Wade M. Sanders*, Carrie A. Karvonen-Gutierrez and Daniel S. McConnell

Assessment for potential bias in multiplexed IL-10 and TNF-α from plex count

https://doi.org/10.1515/cclm-2025-0458
Received April 14, 2025; accepted January 5, 2026;
published online January 19, 2026

Abstract

Objectives: Multiplex arrays offer a high-throughput, cost-effective means for quantifying multiple analytes simultaneously, essential for large-scale biomarker research. However, the shared reactive environment in multiplex assays could lead to variations in results compared to single plex assays, potentially impacting outcomes. This study aimed to explore these differences by examining a "3-plex" cytokine panel vs. single plex assays for tumor necrosis factor-alpha (TNF-α) and interleukin-10 (IL-10).

Methods: Using the R&D Systems Luminex HS Cytokine Panel A, 72 serum samples were tested across several batches, including three "3-plex" and two "1-plex" batches each for IL-10 and TNF-α. We evaluated differences in cytokine values using paired t-tests, comparing within 1-plex, between 3-plex, and between 1-plex and 3-plex assays. Bland-Altman plots visually assessed absolute and percentage differences.

Results: IL-10 values were similar between 1-plex and 3-plex assays, showing an average difference of 4.2 fmol/L (10.7 %), which was less than the within plex differences. Conversely, TNF-α showed a 16.7 % difference in the between plex comparison, compared to a 12 % difference in the within plex comparisons. Statistically significant differences emerged mainly for IL-10 across all comparisons. Bland-Altman analyses indicated pronounced variability at low analyte concentrations.

Conclusions: While the multiplex assays demonstrated variation at low analyte levels, especially for IL-10, such differences might not substantially affect comparability when mixed with single plex assays, particularly in datasets dominated by low concentrations.

Keywords: cytokines; Luminex; multiplex assays; multiplex immunoassay; suspension multiplex

Introduction

Epidemiological studies conducting biomarker research have numerous laboratory analytic tools available to quantitatively link the amount of a given analyte (hormone, marker, target, etc.) in a biospecimen (e.g., serum, plasma, urine) to a signal output. That signal is then interpreted by comparing that signal to the signals from standard specimens with known quantities of the analyte. Within this basic framework there is significant heterogeneity in how the biospecimen is prepared, how signal generation is accomplished, what that signal is, how the analyte is held within a detectable phase, and what that detectable phase is. This heterogeneity begets a multitude of practical considerations when selecting assays for use, particularly in research studies in which accuracy and precision are critically important.

In addition to the importance of accuracy and precision, the pragmatic issues of cost and needed biospecimen volume are important considerations. Assays with excellent sensitivity and repeatability often come at a considerable monetary and labor cost and require high sample volume. As such, many research studies with large biomarker investments are moving from enzyme-linked immunosorbent assay (ELISA) methodologies to multiplex assays. ELISA assays are time and labor intensive, require multiple steps, and allow the measurement of only one analyte at a time. On the other hand, multiplex assays can capture and quantify **multiple** analytes from a single sample volume, allowing for immense scaling up of biomarker research and improving efficiency. While multiplex methodologies have been validated against gold-standard methods for a variety of biomarkers including inflammatory and metabolic markers [1, 2], growth factors [3], genomic tags [4], and auto-antibodies [5] (among many other markers), validation studies of panel variations within the same multiplex methodology have not been published. While it is not feasible to test every possible panel combination against all others, it is critical to understand whether

Presented as a poster at the 2024 Endocrine Society Conference in Boston, MA.

***Corresponding author: Wade M. Sanders**, Department of Epidemiology, School of Public Health, The University of Michigan, 1415 Washington Heights, Ann Arbor 48109, MI, USA, E-mail: wadesand@umich.edu. https://orcid.org/0000-0002-9759-5543

Carrie A. Karvonen-Gutierrez and Daniel S. McConnell, Department of Epidemiology, School of Public Health, The University of Michigan, Ann Arbor, MI, USA

panel variations using multiplex methods may impact the validity of the results so that concentrations of a given analyte can be compared regardless of which other analytes were included in the multiplex.

To evaluate this research question, concentrations of interleukin-10 (IL-10) and tumor necrosis factor alpha (TNF-α) were examined between assays run in single plex (i.e., only one analyte per assay, "1-plex") vs. multiplex (i.e., IL-10, TNF-α, and IL-1β assayed together, "3-plex"). We hypothesized that any differences between the assay methods would not exceed what could be attributed to inter-assay variation.

Materials and methods

This study utilized 72 serum samples, collected from participants at the Michigan site of the Study of Women's Health Across the Nation (MI-SWAN), at the 12th follow-up visit (2010) and stored at −80 °C until thawed for this study. SWAN is a study of health and functioning across the menopausal transition and into late adulthood; the sampling and recruitment details for SWAN are published [6]. Participant's biospecimens were randomly selected from the larger MI-SWAN cohort; women with samples in this analysis had slightly greater prevalence of metabolic syndrome (p=0.0051) but no differences in diabetes status, dyslipidemia, hypertension, waist circumference, race/ethnicity, and body mass index. The SWAN protocol was approved by the University of Michigan's Institutional Review Board, and all participants provided written informed consent at each follow-up visit.

All samples were assayed for IL-10 and TNF-α using the Luminex HS Cytokine Panel A (R&D Systems, Minneapolis, MN, USA; catalog #LHSCM000) according to manufacturer instructions. IL-1β was included in this multiplex panel, but the results are not reported here. Manufacturer reported IL-10 intra-assay coefficients of variation (CV) were 7.3 , 3.8, and 5.2 % at 200, 1,413, and 28,952 fmol/L, respectively, and inter-assay CVs were 12.5 , 9.7, and 10.1 % at 195, 1,413, and 29,342 fmol/L, respectively. Manufacturer reported TNF-α intra-assay coefficients of variation (CV) were 4.9 , 4.1, and 6.8 % at 316, 2,535, and 59,780 fmol/L, respectively, and inter-assay CVs were 10.7 , 8.0, and 10.2 % at 292, 2,379, and 64,499 fmol/L, respectively. Assays were conducted by trained technicians at the Central Ligand Assay Satellite Services (CLASS) Laboratory at the University of Michigan using a Luminex MAGPIX for fluorometric reading. All test plates included a duplicate measure of quality control materials from R&D Systems (Catalog # QC11) to test for accuracy, as well as internal lab serum pools (American Red Cross Pool, male and female) to test for inter-assay variation. Quality control measures were compared to reference assays conducted at CLASS Lab using color-coded (by plex count) scatter plots, plotting the low and mid R&D controls against the other, as well as the male and female serum pools.

Each sample was assayed a total of seven times in singleton: three separate "3-plex" batches, with each plate quantifying all three markers; two separate IL-10 "1-plex" batches, where only IL-10 was quantified on a given plate; and two separate TNF-α "1-plex" batches, where only TNF-α was quantified on a given plate. On the first day of assay, two 3-plex batches and two IL-10 1-plex batches were conducted, alternating between the plex counts with an offset of 1.25 h to allow for the fluorometric read time. On a separate day, an additional, equivalent, and pristine serum aliquot was used to conduct the two TNF-α 1-plex batches and an additional 3-plex batch. Between batch setups, samples were capped and returned to 4 °C and then pulled with sufficient time to warm to room temperature.

There was a high prevalence of low concentrations for IL-10, and therefore an expanded lower limit of detection (LLoD) was defined as per the Clinical and Laboratory Standards Institute (CLSI) EP17 guidelines to determine where left-censoring (i.e., censoring of low values) should occur [7]. Participants with at least one IL-10 result below 13.5 fmol/L (the determined LLoD) were excluded from all IL-10 analyses, leaving 35 samples with consistently readable IL-10 concentrations; the number of usable values in a given batch ranged from 40 to 65. Of the 72 subjects, two did not have sufficient sample volume for the full set of TNF-α 1-plex batches and were therefore excluded from the TNF-α analyses. Differences between paired IL-10 and paired TNF-α values were assessed for normality by Shapiro-Wilk testing and Q-Q plots (included in the Supplementary Materials). Except for the within 1-plex comparison, IL-10 results required log-transformation prior to paired t-test and Bland-Altman analysis due to failing normality testing.

Paired sample differences were evaluated for statistical significance utilizing paired t-tests (α=0.01). For both IL-10 and TNF-α, we compared the paired 1-plex batch results, the paired 3-plex batch results (batch 1 vs. 2, 1 vs. 3, and 2 vs. 3), and the average of 1-plex and 3-plex results to one another. Then, correlations for the aforementioned comparisons were assessed graphically using scatter plots comparing a least squares linear regression with a 1:1 reference line; linear regression equations were used to estimate constant error (y-intercept) and proportional error (1.0-slope) for each scatter plot; differences were examined with Bland-Altman plots of differences and percent differences. Notably, the 3-plex differences or % differences (comparing batches

1–2, 1–3, and 2–3) were averaged together prior to Bland-Altman analysis. Each Bland-Altman plots includes a mean difference, upper limit of agreement, and lower limit of agreement line, each with 95 % confidence interval error bars, and with the 0 of the y-axis serving as the reference zero line.

Statistical analyses were conducted in SAS version 9.4 (SAS Institute, Cary, NC), and the figures were generated in Microsoft Excel version 2312. Linear regressions and correlation coefficients were generated directly in Excel.

Results

Results for the 72 samples from the seven batches of assays are summarized in Table 1. The batchwise averages for 1-plex IL-10 were 32.0 fmol/L (SD 23.4) and 30.9 fmol/L (SD 25.7), and for 3-plex IL-10 were 29.8 fmol/L (SD 23.3), 32.4 fmol/L (SD 24.5), and 31.7 fmol/L (SD 22.0). The batchwise averages for 1-plex TNF-α were 376.3 fmol/L (SD 164.2) and 388.3 fmol/L (SD 167.2), and for 3-plex TNF-α were 333.2 fmol/L (SD 142.6), 316.3 fmol/L (SD 139.4), and 326.8 fmol/L (SD 150.0). Comparisons within 1-plex and 3-plex batches are summarized in Table 2. For IL-10, all comparisons were statistically significantly different, whereas for TNF-α only the within 3-plex (comparing batches 1 and 2) and the between plex comparison were statistically significantly different.

Table 1: Summary of 1-plex and 3-plex results and differences.

		Interleukin 10			Tumor necrosis factor alpha		
		fmol/L			fmol/L		
Assay	Batch/comparison	Mean	SD	n	Mean	SD	n
1-plex	1	32.0	23.4	65	376.3	164.2	70
1-plex	2	30.9	25.7	46	388.3	167.2	70
1-plex	1 & 2 difference	6.8	4.9	46	−12.0	49.5	70
3-plex	1	29.8	23.3	40	333.2	142.6	70
3-plex	2	32.4	24.5	41	316.3	139.4	70
3-plex	3	31.7	22.0	61	326.8	150.0	70
3-plex	1 & 2 difference	−3.7	4.6	35	16.9	35.9	70
3-plex	1 & 3 difference	−8.4	8.5	39	6.4	53.5	70
3-plex	2 & 3 difference	−5.9	6.6	41	−10.5	41.8	70
Both	Between plex difference	4.2	4.3	35	56.9	36.1	70

Correlation plots, Bland-Altman plots of differences, and Bland-Altman plots of % differences are displayed in Figures 1A–I and 2A–I for each comparison (e.g., within 1-plex, within 3-plex, and between plex) for IL-10 and TNF-α, respectively. The correlation coefficients were less than 0.975 for the within 3-plex IL-10 comparison (Figure 1B), the between plex IL-10 comparison (Figure 1C), and for all TNF-α comparisons (Figures 2A–C), indicating our estimates of error may be affected by scatter in those cases. The IL-10 comparisons displayed some constant error (the greatest was 8.9 fmol/L for within 1-plex) and a proportional error of about 14 % for the within 3-plex comparison only. Constant errors for TNF-α were minor in comparison to the result averages, but proportional error was 6.2 % for the within 3-plex comparison, and 15 % for the between plex comparison.

The Bland-Altman plots were consistent with the error estimates from our linear regression. Within plex difference comparisons of IL-10 were heavily affected by proportionally large constant errors at low values, and differences were primarily unidirectional as a result (Figure 1D and E, G and H) – the IL-10 between plex % difference comparison alone formed a diffuse funnel pattern around the proportional error estimate (Figure 1I). The TNF-α Bland-Altman plots of % difference all formed funnels arrayed roughly around the estimated proportional error (Figure 2G–I).

Quality control results on these test plates were generally consistent with those on reference assays conducted by the CLASS Lab in terms of both magnitude and concentration-associated variability (Table 3). Scatter plots of internal laboratory control results (pooled male and

Table 2: Summary of all paired *t*-test results for 1-plex and 3-plex batches.

	Interleukin 10					Tumor necrosis factor alpha				
Comparison	Unit	Mean diff.	99 % CI	T-Stat.	p-Value	Unit	Mean diff.	99 % CI	T-Stat.	p-Value
Within 1-plex	fmol/L	8.13	(6.42, 9.84)	13.0	<0.0001	fmol/L	−12.04	(−27.7, 3.63)	−2.0	0.0457
Within 3-plex (1 vs. 2)	ln(fmol/L)	−0.13	(−0.19, −0.07)	−5.6	<0.0001	fmol/L	16.91	(5.54, 28.3)	3.9	0.0002
Within 3-plex (1 vs. 3)	ln(fmol/L)	−0.32	(−0.41, −0.23)	−10.1	<0.0001	fmol/L	6.43	(−10.5, 23.4)	1.0	0.3184
Within 3-plex (2 vs. 3)	ln(fmol/L)	−0.19	(−0.27, −0.10)	−6.0	<0.0001	fmol/L	−10.48	(−23.7, 2.76)	−2.1	0.0397
Between plexes	ln(fmol/L)	0.07	(0.01, 0.12)	3.4	0.002	fmol/L	56.88	(45.1, 68.7)	12.8	<0.0001

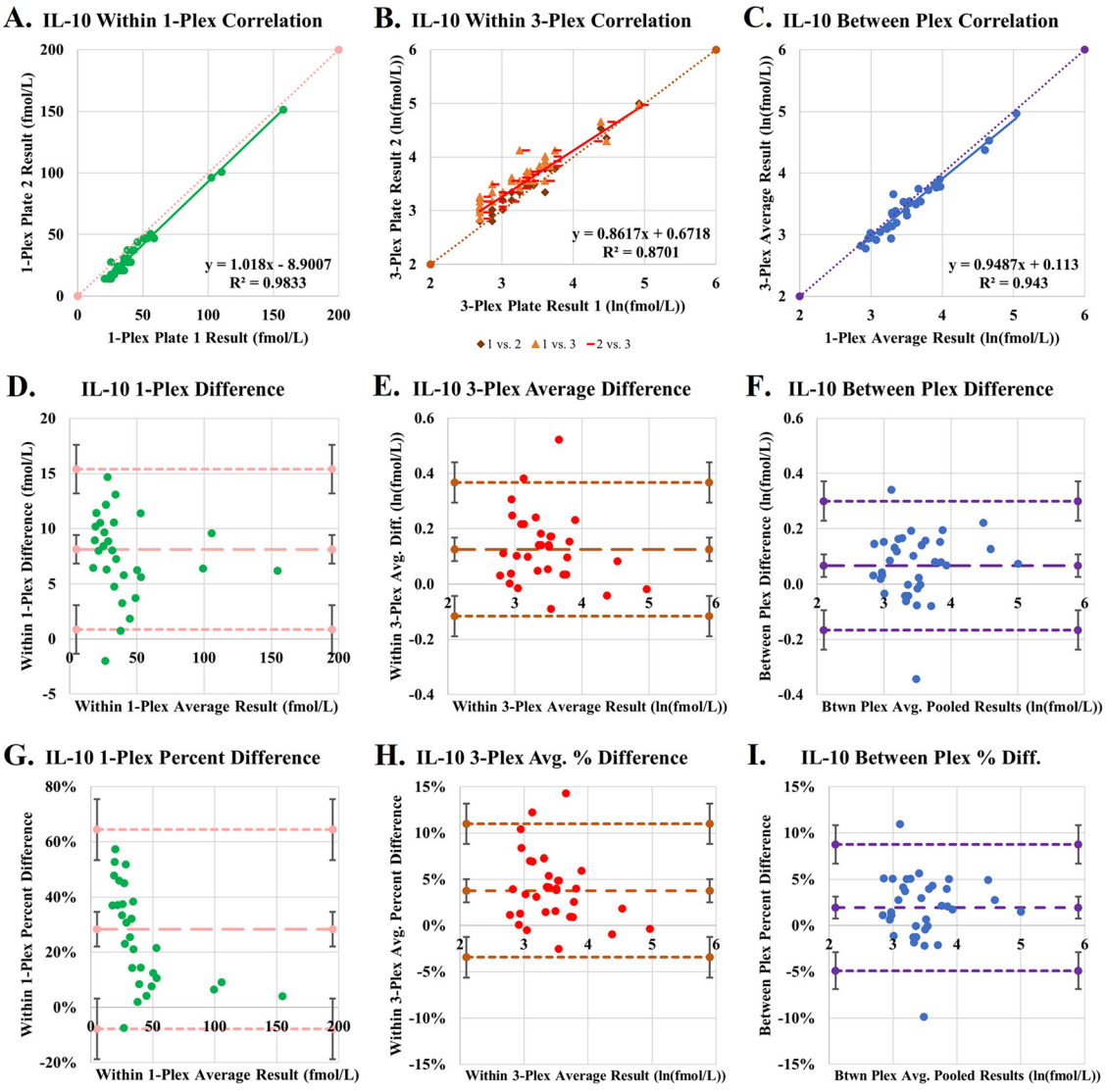

Figure 1: Scatter plots comparing IL-10 results within the 1-plex IL-10 plates (A, D, G), within the 3-plex IL-10 assays (B, E, H), and between the 1-plex and 3-plex assays (C, F, I). Plots A, B, C include a reference line for a perfect 1:1 comparison, and a linear regression line of green (plot A, 1-plex IL-10), red (plot B, 3-plex IL-10), or blue (plot C, between plex IL-10); the linear regression equation for each linear regression is listed on the plot, as well as the associated R2 value. Plots D, E, and F are Bland-Altman plots of difference – scatter plots of the average result of each sample replicate and difference between them. Plots G, H, and I are similar Bland-Altman plots with the difference standardized to the average result and displayed as a percentage. The center line on each Bland-Altman plot is the mean difference which is flanked by two 95 % limit of agreement lines. Each line has a 95 % confidence interval represented as error bars.

female serum), and low and mid control results, shown in Figure 3, demonstrated that their relative concentrations for both IL-10 and TNF-α fell within the distribution of the reference assay results, with one set of male and female pool results on a TNF-α 1-plex plate as outliers ([419.5, 387.3] fmol/L, Figure 3C). Conducting a sensitivity analysis by removing the associated plate from the paired t-test attenuated the between plex difference result (mean difference 0.145 ln(fmol/L)), but it remained statistically significant regardless (95 % CI [0.104, 0.186]).

Discussion

This study evaluated differences in analyte values for IL-10 and TNF-α when evaluated using single- or multi-plex laboratory methods. While limited somewhat by systemic low values of IL-10 and the associated significant inter-assay variation, results suggest that differences in IL-10 concentrations measured on 1- vs. 3-plex are negligible as compared to those limitations. This finding is reinforced by the TNF-α results which did not demonstrate a high proportion of low

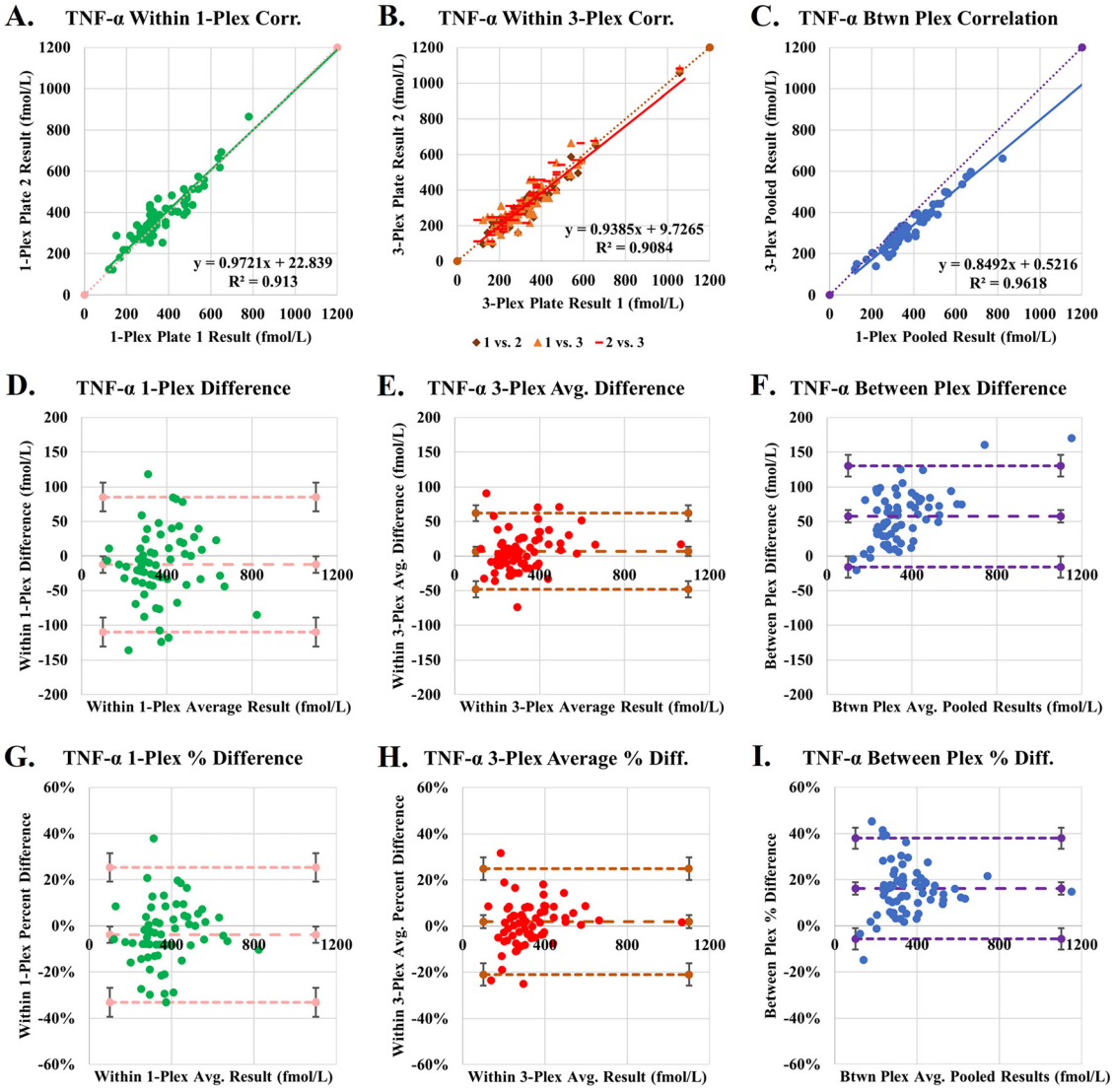

Figure 2: Scatter plots comparing TNF-α results within the 1-plex TNF-α plates (A, D, G), within the 3-plex TNF-α assays (B, E, H), and between the 1-plex and 3-plex assays (C, F, I). Plots A, B, C include a reference line for a perfect 1:1 comparison, and a linear regression line of green (plot A, 1-plex TNF-α), red (plot B, 3-plex TNF-α), or blue (plot C, between plex TNF-α); the linear regression equation for each linear regression is listed on the plot, as well as the associated R2 value. Plots D, E, and F are Bland-Altman plots of difference – scatter plots of the average result of each sample replicate and difference between them. Plots G, H, and I are similar Bland-Altman plots with the difference standardized to the average result and displayed as a percentage. The center line on each Bland-Altman plot is the mean difference which is flanked by two 95 % limit of agreement lines. Each line has a 95 % confidence interval represented as error bars.

values. Differences in 1-plex versus 3-plex TNF-α concentrations were 16.7 % on average, but this magnitude of difference was comparable to differences within-plex but across plates (11–12 %). As such, we conclude that differences in analyte concentrations measured with multiplex methodologies as compared to single plex methodologies are due to random error and are not a systemic source of bias, at least for these analytes measured using this exact methodology.

To our knowledge there is no other published study of the potential bias induced in multiplex assays when the plex count is varied. Reviews of multiplex methods emphasize that suspension multiplexes have compounding potential for cross-reactivity between analyte targets and antibodies specific to those targets [5, 8]. It is recommended that combinations of markers are validated as part of a multiplex panel's development, but these are not data that are shared by commercial manufacturers or otherwise disbursed

Table 3: Summary of quality control values on test plates and reference assays.

	QC low fmol/L	QC mid fmol/L	QC high fmol/L	Male pool fmol/L	Female pool fmol/L
Expected IL-10 value	375.3	3,021.9	23,005.3	N/A	N/A
Average reference IL-10 measure	311.0	2,412.1	20,070.2	20.0	27.8
Aggregate reference CV	8.7 %	6.9 %	7.6 %	23.9 %	17.7 %
1 IL-10 1-plex	352.4	2,849.8	22,738.7	20.5	30.7
2 3-plex	321.2	2,640.2	21,989.1	21.0	27.8
3 IL-10 1-plex	339.2	2,765.0	22,026.1	12.2	20.5
4 3-plex	314.9	2,675.8	21,242.9	16.6	28.3
6 3-plex	308.5	2,530.1	18,618.2	23.9	35.1
Expected TNF-α value	467.9	3,899.5	30,416.5	N/A	N/A
Average reference TNF-α measure	333.8	3,212.8	28,125.9	332.2	302.6
Aggregate reference CV	9.0 %	7.7 %	6.5 %	10.5 %	8.8 %
2 3-plex	361.1	3,734.6	31,252.9	306.9	306.9
4 3-plex	336.5	3,434.7	31,422.9	326.0	300.3
5 TNF-α 1-plex	290.9	3,214.0	27,535.9	341.6	311.6
6 3-plex	298.3	3,262.4	26,112.5	340.4	310.0
7 TNF-α 1-plex	313.5	3,299.4	28,660.1	419.6	387.2

publicly. Therefore, in any case where a multiplex panel is varied, it should not be assumed that the results from it can be directly compared with differently-plexed results without first validating those comparisons.

The most significant use case for similar validation studies of multiplexes is in longitudinal studies where there are different biomarkers of interest across visits of the study. Best practice would be to keep multiplexes separated such that markers measured consistently at visits are insulated from those that may vary from visit to visit, but if cost- and sample-efficiency are a primary concern, then any variations in the multiplex should be validated against each other as if they were distinct assay methodologies in addition to standard precautionary measures, like using an array of internally and externally sourced quality controls. Failure to do so may induce a spurious relationship between sample visit and any analytes measured on the multiplexes.

Further assessments on this topic should consider altering this experiment to minimize the effect of inter-assay variation. The simplest workaround would be to assess analytes placed more central in the assay's range – while we attempted this by isolating TNF-α to its own assay, we were still measuring in the bottom third of standard curves points and saw high variation as a result. Increasing the number of sample measures (i.e., more batches) would also even out random variation in the aggregate measures. Further control in batch-to-batch differences is possible by conducting discrete batches from equivalent frozen aliquots, as there is

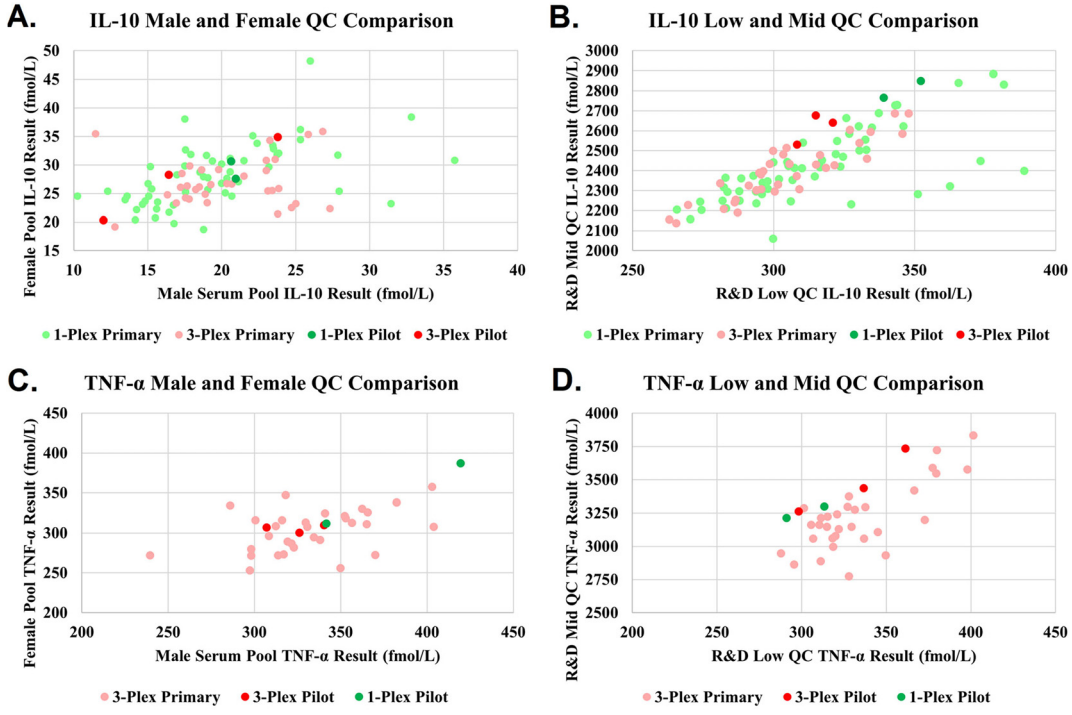

Figure 3: Scatter plots of quality control results for the IL-10 assays (plots A and B), and TNF-α assays (plots C and D). Plots A and C are comparisons between male serum pool results and female serum pool results by batch, and plots B and D are comparisons between R&D Systems low and mid QCs (catalog #QC11). Light green points are 1-plex results and light red points are 3-plex results – the corresponding results from the pilot assays are featured in dark green and dark red, respectively. There are 228 batches plotted for IL-10 and 39 batches plotted for TNF-α.

a possibility we experienced drift in sample integrity over time despite attempts to keep samples chilled.

Research ethics: Not applicable.
Informed consent: Informed consent was obtained from all individuals included in this study, or their legal guardians or wards.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission. W.S. conducted the laboratory work, statistical analysis, and manuscript preparation. C.K. and DM conceptualized the project/experiment and developed the analysis plan.
Use of Large Language Models, AI and Machine Learning Tools: None declared.
Conflict of interest: The authors state no conflict of interest.
Research funding: None declared.
Data availability: Not applicable.

References

1. Liu MY, Xydakis AM, Hoogeveen RC, Jones PH, Smith EO, Nelson KW, et al. Multiplexed analysis of biomarkers related to obesity and the metabolic syndrome in human plasma, using the Luminex-100 system. Clin Chem 2005;51:1102–9.
2. de Jager W, te Velthuis H, Prakken BJ, Kuis W, Rijkers GT. Simultaneous detection of 15 human cytokines in a single sample of stimulated peripheral blood mononuclear cells. Clin Diagn Lab Immunol 2003;10:133–9.
3. Keyes KA, Mann L, Cox K, Treadway P, Iversen P, Chen YF, et al. Circulating angiogenic growth factor levels in mice bearing human tumors using Luminex Multiplex technology. Cancer Chemother Pharmacol 2003;51:321–7.
4. Bortolin S, Black M, Modi H, Boszko I, Kobler D, Fieldhouse D, et al. Analytical validation of the tag-it high-throughput microsphere-based universal array genotyping platform: application to the multiplex detection of a panel of thrombophilia-associated single-nucleotide polymorphisms. Clin Chem 2004;50:2028–36.
5. Tighe PJ, Ryder RR, Todd I, Fairclough LC. ELISA in the multiplex era: potentials and pitfalls. Proteonomics Clin Appl 2015;9:406–22.
6. Sowers M, Crawford SL, Sternfeld B, Morganstein D, Gold EB, Greendale GA, et al. Chapter 11 – swan: a multicenter, multiethnic, community-based cohort study of women and the menopausal transition. In: Lobo RA, Kelsey J, Marcus R, editors. Menopause. San Diego: Academic Press; 2000:175–88 pp.
7. Armbruster DA, Pry T. Limit of blank, limit of detection and limit of quantitation. Clin Biochem Rev 2008;29(Suppl 1):S49–52.
8. Elshal MF, McCoy JP. Multiplex bead array assays: performance evaluation and comparison of sensitivity to ELISA. Methods 2006;38:317–23.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/cclm-2025-0458).