Contents lists available at ScienceDirect

Journal of Immunological Methods

journal homepage: www.elsevier.com/locate/jim

Review

The Immunology Quality Assessment Proficiency Testing Program for CD3⁺4⁺ and CD3⁺8⁺ lymphocyte subsets: A ten year review via longitudinal mixed effects modeling

J. Bainbridge^a, C.L. Wilkening^b, W. Rountree^a, R. Louzao^{a,*}, J. Wong^a, N. Perza^a, A. Garcia^a, T.N. Denny^a

^a Duke Human Vaccine Institute, Duke University Medical Center, Durham, NC, USA

^b Statistical Analysis and Data Center, Boston, MA, USA

ARTICLE INFO

Article history:

Received 6 September 2013

Received in revised form 24 May 2014

Accepted 30 May 2014

Available online xxxx

Keywords:

Proficiency testing

Lymphocyte subset phenotyping

IQA

Mixed effects models

Longitudinal analysis

Flow cytometry

ABSTRACT

Since 1999, the National Institute of Allergy and Infectious Diseases Division of AIDS (NIAID DAIDS) has funded the Immunology Quality Assessment (IQA) Program with the goal of assessing proficiency in basic lymphocyte subset immunophenotyping for each North American laboratory supporting the NIAID DAIDS HIV clinical trial networks. Further, the purpose of this program is to facilitate an increase in the consistency of interlaboratory T-cell subset measurement (CD3⁺4⁺/CD3⁺8⁺ percentages and absolute counts) and likewise, a decrease in intralaboratory variability. IQA T-cell subset measurement proficiency testing was performed over a ten-year period (January 2003–July 2012), and the results were analyzed via longitudinal analysis using mixed effects models. The goal of this analysis was to describe how a typical laboratory (a statistical modeling construct) participating in the IQA Program performed over time. Specifically, these models were utilized to examine trends in interlaboratory agreement, as well as successful passing of proficiency testing. Intralaboratory variability (i.e., precision) was determined by the repeated measures variance, while fixed and random effects were taken into account for changes in interlaboratory agreement (i.e., accuracy) over time. A flow cytometer (single-platform technology, SPT) or a flow cytometer/hematology analyzer (dual-platform technology, DPT) was also examined as a factor for accuracy and precision. The principal finding of this analysis was a significant ($p < 0.001$) increase in accuracy of T-cell subset measurements over time, regardless of technology type (SPT or DPT). Greater precision was found in SPT measurements of all T-cell subset measurements ($p < 0.001$), as well as greater accuracy of SPT on CD3⁺4⁺% and CD3⁺8⁺% assessments ($p < 0.05$ and $p < 0.001$, respectively). However, the interlaboratory random effects variance in DPT results indicates that for some cases DPT can have increased accuracy compared to SPT. Overall, these findings demonstrate that proficiency in and among IQA laboratories have, in general, improved over time and that platform type differences in performance do exist.

© 2014 Elsevier B.V. All rights reserved.

Contents

1. Introduction	0
2. Methods	0
2.1. IQA Proficiency Testing Program overview	0
2.2. Statistical approach	0

* Corresponding author at: 2 Genome Court, MSRB II Building, Room 4019, DUMC 103020, Durham, NC 27710, USA. Tel.: +1 919 684 5861; fax: +1 919 681 8251.
E-mail address: raul.louzao@duke.edu (R. Louzao).

2.3.	Data and analyses	0
2.3.1.	Data characteristics	0
2.3.2.	Accuracy analysis	0
2.3.3.	Platform variability analysis	0
2.3.4.	Mark analysis	0
2.3.5.	Program drop out analysis	0
2.3.6.	Model selection	0
3.	Results	0
3.1.	Laboratory characteristics	0
3.2.	Unadjusted outcome characteristics	0
3.3.	Accuracy analysis	0
3.4.	Platform variability analysis	0
3.5.	Mark analysis	0
3.6.	Program drop out analysis	0
3.7.	Laboratory specific trajectories	0
4.	Discussion	0
	Acknowledgments	0
	References	0

1. Introduction

In the last 15 to 20 years, much of the focus of HIV research globally has been on the development of immunological or virological laboratory markers to determine HIV infection status and monitor a patient's response during treatment or disease progression. These markers are frequently used to monitor patients who are enrolled in multicenter clinical trials assessing new antiretroviral therapies (ARTs) or vaccine-related products. As classification based on these markers often serves as a factor for treatment decisions, enrollment into clinical trials, and clinical prognosis (Calvelli et al., 1993), there exists a critical need for accurate and precise measurements. Laboratories will typically make changes in technology or, more likely, experience changes in staff over multi-year periods. The long-term tracking of proficiency metrics can reflect the laboratory's overall performance. Having access to such information is required for laboratories involved in clinical care settings to meet accreditation requirements. However, it is also critical to have such information to review performance metrics of those laboratories involved in multicenter clinical trials.

Since 1999, the National Institute of Allergy and Infectious Diseases, Division of AIDS (NIAID/DAIDS), has funded the Immunology Quality Assessment (IQA) Program, a continuation of the Flow Cytometry Quality Assessment Program implemented in 1987 and described previously (Kagan et al., 1993; Calvelli et al., 1993). Broadly, the goal of the IQA Program is to provide external quality assessment for laboratories supporting the NIAID DAIDS HIV clinical trial networks. One aspect of the IQA Program is to assess proficiency in basic lymphocyte subset immunophenotyping for all North American laboratories supporting the networks. The goal of this program is to ensure that these laboratories provide consistent, high quality results, with little inter- and intralaboratory T-cell subset measurement variability.

Participating sites in the IQA Program are assessed for their ability to conduct four T-cell subset measurements ($CD3^+4^+$ / $CD3^+8^+$ percentages and absolute counts) six times (sendouts) per year using fresh whole blood samples from different donors and replicate schemes (singletons to quadruplicates) provided by the IQA Program. Using their standard immunophenotyping assays, participating sites assess the samples and report their

results to the IQA Program for proficiency grading. Sites that perform poorly are subject to remediation. Sites that continue to perform poorly may be deemed unable to process samples for HIV clinical trials.

In terms of immunophenotyping assays, there are two types of flow cytometer technology used by laboratories. The first, and most common, method is dual-platform technology (DPT), in which absolute T-cell numbers are calculated from three measurements determined by two different instruments, a hematology analyzer and a flow cytometer. The white blood cell count and the percent lymphocytes are obtained from the hematology analyzer and the percent T-cell markers from the flow cytometer. T-cell lymphocyte subset counts are determined by multiplying these three measurements. The second is single-platform technology (SPT), which is designed to enable determinations of both absolute and percentage T-cell lymphocyte subset values using a single tube and a single instrument (flow cytometer).

The long history of this program provides the unique opportunity to assess the overall site proficiency over time. From January 2003 to July 2012, the IQA Program sent over 23,000 proficiency testing samples to 101 laboratories for the purpose of monitoring their performance and proficiency. For the sake of consistency, this report focuses on the time period that exclusively utilized the same validated software data collection system. For this longitudinal analysis, mixed effects models were used (Fitzmaurice et al., 2004) to estimate the components of measurement error, particularly the change in interlaboratory agreement over time (i.e., accuracy). This statistical model was deemed most appropriate, as mixed effects models: 1) allow separate estimates of accuracy and precision, 2) account for repeated measures in the data, 3) account for missing data while making full use of all the available data, and 4) permit controlling for unobserved laboratory level factors (latent classes) via the random effects. The results from these models allowed us to both graphically present laboratory trajectories and characterize them via area under the curve.

Four mixed effects models were used to assess accuracy for T-cell subset measurements ($CD3^+4^+$ / $CD3^+8^+$ percentages and absolute counts). Additionally, these four models allowed the secondary comparison of platform variability

(SPT vs. DPT), as it was necessary to consider platform type in the modeling process. Finally, the outcome of a pass/no pass, represented as a binary (0/1) metric, was used as a supportive analysis called the mark analysis. This supportive analysis also used four mixed effects models, but the outcome was a (0/1) for each T-cell subset measurement. While this binary outcome is a less sensitive metric for analysis, it is useful for corroboration of the findings of the primary analyses.

2. Methods

2.1. IQA Proficiency Testing Program overview

The IQA Program is designed to provide a proficiency testing scheme that closely mimics how a site would normally handle clinical trial samples. The IQA Program provides DAIDS clinical trial laboratory whole blood samples six times per year. This panel consists of five previously characterized specimens with a range of CD3⁺4⁺ and CD3⁺8⁺ percentages and absolute counts. Consent is obtained from donors according to Duke University IRB, Federal, and State regulations. Upon successful donor safety screening, approximately 450 mL of K3 ethylenediaminetetraacetic acid (K3-EDTA) anti-coagulated blood is collected and aliquoted. Blood samples are shipped without temperature stabilization overnight to all participating laboratories without temperature monitors per ATCG guidelines (www.hanc.info/labs/labresources/procedures/ACTGIMPAACT/%20Lab/%20Manual/ACTN/%20Instructions/%20for/%20Overnight/%20Shipment.pdf). Participating laboratories are required to measure and report the same immunophenotypes that they perform on specimens from patients enrolled in NIAID-sponsored investigations. The IQA Program evaluates both interlaboratory and intralaboratory performance of the CD3⁺4⁺ and CD3⁺8⁺ percentages.

Sendouts occur bi-monthly, and two sendouts are graded each trimester, for a total of three grading evaluations per year. Five donor samples are provided at each sendout, and follow one of two possible replicate schemes that are randomly selected. One scheme has triplicates and two singletons for both sendouts. The other scheme has a quadruplicate and a singleton in one sendout and a duplicate and three singletons in the other. In each trimester, there are a total of six interlaboratory determinations and two intralaboratory determinations.

The interlaboratory analysis uses the following robust statistics: the consensus sample median value (the 50th percentile value for a set of specimens from a given sample as measured by all laboratories), the inter-quartile range (IQR) of a sample (the 75th minus the 25th percentile values), the adjusted IQR (AIQR, 0.75 times the IQR), the residual (the laboratory specimen value minus the consensus sample median value), and the deviate (the residual divided by the AIQR). The absolute values of the residual and deviate are used for grading. The intralaboratory analysis is based upon the range (maximum–minimum value) of the replicate sample for each laboratory.

Sites receive marks for inter- and intralaboratory results as an evaluation of proficiency. The criteria are applied to each specimen, and a sample with replicate specimens only receives one mark regardless of how many specimens meet

the evaluation criteria. The criteria for a 'no pass' interlaboratory mark for the percentage T-cell subset measurements are that the residual must be $\geq 5\%$, and the deviate must be ≥ 2 . The criteria for a 'no pass' intralaboratory mark for CD3⁺4⁺ absolute counts are that the residual must be ≥ 100 , and the deviate must be ≥ 2 . For CD3⁺8⁺ absolute counts, the residual must be ≥ 200 , and the deviate must be ≥ 2 . Certification is based on the number of 'no pass' marks. If a laboratory receives one-third or more 'no pass' marks for CD3⁺4⁺ or CD3⁺8⁺ assessments during a trimester, certification is not granted for the laboratory for that trimester.

If necessary, the IQA Program provides interventions to identify source problems and assist in remediation. Laboratories that have failed the trimester must submit instrument dot plots and acquisition data along with daily QA/QC data for review. These data are reviewed for appropriate instrument set-up, calibration, control runs, and proper acquisition and analysis procedures. Problems with testing identified in the review of site-provided documents are discussed and resolved with the laboratory. If needed, a panel of samples is sent to the laboratory for testing in parallel with the IQA Program, and the results are evaluated by the IQA laboratory. If there is agreement (tolerance limits defined as $\pm 5\%$ for % values and $\pm 10\%$ for absolute count values), no further intervention may be required. The laboratory's performance on subsequent sendouts, however, is closely monitored. If there is no agreement or partial agreement between the parallel testing results, further remediation may include conducting on-site visits to review testing procedures, instrument set-up, performance, and staff proficiency. During the time frame of this analysis, the IQA provided assistance on 61 occasions, assisting 26 laboratories.

2.2. Statistical approach

Longitudinal mixed effects models (Fitzmaurice et al., 2004, 2009) were used for within-sample repeated measures, as well as repeated measures over time by laboratories. The mixed effects consist of fixed effects, random effects, and repeated measures. Fixed effects are used to characterize the 'typical' laboratory (i.e., the model-based standard lab) from all laboratories included in the analysis. The typical laboratory can be described as the statistical reference for all laboratories in the dataset, which are described by the random effects. The random effects provide laboratory specific information, as well as control for unmeasured laboratory factors. Repeated measures evaluate the intralaboratory variance, assuming a compound symmetric covariance structure. Detailed information about the models can be found in the supplemental material, including model specifications and parameter estimates. A laboratory's residual over time was used as the outcome for the primary accuracy analysis. The term "residual" refers to the difference between a laboratory's measured value and the consensus average for a sample. The consensus average was used as a proxy for a constant value, as there is no true known value for each sample.

Through the use of mixed effects models, one can model an individual laboratory's outcomes over time as well as the typical lab. These models can also assess laboratories entering and exiting the program (or switching platform type) by using all available data to generate parameter estimates. Based upon the models, trajectories of the T-cell

subset measurement residuals (accuracy analysis) and the probability of a laboratory receiving a 'no pass' interlaboratory mark (mark analysis) were performed. A trajectory is a description of the model-based values over time of an outcome. Laboratories' trajectories were graphically presented and characterized via the use of the area under the curve (AUC).

2.3. Data and analyses

2.3.1. Data characteristics

The data describing replicates are the absolute values of the residuals (absolute residual) from each laboratory for each T-cell subset measurement (CD3⁺4⁺/CD3⁺8⁺ percentages and absolute counts). These absolute residuals represent the total error in a measurement. The data describing each sample are the consensus (all laboratories, all replicates) median values for the T-cell subset measurements and the receipt of a 'no pass' interlaboratory mark for each of the T-cell subset measurements. Intralaboratory marks were not used for this study due to time and space limitations.

2.3.2. Accuracy analysis

The accuracy analysis was implemented with four linear mixed effects models (SAS PROC MIXED (SAS Institute Inc., 2011)). The four outcomes for the accuracy analysis were the absolute residuals of each T-cell subset measurement with a natural log transformation (residual). The primary predictor variables were the time in years from January 2003 to the month and year of the sendout (time) and time squared (time²). The time squared term permits assessment of change in the direction of laboratory's results over time. The time and time squared terms together allow for the assessment of the primary outcome, a reduction in residuals over time, conditional upon the other terms in the model. These other terms were the sample median with a natural log transformation (median) and the use of SPT or DPT (platform) during a given sendout. These terms are included to control for their effect on proficiency and different levels over time, while assessing the main effects of interest of the time variables. Duration in the program was considered as a predictor variable separate from calendar time, but their collinearity precluded its use. The accuracy analysis models also provide secondary comparisons: the comparison of SPT laboratories versus DPT laboratories, in terms of typical laboratory accuracy, and the platform variability analysis.

2.3.3. Platform variability analysis

For the platform variability analysis, separate estimates of residual intralaboratory sample variances were calculated based on platform type. These variances were then compared as a ratio (SPT/DPT) to estimate the magnitude of the difference in precision for a typical lab by platform type. Interlaboratory variances were calculated for the percent

outcomes only and compared with the ratio of (SPT/DPT) for the beginning and end of the time period examined. These variance estimates are generated using the random effects and repeated measure variance terms from the covariance structure in the accuracy analysis.

2.3.4. Mark analysis

The mark analysis used four generalized linear mixed models with binary outcomes of 'pass/no pass' interlaboratory marks for each T-cell subset measurement (SAS PROC GLIMMIX (SAS Institute Inc., 2011)). Predictor variables were examined using those described for accuracy analysis, with the inclusion of the number of replicates in a sample.

2.3.5. Program drop out analysis

Trimester results were compared between laboratories that were in the program for the entire time period and those that permanently dropped out. Pattern mixture models (Hedeker and Gibbons, 1997; Fitzmaurice et al., 2009) were fit to address the possible difference between laboratories that participated through the end of the time period and those that dropped out. Model fit was assessed by likelihood ratio tests for variable inclusion and by residual analysis to investigate departure from model assumptions.

2.3.6. Model selection

A standard mixed model fitting procedure was used (Fitzmaurice et al., 2004) with a forward stepwise process for the fixed effects. Likelihood ratio tests were utilized to determine the inclusion of any given fixed effect in the model, while using maximum likelihood estimation for the parameter values. The criteria of $p < 0.05$ was used to determine if an effect was significant and warranted inclusion in the model. After the best-fitting set of fixed effect terms was determined for a given T-cell subset measurement, the random effects and repeated measures covariance structures were determined. A similar process was used for the mark analyses.

3. Results

3.1. Laboratory characteristics

There were 101 laboratories that participated in the IQA Program during the time period examined. Of these labs, 68 (67%) used DPT only, 12 (12%) used SPT only, and 21 (21%) switched from DPT to SPT during the study period. Sixty-four (63%) of the 101 labs were participants in the IQA Program from the start of the study program to the end of the time period examined. The number of laboratories in each calendar year broken out by platform is shown in Table 1. There are a decreasing number of DPT laboratories with a corresponding increase in the number of SPT laboratories over time. The number of laboratories that failed a given

Table 1
Number of laboratories by year by platform type.

Platform	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
DPT	74	73	70	68	64	56	51	47	46	43
SPT	11	10	15	18	22	26	26	26	26	25
Total	85	83	85	86	86	82	77	73	72	68

Table 2

Mean absolute residual by year and T-cell subset.

T-cell subset	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
CD3 ⁺ 4 ⁺ %	1.0	1.0	0.9	0.9	0.8	0.9	0.8	0.8	0.9	0.8
CD3 ⁺ 4 ⁺ count	47.4	37.7	25.8	24.2	35.7	42.3	38.4	28.0	27.3	23.9
CD3 ⁺ 8 ⁺ %	1.7	1.5	1.6	1.5	1.2	1.8	1.2	1.2	1.2	1.1
CD3 ⁺ 8 ⁺ count	93.3	61.2	60.2	52.2	56.2	66.9	58.3	53.1	47.7	38.1

trimester ranged from zero to five, with one or two being typical. Only 26 (26%) laboratories failed a trimester over the study period, and the average number of trimesters that failed for these 26 laboratories was approximately two. No laboratory was ever decertified, and they all regained full certification. Of the 26 laboratories, eight (31%) left the program prior to the study end point. Comparatively, 29 of the 75 laboratories (39%) with no failing trimester left the program prior to study end.

3.2. Unadjusted outcome characteristics

Unadjusted T-cell subset measurement average residuals by year are shown in Table 2. There is a downward trend in all T-cell subset measurements, although with some variability. There is a noticeable counter to the trend in 2007–2009 for the CD3⁺4⁺ counts, CD3⁺8⁺ counts, and the CD3⁺8⁺ percentages. The unadjusted percentages of 'no pass' interlaboratory marks for each year in the study are shown in Table 3. There is also a downward trend for three of the four T-cell subset measurements (CD3⁺4⁺ counts, CD3⁺8⁺ counts, and the CD3⁺8⁺ percentages). There is also a countertrend centered on 2008 for the T-cell subset count measurements.

3.3. Accuracy analysis

For the primary analysis, all T-cell subset measurement residuals had a decrease associated with time ($p < 0.001$). The model-based trajectories of CD3⁺4⁺% and CD3⁺8⁺% subsets show a steady decrease in the residuals for the typical laboratory from January 2003 until 2009 (Figs. 1 and 2, left hand axis). The model-based trajectories of the T-cell subset absolute count measurements show a steady decrease in the residuals over the entire time span (Figs. 3 and 4, left hand axis). Model-based percentage changes in residuals from the start of the study to the end of the study for a typical laboratory are shown in Table 4. For the typical lab, these ranged from an 11% decrease to a 43% decrease. Model-based platform effects comparing DPT to SPT for the typical laboratory ranged from no difference to a 12% increase in the residuals as shown in Table 4. Platform fixed effects were found to be significant for the T-cell subset percentage

measurements ($p < 0.05$ for CD3⁺4⁺% and $p < 0.001$ for CD3⁺8⁺%), but not the T-cell subset absolute count measurements. Table 5 shows model-based area under the curve (AUC) summary statistics for CD3⁺4⁺% broken out by platform type. These statistics also factored in a switch from DPT to SPT by a laboratory. Lower values correspond to smaller absolute residuals over the course of the study. The non-switching DPT average AUC is 16% greater than the non-switching SPT laboratories. For switching laboratories the DPT average AUC is 11% greater than the SPT AUC. The range for all DPT labs is 3.54% to 10.23%, while the range for all SPT labs is 4.04% to 5.95%.

3.4. Platform variability analysis

The model-based intralaboratory variance ratios were approximately 0.85 (SPT/DPT) for all four T-cell subset measurements, demonstrating a statistically significant reduction of 15% in a typical SPT intralaboratory variability relative to the typical DPT laboratory ($p < 0.001$, for all comparisons). These p-values were derived from a set of likelihood ratio tests (calculations not presented) comparing the repeated measures covariance structure, grouped versus ungrouped for platform type, over the entire time period.

In contrast, the interlaboratory variance ratios were derived from the random effects. Only T-cell subset percentages were significant in these models. The model-based interlaboratory variance ratios (SPT/DPT) were 0.15 and 0.18 for CD3pt + 4⁺% and CD3⁺8⁺%, respectively in January 2003 and were 0.20 and 0.05 in July 2012 ($p < 0.001$, all comparisons using a likelihood ratio test for the random effects covariance structures).

3.5. Mark analysis

Model-based trajectories of the probability of a 'no pass' interlaboratory mark are shown in Figs. 1 to 4 (right hand axis) for the typical laboratory. The trajectory of the probability of a CD3⁺8⁺% 'no pass' interlaboratory mark shows a steady downward trend for the typical laboratory of either platform type, with a more pronounced decrease for DPT. There is also a steady decrease in the probability over time for 'no pass' intralaboratory marks for T-cell subset

Table 3

Percent bad interlaboratory marks by year and T-cell subset.

T-cell subset	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
CD3 ⁺ 4 ⁺ %	1.8	0.7	0.9	0.9	0.4	1.2	1.1	0.9	1.7	0.9
CD3 ⁺ 4 ⁺ count	7.1	6.8	4.1	3.1	5.1	6.3	5.5	3.8	2.9	2.5
CD3 ⁺ 8 ⁺ %	6.2	3.1	4.6	3.8	3.0	2.3	3.2	2.4	2.5	1.4
CD3 ⁺ 8 ⁺ count	7.9	4.4	4.9	2.8	2.6	4.2	3.4	2.6	2.3	1.6

Fig. 1. CD3⁺4⁺% model-based absolute residual values (left axis) and probability of bad interlaboratory mark trajectory (right axis) for typical laboratories by platform with sample median values of 20.

absolute count without a platform type effect. Table 6 gives the model-based estimates of fold change in odds of a 'no pass' interlaboratory mark over time, taking into account platform type for the typical laboratory. In the table, an estimate of one indicates that there is no statistical significance. The effect of time represents the entire duration of the analysis and is significant in three of the four models. This indicates that a typical laboratory is much less likely to get a 'no pass' interlaboratory mark at the end of the analysis period in comparison to the beginning. For a typical laboratory, this ranges from a five-fold to ten-fold reduction in the odds of a 'no pass' interlaboratory mark. The effect of platform type was found to be significant for the T-cell subset percentage measurements only, with DPT having a three to five fold increase in the odds of a 'no pass' interlaboratory mark relative to SPT for the typical laboratory.

For a typical laboratory across all T-cell subset measurements, higher values of the sample median were associated with increased odds of a 'no pass' interlaboratory mark ($p < 0.001$).

Likewise, a greater number of replicates ($p < 0.05$ to $p < 0.001$) were associated with increased odds of a 'no pass' interlaboratory mark. Use of DPT was found to be associated with higher odds of a 'no pass' interlaboratory mark for CD3⁺4⁺% and CD3⁺8⁺% ($p < 0.05$, $p < 0.001$, respectively). These parameter estimates are provided in the supplemental information, see Supplemental Tables 9 to 12.

3.6. Program drop out analysis

A program drop out analysis was done to check that poor performing laboratories were not biasing our results. A Chi-squared test of independence indicated no relationship between a 'pass/no pass' performance for a trimester and program drop out ($\chi^2 = .52$, $df = 1$, $p = .47$). Additionally, the results of the pattern mixture models found no indication of differences between laboratories that left the program prior to July 2012 and those that were still in the program in July 2012.

Fig. 2. CD3⁺8⁺% model-based absolute residual values (left axis) and probability of bad interlaboratory mark trajectory (right axis) for typical laboratories by platform with sample median values of 50.

Fig. 3. $CD3^{+}4^{+}$ absolute count model-based absolute residual values (left axis) and probability of bad interlaboratory mark trajectories (right axis) for typical laboratory with sample median value of 350 (no model-based differences in platforms).

This was evidenced by the lack of significance of the likelihood ratio tests conducted. P-values for these models ranged from 0.06 to 0.81 with a cutoff of 0.05 determined a priori with no adjustment for multiple comparisons.

3.7. Laboratory specific trajectories

Making full use of the capabilities of mixed effects models to provide laboratory specific information, selected trajectories were generated. These laboratory trajectories were chosen to demonstrate the range of results found between laboratories relative to the trajectory of a typical laboratory (found in Fig. 3). Examples of laboratory specific trajectories can be found in Figs. 5 and 6 for $CD3^{+}4^{+}$ absolute count residuals and probability of a 'no pass' interlaboratory mark. These trajectories are based on a combination of the fixed effects and the laboratory specific best linear unbiased predicted values from the random effects.

4. Discussion

A longitudinal analysis of the IQA Program was performed using 10 years of program data. Mixed effects models were

used to assess accuracy and precision, along with the probability of receiving a 'no pass' mark for four T-cell subset measurements: $CD3^{+}4^{+}$ and $CD3^{+}8^{+}$ percentages and absolute counts. The primary outcome for this analysis was the reduction in the residuals over time (i.e., an increase in accuracy), and this was significant for each T-cell subset measurement analyzed. Furthermore, the use of time squared, in addition to the linear effect of time, permitted identification of potential changes in the direction of trends. (e.g., Fig. 5, Lab 1). No differential laboratory drop out effect was found with respect to performance. The supporting mark analysis showed similar results to the accuracy analysis, with a corresponding decrease in the probability of a 'no pass' interlaboratory mark for the typical laboratory in three of the four T-cell subset measurements. In the single T-cell subset measurement, $CD3^{+}4^{+}$ percentage, where no improvement was found, the initial probability of a 'no pass' interlaboratory mark was so low that further improvement in a typical lab was unlikely (i.e., a floor effect was found).

The secondary comparison of typical DPT to SPT laboratories indicates superior performance by SPT laboratories on T-cell subset percentages for accuracy and all T-cell subset measurements for precision. This supports previous findings

Fig. 4. $CD3^{+}8^{+}$ absolute count model-based absolute residual values and probability of bad interlaboratory mark trajectories for typical laboratory with sample median value of 550 (no model-based differences in platforms).

Table 4

Model-based estimates of percentage change in residual due to time (entire period) and platform by T-cell subset.

T-cell subset	Time	95% CI	DPT	95% CI
CD3 ⁺ 4 ⁺ %	–11	(–15, –7)	7	(4, 11)
CD3 ⁺ 4 ⁺ count	–40	(–47, –31)	0	(NA, NA)
CD3 ⁺ 8 ⁺ %	–16	(–22, –10)	12	(7, 17)
CD3 ⁺ 8 ⁺ count	–43	(–51, –33)	0	(NA, NA)

(Whitby et al., 2013), with both studies finding similar precision estimates by platform. However, there is a high degree of interlaboratory variability in DPT results and some DPT laboratories outperformed all SPT laboratories on T-cell subset percentages. Hultin et al. found similar results for SPT and DPT when testing both platforms within Multicenter AIDS Cohort Study (MACS) laboratories, so we decided to explore this complex phenomenon further by assessing the MACS laboratories in our data set. In these DPT laboratories AUCs and trajectories for CD3⁺4⁺ percentages were reviewed. This subgroup outperformed the typical SPT lab and, in some cases, all SPT laboratories on these measures.

Thus, our results support the findings of both studies (Hultin et al., 2010; Whitby et al., 2013). In particular, the AUC for CD3⁺4⁺ percentages was lower for some DPT laboratories in comparison to all SPT laboratories. This implies that DPT laboratories can outperform SPT laboratories, but in the typical case, the SPT laboratory has greater accuracy (for the T-cell subset percentages) and precision (all T-cell subset measurements).

While a difference is observed between platform types for the percentage measurements, no difference is observed for the absolute counts. This finding may be an artifact of the extended time between blood draw and measurement by a DPT laboratory's hematology instrument. It has been previously reported that testing samples more than six hours after collection falsely increases the lymphocyte count due to degranulation of granulocytes (Calvelli et al., 1993; Nicholson et al., 1984). However, for this program count data from DPT laboratories are not collected to determine if this is the case.

One issue the program faced was elevated residual CD3⁺4⁺ absolute count values during 2007 to 2009, which were a byproduct of higher median IQA Program sample values. This corresponded to a policy decision to use donors with higher CD3⁺4⁺ absolute count values during this time period. It was hypothesized that this would provide greater sensitivity in determining a laboratory's proficiency. However, effective May 2010, lower CD3⁺4⁺ absolute count donor values were selected in order to more closely

Table 5

Area under the curve for CD3⁺4⁺% residual by platform and platform switching status (lower is greater accuracy).

Switched	Platform	N	Mean	Std. dev.	Minimum	Maximum
No	DPT	68	5.99	1.27	4.18	10.23
No	SPT	12	5.17	0.49	4.04	5.66
Yes	DPT	21	5.84	1.43	3.54	8.88
Yes	SPT	21	5.27	0.28	4.77	5.95

Table 6

Model-based estimates for fold change in odds of a bad interlaboratory mark due to time (entire period) and platform by T-cell subset.

T-cell subset	Time	95% CI	DPT	95% CI
CD3 ⁺ 4 ⁺ %	1.0	(NA, NA)	2.9	(1.04, 8.34)
CD3 ⁺ 4 ⁺ count	–6.3	(–3.15, –12.69)	1.0	(NA, NA)
CD3 ⁺ 8 ⁺ %	–10.4	(–4.22, –25.46)	4.8	(2.29, 9.89)
CD3 ⁺ 8 ⁺ count	–5.1	(–2.40, –10.68)	1.0	(NA, NA)

approximate clinically meaningful values. This necessitated controlling for the sample median value in the models to accurately assess the effect the primary analytic variables, namely time and time squared. This provides a better fitting model for the T-cell subset measurements over time and provides better estimates of laboratory performance. Based upon the results of this study, the IQA Program has implemented a procedural change for selection of qualifying donor's samples. In order to minimize variation in the probability of a 'no pass' mark across send-out groups, CD3⁺8⁺ percentage values for the replicate samples should be closely matched across groups in a given sendout. CD3⁺8⁺ percentage values were chosen given the larger range of sample median values relative to the already range restricted CD3⁺4⁺ percentage values and the inherent difficulty of matching on both values. The intent of this evidence-based change is to decrease variability in results as a function of group assignment. Historical results will continue to be examined so that improvements can be made to the program.

In conclusion, this 10-year IQA Program analysis demonstrates improved laboratory performance by reducing overall measurement variability over time. For example, in a hypothetical case where the CD3⁺4⁺ absolute count was 330, the typical value would range between 308 and 356 (i.e., 48 cells/μl) in January 2003 and then tighten to 316 to 344 (i.e., 28 cells/μl) in July 2012 based on a 95% CI. This translates into a 42% reduction of variability and leads to a substantially reduced probability of classifying the sample as above 350 cells/μl when in fact it is below. This reduction in variability suggests overall improved performance of IQA Program monitored laboratories during the period of time represented in this study. Future research could include exploring additional sources of variation in results and further exploration of the effects of procedural changes.

Acknowledgments

This project has been funded with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN27220070054C. We wish to thank all involved in the IQA CD3⁺4⁺/CD3⁺8⁺ program over the years including Su Chun Chen, Bill White, Auguste Badiabo, and Daniella Livnat.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.jim.2014.05.017>.

Fig. 5. $CD3^{+}4^{+}$ absolute count model-based absolute residual values trajectories for five laboratories with sample median value of 350.

Fig. 6. $CD3^{+}4^{+}$ absolute count model-based probability of a bad interlaboratory mark trajectories for five laboratories with sample median value of 350.

References

- Calvelli, T., Denny, T.N., Paxton, H., Gelman, R., Kagan, J., 1993. Guideline for flow cytometric immunophenotyping: a report from the National Institute of Allergy and Infectious Diseases, Division of AIDS. *Cytometry* 14 (7), 702.
- Fitzmaurice, G.M., Laird, N.M., Ware, J.H., 2004. *Applied Longitudinal Analysis*. Wiley-InterScience, Hoboken, NJ.
- Fitzmaurice, G.M., Davidian, M., Verbeke, G., Molenberghs, G., 2009. *Longitudinal Data Analysis*. CRC Press, Boca Raton.
- Hedeker, D., Gibbons, R.D., 1997. Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychol. Methods* 2 (1), 64.
- Hultin, L.E., Chow, M., Jamieson, B.D., O'Gorman, M.R., Menendez, F.A., Borowski, L., Denny, T.N., Margolick, J.B., 2010. Comparison of interlaboratory variation in absolute T-cell counts by single-platform and optimized dual-platform methods. *Cytometry B Clin. Cytom.* 78 (3), 194 (May).
- Kagan, J., Gelman, R., Waxdal, M., Kidd, P., 1993. NIAID Division of AIDS flow cytometry quality assessment program. *Ann. N. Y. Acad. Sci.* 677, 50 (Mar 20).
- Nicholson, J.K.A., Jones, B.M., Cross, G.D., McDougal, J.S., 1984. Comparison of T and B cell analyses on fresh and aged blood. *J. Immunol. Methods* 73 (1), 29 (10/12).
- SAS Institute Inc., 2011. *SAS/STAT 9.3 User's Guide*. SAS Institute Inc.
- Whitby, L., Whitby, A., Fletcher, M., Helbert, M., Reilly, J.T., Barnett, D., 2013. Comparison of methodological data measurement limits in $CD4^{+}$ T lymphocyte flow cytometric enumeration and their clinical impact on HIV management. *Cytometry B Clin. Cytom.* 84B (4), 248.