Accuracy, Precision, and Consistency of Expert HIV Type 1 Genotype Interpretation: An International Comparison (The GUESS Study)

Andrew R. Zolopa, Laura C. Lazzeroni, Alex Rinehart, François Brun Vezinet, François Clavel, Ann Collier, Brian Conway, Roy M. Gulick, Mark Holodniy, Carlo-Frederico Perno, Robert W. Shafer, Douglas D. Richman, Mark A. Wainberg, and Daniel R. Kuritzkes

1Stanford University School of Medicine and 2Veterans Affairs Palo Alto Healthcare System, Palo Alto, and 3Veterans Affairs San Diego Healthcare System and 4University of California, San Diego, California; 5University of Washington School of Medicine, Seattle; 6Weill Medical College of Cornell University, New York; 7Section of Retroviral Therapeutics, Brigham and Women’s Hospital and Division of AIDS, Harvard Medical School, Boston, Massachusetts; 8Tibotech-Virco, Mechelen, Belgium; 9Bichat Hospital University and 10Inserm U552, Hospital Bichat-Claude Bernard, Paris, France; 11McGill University AIDS Center, Montreal, Quebec, and 12University of British Columbia, Canada; and 13University of Rome Tor Vergata, Italy

Background. Resistance testing is considered standard of care in HIV medicine, but there is no standard interpretation system for genotype tests. We sought to determine how much agreement exists within a group of experts in the interpretation of complex genotypes.

Methods. Genotypes from clinical specimens were sent to an international panel of 12 resistance experts. Phenotypic susceptibility testing of these clinical isolates was performed with antivirogram. Experts predicted phenotype fold change category (<2.5-fold change, 2.5–4.0-fold change, >4.0- to 7.0-fold change, >7.0- to 10-fold change, >10- to 20-fold change, or >20-fold change) and predicted expected drug activity for each of 16 antiretroviral drugs. Experts were also asked to make treatment recommendations on the basis of the genotype.

Results. The experts predicted the exact phenotype fold change category correctly 44% of the time, but they varied widely by antiretroviral drug (range, 25%–74%). The highest accuracy was observed for lamivudine (74%) and the nonnucleoside reverse transcriptase inhibitors (66%–69%). Experts generally predicted higher levels of resistance to the remaining nucleoside reverse transcriptase inhibitors than what was found by phenotypic testing. Agreement among experts in predicting phenotype fold change category ranged widely depending on the drug (median agreement, 42% [range, 28%–74%]); the same pattern was observed in predicting expected drug activity (median agreement, 45% [range, 32%–87%]). Experts agreed on treatment recommendations in a median of 79% of instances, and recommendations were consistent over time, with blinded retesting.

Conclusions. Although their ability to predict phenotype from a genotype varied for individual antiretroviral drugs, this expert panel had a high degree of agreement in deriving treatment recommendations from the genotype.

Drug resistance testing is now considered the standard of care in the clinical management of cases of HIV infection [1–5]. Despite the widespread use of drug resistance testing, there is still no standard method by which test results are interpreted. For genotypic testing, there are several rules-based algorithms that rely, to various degrees, on expert opinion or consensus statements from panels of experts [6–9]. The amount and quality of interpretive information provided with genotypic test reports differs considerably among diagnostic laboratories. Consequently, many practitioners still find the results of genotypic resistance tests to be confusing, and they frequently seek expert guidance in interpreting these results [8–10].

Randomized, controlled trials have demonstrated improved virologic outcome when genotypic resistance testing was used to guide changes in antiretroviral therapy for patients in whom the current regimen was failing. Some of this benefit was related to the expert advice that often accompanied the test result in these studies.
Although expert advice can improve virologic outcome, it is not clear how well experts agree in their interpretations of genotypes.

We systematically evaluated genotype interpretations from clinical specimens made by an international panel of experts. We tested the ability of these experts to predict measured phenotype, their agreement in predicting phenotype and expected antiretroviral drug activity, and the consensus of their treatment recommendations. We also tested the consistency of their interpretations with blinded retesting.

**METHODS**

**Genotype specimens.** We randomly selected 45 unique genotypes from clinical specimens obtained in 2000 and 2001 for which both genotype and phenotype data were available from the Virco laboratory (Mechelen, Belgium). These clinical specimens were obtained from patients in care in North America and western Europe and were sent to the Virco laboratory for standard resistance testing. The genotypes selected from the Virco database had to have corresponding phenotype results for resistance or susceptibility to 16 antiretroviral drugs, including lopinavir/ritonavir and tenofovir, to be included in the study. Furthermore, the clinical isolates selected had to demonstrate resistance to at least 1 drug on the basis of the phenotype (i.e., the IC\textsubscript{50} for at least 1 antiretroviral drug had to be \(4\)-fold above that for the control). The specimens were not selected on the basis of the actual genotype result. The individuals involved in selecting the specimens from the database did not participate as experts in the GUESS study.

**Web-based genotype reports.** Genotype results for the clinical specimens were provided to the experts as amino acid changes from a reference clade B strain (HXB2) for both the reverse-transcriptase and protease sequences. All changes from the reference strain were listed without any further categorization or interpretation. Experts were then asked to predict the genotypic susceptibility for 16 antiretroviral drugs as a categorical fold change in IC\textsubscript{50}. The categories included: \(<2.5\)-fold change, 2.5- to 4-fold change, \(>4\)- to 7-fold change, \(>7\)- to 10-fold change, and \(>10\)- to 20-fold change, compared with the control. The phenotype categories were chosen somewhat arbitrarily but were intended to be narrow enough to capture important resistance categories for drugs with relatively narrow phenotype dynamic ranges (e.g., stavudine) and drugs for which clinical cutoff values had been established (e.g., abacavir and lopinavir/ritonavir) without requiring the expert to predict the exact phenotype fold change. Experts were also asked to predict expected drug activity on a 6-point scale (ranging from 0 for no activity to 5 for full activity). Finally, experts were asked to make treatment recommendations.

The experts received the genotypes as 2 sets of 25 genotypes, given 4 weeks apart. In each of the 2 sets, 5 genotypes were duplicated to test for consistency over time. The duplicate genotypes were not identified as such to the experts in any way. Experts were free to use whatever resources they wished to make their predictions and complete the forms. Although free to do so, experts did not consult with one another in arriving at their responses.

**Experts.** Experts were selected by 2 of the authors (A.R.Z. and D.R.K.) and were identified on the basis of their substantial contributions to the field of HIV treatment and drug resistance. We included expert virologists and clinician scientists. Twelve of 20 experts recruited to the study provided completed results; the others declined to participate because of time constraints.

**Statistical methods.** Measured phenotypes, expressed as fold change in IC\textsubscript{50} compared with the control, were converted to the 6-point ordinal scale used by the experts. A combined rating for each genotype was computed by taking the median expert rating for that genotype. When the median rating fell between 2 ordinal values, the measurement was rounded down to the nearest ordinal value. For each drug, the overall correlation between the phenotypes predicted by individual experts and the measured phenotypes was calculated using Spearman correlation coefficient. Correlations were also determined for the median fold change in phenotype predicted by the group and the measured phenotype. In addition, the correlation between experts across all 66 possible pairs of experts was determined for predicted phenotype and for predicted drug activity.

For each drug, a permutation procedure was applied to test whether correlations with the measured phenotype differed from zero. Two thousand random data sets were generated under the null hypothesis by taking the measured phenotypes and randomly assigning them to the 50 observed genotypes. For each such data set, a random correlation was computed for each drug. One-sided \(P\) values were calculated by counting how frequently randomly generated correlations for a drug exceeded the observed correlation. A similar procedure was used to test whether interexpert correlations for phenotype and activity differed from zero. In this case, data sets were created under the null hypothesis by randomizing each expert’s actual ratings relative to the genotypes separately for each expert.

Bootstrap analysis was used to test whether individual expert predictions differed from the median score in terms of agreement with the measured phenotype. Five thousand bootstrap data sets were created by randomly resampling genotypes with replacement. The empirical bootstrap 95% CIs were calculated for the difference in percent agreement with the measured phenotype. \(P\) values for the differences were obtained by finding the highest confidence level at which the 95% CIs excluded a
difference of zero. The confidence level was subtracted from 1, and the result was doubled to obtain a 2-sided $P$ value.

RESULTS

An international panel of 12 experts interpreted 50 genotypes each (including 5 repeated genotypes per expert), resulting in a total of 600 evaluations for each antiretroviral agent. The genotypes from the clinical samples differed from the reference amino acid sequence at a median of 22 positions (range, 11–36 positions) in the reverse-transcriptase sequence and 11 positions (range, 5–20 positions) in the protease sequence. All samples had at least 1 drug-resistance mutation for reverse-transcriptase inhibitors and at least 1 drug-resistance mutation for protease inhibitors. Mutations associated with resistance to nonnucleoside reverse-transcriptase inhibitors were present in 19 of the 45 unique genotypes; the M184V mutation was present in 29 unique genotypes (including 4 M/V mixtures at codon 184) (tables A1 and A2 in the Appendix list complete genotypes and phenotypes [online only]).

Predicting phenotype from genotype. Figure 1 illustrates the accuracy with which the expert panel predicted the phenotype fold change category for each of the 16 antiretroviral drugs on the basis of a genotype result. Figure 1 displays the percent of expert predictions that were correct (i.e., matched the measured phenotype fold change category), as well as the distribution of these phenotype predictions around the measured phenotype category. For nucleoside reverse-transcriptase inhibitors, accuracy of these predictions varied from 25% for abacavir to 74% for lamivudine; for protease inhibitors, accuracy varied from 26% for nelfinavir to 40% for lopinavir/ritonavir. Experts were able to predict the measured phenotype category correctly for each of the nonnucleoside reverse-transcriptase inhibitors approximately two-thirds of the time.

The experts had an overall accuracy of 44% (median range by expert, 32%–54%) for predicting the correct phenotype category. For all drugs except for didanosine, the ability of the experts to predict phenotype category was statistically significant when compared with a random prediction ($P = .42$).

The predictions of the experts were within 1 category of the measured phenotype category a median of 71% of the time, with a range of 53% (for abacavir) to 90% (for lamivudine) (figure 1). The experts generally predicted a higher degree of reduced susceptibility for the nucleoside analogues (the only exception was for lamivudine). For the protease inhibitors,
Figure 2. Agreement among experts in predicting phenotype fold change category, by drug. Dark bars indicate the percentage agreement between all possible pairs of experts, and light bars indicate the percentage agreement between pairs of experts within 1 fold change category of one another. ABC, abacavir; APV, amprenavir; AZT, zidovudine; DDI, didanosine; DDC, zalcitabine; DLV, delavirdine; D4T, stavudine; EFV, efavirenz; IDV, indinavir; LPV/R, lopinavir/ritonavir; NFV, nelfinavir; NVP, nevirapine; RTV, ritonavir; SQV, saquinavir; TDF, tenofovir; 3TC, lamivudine.

~~75% of the expert predictions were within 1 category above or below the measured phenotype category.

We also evaluated whether a consensus estimate of phenotype fold change category, based on the group median for all 12 experts, agreed more closely with the measured phenotype than did the individual predictions. For most of the antiretrovirals, the consensus estimate was not more accurate than the individual predictions given in figure 1 (data not shown).

Interest in agreement on predicting phenotype and antiretroviral activity. The percentage agreement between all possible pairs of experts and the percentage of all pair-wise comparisons that were within 1 category of each other are shown in figure 2. For the nucleoside and nucleotide reverse-transcriptase inhibitors, the percentage agreement between predictions of phenotype category ranged from 28% for abacavir to 73% for lamivudine, with a median agreement of 42% for the

Figure 3. Agreement among experts in predicting drug activity category with use of a 6-point scale. Dark bars indicate the percentage agreement between all possible pairs of experts, and light bars indicate the percentage agreement between pairs of experts within 1 fold change category of one another. ABC, abacavir; APV, amprenavir; AZT, zidovudine; DDI, didanosine; DDC, zalcitabine; DLV, delavirdine; D4T, stavudine; EFV, efavirenz; IDV, indinavir; LPV/R, lopinavir/ritonavir; NFV, nelfinavir; NVP, nevirapine; RTV, ritonavir; SQV, saquinavir; TDF, tenofovir; 3TC, lamivudine.
class. For the nucleoside and nucleotide analogues, the experts predicted phenotype within 1 category of each other 80% of the time. Percentage agreement between experts for predicted phenotype fold change category for the protease inhibitors ranged from 38% for nelfinavir and indinavir to 43% for amprenavir. Experts were able to predict within one category ~75% of the time for the protease inhibitors. The highest levels of agreement occurred for the nonnucleoside reverse-transcriptase inhibitors and ranged from 69% agreement for efavirenz to 79% agreement for nevirapine.

The level of agreement among experts in predicting drug activity on a 6-point scale was similar in range and pattern to that seen in predicting phenotype fold change category (figure 3). However, there was somewhat less agreement among experts in predicting drug activity for didanosine, zalcitabine, stavudine, and tenofovir, compared with predictions of phenotype category for these drugs.

**Treatment recommendations.** Experts were asked whether they recommended each of 16 antiretroviral agents on the basis of the genotype result. Consensus among experts on treatment recommendations ranged from 62% agreement for zalcitabine to 90% agreement for lamivudine, with a median consensus of 79% (figure 4). Experts agreed 77% of the time to use lopinavir/ritonavir and 77% of the time to use tenofovir, and they agreed 75% of the time not to use zidovudine and 74% of the time not to use nelfinavir. Interestingly, these experts agreed 69% of the time not to recommend lamivudine. When the M184V mutation was present, lamivudine was only recommended 1.3% of the time by this panel of experts. In contrast, when no M184V mutation was present, experts recommended lamivudine 76% of the time. Agreement between experts on treatment recommendations was >80% for zidovudine, lamivudine, the nonnucleoside reverse-transcriptase inhibitors, saquinavir, nelfinavir and lopinavir/ritonavir. Levels of disagreement were higher for recommending abacavir (32% disagreement), amprenavir (32% disagreement), didanosine (31% disagreement), stavudine (30% disagreement), indinavir (30% disagreement), and ritonavir (29% disagreement) (figure 4).

**Intrarater consistency.** Experts’ rating of expected drug activity was highly correlated with prediction of phenotype fold change. Correlations ranged from 0.55 for stavudine to 0.99 for neviripine and delavirdine, with a median correlation of 0.90. The correlation was ≥0.75 for all antiretrovirals except stavudine. Predicted phenotype categories and treatment rec-
Figure 5. Consistency of expert predictions over time, by drug. Bars indicate the mean percentage agreement between initial and second expert predictions, based on analysis of 5 duplicate genotypes, of drug activity category, phenotype fold change category, and treatment recommendations. 3TC, lamivudine; ABC, abacavir; APV, amprenavir; AZT, zidovudine; D4T, stavudine; DDI, didanosine; DDC, zalcitabine; DLV, delavirdine; EFV, efavirenz; IDV, indinavir; LPV/R, lopinavir/ritonavir; NFV, nelfinavir; NVP, nevirapine; RTV, ritonavir; SQV, saquinavir; TDF, tenofovir.

Several studies have compared the various interpretation systems for drug resistance in HIV-1 that are available to clinicians, but there has been no systematic evaluation of experts’ interpretations of HIV-1 genotypes [13–17]. In this comparison of genotype interpretations, an international panel of experts was generally able to predict the measured phenotype fold change category on the basis of a genotype 30%–40% of the time for most antiretroviral drugs. The exceptions were lamivudine (for which experts correctly predicted the measured phenotype 75% of the time) and the nonnucleoside reverse-transcriptase inhibitors (for which the experts correctly predicted the phenotype category two-thirds of the time). Furthermore, the experts in this study were able to predict phenotype fold change within 1 category 70% of the time. It is important to note that the phenotype fold categories that the experts were asked to predict were relatively narrow (see Methods). When we collapsed the phenotype fold categories into 3 broader categories (<4-fold change, 4–10-fold change, and >10-fold change), overall accuracy improved to a median of 66%, with a range of 45% for abacavir to 86% for tenofovir (data not shown). However, in the case of certain nucleoside reverse-transcriptase inhibitors, even the relatively narrow categorical ranges used in this study might encompass values above and below recently established clinical cutoff values that are based on clinical response data [18–21].

From these results, it is clear that predicting the susceptibility of HIV-1 to drugs for which a single-point mutation confers...
high-level resistance is relatively straight forward, and agreement among experts tends to be high. However, this group of experts achieved accurate predictions of susceptibility less often for drugs that require multiple mutations for resistance to develop. Predictions of phenotype on the basis of genotype and the reliability of expert interpretation of genotypes are further complicated by variability in the phenotype assay, the relatively narrow dynamic range of phenotypes seen with certain nucleoside reverse-transcriptase inhibitors (e.g., stavudine), and incomplete understanding of resistance and cross-resistance patterns (particularly for newer drugs) [18–21].

The ability to predict the phenotype fold change category was not significantly improved for most of the antiretrovirals when the expert predictions were averaged together and then compared with the measured phenotype, although the median prediction was somewhat more accurate for lamivudine and nevirapine than for other drugs. We had anticipated that an averaged prediction might, in fact, be more accurate than individual predictions, but this was not the case.

Experts tended to overestimate the phenotype fold change category for the nucleoside reverse-transcriptase inhibitors. The overestimation of phenotype fold change category does not necessarily mean that the experts are wrong in their predictions of the clinical activity of these drugs. It may be that, for certain drugs for which phenotypic resistance is difficult to measure, genotypic information provides a more accurate predictor of drug activity. However, we were not able to test this hypothesis in the current study because of the lack of outcome data. Studies that compare expert predictions of outcome made on the basis of genotype versus predictions made on the basis of phenotype would be important next steps in improving our understanding of the relative predictive capacity of genotype, expert interpretation, and phenotype testing.

This panel of international experts tended to agree with one another to a significant extent. IntereXpert agreement in predicting phenotype fold change category was greater than individual or group agreement between predicted and measured phenotype (figures 1 and 2). This level of agreement occurred despite the fact that the 45 genotypes used in this study were highly complex and that no guidance regarding interpretation was provided to the expert panel. These findings suggest that a consensus is emerging within the expert community regarding the interpretation of genotype data.

Furthermore, experts tended to agree with one another when predicting drug activity on the basis of viral genotype. The lack of data on the virologic outcomes of the patients from whom the viral sequences were obtained prevented us from determining the extent to which predicted drug activity correlated with treatment success. It was somewhat surprising that the experts were no more likely to agree on drug activity than on phenotype fold change category. For didanosine, stavudine, abacavir, and tenofovir, there still appears to be a fair amount of disagreement, even among experts, regarding drug activity given a particular set of mutations. This observation most likely reflects the difficulty of identifying clinical cutoff values for didanosine and stavudine, as well as the fact that most experts had more-limited clinical experience with abacavir and tenofovir, compared with other drugs, at the time of this study.

This expert panel tended to agree on treatment recommendations to a large extent. Overall, the panel agreed nearly 80% of the time on treatment recommendations made on the basis of the genotype result. Experts tended to agree on using lopinavir/ritonavir and tenofovir for those genotypes that tended to have a relatively large number of mutations. Levels of disagreement about treatment were higher for abacavir, amprenavir, didanosine, and stavudine; in part, this likely reflects the difficulties with resistance interpretation discussed above.

The correlations of predicted phenotype and drug activity with treatment recommendations were generally in the 70%–80% range. The major exception was tenofovir. In general, the experts recommended the drug without much regard to the predicted activity or phenotype. This finding may reflect the experts’ relative lack of experience with a drug that was newly approved at the time of this study. Data relating activity of tenofovir to specific mutational patterns in HIV-1 became available only after the conclusion of this study [18].

In summary, we found that this panel of international experts correctly predicted phenotype from complex genotypes within a fairly narrow range 30%–40% of the time. There also appeared to be a general consensus among the panel members in terms of their prediction of phenotype and drug activity on the basis of genotype. Finally, the experts displayed a high degree of concordance in treatment recommendations made solely on the basis of genotype. Treatment recommendations were also very consistent over time. We are encouraged by these results, because it appears that there are reasonably good levels of agreement in the expert community regarding the interpretation of genotype for making treatment recommendations, and the recommendation of treatment is the critical component of resistance test interpretation that is sought by the average clinician.

Acknowledgments

Financial support. Stanford AIDS Clinical Trials Group (A127666); Cornell University K24 (AI 051966); National Institutes of Health (AI 27670, AI 38858, AI 43638, and AI 29164); University of California, San Diego, Center for AIDS Research (AI 36214); and the Research Center for AIDS and HIV Infection of the San Diego Veterans Affairs Healthcare System.

Potential conflicts of interest. A.R. is an employee of Tibor-Virco. All other authors: no conflicts.
References