Clinical Pharmacology Quality Assurance (CPQA) Program: Models for Longitudinal Analysis of Antiretroviral (ARV) Proficiency Testing for International Laboratories

Robin DiFrancesco, MBA, MT[1], **Susan L. Rosenkranz, PhD**[2], **Charlene R. Taylor, BS, MT**[1], **Poonam G. Pande, PhD**[3], **Suzanne M. Siminski, MBA**[2], **Richard W. Jenny, PhD**[4], and **Gene D. Morse, PharmD**[1]

[1]HIV Clinical Pharmacology Research Program, Translational Pharmacology Research Core, Center of Excellence in Bioinformatics and Life Sciences; School of Pharmacy and Pharmaceutical Sciences, University at Buffalo, Buffalo, NY USA

[2]Frontier Science Technology and Research Foundation, Amherst, NY USA

[3]RTI International, Research Triangle Park, NC USA

[4]New York State Department of Health, Albany, NY USA

Abstract

Among National Institutes of Health (NIH) HIV Research Networks conducting multicenter trials, samples from protocols that span several years are analyzed at multiple clinical pharmacology laboratories (CPLs) for multiple antiretrovirals (ARV). Drug assay data are, in turn, entered into study-specific datasets that are used for pharmacokinetic analyses, merged to conduct cross-protocol pharmacokinetic analysis and integrated with pharmacogenomics research to investigate pharmacokinetic-pharmacogenetic associations. The CPLs participate in a semi-annual proficiency testing (PT) program implemented by the Clinical Pharmacology Quality Assurance (CPQA) program. Using results from multiple PT rounds, longitudinal analyses of recovery are reflective of *accuracy* and *precision* within/across laboratories. The objectives of this longitudinal analysis of PT across multiple CPLs were to develop and test statistical models that longitudinally: (1)assess the precision and accuracy of concentrations reported by individual CPLs; (2)determine factors associated with round-specific and long-term assay accuracy, precision and bias using a new regression model. A measure of absolute recovery is explored as a simultaneous measure of accuracy and precision.

Overall, the analysis outcomes assured 97% accuracy (±20% of the final target concentration of all (21)drug concentration results reported for clinical trial samples by multiple CPLs).Using the CLIA acceptance of meeting criteria for 2/3 consecutive rounds, all ten laboratories that participated in three or more rounds per analyte maintained CLIA proficiency. Significant associations were present between magnitude of error and CPL (Kruskal Wallis [KW]p<0.001), and ARV (KW p<0.001).

CORRESPONDANCE/REPRINTS: Robin DiFrancesco, 701 Ellicott Street, Buffalo, NY, 14203, Phone: 716-881-7465, Fax: 716-849-6890, rda@buffalo.edu.

CONFLICTS OF INTEREST
No conflicts of interest are declared.

Keywords

longitudinal analysis model; antiretroviral proficiency testing; inter-laboratory performance

INTRODUCTION

The quality of clinical pharmacology in HIV research is of global concern. The National Institutes of Health support multiple HIV clinical research networks and investigator-initiated clinical trials to sponsor the investigation of treatment paradigms in diverse international populations of HIV+ and HIV− individuals. Most of these investigations involve the use of drug treatment for infected patients as well as the use of drugs for the prevention of infection in those at risk. The pharmacology objectives of these research studies require extensive bioanalysis for drug concentrations. Furthermore, cross-protocol analyses frequently employ these drug concentrations to address additional questions about pharmacokinetics, pharmacodynamics and pharmacogenetics. Because these studies impact the strategies used for treatment and prevention, the quality of the drug measurements is critical.

Through the Clinical Laboratory Improvement Act (CLIA) of 1988, the US federal government mandated that clinical laboratories participate in and demonstrate satisfactory performance in proficiency testing (PT) programs [1]. However, for clinical pharmacology laboratories, unless the drug concentration is used clinically to treat the patient, the assay of drug concentrations in plasma and other matrices performed for drug development and research purposes falls outside the CLIA mandate. The European proficiency testing program *International Inter laboratory Quality Control Program for Measurement of Antiretroviral Drugs in Plasma*, has been conducted for more than a decade. The program was established by Radboud University Nijmegen Medical Center, The Netherlands, and later continued in collaboration with the Dutch Association for Quality Assessment in Therapeutic Drug Monitoring and Clinical Toxicology. This program was needed in Europe as therapeutic drug monitoring (TDM) for antiretrovirals (ARV) was a standard in clinical treatment for HIV+ patients receiving ARV [2, 3]. The European program analyzed laboratory PT results and published results in 2002, 2003 and 2010[1][4–6]. The most recent report was for ten years of PT results, exhibiting 84% accuracy, and somewhat lower accuracy(78%) in samples at lower prepared concentrations. In addition, accuracy varied by individual ARV. A multivariate model of these two factors was not fit; however, it was apparent that those ARVs with higher therapeutic concentration ranges (efavirenz, lopinavir) were more likely to have greater accuracy than those with lower therapeutic concentration ranges (saquinavir, atazanavir). A similar program has been implemented by this organization for antifungals [7].

A US-based ARV PT program, operated by the ACTG/IMPAACT from 2000–2008, has also published PT program results [8, 9]. This program was a collaboration of multiple university CPLs analyzing ARV concentrations in pediatric and adult clinical trials. Two of the participating universities prepared and shipped PT samples with one of the two coordinating the compilation and distribution of reports. The program maintained an overall accuracy of 96%.The program also concluded that the measure of accuracy was dependent on the ARV and laboratory. Initial data indicated that the concentration of the ARV, classified as "low""medium" or "high" spike per ARV, was a significant variable. However after additional PT results were obtained, concentration was no longer found to be a significant variable. This program was discontinued in 2008.Table 1 summarizes some key

---

[1]Dutch Association for Quality Assessment in Therapeutic Drug Monitoring and Clinical Toxicology.

characteristics, methods and performance results reported in the literature for the US and European ARV PT programs. Many fundamental differences exist between the two programs, such as number of samples per PT event (herein referred to as "round")and the statistical methods employed in the analyses.

The Clinical Pharmacology Quality Assurance and Quality Control (CPQA) program was established in 2008 and provides multiple quality-centered activities for all NIAIDHIV Clinical Trial Networks globally. Quality strategies include: (1)training curricula for clinical sites conducting studies and for laboratories analyzing specimens; (2)confidential, anonymous, peer-review of assay validation reports submitted by participating CPLs for approval to assay clinical trial pharmacology samples ;(3)a phased laboratory assessment system, and; (4)a PT program for ARV measurements. This program was established in May 2008 and is currently ongoing. The CPQA PT program includes multiple CPLs that report ARV concentrations; for each ARV 2 to 10 CPLs reported results during the first 2 years of testing. Although all CPLs use liquid chromatography, the reagents, calibrators, equipment, consumables and detection method differ. Recognizing the lack of statistical models used to quantify variances in laboratory PT programs such as this, CPQA developed new models testing differences in recovery among CPLs and analytes, to contrast two alternate ways of assigning target value and to compare regulatory and model-based statistical limits of acceptance.

Most often, the drug quantitation tests to which the CLIA PT mandate applies include "high volume" tests such as phenytoin and digoxin. For such tests, the number of participating labs is large enough that PT performance assessments are applied separately to different "peer groups," that is, labs that use common (or very similar) measurement procedures [1, 10]. In simulation studies, researchers have evaluated both the original CLIA rules for PT evaluation and newly proposed rules for diagnostics that individual labs might apply to their own PT data to assess and track systematic and random error in their tests. Such rules are defined by (1) the assignment of target values, (2) the acceptance criterion used (eg, regulatory, statistical or clinical) and (3) how scores for the 5 individual PT samples are reduced to the single test result of "satisfactory" or "unsatisfactory". These simulation studies have yielded important findings regarding which rules work best in situations of high- and low-noise tests; however, these studies have assumed a large enough number of participating labs that group means, medians and standard deviations are not affected by a single lab whose measurement procedure exhibits high systematic or random error [1, 11, 12].

The CPQA primary goal for PT is to assure the accuracy of CPL reported concentrations (RCs); PT reports flag deviations from target concentrations in the current round and assign each CPL a score based on performance in the current and 2 prior rounds. However, goals of the program and the performance of the CPLs may be enhanced by a comprehensive, longitudinal analysis of ARV PT data, the analysis of which may suggest ways to improve the policies and practices of both laboratories and the PT program. The objectives of this longitudinal analysis of PT results across multiple CPLs and rounds were to develop and test statistical models, applied to longitudinal data, in order to (1)assess the precision and accuracy of concentrations reported by individual CPLs; and (2) determine factors associated with round-specific and long-term assay accuracy, precision and bias. In addition, the metric of absolute recovery was explored as a simultaneous measure of accuracy and precision. The outcomes achieved by the proposed models were interpreted for PT relevancy and impact.

METHODS

Every six months the CPQA PT program offered prepared plasma samples containing pre-specified concentrations (unknown to CPLs) of up to 21 ARV analytes: abacavir (ABC), amprenavir (APV), atazanavir (ATV), darunavir (DRV), didanosine (DDI), efavirenz (EFV), emtricitabine (FTC), etravirine (ETR), indinavir (IDV), lamivudine (3TC), lopinavir (LPV), maraviroc (MVC), nelfinavir (NFV), nevirapine (NVP), raltegravir (RGV), ritonavir (RTV), saquinavir (SQV), stavudine (D4T), tenofovir (TFV), tipranavir (TPV), zidovudine (ZDV). In each round and for each ARV, 5 concentrations, spanning an expected therapeutic range of each ARV, as well as occasional concentrations below or above, were provided. Samples are prepared by an outside subcontractor and tested by the CPQA lab prior to distribution. PT samples were stored at $-70 \pm 15°C$ and then shipped on dry ice to participating laboratories with detailed instructions. Upon arrival, each laboratory confirmed sample integrity and indicated planned reporting of specific analytes. Results were reported either through an online Laboratory Data Management System (LDMS) or via a template which was then uploaded into the LDMS database. At the end of the submission period, a completeness evaluation was performed to confirm that all planned results were received; discrepancies were queried for resolution. To summarize the proficiency of individual labs, a pre-specified scoring algorithm was applied to the RCs (see next paragraph). The scoring algorithm reflects US Clinical Laboratory Improvement Act (CLIA) PT regulations[13]. After review and approval by the CPQA advisory board chair, a final report was sent to the participating laboratories (with laboratories de-identified) and key leadership (laboratories identified per network leader).

An individual RC is deemed Acceptable provided a concentration is present where expected, and the concentration is within 20% of the final target (FT)[14]. (If a concentration is reported as below the lower limit of quantification (BLQ), and the run lower limit was below 80%*FT, the concentration was labeled Unacceptable.) For a given prepared sample, if the number of labs reporting for that sample is large enough, the variability between CPLs is small enough (15%) and the percent deviation of the group mean (GM, determined after removal of outliers, if any)from the weighed-in value (WIV) is >5%, the FT is set to the GM. Otherwise, FT is set to the WIV. At the analyte level, a CPL'sperformance is deemed Satisfactory for the round provided at least 80% of RCs are Acceptable. If the CPL score is <80% for an analyte, the CPL submits a corrective action plan to reestablish accuracy; a root cause is requested. Finally, in accordance with CLIA rules, a lab is classified as successful for an analyte provided the round-specific score was Satisfactory in at least 2 of the last 3 rounds (including the current).

Enumeration

The results of the first four CPQA-operated PT rounds were tabulated to characterize participation and analyte inclusion. Characteristics included: number of CPLs per analyte tested, the frequency with which FT=WIV (as opposed to GM), the number of results per analyte per CPL, and overall percentage of Acceptable results per laboratory. The number of analyte scores and the number of CPLs associated with Unsatisfactory scores were determined; root causes of Unsatisfactory performance, as reported by CPLs, were also summarized.

Statistical analysis

The primary measures analyzed were log recovery and absolute recovery, where log recovery is defined as the natural (base e) log of the ratio of RC to nominal concentration (NC). Two versions of NC were considered: WIV and FT (where FT is either WIV or GM, determined as described above) and absolute recovery is defined as $100*[NC + abs(NC-$

RC)]/NC. Use of(log or absolute) recovery metric allowed pooling of data across CPLs, analyte and NC. Log recovery was also approximately normally distributed, allowing application of least-squares linear regression and related procedures analysis of variance (ANOVA) and covariance (ANCOVA).

Bias (systematically high or low recovery) was assessed via linear regression models of $\log(RC_i) = \beta_0 + \beta_1 \log(NC_i) + \epsilon_i$, where i indexes a given RC, β_0 represents the intercept (expected difference between RC and NC on the log scale), β_1 represents the slope (capturing how RC may vary from NC as NC changes), and ϵ_i is random error, assumed distributed $N(0, \sigma^2)$. In the absence of preparation and assay error (and with NC=WIV), β_0 and β_1 would take values 0.0 and 1.0. We refer to the regression models when NC=WIV and NC=FT as Models 1A and 1B, respectively. As an alternate strategy for classifying results as Acceptable (or not), we calculated 95% prediction limits around estimated regression equations; RCs falling outside these limits were deemed Unacceptable.

To assess factors including the ARV, within-ARV concentration category (low/medium/ high for the ARV), across-ARV concentration category (WIV in ranges: 30–200, 200–500, 500–1200, 1200–3200, 200–8000 ng/mL), CPL (lab identity) and round, one factor at a time, the regression model was extended to ANCOVA: $\log(RC_i) = \beta_0 + \beta_1 \log(NC_i) + \beta_{2,i}X_{2,i} + \epsilon_i$ where i indexes a given RC, β_{2i} represents the expected displacement from β_0 when $X_{i,2} = 1$, $X_{i,2}$ represents an indicator variable (0,1) for the ith level of explanatory factor X_2 and β_0 β_1 and ϵ_i are as above. We refer to this ANCOVA model when NC=WIV as Model 2A, and where NC=FT as Model 2B. If the overall F-test for X_2 indicated a statistically significant association with bias ($p<0.05$), Tukey-Kramer adjusted estimates of all pairwise differences between factor levels were calculated. 95% confidence intervals (CIs) around recovery for a given factor level were calculated as ($100*\exp(b_L)$, $100*\exp(b_U)$), where b_L and b_U represent lower and upper bounds of the 95% CI around factor level parameter estimates b_{2i} of β_{2i}.

To assess the magnitude of error, the metric was absolute recovery. Since a regression model is not appropriate for unsigned deviations, the Kruskal-Wallis (KW) procedure was used to test the equality of medians. If the KW test indicated there was a statistically significant difference among factor levels, to identify which differences were statistically significant, the following model was fit: $Y_i = \beta_0 + \beta_1 \log(NC_i) + \beta_2 R_i + \epsilon$ where Y_i and R_i are the predicted and observed ranks of absolute recovery, respectively, and other symbols are as defined above. We refer to this rank-based ANOVA model as Model 3A when NC=WIV, and as Model 3B, when NC=FT. If KW indicated a factor was associated with absolute recovery, Tukey-Kramer adjusted tests were applied to ranks of recoveries to identify statistically significant pairwise differences between factor levels.

RESULTS

Table 2 summarizes the results of the algorithm used to determine FTs. For 47% of unique samples, the WIV was chosen as the FT. The percent of samples with FT=WIV varied across analytes, and ranged from 0 to 100%. For analytes 3TC, DDI, DRV, NVP, RGV and TPV, the WIV was used for 20% or fewer samples; for all other analytes, the WIV was used for at least 40% of samples. In the case of DDI, the selection of GM as the FT was driven by a small number of RCs (DDI was offered in only one round). In the case of the more frequently reported DRV and NVP, choice of GM was driven by large deviations between RCs and the respective WIVs.

Table 3 enumerates the number of results per analyte per CPL and also provides totals by analyte and CPL. Ten laboratories participated in the PT program, and reported results for 6

to 20 ARV. Six CPLs participated in all four rounds. CPLs 1, 3, 8 and 9 began participation at later dates. CPL participation per ARV ranged from 20% (2 CPLs reporting for ETR) to 100%(many CPLs). ATV, DRV, EFV, FTC, LPV, NFV, NVP, RTV, and TFV were reported more frequently than the other ARV (with 90 results reported and at least 7 CPLs reporting for each ARV). For each drug, 95% of RCs were Acceptable except for FTC (90%), NFV (89%) and TFV (92%). Pooling RCs over all rounds, 97% were Acceptable. All RCs for CPLs 5, 8 and 10 were Acceptable; for CPLs 3 and 6, 91% and 94% of RCs were Acceptable; and for the remaining CPLs, at least 97% of RCs were Acceptable.

Satisfactory scores were achieved for 97% of CPL analytes/rounds. All scores for CPLs 1, 5, 8 and 10 were Satisfactory; for CPL 3, 91% of scores were Satisfactory; and for the remaining CPLs, at least 95% of scores were Satisfactory. Similarly to RCs, Satisfactory scores of <95% were seen for the analytes FTC (86%), NFV (83%) and TFV (83%).

Six of the ten CPLs failed to meet criteria for one or more analyte in one or more rounds (not tabulated). These CPLs submitted nine analyte-specific corrective action plans for failure to meet proficiency standards during a round. Root cause analyses by CPLs indicated that three main sources of error occurred: technical/human error (40%), assay calibration problem (33%), and failure to follow assay standard operating procedure (27%). Laboratories that participated in 3 or more rounds per analyte maintained CLIA proficiency[2].

Of the 1,706 results, ten were omitted from the statistical analyses due to errors unrelated to assay performance. These errors included clerical errors such as incorrect or reversed transcription and extrapolation over the calibration range limit. Thus, 1,696 concentrations were included in analyses.

Models to assess bias

Upon fitting regression Model 1A, the estimated intercept is positive but small ($b_0 \pm SE(b_0)$ 0.0039±0.0118, not statistically significant than zero). The slope estimate was significantly greater than 1.0 (Wald p-value < 0.001) and supports the general tendency of CPLs to report concentrations to be above the WIV. Model estimates of recovery for the lowest and highest WIVs, 30.6 and 84, 308 ng/mL, were 102% and 105%, respectively. When Model 1B is fit to the data, the estimated intercept is slightly lower (0.0036±0.0101, not statistically significant from zero), and the slope is closer to one (1.001 ± 0.0014) but still significantly different than 1.0 (Wald p-value <0.001). Model estimates of recovery for the extreme WIVs, 30.6 and 84, 308 ng/mL, were 100% and 101%, respectively.

Using Model 1A to assess acceptability of PT results, results shifted slightly from the criteria used by the PT program: 0.5% of the results (8 results) that were Unacceptable were inside the prediction limits, and 3.2% of the results that were Acceptable were outside the limits. As expected, compared to criteria used by the PT program, classifications based on Model 1B were in better agreement with the PT program criteria: 2% of Acceptable results were outside the prediction limits, and there were no reclassifications in the other direction.

Factors identified as significant in ANCOVA models varied between Models 2A and 2B. With Model 2A (NC=WIV), the variables ARV, laboratory identity, and WIV category were significant (F-test p 0.005 for all factors). Among ARVs, as shown in Table 4a, a number of statistically significant pairwise differences were identified. The majority of point estimates for recovery (relative to WIV; also shown in Table 4a) were above 100%, indicating a

---

[2]CPLs are required to maintain CLIA proficiency for the analytes reported for clinical trial specimens during the time frame of bioanalysis.

tendency to overestimate ARV concentrations. Among CPLs: estimated recovery for CPLs 6 and 7 were significantly higher than those of CPL2 and 5; CPL 7 recovery was also higher than those of CPLs 1, 3, 4, 8 and 9; and CPL 10 recovery was higher than that of CPL 2. In Figure 1a, it can be seen that median recoveries for CPLs 6 and 7 were well above 100%. The "high" WIV category (high within ARV) was associated with lower recoveries relative to the "medium" and "low" WIV categories (not shown).

Notably, using Model 2B (NC=FT), only laboratory identity was associated with biased recovery, with CPLs 6 and 7 exhibiting significantly higher recoveries than CPLs 2, 4, and 5. Differences between CPL 7's recovery and recoveries for CPLs 1, 8, and 9 were also statistically significant. Both CPL 6 and 7 had medians that were the higher than other CPLs (Figure 1b). In this figure, CPL 3 has the widest interquartile range (box), CPL 6 has the widest spread with a number of points outside the whiskers, and CPL 7 exhibits the highest median recovery.

Models to assess error

Using Model 3A (NC=WIV) to test associations of ranks of absolute recoveries with potential covariates, significant associations were present between magnitude of error and CPL (KW $p<0.001$), with CPLs 3, 6 and 7 exhibiting larger errors than the remaining labs (Figure 1c). Tukey-Kramer-adjusted p-values for all pairwise comparisons indicated that errors were larger for CPL 3, 6, and 7 compared to the other 6 CPLs. Significant associations were also present between magnitude of error and ARV (KW $p<0.001$). Figure 1d displays the relationship between ARV and magnitude of error. Median absolute recoveries were highest for DDI, DRV, IDV and RGV; for these ARVs, all medians were near 110% (not shown). Many pairwise differences (in ranks of absolute recovery) were statistically significant: DRV ranks were higher than those of MVC, ZDV, ETR, SQV, LPV, FTC, TPV, D4T, RTV, EFV and TFV; and RGV ranks were higher than those of MVC, ZDV, SQV, LPV, FTC, TPV and RTV. (Others pairs were significantly different; in the interest of brevity they are not listed.) Higher (ranks of) errors were seen in round 23 relative to rounds 24 and 25. Magnitude of error varied significantly by round (round 23 errors larger than those of rounds 24 and 25); errors did not differ significantly by (within- or across-ARV) concentration category.

Using Model 3B(NC=FT) to test associations of ranks of absolute recoveries with potential covariates, significant associations were again present between magnitude of error and CPL (KW $p<0.001$). CPL3, 6 and 7 again exhibit larger errors that the remaining labs when NC=FT(Figure 1d). In addition to these, CPL 3 exhibited larger errors than CPL 6.Significant associations were also present between magnitude of error and ARV (KW $p<0.001$): errors for DRV and NVP were larger than errors for RGV, RTV and ZDV; in addition, errors for DRV were larger than errors for MVC (not shown). Magnitude of error did not exhibit significant differences by round, or within or between ARV concentrations.

Table 5 summaries the associations of bias and magnitude of error with the covariates tested in the regression models. Across all 4 models, lab identity was a significant factor in both bias and error. ARV was a significant factor in bias when NC=WIV (model 2A), and in error (both models 3A and 3B).For three of the models, ARV was a significant factor whereas no bias was detected by ARV when NC=FT (model 2B).For the magnitude of error models, significant associations between bias and within-analyte concentration category were only noted when the WIV was used as the NC. The factor "round" was significantly associated with magnitude of error only, and only when NC=WIV.

DISCUSSION

The overarching aim of the PT program is to provide a means by which the stakeholders in HIV research (funding agency, investigators, patients, and health care providers) can be assured that pharmacology assays conducted for clinical trials are sufficiently accurate and reproducible so that minimal bias and minimal variability are introduced into ultimate findings and conclusions of these clinical trials. Currently, a twice-yearly PT program addresses overall and analyte-specific performance of individual labs conducting pharmacology assays, providing warnings for each round as needed if performance is unsatisfactory. Performance over three rounds is used to determine CLIA status and serves as a longitudinal view.

Without employing some type of statistical analysis, the general bias over four rounds, or two years, is not evident. In models examining the potential for bias, recovery relative to the nominal concentration, on the log scale, gave values that were approximately normally distributed. A linear regression model fit to these data identified an overall tendency of slight overestimation of concentrations (positive intercept estimate) and a global tendency for larger overestimation at larger nominal concentrations (positive slope estimate). Using a prediction interval around the estimated regression line provided an alternate to the current acceptance criteria. By this procedure an additional 54 (3.2% of 1706 reported concentrations) would be considered unacceptable, and 8 concentrations considered unacceptable by the current criteria would be deemed acceptable, changing the overall rate of acceptable concentrations from 97.6% to 94.9%. Using the group mean (mean over labs) as the nominal concentration (for about half of the prepared samples, as was done by the PT program) attenuated estimates of bias overall and as a function of concentration magnitude. Prediction intervals from the regression of recovery on final targets (i.e., when the group mean is used when use of the WIV is not indicated) would classify 95.6% of concentrations as acceptable (compared to 97.6% from the existing PT program criteria). These results suggest that a smaller window of acceptance should be considered. Continuous monitoring the PT data longitudinally and/or simulation studies will assist the CPQA in making this decision.

Pooling results over four rounds (but not yet applying statistical models), it appeared that CPL 3 and 6 had more difficulties than the other CPLs, as evidenced by lower rates of acceptable reported concentrations and of satisfactory scores (95%). Using regression models to examine associations of CPLs identified with bias indicated higher recoveries for CPL 6 and 7 (but not 3) compared to the other labs and CPL 2 was identified as having lower recoveries as compared to other CPLs. This finding held regardless of which nominal concentration was used (i.e., held for both models 2A and 2B). CPLs 3, 6 and 7 (but not CPL2) were also deemed to exhibit higher error (as measured by absolute recovery) compared to the other labs, regardless of which NC was used (models 3A and 3B).Pooling and tabulating results of the four rounds, ARVs TFV, FTC, and NFV appeared to be most challenging to assay, as evidenced by lower rates of acceptable reported concentrations and of satisfactory scores (<95%). However, using regression models to examine associations of analyte with bias, reported concentration for DRV and RGV were significantly higher than WIVs but not the final targets. Since recoveries using weighed-in value (Model 3A) were high, (ranks of) absolute recovery (errors) were noted to be largest for these ARVs as well. And although the final target was adjusted to the group mean for 90% of DRV and RGV specimens, the adjustment did not correct for error and DRV and RGV remained high in the ranks (1 and 3, respectively) of absolute error. NVP was identified as second in the rankings for model 3B inferring that in the absence of bias, error can still be large. If the ARV exhibits both positive and negative deviations from the final target, bias tends to cancel out, but persist in absolute recovery, indicating variability. In the presence of bias, errors can also

be large when using WIV or FT in which case the contribution of variability to error is difficult to tease out.

A discussion of the differences in findings by analysis approach is warranted. All three methods (pooling and tabulating, and regression-based assessments of bias in recovery and error (regardless of which NC was used) identified CPLs 3, 6 and 7 as having somewhat poorer performance than the other labs. However, conclusions about analytes differed between methods. Pooling and tabulating over the four rounds pointed to problems with the assay of FTC, NFV and TFV, whereas models of bias (recovery deviating from 100%) indicated problems with the assay of DRV and RGV when WIV was used as the NC, and no problems when FT was used as the NC. Models of error (ranks of absolute recovery) also pointed to problems with DRV and RGV assay when weighed-in value was used as the nominal concentration, and problems with DRV and NVP when FT was used at the NC. The large discrepancy between reported DRV concentrations and weighed-in values could reflect either preparation error or an error in assay methodology that is common across some but not all labs. While errors common across labs may seem less feasible, it was known that all labs use the same source for reference powders and prepare calibration standards in a similar fashion. Near the last of the four rounds included here, multiple laboratories (using ultraviolet detection) have reported secondary peaks emerging from their standard calibration stocks as a potential source of error; this was reported on CPQA-held cross-network conference calls of network CPL supervisors held every other month. For 90% of DRV samples, the group mean was used as the target, which "protected" labs from failing. In the case of preparation error, this protection is desirable. For the DRV results included here, this correction may be considered inappropriate. Hence, the choice of nominal concentration can be problematic. In contrast, although reported concentrations for FTC were near the WIV (unbiased), some labs did fail for the ARV. However, agreement among the majority of labs supported the accuracy of sample preparation and laboratory methodologies.

Other factors examined were round and 2 categorizations of ARV concentration. In most cases, these variables did not exhibit significant associations with bias and/or error, except for Model 2A which showed significant association with concentration (both categorizations) and Model 3A which showed significant association with round. In the latter case, the source or error most likely from arises from preparation of PT samples, where weighing for solutions to prepare PT samples are new from round to round.

While the regression models employed here may be valuable tools to diagnose potentially problematic assay factors, results from the model depend on which quantity is used as the nominal concentration. Use of the WIV as the nominal concentration has the desirable statistical property that it is independent of the dependent variable "reported concentration." However, in the event of preparation error, the WIV itself is biased from the true but unknown concentrations in the PT samples. When the number of CPLs reporting for an analyte is large enough (4 or more), when the variability among labs in reported values is small (15%) and when the deviation of the group mean (mean reported concentration across labs) is large (>5%), the group mean is used as the target concentration – the determination of acceptable/unacceptable is relative to this group mean, not to the WIV concentration. Substituting a data-driven target for the prepared concentration, in these instances, is designed to prevent application of the "unacceptable" label when there was preparation error. However, a drawback in using this final target as the so-called independent variable is that, when the group mean is substituted, the independent variable is not statistically independent from but is based on (subsets of) reported concentrations (for some samples), the dependent variable. Furthermore, this substitution could give an advantage to some labs. For example, given the discussion of DRV PT above, if multiple

labs used the problematic reference powder and a single lab used a no problematic reference powder, the single lab could be penalized by this choice of nominal concentration. Future research will investigate the use of the median reported concentration as the independent variable; this is the approach taken by the DAIDS Virology Quality Assurance program [15].

The data at hand are unbalanced, in the sense that labs participated for different subsets of analytes and in a different numbers of rounds. If recovery truly varies by analyte, then differential participation by labs (some labs reporting for "problematic" analytes and others not) would tend to unfairly penalize the former type of lab when using a statistical model that does not simultaneously estimate the effects of analyte and lab. In future research, we will examine models that account for multiple factors (e.g., both analyte and lab) simultaneously. These models will need to use appropriate estimation and inference techniques in the situation of unbalanced data.

CONCLUSIONS

Given that a multiplicity of variables both within and outside of laboratory control can affect results, proficiency testing in the context of drug assays for HIV clinical trials is not a pure measure of laboratory-specific accuracy or reproducibility. Application of statistical models to multiple rounds of proficiency testing data sheds light on the potentially inaccurate conclusions made by observing only "Acceptable" or "not Acceptable," and focuses attention on bias and error as sources of problematic laboratory (or laboratory method) performance. Using two different measures of nominal concentration (WIV or FT as determined by a data driven algorithm), the complexity of identifying the relationships between bias and error, as well as laboratory and ARV, is apparent.

With respect to the current information these models have provided, laboratory and analyte remain the most significant variables whether they are associated or not. It is noteworthy that the CPLs maintained CLIA proficiency for ARV quantitation with excellent scores. This approach to longitudinal analysis of proficiency testing provides novel insights into inter-laboratory characteristics that are key components to the crucial role of the CPLs in improving the quality of assay performance and for using data from different studies to accomplish the clinical pharmacology research goals of the NIAID HIV research networks.

Acknowledgments

REFERENCES

1. Miller WG, Jones GR, Horowitz GL, Weykamp C. Proficiency testing/external quality assessment: current challenges and future directions. Clin Chem. 2011 Dec; 57(12):1670–1680. [PubMed: 21965556]

2. Droste JA, Koopmans PP, Hekster YA, Burger DM. TDM: therapeutic drug measuring or therapeutic drug monitoring? Ther Drug Monit. 2005 Aug; 27(4):412–416. [PubMed: 16044095]

3. Aarnoutse RE, Schapiro JM, Boucher CA, Hekster YA, Burger DM. Therapeutic drug monitoring: an aid to optimising response to antiretroviral drugs? Drugs. 2003; 63(8):741–753. [PubMed: 12662123]

4. Aarnoutse RE, Verweij-van Wissen CP, van Ewijk-Beneken Kolmer EW. International interlaboratory quality control program for measurement of antiretroviral drugs in plasma. Antimicrob Agents Chemother. 2002 Mar; 46(3):884–886. [PubMed: 11850280]

5. Droste JA, Aarnoutse RE, Koopmans PP. Evaluation of antiretroviral drug measurements by an interlaboratory quality control program. J Acquir Immune Defic Syndr. 2003 Mar 1; 32(3):287–291. [PubMed: 12626888]

6. Burger D, Teulen M, Eerland J. The International Interlaboratory Quality Control Program for Measurement of Antiretroviral Drugs in Plasma: a global proficiency testing program. Ther Drug Monit. 2011 Apr 33.(2):239–243. [PubMed: 21383652]

7. Brüggemann RJ, Touw DJ, Aarnoutse RE, Verweij PE, Burger DM. International interlaboratory proficiency testing program for measurement of azole antifungal plasma concentrations. Antimicrob Agents Chemother. 2009 Jan; 53(1):303–305. [PubMed: 19015363]

8. Holland DT, DiFrancesco R, Stone J. Quality assurance program for clinical measurement of antiretrovirals: AIDS clinical trials group proficiency testing program for pediatric and adult pharmacology laboratories. Antimicrob Agents Chemother. 2004 Mar; 48(3):824–831. [PubMed: 14982771]

9. Holland DT, DiFrancesco R, Connor JD. Quality assurance program for pharmacokinetic assay of antiretrovirals: ACTG proficiency testing for pediatric and adult pharmacology support laboratories, 2003 to 2004: a requirement for therapeutic drug monitoring. Ther Drug Monit. 2006 Jun; 28(3): 367–374. [PubMed: 16778721]

10. Steele BW, Wang E, Palomaki GE, Klee GG, Elin RJ, Soldin SJ, Witte DL. An evaluation of analytical goals for assays of drugs. Arch Pathol Lab Med. 2001; 125:729–735. [PubMed: 11371222]

11. Carey RN, Cembrowski GS, Garber CC, Zaki Z. Performance characteristics of several rules for self-interpretation of proficiency testing data. Arch Pathol Lab Med. 2005; 129:997–1003. [PubMed: 16048412]

12. Cembrowski GS, Hackney JR, Carey N. The detection of problem analytes in a single proficiency test challenge in the absence of the Health Care Financing Administration rule violations. Arch Pathol Lab Med. 1993; 17:437–443. [PubMed: 8466413]

13. Medicare, Medicaid and CLIA programs; regulations implementing the Clinical Laboratory Improvement Amendments of 1988 (CLIA). Fed Regist. 1992 Feb 28.57:7002–7186. [PubMed: 10170937]

14. Jenny RW, Jackson-Tarentino KY. Causes of unsatisfactory performance in proficiency testing. Clin Chem. 2000 Jan; 46(1):89–99. [PubMed: 10620576]

15. [Accessed December 4, 2012] VQA Proficiency Testing Scoring Document for Quantitative HIV-1 RNA [HANC HIV/AIDS Network Coordination Web site]. 2006 Mar 15. Available at: https://www.hanc.info/labs/labresources/vqaResources/ptProgram
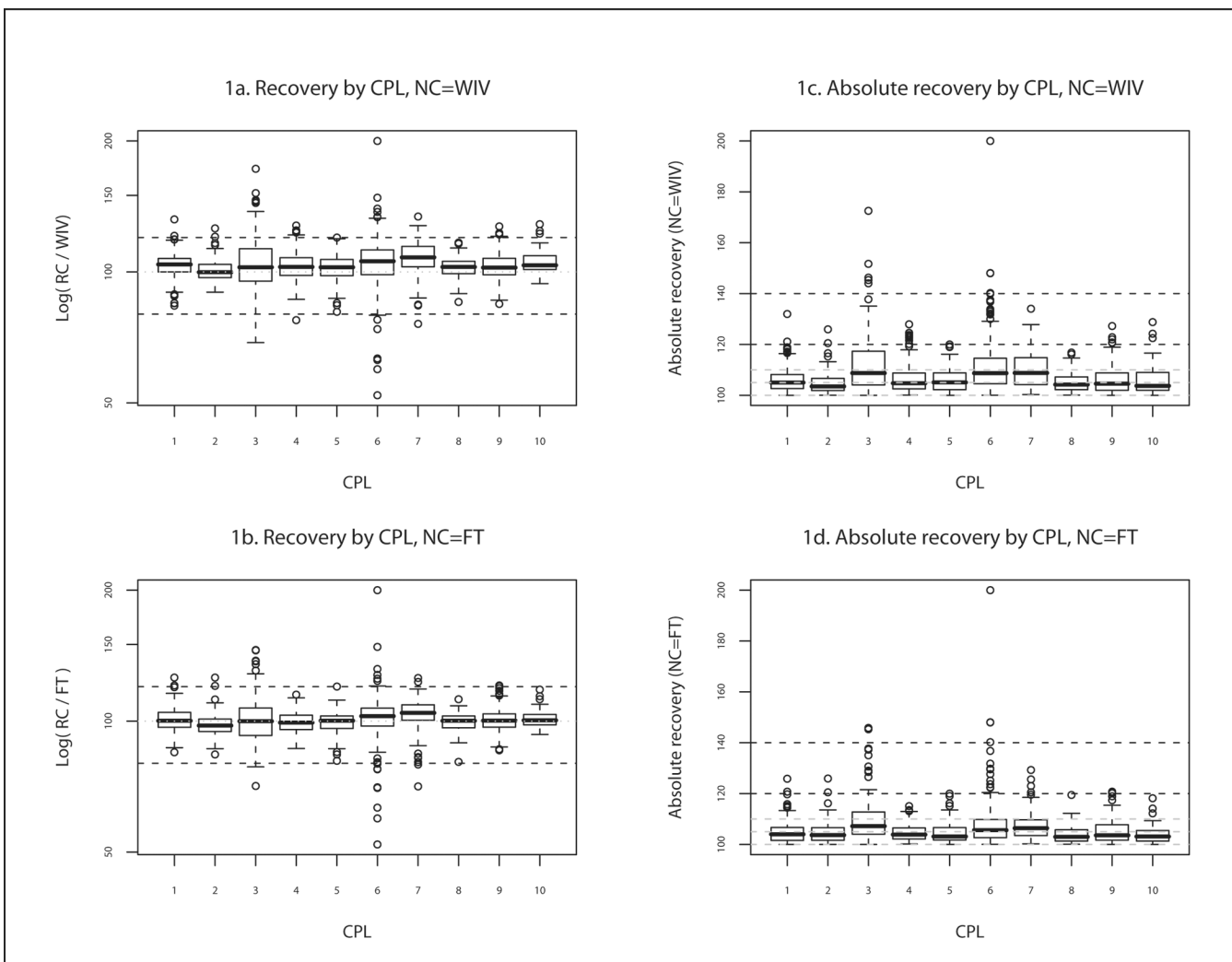
Figure 1. Box plots of log recoveries and absolute recoveries by CPL, for two different types of nominal concentrations

Upper and lower left (1a and 1b): Box plots of log recovery, defined as natural log of reported/nominal concentration (RC/NC), where NC is defined as the weighed-in value (WIV) and the final target (FT) in the upper and lower plots, respectively. Dotted gray line at 100% indicates perfect agreement between RC and NC. Dashed black lines indicate recoveries of 80% and 120%.

Upper and lower right (1c and 1d): Box plots of absolute recovery, defined as 100* [NC + abs(NC-RC)]/NC . Dashed gray lines indicate absolute recoveries of 100%, 105% and 110%; dashed black lines indicate absolute recoveries of 120% and 140%. The box's bold center line is the median. The upper and lower edges of box represent the 25th and 75th percentiles. The whiskers extend to either: (1) 1.5 times the interquartile range (the distance between 25th and 75th percentiles) or the (2) most extreme point, whichever is closer to the box. Points outside the whiskers are displayed individually. These points are outliers relative to the box plot, but without positing a particular distribution for the concentrations, no probabilistic statement is inferred.

Table 1

Prior Antiretroviral (ARV) Proficiency Testing Programs, Characteristics and Findings

	European (ref 1, 2002)	European (ref 2, 2003)	US (ref 4,2004)	US (ref 5,2006)	European (ref 3, 2011)
Enumerations					
ARV (PI, NNRTI, NRTI)	4 (4,0,0)	8 (6,2,0)	8 (6,2,0)	14 (7,2,5)	8 (6,2,0)
levels /ARV	3	3	6	6	3
labs (rounds)	9 (1)	30 (3)	8 (3)	9 (4)	56 (20)
results	88	193	448	1688	12798
acceptability	2/3 WIV	2/3 WIV	5/6 FTV	5/6 FTV	2/3 WIV
% results passing	36–80%	82%	96%	96%	84%
Statistically Significant variables					
Laboratory	-	yes	yes	yes	-
ARV	-	yes	yes	yes	yes
Concentration	-	yes	yes	no	yes
Round/event	-	-	-	no	-
HPLC-UV vs -MS	yes	yes	-	yes	no
Preparation for sample analysis	no	yes	-	yes	-
Statistical Tests Used	ANOVA	MANOVA	ANOVA/ Bonferroni	• General linear model (with backward stepwise elimination)	• ANOVA • Pearson's Chi-square test • Independent sample T-tests

PI = protease inhibitors, N/NRTI = non/nucleoside reverse transcriptase inhibitor

WIV = weighed in value

FTV = final target value

HPLC-UV, -MS = High Performance Liquid Chromatography-Ultraviolet, -Mass Spectrometer

[1]Two newer PI not included in analyses (DRV,TPV);

[2]FTV based on algorithm described in publications;

[3]tested for ARV class also

Table 2

Characteristics of samples by analyte.

ARV	Number of unique samples	% where PD 5%	% where CV 15%	% where FT=WIV	% where number of participating labs...		
					… 2	… 3	… 4
3TC	20	40.0	100.0	15.0	0.0	50.0	50.0
ABC	15	73.3	93.3	46.7	33.3	66.7	0.0
APV	15	60.0	93.3	40.0	0.0	33.3	66.7
ATV	20	40.0	100.0	40.0	0.0	0.0	100.0
D4T	15	73.3	100.0	80.0	66.7	33.3	0.0
DDI	5	20.0	100.0	0.0	0.0	100.0	0.0
DRV	20	0.0	100.0	10.0	0.0	25.0	75.0
EFV	20	60.0	100.0	60.0	0.0	0.0	100.0
ETR	15	53.3	100.0	100.0	100.0	0.0	0.0
FTC	20	65.0	80.0	55.0	0.0	25.0	75.0
IDV	15	53.3	100.0	53.3	0.0	0.0	100.0
LPV	20	90.0	95.0	95.0	0.0	0.0	100.0
MVC	10	70.0	100.0	50.0	50.0	50.0	0.0
NFV	20	50.0	70.0	45.0	0.0	0.0	100.0
NVP	20	10.0	100.0	10.0	0.0	0.0	100.0
RGV	15	13.3	100.0	13.3	0.0	6.7	93.3
RTV	20	55.0	100.0	50.0	0.0	0.0	100.0
SQV	20	80.0	90.0	85.0	0.0	10.0	90.0
TFV	20	45.0	80.0	50.0	0.0	25.0	75.0
TPV	15	53.3	93.3	20.0	0.0	66.7	33.3
ZDV	20	85.0	95.0	45.0	25.0	50.0	25.0
Total	360	52.5	94.2	47.2	11.1	21.7	67.2

CV = Coefficient of variation across labs (standard deviation/group mean [GM]), FT = Final target (GM or weighed-in value [WIV]). PD = absolute percent deviation (100*abs(GM – WIV)/WIV).

Table 3

Overview of First Four CPQA Proficiency Testing Rounds
Counts of reported concentrations (RCs) by laboratory and analyte; Percent of RCs that were Acceptable; Percents of scores that were Satisfactory.

ARV (WTIV range, ng/mL)	Number of Reported Concentrations by Laboratory (CPL)										Percent of labs reporting for this analyte	Total number of concentrations reported (Percent of concentrations that were Acceptable)	Number of scores determined (Percent of scores that were Satisfactory)
	1	2	3	4	5	6	7	8	9	10			
3TC (60 to 1,500)	15			5	20	20		5	10		60%	75 (100%)	15 (100%)
ABC (180 to 6,000)	10			5		15			10		40%	40 (98%)	8 (100%)
APV (225 to 12,000)	15		10		5	15			10		50%	55 (96%)	11 (100%)
ATV (150 to 12,000)	15	20	10	20	20	20		5	10	20	90%	140 (96%)	28 (96%)
D4T (30 to 1,000)	10				10				10		30%	30 (100%)	6 (100%)
DDI (150 to 2,500)	5					5			5		30%	15 (100%)	3 (100%)
DRV (150 to 7,519)	15	5	10	20	5	20		5	10	20	90%	110 (97%)	22 (95%)
EFV (450 to 12,000)	15	10	10	20	20	20	20	5	10	20	100%	150 (100%)	30 (100%)
ETR (75 to 1,500)	15	5									20%	20 (100%)	4 (100%)
FTC (90 to 3,200)	15	20	14	20		20		5	10		70%	104 (90%)	21 (86%)
IDV (300 to 12,000)	15		10		15	15			10		50%	65 (100%)	13 (100%)
LPV (300 to 15,000)	15	20	10	20	20	20	20	5	10	20	100%	160 (99%)	32 (100%)
MVC (80 to 1,000)			5			10	10				30%	25 (100%)	5 (100%)
NFV (252 to 14,000)	15		5	5	15	20	20		10		70%	90 (89%)	18 (83%)
NVP (450 to 12,000)	15	5	10	20	20	20	20		10		80%	120 (99%)	24 (100%)

ARV (WIV range, ng/mL)	Number of Reported Concentrations by Laboratory (CPL)										Percent of labs reporting for this analyte	Total number of concentrations reported (Percent of concentrations that were Acceptable)	Number of scores determined (Percent of scores that were Satisfactory)
	1	2	3	4	5	6	7	8	9	10			
RGV (225 to 6,000)	15		9	15			15	5			50%	59 (100%)	12 (100%)
RTV (100 to 3,200)	15	20	10	20	20	20	20	5	10	20	100%	160 (99%)	32 (100%)
SQV (225 to 7,549)	15		10	5	20	20		5	10	3*	80%	88 (98%)	18 (100%)
TFV (38 to 1,000)	15	20	15	20		5		5	10		70%	90 (92%)	18 (83%)
TPV (3,000 to 80,000)	15		10			15			10		40%	50 (98%)	10 (100%)
ZDV (60 to 2,500)	15		10	5	15			5	10		60%	60 (98%)	12 (100%)
All Drugs: Total number of concentrations reported (Percent of concentrations that were Acceptable)	280 (99%)	125 (98%)	158 (91%)	200 (98%)	205 (100%)	280 (94%)	125 (97%)	55 (100%)	175 (98%)	103 (100%)	--	1706 (97%)	--
Number of scores determined (Percent of scores that were Satisfactory)	53 (100%)	25 (96%)	32 (91%)	40 (95%)	41 (100%)	56 (95%)	25 (96%)	11 (100%)	35 (97%)	21 (100%)	--	--	342 (97%)

Abbreviations: 3TC lamivudine, ABC abacavir, APV amprenavir, D4T stavudine, DDI didanosine, DRV darunavir, EFV efavirenz, ETR etravirine, FTC embtriciabine, IDV indinavir, LPV lopinavir, MVC maraviroc, NFV nelfinavir, RGV raltegravir, RTV ritonavir, SQV saquinavir, TFV tenofovir, TPV tipranavir, ZDV zidovudine.

Black shading indicates no values reported by CPL; other shading in cells of table indicates percent of reported concentrations that were acceptable:

100%	80% or above but less than 100%	Below 80%

*
For CPL 10/SQV: All reported values were Acceptable; two values could not be reported due to blank-specific interference.

Table 4

a. Statistically significant pairwise differences (on the log scale) in estimated bias (Model 2A) among analytes. Bias is represented by estimates of β_3 in the ANCOVA model: $\log(RC_i) = \beta_0 + \beta_1\log(NC_i) + \beta_{2,i}X_{2,i} + \beta_3$ (Model 2A, where NC=WIV). No pairwise differences were seen for Model 2B (where NC=FT).

	DRV 112.1% (n=110)	RGV 111.7% (n=59)	NVP 109.4% (n=120)	DDI 109.3% (n=30)	NFV 108.4% (n=90)	ATV 108.1% (n=90)	EFV 107.3% (n=150)	TPV 107.2% (n=50)	3TC 107.1% (n=75)	ABC 106.7% (n=40)	TFV 105.4 (n=90)	RTV 104.8% (n=160)	D4T 104.2% (n=30)	LPV 103.9% (n=160)	SQV 102.7% (n=88)	APV 101.7% (n=55)	ETR 100.8% (n=20)	FTC 100.5% (n=104)	MVC 99.3% (n=25)	ZDV 98.4% (n=60)	IDV 96.6% (n=65)
DRV	DRV						0.016				0.062	0.067	0.072	0.076	0.087	0.097	0.106	0.110	0.121	0.114	0.148
RGV		RGV									0.059	0.064		0.073	0.084	0.094	0.103	0.106	0.118	0.111	0.145
NVP			NVP									0.043		0.052	0.064	0.073	0.082	0.086	0.090	0.097	0.124
DDI				DDI																	0.123
NFV					NFV									0.043	0.054	0.064		0.076	0.088	0.081	0.115
ATV						ATV								0.040	0.052	0.061		0.074	0.078	0.085	0.112
EFV							EFV								0.044	0.054		0.066	0.078	0.071	0.105
TPV								TPV										0.065		0.069	0.103
3TC									3TC									0.064	0.075	0.069	0.103
ABC										ABC								0.060		0.065	0.099
TFV											TFV							0.048			0.086
RTV												RTV						0.043			0.081
D4T													D4T								0.076
LPV														LPV							0.061
SQV															SQV						0.061
APV																APV					
ETR																	ETR				
FTC																		FTC			
MVC																			MVC		
ZDV																				ZDV	
IDV																					IDV

Table 4b. Statistically significant pairwise differences in ranks of estimated error (Model 3A) among ARV.

Error is represented by estimates of γ_2, in the ANCOVA model $Y_i = \gamma_0 + \gamma_1 \log(NC_i) + \gamma_2 R_i + \epsilon$, where Y_i and R_i are the predicted and observed ranks of absolute recovery, respectively, and other symbols are as defined in the text(Model 3A, where NC=WIV).
Statistically significant pairwise differences seen for Model 3B (where NC=FT); ranks of absolute deviations for DRV and NVP are higher than those of RGV, RTV and ZDV; and for DRV, also higher than that for MVC.

	DRV (n=110)	RGV (n=59)	NVP (n=120)	DDI (n=30)	NFV (n=90)	ATV (n=90)	EFV (n=150)	TPV (n=50)	3TC (n=75)	ABC (n=40)	TFV (n=90)	RTV (n=160)	D4T (n=30)	LPV (n=160)	SQV (n=88)	APV (n=55)	ETR (n=20)	FTC (n=104)	MVC (n=25)	ZDV (n=60)	IDV (n=65)
DRV	DRV						276	378			246	317	366	414	431		446	392	486	478	
RGV		RGV						348				287		384	401			361	455	447	
NVP			NVP											282	298			259		323	
DDI				DDI																	
NFV					NFV									286	303			264		350	
ATV						ATV								217	234					281	
EFV							EFV														
TPV								TPV													
3TC									3TC											295	
ABC										ABC											
TFV											TFV										
RTV												RTV									
D4T													D4T								
LPV														LPV							
SQV															SQV						
APV																APV					
ETR																	ETR				
FTC																		FTC			
MVC																			MVC		
ZDV																				ZDV	
IDV														305	321			282		368	IDV

Table 5

Summary of associations between bias and the 6 factors, and between error and the 6 factors.

| | Bias | | | | Error | | | |
| | Model 2A: NC=WIV | | Model 2B: NC=FT | | Model 3A: NC=WIV | | Model 3B: NC=FT | |
Factor	Overall F-test	Summary	Overall F-test p-value	Summary	KW test p-value	Summary	KW test p-value	Summary
ARV	<0.001	Highest estimated bias for DRV and RGV; many statistically significant pairwise differences.	0.253	--	<0.001	Many; highest (ranks of) errors for DRV and RGV.	<0.001	{DRV,NVP} > {RGV,RTV,ZDV}, and DRV>MVC.
Lab	<0.001	Labs 6 and 7 have higher recoveries than most other labs. Lab 2 has lowest recovery(below 100%).	<0.001	Labs 6 and 7 have higher recoveries than most other labs. Lab 2 has lowest recovery (below 100%).	<0.001	{3, 6, 7} >all others.	<0.001	{3, 6, 7} >all others, and 3>6.
WIV category, L/M/H within analyte	<0.001	Low & Medium > High	0.099	--	0.872	--	0.563	--
WIV category, "power of ten"	0.005	Despite p<0.05, no pairwise differences were statistically significant.	0.907	--	0.125	--	0.941	--
Round	0.093	--	0.413	--	<0.001	Round 23 > 24 & 25.	0.094	--

KW = Kruskal-Wallis test. ARV = antiretroviral drug, WIV = weighed-in value. L/M/H = low/medium/high. NC = nominal concentration. FT = final target.