following procedure was used to select the mutations most strongly associated with the virological response but also to assess the strength of the evidence that an identified mutation is a ‘true’ independent predictor.

- First, at step 0, all mutations at all positions were included in the model as well as the adjustment variables described above. Then, a backward selection technique \((\alpha=0.05)\) was used to retain those that were independently associated with the virological response in the data set. Only main factors were considered, and potential interactions were not taken into account (step 1).
- Secondly, a backward elimination method was applied to 100 bootstrap samples of the data set, and all mutations between position 20 and 245 that were selected in >75% of the models were retained (step 2).
- We then repeated the second step with the mutations selected at step 1, forcing those retained at step 2 into the model (step 3).

The final set of mutations was the mutations retained after the three steps.

Method 3. We tested the association between each possible mutation at each position among positions 20–245 of the reverse transcriptase gene and the binary outcome measure by using Fisher’s exact test.
Mutations with \( P \) values <0.20 were retained. We used the Cochran–Armitage removal procedure to select the final set of mutations most strongly associated with the virological response. The removal procedure is similar to the one described in method 1 except that the Cochran–Armitage test is used instead of the Jonckheere–Terpstra test. The Cochran–Armitage test is also a test for ordered alternatives appropriate for categorical response.\(^{16}\)

Method 4. As for method 2, logistic regression models of the binary outcome measure were adjusted for the adjustment variables described above. We looked at all possible mutations at each position among positions 20–245 of the reverse transcriptase gene.

We used the same procedure as that described in method 2 to select the final set.

In our work the variability due to baseline characteristics (from the country-based data sets) is assessed by the percentage of mutations retained only once using a given method applied to the four country-based data sets. A high percentage indicates that only a few mutations were selected in at least two genotypic scores. The variability due to the statistical methods is assessed by the percentage of mutations retained only once in a given country-based data set using the four distinct methods.

We used Fleiss’ kappa statistic to assess the agreement within the four statistical methods in each data set and the agreement within the four data sets given that the same method is used over the four data sets.\(^{17}\) The following classification was used for interpreting kappa values: <0, poor agreement; 0.0–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, almost perfect agreement.\(^{18}\)

We built a genotypic resistance score by scoring a mutation associated with a poorer virological response as (+1) and those associated with a better virological response as (–1). To illustrate further the variability induced by the data set, the score built with method 1 on the France data set was applied to the USA/Canada data set, and vice versa.

Analyses were performed with the SPSS software package version 15.0 for Windows (SPSS Inc., Chicago, IL, USA) and the SAS software package version 9.1 for Windows (Cary, NC, USA).

### Results

#### Baseline characteristics of the study population

Patients were enrolled between January 2002 and February 2003, with a median in April 2002 in France; between October 1999 and May 2004, with a median in September 2002 in Italy; between May 1993 and November 2002, with a median

#### Table 1. Baseline characteristics of the patients in the four data sets

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>France ((n=474)^a)</th>
<th>Italy ((n=440)^b)</th>
<th>USA/Canada ((n=219)^c)</th>
<th>Others ((n=320)^d)</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, years, median (IQR)</td>
<td>40 (35–46)</td>
<td>40 (35–44)</td>
<td>39 (34–44)</td>
<td>40 (34–46)</td>
<td>0.133</td>
</tr>
<tr>
<td>Male, n (%)</td>
<td>398 (84%)</td>
<td>282 (66%)</td>
<td>152 (87%)</td>
<td>252 (82%)</td>
<td>0.285</td>
</tr>
<tr>
<td>Transmission group, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>men having sex with men</td>
<td>82 (55%)</td>
<td>25 (21%)</td>
<td>41 (53%)</td>
<td>158 (60%)</td>
<td></td>
</tr>
<tr>
<td>IVDU</td>
<td>21 (14%)</td>
<td>26 (21%)</td>
<td>7 (9%)</td>
<td>26 (10%)</td>
<td></td>
</tr>
<tr>
<td>heterosexual</td>
<td>19 (13%)</td>
<td>52 (43%)</td>
<td>3 (4%)</td>
<td>65 (25%)</td>
<td></td>
</tr>
<tr>
<td>other</td>
<td>26 (18%)</td>
<td>18 (15%)</td>
<td>26 (34%)</td>
<td>14 (5%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>HIV RNA, log(_{10}) copies/mL, median (IQR)</td>
<td>4.17 (3.67–4.71)</td>
<td>4.49 (3.89–5.00)</td>
<td>4.38 (3.87–4.92)</td>
<td>4.12 (3.53–4.85)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>CD4 cell count, cells/mm(^3), median (IQR)</td>
<td>320 (202–460)</td>
<td>310 (162–443)</td>
<td>231 (140–331)</td>
<td>240 (124–382)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Previous exposure to didanosine, n (%)</td>
<td>226 (48%)</td>
<td>79 (18%)</td>
<td>35 (16%)</td>
<td>116 (36%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Number of previous ARVs, median (range)</td>
<td>5 (1–12)</td>
<td>3 (1–10)</td>
<td>3 (1–11)</td>
<td>4 (1–11)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Number of baseline drugs, including didanosine, median (IQR)</td>
<td>3 (2–3)</td>
<td>3 (2–3)</td>
<td>3 (2–3)</td>
<td>3 (2–3)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Number of TAMs, median (IQR)(^e)</td>
<td>3 (1–3)</td>
<td>1 (0–3)</td>
<td>2 (0–3)</td>
<td>1 (0–3)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>The ANRS genotypic sensitivity score, median (range)</td>
<td>0 (0–2.5)</td>
<td>1 (0–3)</td>
<td>1 (0–4)</td>
<td>1 (0–4.5)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

IQR, interquartile range; IVDU, intravenous drug user; ARVs, antiretrovirals; TAMs, thymidine analogue mutations.

\(^a\)France: age was available for 474 patients (100%), gender for 474 (100%), transmission group for 148 (31%) and the CD4 cell count for 466 (98%).

\(^b\)Italy: age was available for 390 patients (89%), gender for 428 (97%), transmission group for 121 (28%) and the CD4 cell count for 120 (27%).

\(^c\)USA/Canada: age was available for 173 patients (79%), gender for 174 (79%), transmission group for 77 (35%) and the CD4 cell count for 170 (78%).

\(^d\)Others (Spain/UK/Switzerland): age was available for 141 patients (44%), gender for 308 (96%), transmission group for 263 (82%) and the CD4 cell count for 290 (91%).

\(^e\)TAMs = M41L + D67N + K70R + L210W + T215YF + K219QE.
in December 1997 in the USA/Canada; and between February 1997 and August 2004, with a median in September 2000 in the others data set.

Baseline characteristics (Table 1) differed somewhat between the four country-based data sets. The median number of drugs to which the patients had previously been exposed was, respectively, 5 (range 1–12), 3 (1–10), 3 (1–11) and 4 (1–11) in the France, Italy, USA/Canada and others data sets (P < 0.001). A larger proportion of patients had been exposed to didanosine in the France (48%) and others (36%) data sets than in the Italy and USA/Canada data sets (18% and 16%, respectively) (P < 0.001). The median number of active drugs in the new regimen, in addition to didanosine, was 0 (range 0–2.5), 1 (0–4) and 1 (0–4.5) in the France, Italy, USA/Canada and others data sets (P < 0.001).

Baseline prevalence of mutations

Figure 1 shows the prevalence of reverse transcriptase mutations at baseline in the four country-based data sets, according to the International AIDS Society (IAS)-USA panel (updated in April 2008). The median number of thymidine analogue mutations (TAMs) was, respectively, 3 (range 0–5), 1 (0–5), 2 (0–5) and 1 (0–4) in the France, Italy, USA/Canada and others data sets (P < 0.001). The respective prevalence rates of didanosine resistance mutations in the IAS–USA system were, respectively, 1%, 3%, 1% and 3% for K65R (P = 0.048), and 7%, 2%, 2% and 5% for L74V (P < 0.001).

Week 8 virological response

The median decline in viral load at week 8 was, respectively, 0.82 log10 copies/mL [interquartile range (IQR) 0.14–1.72], 1.22 log10 copies/mL (0.28–2.03), 1.35 log10 copies/mL (0.50–1.80) and 1.72 log10 copies/mL (1.09–2.24) in the France, Italy, USA/Canada and others data sets (P < 0.001). The respective proportions of patients whose viral load fell by ≥0.6 log10 at week 8 were 58%, 67%, 73% and 83% (P < 0.001).

Genotypic resistance mutations for didanosine

Sixty-one distinct mutations were selected by at least one of the four statistical methods in at least one of the four country-based data sets (Table 2). By applying method 1 to each of the four geographic data sets, we assessed the variability due to baseline characteristics in providing four different genotypic scores [the first column (M1) of each country in Table 2]. Thirty-one distinct mutations were selected and 81% of them were selected only in a single country base data set. A large variability was also obtained with the other three methods (79%–88%) (Figure 2a). The percent agreement statistics assessed with Fleiss' kappa statistic confirms the poor agreement within data sets (range 0.0302–0.1090) (Table 3).

By applying the four methods to the France data set, we again obtained four different scores providing an assessment of the variability due to the statistical methodology (Table 2: column M1–M4 for France). Twenty-nine distinct mutations were selected and 48% of them were selected only by one statistical method. Similar results were obtained with the others three data sets (50%–59%) (Figure 2b). Fleiss' kappa statistics also confirm a better agreement of the methods (range 0.3146–0.4897) compared with the agreement provided by the country-based data sets (Table 3).

The proportion of mutations common to methods 1 and 2, using the continuous outcome measure, was 23%, 38%, 23% and 31% in the France, Italy, USA/Canada and others data sets, respectively, while the proportion of mutations common to methods 3 and 4, using the binary outcome measure, was 57%, 45%, 40% and 45% in the France, Italy, USA/Canada and others data sets, respectively.

To illustrate further the variability induced by the data set, the score built with method 1 on the France data set was applied to the USA/Canada data set (Figure 3a). From the France data set, 67%, 29% and 4% of patients were considered as having a susceptible virus (scores between −2 and 0), intermediate virus (score between 1 and 2) and resistant virus (score between 3 and 4). When this scoring system was applied to the USA/Canada data set, 79%, 20% and 1% of viruses were predicted to be...
<table>
<thead>
<tr>
<th></th>
<th>France</th>
<th>Italy</th>
<th>USA/Canada</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td>M1</td>
<td>M2</td>
<td>M3</td>
</tr>
<tr>
<td>V35I</td>
<td>3.8</td>
<td>9.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V35L</td>
<td>1.1</td>
<td>1.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T39A</td>
<td>10.1</td>
<td>6.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T39E</td>
<td>0.8</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E40F</td>
<td>1.5</td>
<td></td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>M41L</td>
<td>61.6</td>
<td>33.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>M1</td>
<td>M2</td>
<td>M3</td>
</tr>
<tr>
<td>V35I</td>
<td>3.8</td>
<td>9.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V35L</td>
<td>1.1</td>
<td>1.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T39A</td>
<td>10.1</td>
<td>6.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T39E</td>
<td>0.8</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E40F</td>
<td>1.5</td>
<td>1.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M41L</td>
<td>61.6</td>
<td>33.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>M1</td>
<td>M2</td>
<td>M3</td>
</tr>
<tr>
<td>V35I</td>
<td>3.8</td>
<td>9.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V35L</td>
<td>1.1</td>
<td>1.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T39A</td>
<td>10.1</td>
<td>6.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T39E</td>
<td>0.8</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E40F</td>
<td>1.5</td>
<td>1.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M41L</td>
<td>61.6</td>
<td>33.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Continued
susceptible, intermediate and resistant, respectively, to didanosine. Patients with a score of 0 had a better virological response than patients with a score of −1. The observed mean viral load reductions at week 8 were 1.23 log<sub>10</sub> copies/mL (standard deviation: 0.94) in susceptible viruses, 1.04 log<sub>10</sub> copies/mL (1.20) in intermediate resistant viruses (P = 0.292 versus susceptible) and 0.34 log<sub>10</sub> copies/mL (0.83) in resistant viruses (P = 0.081 versus susceptible).

Similarly, the score built with the USA/Canada data set was applied to the France data set (Figure 3b). Regarding Figure 3(b), from the USA/Canada data set, we considered viruses with scores between −1 and 0 as susceptible, a score of 1 as intermediate, and scores between 2 and 3 as resistant. In the USA/Canada data set used to derive this score, 83% of cases were susceptible to didanosine, 16% were intermediate and 2% were resistant, while in the France data set the respective proportions were 86%, 13% and 1%. In this data set (France) and with this scoring system, the mean viral load reduction in patients with susceptible, intermediate and resistant viruses was 1.05 log<sub>10</sub> copies/mL (1.08), 0.80 log<sub>10</sub> copies/mL (1.12) (P = 0.050 versus susceptible) and 0.87 log<sub>10</sub> copies/mL (0.88) (P = 0.773 versus susceptible), respectively.

### Discussion

One difficulty with data set-based genotypic resistance scores is that they are dependent on the resistance mutations and polymorphisms present in the study population. Here we show that the data set used to develop a genotypic resistance scoring system contributes more than the choice of statistical method to the observed variability in the selection of the important mutations for resistance to a given drug. The four data sets used here yielded similar variability, even though the USA/Canada data set was half the size of the other data sets; this suggests that score variability is mainly due to the composition of the data set rather than its size.

When the same statistical procedure was used to build genotypic resistance scores for data sets from different countries, the proportion of mutations selected in a single data set was between 79% and 88%. Similar results were obtained when a binary outcome is defined as a fall in viral load of ≥1 log<sub>10</sub> at week 8 instead of a fall of 0.6 log<sub>10</sub>. This large variability may be explained in part by differences in the treatment history of the different populations, which influences the baseline viral genotypes. In addition, antiretroviral treatment strategies differ from one country to another. For example, most patients start with two nucleoside analogue reverse transcriptase inhibitors plus one protease inhibitor in France, and with two nucleoside analogue reverse transcriptase inhibitors plus one non-nucleoside reverse transcriptase inhibitor in the USA. In addition, many patients are still diagnosed and treated late in the course of the disease. In this situation there may be no time to tailor treatment to the individual patient, and this may lead to an inappropriate choice of drugs and thus to a higher risk of selecting resistant mutants.

A constant difficulty in this field is to derive a genotypic score for a drug that is the most frequently included in a triple-drug regimen. The use of two nucleoside analogue reverse transcriptase inhibitors in most combination regimens makes it a challenge to determine the impact of mutations in the reverse transcriptase gene to the specific nucleoside analogue reverse transcriptase inhibitor used, although this is also true for the third compound of the regimen. The ideal situation where only one drug is added to the current regimen arises infrequently. The problem is worse in heavily pre-treated patients receiving new drugs from new classes, such as raltegravir or maraviroc. Nevertheless, most of the genotypic scores have been developed in such conditions.

Although we had no information on the HIV subtypes for this study, they probably played an important role in the different patterns of mutations found in the four data sets, as the subtype distribution varies across countries. Moreover, it was recently shown that the HIV subtype can affect the performance...
of resistance algorithms for many drugs.\textsuperscript{19} Adherence to therapy is another potential source of variability. Patients with poor adherence (as shown by drug trough assay) should be removed from the data set used to develop resistance scores. Indeed, such patients may have a poor virological response due to poor adherence, and this may influence the interpretation of the effect of baseline resistance mutations on virological outcome. In addition, a recent study suggests that a poor adherence to therapy may be indirectly linked to non-B subtype infection.\textsuperscript{19}

Another drug could give a lower variability than that obtained with didanosine. Several studies have highlighted the difficulty of developing a reliable system for genotypic resistance to didanosine,\textsuperscript{8,20} particularly because patterns of mutations conferring resistance to didanosine are not fully identified.\textsuperscript{21} Likewise, clinically relevant cut-offs are difficult to establish even with phenotypic assay.\textsuperscript{22} For example, the lack of benefit of the phenotypic arm in the CCTG 575 trial may have been partly due to the use of inaccurate clinical cut-offs for a number of drugs, including didanosine.\textsuperscript{23}

One way to improve genotypic resistance algorithms or interpretation systems is to use better statistical tools to assess the association between the genotype and virological outcome. Moreover, the use of standardized methods reduces variability between interpretation systems. When we used four different statistical procedures to build genotypic resistance scores based on the same data set, 48\%–59\% of mutations were selected in only one statistical method. It was largely due to the statistical procedure used for the continuous outcome measure compared with the statistical procedure used for the binary outcome measure. The variability observed with the binary outcome measure was lower, because this measure is more appropriate when the viral load reduction is large. In addition, variability was lower when the proportions of patients with treatment success and failure were well balanced. The Fleiss' kappa statistic confirms a poor agreement when each method is compared across data sets (0.0302–0.1090) and a moderate agreement when comparing methods within each data set (0.3146–0.4897).

The score derived from the France data set failed to predict the short-term virological response when applied to the

\begin{table}
\centering
\begin{tabular}{lcc}
\hline
Agreement within methods & Kappa (standard error) & Kappa (standard error) \\
& (only the 61 distinct mutations selected)\textsuperscript{a} & (all mutations tested, n=185) \\
\hline
in France data set & 0.2182 (0.0456) & 0.3146 (0.0266) \\
in Italy data set & 0.4435 (0.0455) & 0.4897 (0.0264) \\
in the USA/Canada data set & 0.4068 (0.0440) & 0.4483 (0.0255) \\
in the others data set & 0.4195 (0.0445) & 0.4650 (0.0258) \\
\hline
Agreement within data sets & & \\
when method 1 is used over the data sets & \textsuperscript{-0.0439 (0.0478)} & \textsuperscript{0.0565 (0.0277)} \\
when method 2 is used over the data sets & \textsuperscript{0.0037 (0.0429)} & \textsuperscript{0.0980 (0.0250)} \\
when method 3 is used over the data sets & \textsuperscript{-0.0679 (0.0459)} & \textsuperscript{0.0302 (0.0267)} \\
when method 4 is used over the data sets & \textsuperscript{0.0390 (0.0442)} & \textsuperscript{0.1090 (0.0257)} \\
\hline
\end{tabular}
\caption{Agreement within the four statistical methods in each data set and agreement within the four country-based data sets when the same method was applied over the four data sets (Fleiss’ kappa statistic)\textsuperscript{a}}
\end{table}

\textsuperscript{a}The following classification was used for interpreting kappa values: <0, poor agreement; 0.0–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, almost perfect agreement.

\textsuperscript{a}Mutations selected by one of the four statistical methods in at least one of the four data sets.

USA/Canada data set, and vice versa. In addition, the models may overfit the data set since the model was trained for a certain data set and therefore may provide a poor fit to a different data set. Recently, we studied the performance of six genotypic interpretation systems and confirmed their high variability.

In conclusion, this study shows that the data set used to construct a genotypic resistance score introduced more variability than the choice of statistical methods. This partly explains why marked variability is found even when the same method is used to construct genotypic resistance scores from different data sets, underlining the need both for standardized methodological approaches and for large continuously updated data sets, in particular for drugs such as didanosine that are mainly used in failing patients.

Figure 3. Performance of the genotypic resistance score constructed with the France data set and the USA/Canada data set using the procedure based on the Jonckheere–Terpstra test (method 1). (a) Performance of the genotypic resistance score constructed with the France data set when applied to the USA/Canada data set. (b) Performance of the genotypic resistance score constructed with the USA/Canada data set when applied to the France data set.

Acknowledgements
These data have been presented at the Seventeenth International HIV Drug Resistance Workshop, Sitges, Spain, 2008 (Abstract 93; Antivir Ther 2008; 13 Suppl 3: A103).

Members of the Standardization and Clinical Relevance of HIV Drug Resistance Testing Project for the Forum for Collaborative HIV Research are listed below.

The data providers included: H. Ribaudo, P. Bohlin and M. Wantman, ACTG Trials, Statistical and Data Analysis Center, Boston, MA, USA; B. Masquelier, Aquitaine Cohort, France; M. Zazzi, ARCA, University of Siena, Italy; B. Hogg and R. Harrigan, Homer Cohort, British Columbia, Canada; J. Lundgren and J. Kjaer, EuroSIDA Cohort, Europe; A. d’Arminio Monforte, ICoNA, Italy; J. Gatell, IDIBAPS Barcelona Hospital Clinic Cohort, Spain; M. J. Pérez, Raman y Cajal Madrid Hospital clinical
Funding
This work was supported by the Institut National de la Sante´ et de la Recherche M´edicale (INSERM).

Transparency declarations
None to declare.

References
14 Jonckheere AR. A distribution-free k-sample test against ordered alternatives. Biometrika 1954; 41: 133–45.