Choosing a new CD4 technology : can statistical method comparison tools influence the decision?

Reference:
Scott Lesley E., Kestens Luc, Pattanapanyasat Kovit, Sukapirom Kasma, Stevens Wendy S..- Choosing a new CD4 technology : can statistical method comparison tools influence the decision?
Cytometry: part B: clinical cytometry - ISSN 1552-4949 - 92:6(2017), p. 465-475
Full text (Publisher's DOI): https://doi.org/10.1002/CYTO.B.21522

Choosing a new CD4 technology: Can statistical method comparison tools influence the decision?

Scott. L.E[1], Kestens. L[2], Pattanapanyasat. K[3], Sukapirom. K[3], Stevens. W.S.[1,4]

[1]Department of Molecular Medicine and Haematology, University of Witwatersrand, Faculty of Health Sciences, School of Pathology. [2]Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium and Laboratory of Immunology, Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium. [3]Center of Excellence for Flow Cytometry, Office for Research and Development, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok, Thailand. [4]The National Health Laboratory Service, Johannesburg, South Africa

Corresponding author:

Lesley E Scott

7 York road Parktown, Johannesburg 2000, South Africa.

+27 11 489 8567 (tel), email: lesley.scott@nhls.ac.za

Author emails:

Kovit Pattanapanyasat: grkpy@mahidol.ac.th

Kasama Sukapirom: siksc@mahidol.ac.th

Luc Kestens: LKestens@itg.be

Wendy. S. Stevens: wendy.stevens@nhls.ac.za

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi: 10.1002/cyto.b.21522

Abstract

Background: Method comparison tools are used to determine the accuracy, precision, agreement and clinical relevance of a new or improved technology versus a reference technology. Guidelines for the most appropriate method comparison tools as well as their acceptable limits are lacking and not standardised for CD4 counting technologies.

Methods: Different method comparison tools were applied to a previously published CD4 data set (n= 150 data pairs) evaluating five different CD4 counting technologies (TruCOUNT, Dual Platform, FACSCount, Easy CD4, CyFlow) on a single specimen. Bland-Altman, percentage similarity, percent difference, concordance correlation, sensitivity, specificity and misclassification method comparison tools were applied as well as visualization of agreement with Passing Bablock and Bland-Altman scatter plots.

Results: The FACSCount (median CD4 = 245cells/µl) was considered the reference for method comparison. An algorithm was developed using best practices of the most applicable method comparison tools, and together with a modified heat map was found useful for method comparison of CD4 qualitative and quantitative results. The algorithm applied the concordance correlation for overall accuracy and precision, then standard deviation of the absolute bias and percentage similarity coefficient of variation to identify agreement, and lastly sensitivity and misclassification rates for clinical relevance.

Conclusion: Combining method comparison tools is more useful in evaluating CD4 technologies compared to a reference CD4. This algorithm should be further validated using CD4 external quality assessment data and studies with larger sample sizes.

Word count: 228

Introduction

Criteria suggested for evaluating new CD4 technologies include selecting an appropriate comparator reference technology, performing population relevant (and adequate sample size) prospective evaluations with fresh patient specimens , including stabilised blood control/reference/external quality assessment material, evaluating site training and readiness, and suitable data analysis with on-going monitoring and surveillance (1). Qualitative data analysis of new CD4 technology includes: ease of use, time to reportable result, error rate, cost, foot print, electrical requirements, additional consumables, storage and waste disposal requirements to name a few. Quantitative data analysis involves determining the accuracy and precision of the new CD4 technology compared to one or more reference technologies. This method comparison relies on the outputs generated by a number of statistical tools (2). Several reference documents are available to guide the method validation process (3-6), but these are written in general terms, open to interpretation and not always applicable to CD4 data, especially since CD4 data has a broad range (<10cells/µl to >1000cells/µl). This will result in outliers influencing a correlation coefficient and therefore CD4 method comparison also based on linear regression is unsuitable when the agreement is not close to perfect (e.g. $r<0.975$), and evaluation on only 40 data pairs is rather low for the broad range of CD4 counts (7). In addition an absolute difference in data pairs changing over this range, which does not concur with clinical relevance. Table 1 lists some of the methods commonly reported in the literature for CD4 method comparison. This now also includes determining the sensitivity, specificity and misclassification of a new CD4 technology result compared to a reference at various clinical thresholds (targets)(8) , which also has become more relevant with the advances made with new point of care (POC) CD4 tests (9). POC CD4 tests now applied more for antiretroviral therapy (ART) initiation at specific clinical threshold and applied less for monitoring also reflects these changes in analytical approaches of method comparison. Over the years, the CD4 clinical thresholds have changed (10). Thresholds for ART initiation in many low and middle income countries have changed with time beginning at CD4 counts <200 cells/µl(11) , which increased to <350 cells/µl(12) and was

further raised to <500 cells/µl in the WHO 2013 guidelines(13). Table 1 also highlights limitations to some of the method comparison tools. We therefore applied various method comparison tools to an existing CD4 clinical study data set (14) to investigate the following: (i) the value of method comparison tools; (ii) the need for new tools with changing clinical needs; (iii) best practices for applying these tools and (iv) best algorithm (order) to using these tools to better perform method comparison on CD4 data. Additional aspects to evaluating CD4 technology such as precision (reproducibility and variability), type of technologies, ease of use and test components, appropriate settings for the technology use, interchangeability with existing technologies and effectiveness are not investigated in this study, but only aspects concerning statistical method comparison for measures of agreement (accuracy). The original data set (14) is chosen purely for the data, and it is not intended to repeat the main study objectives.

Methods

Clinical data set and CD4 technologies

A total of 150 peripheral K_3EDTA blood samples were obtained from patients sent for routine lymphocyte immunophenotyping to the Department of Immunology, Faculty of Medicine Siriraj Hospital, Bangkok, Thailand as previously described (14). Although this previously published study evaluated CD4 technology performance, we reanalysed the data with a different method comparison approach. The five CD4 counting technologies previously described (14) were: (i) TriTEST TruCOUNT performed on a FACSCalibur system (all Becton Dickinson, San Jose, CA, USA); (ii) FACSCount (Becton Dickinson, San Jose, CA, USA); (iii) Guava Personal Cell Analyser (PCA) Easy CD4 (Millipore Corporation, Billerica, MA, USA); (iv) CyFlow (Partec GmbH, Munster, Germany); (v) TriTEST dual platform DP (generated using the absolute lymphocyte count from a haematology analyser) also performed on the FACSCalibur (Becton Dickinson, San Jose, CA, USA) and a

haematology analyser. These are referred in the text as: TruCOUNT, FACSCount, Easy CD4, CyFlow and dual platform . As per guidelines (1) and also used in the original study (14), the FACSCount was considered the gold standard/reference technology against which the other four technologies were compared. It is important to note the FACSCount has an accurate reporting range between 50 – 5000 cells/uL(15), and consideration is therefore required when comparing this range across technologies, to exclude values <50cells/µl before performing the analysis. In this data set 23/150 values reported <50cells/µl on the FACSCount, but seeing as this was an exercise of data and not a focus on technology, these values were not excluded from the analysis in this study. Standard operating procedures and project design for the immunophenotyping methodologies as well as quality control are detailed elsewhere(14).

Method comparison tools

Table 1 describes the method comparison tools applied to this data set, which include the Bland-Altman, percentage similarity, percent difference, concordance correlation, sensitivity and specificity, misclassification and describes the relevant visualisation tools. Special mention is made of the concordance correlation coefficient *Pc*(16,17), which has only recently been introduced for CD4 method comparison(18). The concordance correlation coefficient *Pc* is a measure of combined precision (how far each observation deviates from a best-fit-line) and accuracy (how far the best-fit-line deviates from the absolute perfect 45° line). The scale provided for interpreting the *Pc* is as follows: *Pc*<0.9 = poor; *Pc* between 0.9 -0.95 = borderline (borderline to nearly good); *Pc*>0.95=good and *Pc*>0.99=excellent (ref 16, 17). Analyses were performed using Stata 13 (StataCorp, College Station, TX) and MedCalc software version 15.6.1 (MedCalc Software bvba, Ostend, Belgium) (19) for visualization tools. The CD4 thresholds applied for sensitivity, specificity and misclassification were:

100cells/µl, relevant for reflex testing to screen *Cryptococcal* meningitis (13), at 350cells/µl based on early ART initiation guidelines (12) and 500 cells/µl based on the latest WHO 2013 guidelines(13).

Results

Commonly used method comparison tools

The median CD4 counts observed from all technologies are similar: FACSCount = 245cells/µl; TruCOUNT = 265cells/µl; EasyCD4 = 233cells/µl; CyFlow = 250cells/µl and Dual Platform = 243cells/µl, with the absolute data range <50cells/µl to 1137cells/µl. Table 2 presents the method comparison result outputs for the four CD4 technologies compared to the FACSCount reference. The measures of agreement such as the Bland-Altman bias show the only CD4 technology that has a negative bias (on average generates lower CD4 counts) against FACSCount is the EasyCD4. This is not reflected by the mean percentage similarity and the mean percent bias, which shows all four CD4 technologies generate higher CD4 values than FACSCount.

Additional method comparison tools

Table 2 further lists misclassification rates across the three clinical thresholds, which are varied within and between each CD4 technology compared to the FACSCount reference. The greatest range in misclassification occurs at the 350cells/µl threshold. This is similarly shown with the sensitivity and specificity calculations and all CI overlap, except for one technology (CyFlow) at the 350cells/µl.

Best practices for performing method comparison

When the percentage similarity and the percent bias method comparison tools are applied in Table 2, using best practices (as highlighted in Table 1), they both identify a negative bias between FACSCount and EasyCD4. In addition, the percentage similarity CV calculated on all the data (n=150)

shows the EasyCD4 technology with the greatest lack of overall agreement with FACSCount, yet the sensitivity of the EasyCD4 at the 350cells/µl threshold is significantly (non-overlapping CI) more than the sensitivity of the CyFlow technology at this threshold, and with more specificity. The mean percent bias, however, identifies CyFlow with the greater bias against FACSCount, but EasyCD4 with the largest percent bias SD (increased variability, lack of precision) against FACSCount. Once the percent similarity and percent bias method comparison tools are applied using best practise, both identify CyFlow with the least overall agreement against FACSCount. This gap between the TruCOUNT, dual platform, EasyCD4 technologies and the CyFlow technology is also illustrated by the weaker concordance correlation. The CI does not overlap with the other technologies'; showing this agreement between CyFlow and FASCount is significantly weaker. This is further visualised in Figure 1 with the Passing Bablock plots, which clearly indicates the broader scatter (and greater distance between the regression and identity line) between CyFlow and the reference FACSCount. These plots also concur with the negative Bland-Altman bias of the Easy CD4 compared to FACSCount, since the regression (fitted) line is below the identity (ideal) line. Figure 2 illustrates the Bland-Altman difference plots, which show the limitation of this method if used in isolation, in that the bias changes over the range in CD4 counts. These plots do show the negative bias of the EasyCD4, the broad scatter of the CyFlow, and fewer outliers with the Dual platform technology. In addition the Easy CD4 vs FACSCount plot shows CD4 counts <100cells/µl appear to generate higher values using Easy CD4, yet >100cells/µl generate lower CD4 counts than FACSCount, which may indicate difference with this technology based on capillary action not hydrodynamic focussing in flow cytometry.

An algorithm approach to method comparison

Figure 1 outlines a process flow diagram recommending an approach to performing method comparison, starting with concordance correlation (*Pc)*, after which the process is divided into a qualitative and quantitative component, and finally followed by an additional component of

visualising ones comparison with Passing Bablok and agreement scatter plots. Table 3 further summarises the main outputs from the various method comparison tools in this novel algorithm approach using a modified heat map to represent not only the individual method comparison output values, but also their order of "acceptability", which would align with current guidelines such as WHO, ISO15189 and Westgard (2,20-22). The colours were assigned based on their relative performance compared to the FACSCount reference CD4 technology, in an attempt to grade the technologies among themselves from higher values to lower values relating good (green) to poor (red) performance. This illustrates a definite distinction between similar technologies such as the TruCOUNT, Dual platform, EasyCD4 and the more varied (less overall agreement) Cyflow. The method comparison indices show clearly that for the CyFlow compared to FACSCount, the trend is of a moderate Pc, a lower sensitivity and specificity, higher percentage similarity SD, higher bias SD, and greater overall misclassification than any other CD4 comparison with FACSCount. Among the remaining three technologies (TruCOUNT, dual platform and EasyCD4), there is some overlap with the method comparison outcomes, in that: a higher misclassification rate is reflected by a reduced sensitivity; a high concordance correlation is reflected by a small absolute bias and low percentage similarity CV; but that the percent bias is less predictable of other companion output indicators.

Discussion

Numerous CD4 counting technologies have been developed and implemented, ranging from manual microscope counting methods to high throughput flow cytometry(1), with the promise of new point of care (POC) technologies(9). Some POC technologies report CD4 as qualitative values not absolute counts, which will require a different method comparison analysis. In addition CD4 counts are used more for ART initiation at specific thresholds, which also requires a different approach to method comparison than traditional bias for example. A recent systematic review of thirty two studies

evaluating the accuracy and precision of fifteen CD4 technologies showed <16 studies presented data on bias and misclassification (21). The range in these performance parameters is also broad and therefore difficult to determine the most suitable CD4 technology without knowing the field in depth. The lack of standardization in method comparison was reported as one of the barriers to impact on decision for policy makers to introduce new CD4 technologies, especially for point of care testing. Guidance on method comparison best practice for CD4 data pair evaluation is clearly a need. The use of commercial reference material (i.e. fixed blood) is also well described for evaluating new CD4 technologies (not included in this analysis as the initial study focussed on clinically relevant fresh blood specimens for result comparison). In addition evaluation of external quality assessment panel results will also contribute to method comparison (22), and perhaps better identify accuracy between technologies using an aggregated mean(23) (but again not performed on fresh clinical specimens).

The decision for acceptable agreement between two technologies is not defined or restricted by any statistical rule, but ultimately is a matter of clinical judgment (24). In the case of CD4 counting, the clinical relevance of a change in the CD4 result generated from a new technology compared to a reference (or predicate) technology is the deciding factor. For example if a CD4 count from a reference technology is determined as 100cells/µl and the new technology yields 120cells/µl, this will not alter the clinical management of the patient, and therefore the technology is considered to have good accuracy, with an absolute difference of 20cells/µl. Similarly if a CD4 count from the reference technology is 1200cells/µl, and from the new technology is 1000cells/µl, this difference of 200cells/µl is also considered acceptable even though the absolute difference is 200cells/µl. It is this broad range in data (<10 to >1000cells/µl) that makes the method comparison analysis of CD4 counting unique from other medical tests, and should include stratification based on clinical thresholds. This was a limitation of sample size in this study, but did apply the threshold of 100cells/µl in the heat map. This is also evident from the several method comparison tools that are

applied to evaluating CD4 technologies, and the fact that studies report using more than one method comparison tool to draw conclusions. This study highlights the limitations of these more commonly used method comparison tools, and also provides guidance on best practices.

These method comparison tools applied to the overall data set (n=150) identified accurate CD4 technologies such as TruCOUNT and Dual platform, yet varied in their decision to select EasyCD4 or CyFlow as acceptable replacements for the FACSCount reference. Once these method comparison tools were applied using best practice, did they concur in their selection of suitable CD4 technologies. It is important, however, to note that method comparison is only one component of overall method evaluation (as mentioned in the introduction), and in-depth knowledge of the CD4 technology is required in combination with method comparison outcomes. In this regard, it is important to stress that using a lymphocyte count in a dual platform system for generation of absolute CD4+ T lymphocytes will produce highly variable results (22,25). This is due to the fact that on an inter-laboratory basis the differential count is extremely difficult to quality control or test using proficiency panels because of the lack of suitable material that controls differentials between different hematology analyzers. The CVs produced will be significantly higher than if the total leukocyte count was used and the percentage lymphocyte count from the flow cytometer. Globally this practice of using an absolute lymphocyte count should be avoided.

Although the objectives of the original study(14) were to determine the performance of the EasyCD4 and CyFlow in comparison to predicate technologies (TruCOUNT, Dual platform and FACSCount), our re-analysis of the data compared all technologies to the FACSCount as example of the influence of method comparison tools on one's decision. It is, however, worth commenting on some of the findings from the original study, and how these may be improved using the new method comparison algorithm approach described here. The original study emphasizes correlation, R^2 and bias. The latter is reported similarly, but the original study concluded with "an order of bias" (CyFlow< Dual

platform< TruCOUNT< FACSCount< EasyCD4), as well as reporting the EasyCD4 and CyFlow as being "substantially different", yet unable to quantify this. In addition the original study reports the EasyCD4 and CyFlow are "unlikely to affect monitoring as long as the systems are not used interchangeably", yet are unable to quantify this in values, all of which are now possible with our suggested algorithm.

Each method comparison tool has its limitation, but combining their outputs shows some complementarity. Misclassification rate is linked to sensitivity, and the Bland-Altman bias is linked to the percentage similarity CV, and both these indicators are linked to the concordance correlation. Based on this study, a recommended order in method comparison tools would be: (i) perform the concordance correlation first using the entire data set to measure overall accuracy and precision; (ii) interrogate the level of agreement using the Bland-Altman and percentage similarity tools based on their strengths (SD of the absolute bias <100cells/µl and percentage similarity CV >100cells/µl); (iii) determine clinical relevance using sensitivity and link this to misclassification rates, without forgetting that in CD4 testing, the specificity is a useful indicator of expected increases in program costs if more patients are identified eligible for ART. Less helpful method comparison parameters therefore are the Bland-Altman limits of agreement, and even though CD4 count log transformation has been suggested, it would not be easy to translate into clinical utility (24,26,27). This is similarly noted for the percent bias, and in this study (which has a limited sample size)(8) did not appear to add any additional value over the absolute bias or percentage similarity CV.

Visual inspection of plots is also an essential component of method comparison to identify outliers and trends (28), which was addressed in this study using the Passing Bablock and Bland-Altman plots. These plots complement the concordance correlation as well as bias measurements, and do give an overall sense of outliers and trends. It is, however, worth mentioning that some modifications to existing plots may also be more suitable to CD4 data, based on our findings. One of the most widely used plots is the Bland-Altman plot(29), yet it will illustrate a funnel shape due to

the difference between data pairs changing over the range in data as previously discussed (30). It may therefore be more useful to apply best practices as described above to the Bland-Altman and percentage similarity scatter plots. In addition it may also be more suitable to plot the absolute CD4 count of the reference (or predicate) technology on the horizontal axis, and not the average ([new test +reference]/2) since the impact of switching to a new or improved CD4 technology for clinical use can really only be assessed if the absolute and not the average CD4 value is presented since a clinician in the field will only ever receive one CD4 result, not the average to make a decision on treatment management.

In addition to evaluating new CD4 technologies for field use, similar data analysis for method comparison affects external quality assessment programs for CD4 counting (31). It would therefore be advantageous to apply the recommendations described in this study to EQA panel evaluations, and larger data sets of multiple CD4 technology comparisons to further establish acceptable limits for *Pc*, absolute bias SD, percentage similarity CV, sensitivity, specificity and misclassification rates.

Acknowledgements and support for the study

This publication was made possible by the financial support of the Thailand Research Fund (TRF) – Senior Research Scholar (KP and KS), the TRF Distinguished Research Professor Grant DPG5980001 and Grand Challenges Canada, grant 0007-02-01-01-01 (WSS and LES). The authors confirm they do not have any conflict of interest.

References

1. Stevens W, Gelman R, Glencross DK, Scott LE, Crowe S, Spira T. Evaluating new CD4 enumeration technologies for resource-contrained countries. Nature Reviews Microbiology 2008;S29-S38.
2. http://www.westgard.com/mvtools.htm. The data analysis tool kit. Westgard QC; Accessed on July 2016.
3. CLSI. Evaluation of Precision Performance of Quantitative Measurement Methods; Approved Guideline—Second Edition EP5-A2 (ISBN 1-56238-542-9) NCCLS, 940 West Valley Road, Suite

1400, Wayne, Pennsylvania 19087-1898 USA. Tholen DW, Kallner A, Kennedy J, Krouwer J, Meier K, editors: NCCLS; 2004.
4. Bergeron M, Lustyik G, Phaneuf S, Ding T, Nicholson JK, Janossy G, Shapiro H, Barnett D, Mandy F. Stability of currently used cytometers facilitates the identification of pipetting errors and their volumetric operation: "time" can tell all. Cytometry 2003;52B:37-9.
5. Barnett D, Bird AG, Hodges E, Linch DC, Matutes E, Newland A, Reilly JT. Guidelines for the enumeration of CD4+ T lymphocytes in immunosuppressed individuals. Clin Lab Haematol 1997;19:231-241.
6. CLSI. Method comparison and bias estimation using patient samples; Approved guideline - Second Edition (Interim Revision). . 940 West Valley Road, Suite 1400, Wayne, Pennslylvania 19087-1898 USA.; 2010.
7. CLSI. Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline-Second Edition.; 2002.
8. Scott LE, Campbell J, Westerman L, Kestens L, Vojnov L, Kohastsu L, Nkengasong J, Peter T, Stevens W. A meta-analysis of the performance of the Pima CD4 for point of care testing. BMC Med 2015;13:168.
9. http://www.unitaid.eu/images/marketdynamics/publications/UNITAID_HIV_Nov_2015_Dx_Landscape.PDF. HIV/AIDS diagnostics technology landscape - 5th edition. . UNITAID; Accessed on July 2016.
10. Stevens WS, Ford N. Time to reduce CD4+ monitoring for the management of antiretroviral therapy in HIV-infected individuals. S Afr Med J 2014;104:559-60.
11. WHO. Antiretroviral therapy for HIV infection in adults and adolescents: towards universal access. Recommendations for a public health approach. Volume 2007. Geneva: World Health Organization; 2006.
12. http://www.who.int/hiv/pub/arv/adult2010/en/. Antiretroviral therapy for HIV infection in adults and adolescents. Recommendations for a public health approach: 2010 revision.; 2010.
13. http://www.who.int/hiv/pub/guidelines/arv2013/en/. Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection. World Health Organisation; 2013.
14. Pattanapanyasat K, Phuang-Ngern Y, Sukapirom K, Lerdwana S, Thepthai C, Tassaneetrithep B. Comparison of 5 flow cytometric immunophenotyping systems for absolute CD4+ T-lymphocyte counts in HIV-1-infected patients living in resource-limited settings. J Acquir Immune Defic Syndr 2008;49:339-47.
15. http://www.bdbiosciences.com/ds/ab/others/339011.pdf. BD FACSCount System User's Guide for use with BD FACSCount CD4 Reagents. Becton Dickinson; Last accessed July 2016.
16. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255-68.
17. Lin LK. A note on the concordance correlation coefficient. Biometrics 2000;56:324-325.
18. Dieye TN, Diaw PA, Daneau G, Wade D, Sylla Niang M, Camara M, Diallo AA, Toure Kane C, Diop Ndiaye H, Mbengue B and others. Evaluation of a flow cytometry method for CD4 T cell enumeration based on volumetric primary CD4 gating using thermoresistant reagents. J Immunol Methods 2011;372:7-13.
19. https://www.medcalc.org/. MedCalc statistical software.
20. http://who.int/diagnostics_laboratory/evaluations/cd4/en/. Summary of the study for Multicentre Evaulation of CD4 technologies as part of the WHO Prequalification of Diagnostics Programme (PQDx). Geneva: WHO: World Health Organisation; Accessed July 2016.

21. Peeling RW, Sollis KA, Glover S, Crowe SM, Landay AL, Cheng B, Barnett D, Denny TN, Spira TJ, Stevens WS and others. CD4 enumeration technologies: a systematic review of test performance for determining eligibility for antiretroviral therapy. PLoS ONE 2015;10:e0115019.
22. Whitby L, Whitby A, Fletcher M, Barnett D. Current laboratory practices in flow cytometry for the enumeration of CD 4(+) T-lymphocyte subsets. Cytometry B Clin Cytom 2015;88:305-11.
23. Gossez M, Malcus C, Demaret J, Frater J, Poitevin-Later F, Monneret G. Evaluation of a novel automated volumetric flow cytometer for absolute CD4+ T lymphocyte quantitation. Cytometry B Clin Cytom 2016.
24. Hollis S. Analysis of method comparison studies. Ann Clin Biochem 1996;33 (Pt 1):1-4.
25. Higgins J, Hill V, Lau K, Simpson V, Roayaei J, Klabansky R, Stevens RA, Metcalf JA, Baseler M. Evaluation of a single-platform technology for lymphocyte immunophenotyping. Clin Vaccine Immunol 2007;14:1342-8.
26. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
27. Scott LE, Galpin JS, Glencross DK. Multiple method comparison: Statistical model using percentage similarity. Cytometry 2003;54B:46-53.
28. Petersen PH, Stockl D, Blaabjerg O, Pedersen B, Birkemose E, Thienpont L, Lassen JF, Kjeldsen J. Graphical interpretation of analytical data from comparison of a field method with reference method by use of difference plots. Clin Chem 1997;43:2039-46.
29. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 1995;346:1085-7.
30. Scott LE, Galpin JS, Glencross DK. Multiple Method Comparison: Statistical Model Using Percentage Similarity. Cytometry: Clinical Communication 2003;54B:46-53.
31. Glencross DK, Aggett HM, Stevens WS, Mandy F. African regional external quality assessment for CD4 T-cell enumeration: development, outcomes, and performance of laboratories. Cytometry B Clin Cytom 2008;74 Suppl 1:S69-79.
32. Pollock MA, Jefferson SG, Kane JW, Lomax K, MacKinnon G, Winnard CB. Method comparison--a different approach. Ann Clin Biochem 1992;29 (Pt 5):556-60.
33. Lin L, Torbeck LD. Coefficient of accuracy and concordance correlation coefficient: new statistics for methods comparison. PDA J Pharm Sci Technol 1998;52:55-9.
34. Passing H, Bablok. A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, Part I. J Clin Chem Clin Biochem 1983;21:709-20.
35. Deming SN, Morgan SL. The use of linear models and matrix least squares in clinical chemistry. Clin Chem 1979;25:840-55.

Table 1: Description of method comparison tools relevant to evaluating new CD4 enumeration technologies.

Method comparison	Principle	Advantage	Limitation	Best practise
Bland-Altman(26)	Measure of agreement through the absolute difference (t-r); the bias is a measure of accuracy; the standard deviation (SD) of the bias is a measure of precision.	The bias reports the same units as CD4 and easy to interpret (positive or negative). This method is robust at CD4 counts <100cells/µl.	The bias is meaningless without the confidence interval (CI), the SD and the median CD4 to report amount of variability in the absolute bias. The absolute difference changes over the range of data, and the difference plot displays a funnel shape with increasing range of CD4. This influences the limits of agreement (LOA) and the absolute bias.	Include the confidence intervals (CI) for the mean difference. Report the mean difference for CD4 counts <100cells/µl. Report the standard deviation (SD) of the bias to determine a measure of precision (variability) of the absolute difference.
Percentage similarity(30)	Measure of agreement through transforming the relationship between the data pairs into a percentage value using the formula: [((t+r)/2) /r] x 100. Data pairs with the same value will be 100% similar and data pairs where the (new) method is greater than the reference will be > 100%, and	The mean percentage similarity indicates accuracy and the standard deviation indicates precision between the two methods. There is no negative scale and the transformed percentage value is comparable between studies using different specimens.	The percentage similarity is not robust at CD4 counts<100cells/µl, and although outliers may not be clinically relevant (t=50cells/µl; r=25cells/µl; percentage similarity mean is 150%, yet the absolute difference is 25cells/µl), they influence the overall agreement.	The percentage similarity mean (accuracy) and the percentage similarity SD (precision) are represented as a single unit, percentage similarity coefficient of variation (CV), which describes overall agreement between two methods. Report the %CV for CD4 counts >100cells/µl.

	conversely <100% if the (new) method has a value smaller than the reference.			
Percent difference(32)	The percentage difference determines the difference between data pairs and represents this as a percentage of the data pair average: [(t-r)/((t+r)/2)] x 100. A 0% difference shows equality between the data pairs.	This is almost a combination between the Bland-Altman and the percentage similarity methods, and as for the percentage similarity, the transformed data pair can be compared between studies.	This tool was first reported only as a graphical output and the bias (especially negative) may be difficult to translate into absolute difference for CD4 counting. The method also appears less robust <100cells/µl.	The mean bias difference and standard deviation >100cells/µl can be calculated when deciding to replace one method with another.
Concordance correlation(16,33)	The concordance correlation coefficient (*Pc*) determines the degree (°) to which data pairs align along the 45° line through the origin, when data pairs are represented on an x,y scatter plot: *Pc = p x Cb. [Pc* contains a measure of precision *P* and accuracy *Cb].*	Precision (*p*) is measured by the distance that each observation deviates from the best-fit-line, and accuracy is measured by how far (°) the best-fit-line deviates from the 45° line. The strength of this overall agreement (*Pc*) ranges from <0.9 (poor) to >0.99 (almost perfect).	*Pc* has not been applied to CD4 data, and must not be confused with correlation (r) or linear regressions (R) which are applied more to measure association [10, 12-15]. The strength of agreement is dependent on the data pairs and therefore may not be easily translated between studies using different specimen ranges.	Determine *Pc* and compare between technologies. Do not confuse with correlation (r) or linear regression R.

Sensitivity, specificity	Sensitivity and specificity are measures of accuracy between two tests in making a correct diagnosis (or discrimination between positive and negative).	In regards to CD4 counting, the "diagnostic" thresholds mostly applied are: 100cells/µl, 350cells/µl and 500cells/µl, based on clinical needs [7] [8]. The absolute CD4 counts are represented in binary (0, 1) based on which threshold is being investigated. The fraction of true and false positive and negative results is reported.	This has only recently been applied to evaluating CD4 counting technologies, and is dependent on sample size.	The sensitivity has relevance for clinical utility of CD4 counting at various thresholds, and the specificity can indicate changes expected to current practice and potential impact on outcomes. Confidence intervals are critical to compare between technologies on a single data set.
Misclassification	False positive and false negative rates calculated at a clinical threshold (as determined above for sensitivity and specificity) are combined to report total rate of specimens misclassified compared to a reference.	Misclassification is useful for interpreting clinical and programmatic impact that a new technology will have if it replaces a reference technology.	Total misclassification is dependent on the overall sample size. Upward and downward misclassification, which is similar to positive and negative predictive values (percentage of patients requiring [or not] treatment incorrectly identified at a specific threshold) is not easily interpreted for CD4 data method comparison.	Report the false positive and false negative rates at a clinically relevant threshold to further identify the cause for misclassification, to interpret clinical impact of replacing existing CD4 reference technologies.
Visualization tools	The Bland-Altman difference plot and the percentage similarity scatter plot (or histogram) are useful to identify outliers and general trends of agreement.	Passing-Bablok is a non-parametric plot which is not affected by outliers from variables with a linear relation. It plots the ideal (or identity) (45°) line and the fitted (or regression) line which gives an indication of direction of the bias (distance from °45) and precision (spread around	Passing-Bablok and Deming regression plots are strongly discouraged in their use of simple linear regression (as in CLSI EP9(7)) as the latter are too sensitive to outliers in CD4 data, which as mentioned above; the bias changes over the range.	Passing-Bablok plot complements the concordance correlation.

	Two other useful plots are the Passing-Bablok (34) and the Deming regression (35) plots.	the regression line). The Deming regression is ideal when repeat measurements are performed as it gives weight to the data points as a function of %CV around each measurement.		

t = new test method, r = reference method; *Typically a difference plot uses the average [(t-r)/2] on the horizontal plot, but for clinical interpretation one only has the t method result.

Table 2: Method comparison data analysis.

	TruCOUNT	EasyCD4	CyFlow	Dual Platform
Bland-Altman Agreement				
Mean bias: test- Reference (CI)	18 (12,23)	-10 (-16, -4)	24 (13, 35)	4 (-2, 9)
LOA	(-51, 87cells/µl)	(-83, 63cells/µl)	(-108, 157cells/µl)	(-67, 74cells/µl)
SD bias (<100cells/µl, n=33)	35 (9cells/µl)	37 (7.8 cells/µl)	68 (20cells/µl)	36 (10 cells/µl)
Percentage similarity agreement				
Mean % (SD), CV	106% (14%), 13%	109% (53.5%), 49%	114.9% (38.1%), 33%	105.6% (18.5%), 18%
Mean % (SD), CV >100cells/µl, n=117	103% (5.0%), 4.7%	98% (5.7%), 5.2%	105% (10.0%), 8.7%	101% (5.6%), 5.3%
Percent difference agreement				
Percent mean bias, SD	12.3% (28%)	18.3% (107%)	29.9% (76%)	11.1% (37%)
Percent mean bias, SD (>100cells/µl), n=117	6%, (10%)	-3.9% (11.8%)	10% (20%)	2% (11.2%)
Concordance correlation (*Pc*)				
Strength of Agreement	0.988 (0.984, 0.991)	0.987 (0.983, 0.991)	0.959 (0.947, 0.972)	0.989 (0.986, 0.993)
Sensitivity				
100 cells/µL	97% (84.2%-99.9%)	90.9% (75.7%-98.1%)	81.8% (64.5%-93%)	93.9% (79.8%-99.3%)
350 cells/µL	95.3% (89.5%-98.5%)	99.1% (94.9%-100%)	84.9% (76.7%-91.1%)	95.3% (89.5%-98.5%)
500 cells/µL	99.2% (95.7%-100%)	99.2% (95.7%-100%)	96% (91.1%-98.7%)	99.2% (95.75-100%)
Specificity				
100 cells/µL	100% (96.9%, 100%)	98.3% (94%, 99.8%)	100% (96.9%, 100%)	100% (96.1%, 100%)

350 cells/µL	97.7% (89%, 99.9%)	97.7% (88%, 99.9%)	93.2% (81.3%, 98.6%)	95.5% (84.5%, 99.4%)
500 cells/µL	100% (85.8%, 100%)	83.3% (62.6%, 95.3%)	95.8% (78.9%, 99.9%)	95.8% (78.9%, 99.9%)
Misclassification				
100 cells/µL				
False positive	0%	1.33%	0%	0%
False negative	0.67%	2.0%	4.0%	1.33%
Total misclassification	0.67%	3.33%	4.0%	1.33%
350 cells/µL				
False positive	1.5%	0.6%	2.0%	1.33%
False negative	3.33%	0.6%	10.6%	3.33%
Total misclassification	4.83%	1.2%	12.6%	4.66%
500 cells/µL				
False positive	0%	2.66%	0.6%	0.6%
False negative	0.66%	0.6%	3.33%	0.6%
Total misclassification	0.66%	3.26%	3.93%	1.2%

Confidence intervals quoted are at 95%

Table 3: A modified heat map approach to summarising the method comparison outputs. The colours indicate: Green (good performance = highest values), orange (acceptable/borderline performance= mid-range values), red (weaker performance = lower range values) compared to the FACSCount reference CD4 technology.

	CD4 technologies comparted to FACSCount			
Parameters	**TruCount**	**DualPlatform**	**EasyCD4**	**CyFlow**
Pc	0.988	0.989	0.987	0.959
Sensitivity at 350	95.3	95.3	99.1	
Specificity at 350	97.7	95.5	97.7	
Total misclassification at 350	4.8	4.66	1.2	
Bland-Altman bias SD <100c/µl	9	10	7.8	
Percentage similarity CV>100c/µl	4.7%	5.3%	5.2%	8.7%
Percent mean bias >100c/µl	6%	2%	-4%	10%

Figure 1: Process flow diagram recommending an algorithm for performing method comparison

Figure 2: Passing Bablock plots of all CD4 technologies compared to the reference technology (FACSCount). The horizontal axis is the reference method and the vertical axis the new method under evaluation. The dotted lines are the 95% confidence interval of the regression line (solid blue line) away from the identity line (x=y). The legend shows the concordance correlation (*Pc*).

Figure 3: Bland-Altman difference plots. The vertical axis is the difference between new technologies and the reference FACSCount technology, and the horizontal axis is the reference technology. The limits of agreement as well as the mean bias with CI are included. Two additional lines are shown on the plot of CyFlow vs FACSCount to illustrate the typical funnel shape that occurs with differences in CD4 data pairs changing over the range of CD4 counts.

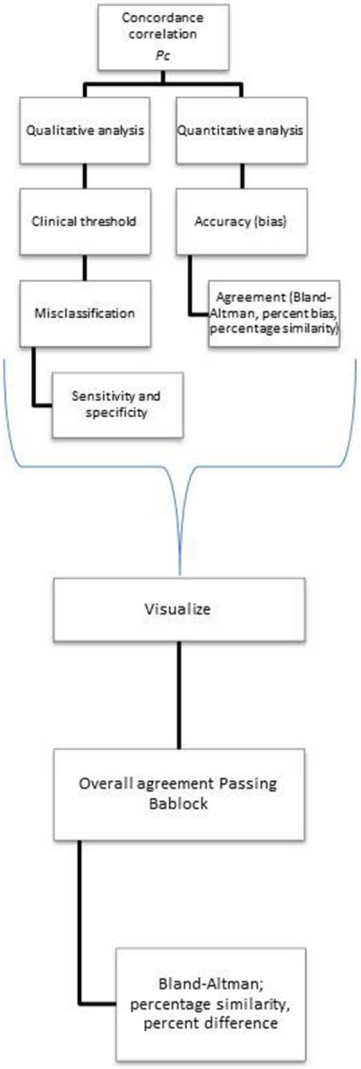

Figure 1: Process flow diagram recommending an algorithm for performing method comparison

104x235mm (96 x 96 DPI)

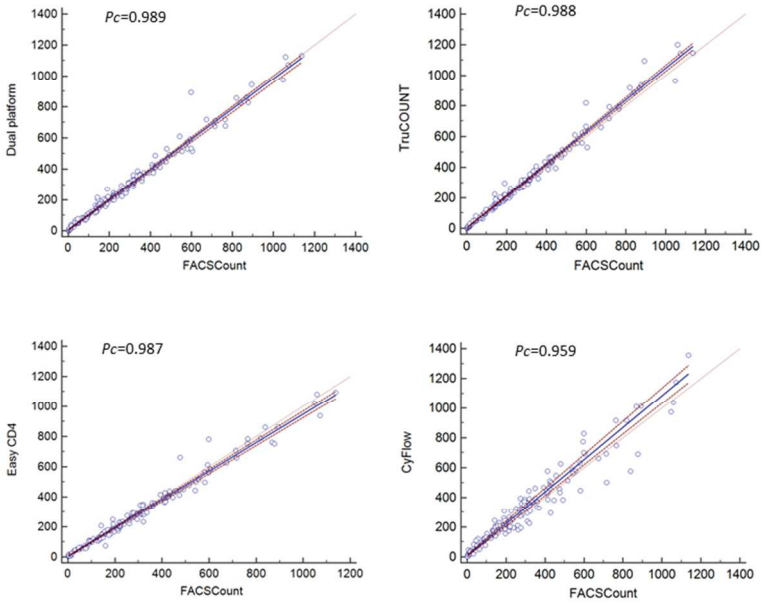

Figure 2: Passing Bablock plots of all CD4 technologies compared to the reference technology (FACSCount). The horizontal axis is the reference method and the vertical axis the new method under evaluation. The dotted lines are the 95% confidence interval of the regression line (solid blue line) away from the identity line (x=y). The legend shows the concordance correlation (Pc).

254x190mm (96 x 96 DPI)

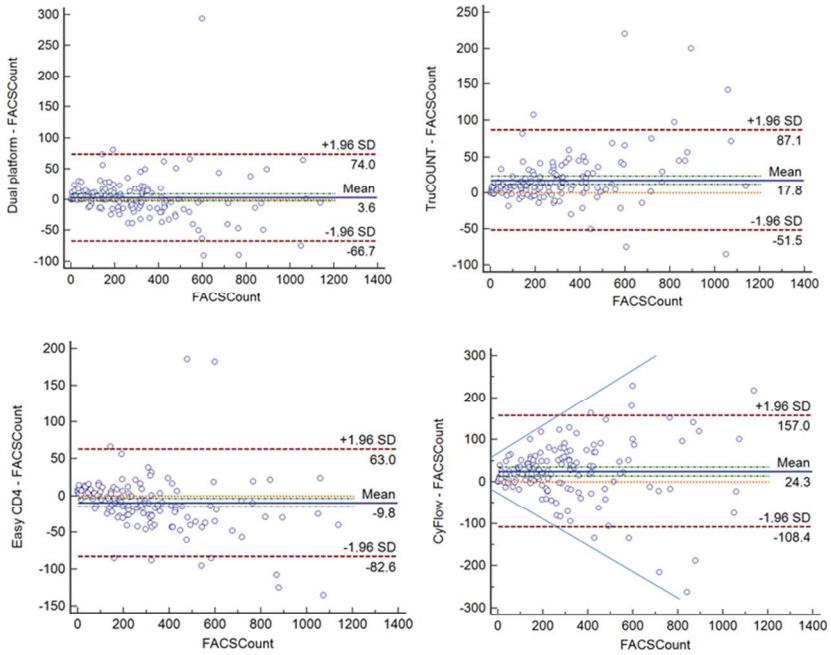

Figure 3: Bland-Altman difference plots. The vertical axis is the difference between new technologies and the reference FACSCount technology, and the horizontal axis is the reference technology. The limits of agreement as well as the mean bias with CI are included. Two additional lines are shown on the plot of CyFlow vs FACSCount to illustrate the typical funnel shape that occurs with differences in CD4 data pairs changing over the range of CD4 counts.

254x190mm (96 x 96 DPI)