Variability in the Laboratory Measurement of Cytokines
A Longitudinal Summary of a College of American Pathologists Proficiency Testing Survey

Vijaya Knight, MD, PhD; Thomas Long, MPH; Qing H. Meng, PhD; Michael A. Linden, MD, PhD; Daniel D. Rhoads, MD

• Context.—The measurement of cytokines in clinical laboratories is becoming an increasingly routine part of immune monitoring when administering biologic and cell-based immunotherapies and also for clinical assessment of inflammatory conditions. While a number of commercial assays and platforms are available for cytokine measurement, there is currently little standardization among these analytical methods.

Objective.—To characterize the variability and comparability among cytokine testing platforms that are commonly used in clinical laboratories.

Design.—We analyzed data for 4 cytokines (interleukin [IL]-1, IL-6, IL-8, and tumor necrosis factor-alpha [TNF-α]) from 6 College of American Pathologists surveys administered from 2015 to 2018. Analyses interrogated variability between testing methods and variability within each laboratory across the mailings.

Results.—Significant variability was noted across methods with analysis of IL-1 showing the least variability and IL-6, IL-8, and TNF-α varying between methods to a greater extent. Intralab variability was also significant with TNF-α measurements again showing the greatest variability.

Conclusions.—This retrospective analysis of College of American Pathologists proficiency testing data for cytokine measurement is the largest method comparison to date, and this study provides a description of the variation of cytokine measurement across methods, across laboratories, and within laboratories. Serial monitoring of cytokines should preferentially be performed by the same method within the same laboratory.

Arch Pathol Lab Med. doi: 10.5858/arpa.2019-0519-CP

Cytokines are small proteins or glycoproteins that are secreted by a variety of immune and nonimmune cells. Cytokines are responsible for a variety of pleiotropic effects, including, but not limited to, regulation of innate and adaptive immunity, communication between and among various cell types, and regulation of inflammation.1 Serum cytokines are being measured and monitored more routinely because of their increased use as (1) therapeutics (eg, Granulocyte Colony Stimulating Factor for treatment of neutropenia and for stem cell mobilization,2,3 interleukin (IL)-2 for cancer immunotherapy,4 (2) as targets for modulation of undesirable inflammation (eg, tumor necrosis factor-alpha[TNF-α]5 and IL-5,6 (3) as biomarkers of inflammation (eg, IL-6),7 and (4) as diagnostic markers (eg, cytokine measurement to support the diagnosis of autoinflammatory syndromes, such as Familial Mediterranean Fever).8 Monitoring cytokine levels is used in part to detect and manage adverse effects, such as cytokine storm or cytokine release syndrome,9,10 and cytokines are also measured during early phases of drug development to assess the potential inflammatory effects of investigational new drugs.11 As clinicians and researchers increasingly desire to measure and monitor cytokine concentrations with a goal of drawing meaningful conclusions, it becomes increasingly important for laboratories to recognize and communicate the accuracy and reproducibility with which these analytes can be measured. This information is needed in order for laboratories to accurately interpret changes in cytokine measurements from a single person over time and to make meaningful conclusions when comparing measurements performed at different sites and in different studies.

Multiple methods are currently used to measure cytokines. Clinical laboratories measure individual cytokines separately or in panels. For example, IL-6 is widely used as a surrogate marker for inflammation, or a panel of cytokines may be used to broadly assess the overall proinflammatory state of the individual, or to assess potential skewing toward a T-helper 1 or 2 phenotype. Measurement of a single
cytokine is commonly performed using enzyme-linked immunosorbent assay or enzyme immunoassay, chemiluminescent assays, and multiplex bead-based assays are often used to interrogate multiple cytokines using a single test.

Accuracy and precision within and between assays can be variable. Even if assays perform with good precision within an individual laboratory, the potential for interlaboratory variability is high because of variability in antibodies, calibration standards, detection reagents, detection methods, and data analysis methods.\(^1\),\(^2\) Proficiency testing (PT) programs provide laboratories with well-defined samples designed to enable interlaboratory comparison, but the PT materials remained stable in lyophilized form throughout the duration of the analysis period. Data across the mailings for each analyte and level, therefore suggesting the PT was no significant difference in results across the mailings for each analyte and level, because it was desired to combine data across the mailings for further statistical comparisons. Differences between methods were considered to be statistically significant if \(P < .05\).

Intralab variability (coefficient of variation [CV]) was calculated for each laboratory with at least 5 mailings of results for each analyte and level. Analysis of variation was also employed to test for significant differences between the method groups for each analyte and level. Testing for intralab CV differences between the method groups could not be conducted because of the low sample sizes.

**RESULTS**

**Study Participation**

A summary of the data, including the technical method groups used for measuring each cytokine and the number of participating laboratories, is shown in Table 1. IL-6 was the most frequently analyzed analyte and enzyme immunoassay or enzyme-linked immunosorbent assay was the most commonly used method. Standard quality assurance and employment of analysis of variation indicated there was no significant difference in results across the mailings for each analyte and level, therefore suggesting the PT materials remained stable in lyophilized form throughout the duration of the analysis period. Data across the mailings were therefore aggregated for further statistical analyses.

**Intermethod Comparisons**

Five methods (enzyme-linked immunosorbent assay or enzyme immunoassay, chemiluminescence, electrochemiluminescence, multiplex bead immunoassay, and magnetic bead–based multiplex immunoassay) were compared. Distributions of results between each method group are categorized into 3 to 5 technical method groups, depending upon the analyte.

Two methods were used to remove a small number of outliers as follows: (1) unreasonable values clearly outside of the distribution of results, and (2) 2 pass 3 SD outlier screen. An analysis of variation model was run to test if there were differences between the mailings for each analyte and level, because it was desired to combine data across the mailings for further statistical comparisons. Differences between methods were considered to be statistically significant if \(P < .05\).

**METHODS**

PT survey results reported to the College of American Pathologists in response to the cytokine 2015B, 2016A, 2016B, 2017A, 2017B, and 2018A surveys were analyzed. Reportable analytes in the surveys included interferon–gamma, IL-1 beta, IL-2, IL-6, IL-8, IL-10, TNF-\(\alpha\), and vascular endothelial growth factor. Each of the 6 surveys rotated the same 3 sample lots as low, medium, and high levels for each analyte. These samples were lyophilized human sera that were reconstituted by each participating laboratory. Testing methodology was reported by each participating laboratory.

Statistical analysis was performed using SAS, Inc (Cary, North Carolina). Four analytes with the greatest number of responses were analyzed in the study, including IL-1, IL-6, IL-8, and TNF-\(\alpha\). The 2 testing methodologies reported by participants were method and instrument. Because of sample size considerations, results were combined data across the mailings for each analyte and level. Analysis of variation was also employed to test for significant differences between the method groups for each analyte and level. Testing for intralab CV differences between the method groups could not be conducted because of the low sample sizes.

A summary of the data, including the technical method groups used for measuring each cytokine and the number of participating laboratories, is shown in Table 1. IL-6 was the most frequently analyzed analyte and enzyme immunoassay or enzyme-linked immunosorbent assay was the most commonly used method. Standard quality assurance and employment of analysis of variation indicated there was no significant difference in results across the mailings for each analyte and level, therefore suggesting the PT materials remained stable in lyophilized form throughout the duration of the analysis period. Data across the mailings were therefore aggregated for further statistical analyses.

**Intermethod Comparisons**

Five methods (enzyme-linked immunosorbent assay or enzyme immunoassay, chemiluminescence, electrochemiluminescence, multiplex bead immunoassay, and magnetic bead–based multiplex immunoassay) were compared. Distributions of results between each method group are categorized into 3 to 5 technical method groups, depending upon the analyte.
displayed as boxplots for each analyte (Figure). No statistically significant difference in means between method groups was observed for IL-1 (low, \( P = .09 \); medium, \( P = .69 \); and high, \( P = .28 \); Figure A) and for the high level of IL-8 (\( P = .19 \)). IL-6, IL-8, and TNF-\( \alpha \) demonstrated statistically significant differences in the method means for each of the 3 concentrations tested, with the exception of the high level of IL-8 (Figure B–D). The greatest variability among methods was noted for TNF-\( \alpha \) measurement. IL-8 analyzed by the magnetic bead–based multiplex immunoassay method (Luminex; Millipore Corporation) showed consistently lower measured concentrations of IL-8 than other methods. Bead multiplex assays (other than magnetic bead–based multiplex immunoassay) measured consistently lower concentrations of TNF-\( \alpha \) than other methods.

**Intralab Variability**

Because identical aliquots of PT material are sent for multiple mailings, intralab precision was determined by calculating the CV for results obtained by a single lab for an individual analyte and level. Intralab CV was calculated only when a participating lab had performed at least 5 PT surveys (Table 2). Additional data for variability (\%CV) within individual laboratories participating in the surveys is presented in Supplemental Table 1 (see supplemental digital content). Intralab variability was greatest for TNF-\( \alpha \), regardless of method and ranged from 6.7% to 102%. Intralab CVs for IL-1 measurement were less variable (4.3%–53.7%) and the least variability was noted for IL-6 and IL-8 regardless of method. However, intralab CVs for these analytes still ranged from 3.7 to 39.3, indicating that intralab variability for cytokine measurement in general tends to be significant.

**DISCUSSION**

This study describes the variability within and between methods and variability within an individual laboratory for measurement of IL-1, IL-6, IL-8, and TNF-\( \alpha \). As investigators and medical practitioners increasingly use cytokine monitoring for evaluating immune response, it is important to recognize the variability in laboratory measurement that currently exists. Our data indicate variability between methods differs for each analyte; less variability was noted for IL-1, whereas variability between methods was much greater for TNF-\( \alpha \). We also noted that variability of cytokine measurements within a single laboratory, regardless of the...
When performing serial testing to monitor cytokine levels in a patient or subject, the same testing method should be used, ideally within the same laboratory. Using different methods for serial monitoring should not be used without careful comparison studies to validate this approach. Results reported by different laboratories whether using the same or a different method are not directly commutable and must be interpreted with caution. Even intralaboratory testing will routinely produce variability of up to 20%, and those interpreting changes in these results should recognize this limitation in precision when considering the clinical significance of serial cytokine measurements. Given the significant variability among cytokine assays and between laboratories, further efforts are required in order to standardize cytokine testing. This study was limited by the sample size and the potential role of using lyophilized samples in these analyses across multiple years. Additionally, the lyophilized material used in this study was initially analyzed by a single method by the supplier prior to distribution by College of American Pathologists through the PT testing program. However, this is the largest comparison study published to date, and we hope this study serves to inform physicians, laboratorians, and investigators of the variability that is currently present in the laboratory measurement of cytokines.

The authors wish to thank Christine Bashleben for administrative assistance during development of this manuscript.

References