

A Critical Evaluation of Enzyme Immunoassay Kits for Detection of Antinuclear Autoantibodies of Defined Specificities. III. Comparative Performance Characteristics of Academic and Manufacturers' Laboratories

MARVIN J. FRITZLER, ALLAN WIIK, ENG M. TAN, JOSEF S. SMOLEN, J. STEVEN McDOUGAL, EDWARD K.L. CHAN, THOMAS P. GORDON, JOHN A. HARDIN, JOACHIM R. KALDEN, ROBERT G. LAHITA, RAVINDER N. MAINI, WESTLEY H. REEVES, NAOMI F. ROTHFIELD, YOSHINARI TAKASAKI, MERLIN WILSON, MARTHA G. BYRD, LLOYD SLIVKA, and JAMES A. KOZIOL

ABSTRACT. Objective. To analyze the performance of different commercial enzyme immunoassay (EIA) kits for measuring antinuclear antibodies (ANA) specific for dsDNA, SSB/La, Sm, and Scl-70.

Methods. EIA kits for detection of ANA from 9 commercial manufacturers were evaluated. The manufacturers were advised that they would be sent coded sera containing mixtures of the Arthritis Foundation/Centers for Disease Control reference reagents, and that they were to use their own test kits to analyze the antibody specificities of these sera and to report the data, in optical density (OD) units or their equivalent. Independently, 12 investigators in academic institutions who have done research in this field agreed to participate in a parallel study. The concentration of the antibodies and the specificities were blinded to the analysts and the coefficients of variation (CV) were computed for each participant.

Results. There were statistically significant differences between laboratories in terms of CV for all 9 kits tested. With the exception of one kit, there were no significant CV differences between the various autoantibody kits provided by each manufacturer and, with the exception of kits from 2 manufacturers, there were no significant differences between the various antibody kits in terms of reproducibility (CV). From the point of view of interlaboratory variability, manufacturers could be separated into either a high or low performance group.

Conclusion. We found a disconcertingly large range of performance characteristics in the various laboratories, which could be quite detrimental in routine utilization of EIA ANA kits. Clinicians should be aware of the performance issues raised in our study, and should know and be involved in how their service laboratory assesses its own performance and the performance of commercial testing systems utilized. Manufacturers and clinical laboratories need to exercise constant quality assurance and surveillance of kit performance in the hands of medical laboratory technologists involved in routine testing. (J Rheumatol 2003;30:2374-81)

Key Indexing Terms:

ELISA
DIAGNOSIS

AUTOANTIBODIES
DIAGNOSTIC KITS

AUTOIMMUNITY
AUTOANTIBODY

Human autoantibodies have a significant place in the history of clinical and molecular medicine. Dating from original observations of the LE cell in 1948¹ to the present day applications of microarray analyses², their use as diagnostic and

prognostic markers of disease has been a valuable adjunct to clinical medicine³⁻⁵. The detection and analysis of autoantibodies in human serum has become a valuable clinical tool that serves to confirm a diagnosis, in some instances predict

From the Standardization Committee in Rheumatic and Related Disorders of the International Union of Immunological Societies; The World Health Organization; The Arthritis Foundation and Centers for Disease Control and Prevention (CDC), Atlanta, Georgia, USA.

M.J. Fritzler, PhD, MD, University of Calgary, Calgary, Canada; A. Wiik, MD, State Serum Institute, Copenhagen, Denmark; E.M. Tan, MD, Scripps Research Institute, La Jolla, CA, USA; E.K.L. Chan, PhD, University of Florida, Gainesville, FL, USA; T.P. Gordon, MD, Flinders University, Adelaide, Australia; J.A. Hardin, MD, Einstein Medical College, New York, NY, USA; J.R. Kalden, MD, University of Erlangen, Erlangen, Germany; R.G. Lahita, MD, St. Vincent Hospital, New York,

NY, USA; R.N. Maini, MD, Kennedy Institute, London, UK; J.S. McDougal, MD, CDC, Atlanta, GA, USA; W.H. Reeves, MD, University of Florida; N.F. Rothfield, MD, University of Connecticut, Farmington, CT, USA; J.S. Smolen, MD, University of Vienna, Vienna, Austria; Y. Takasaki, MD, Juntendo University, Tokyo, Japan; M. Wilson, MD, Immunotek Reference Laboratory, New Orleans, LA, USA; M. Byrd, BS, MT(ASCP), CDC; L. Slivka; J.A. Koziol, PhD, University of Calgary.

*Address reprint requests to Dr. M.J. Fritzler, Faculty of Medicine, 3330 Hospital Dr. NW, Calgary, AB, Canada T2N 4N1.
E-mail: fritzler@ucalgary.ca*

Submitted August 22, 2002; revision accepted March 28, 2003.

disease course, and in isolated instances, provide a guide to preventive therapies⁴. Over the past 2 decades the detection of autoantibodies has captured the interest of commercial vendors that now market a variety of diagnostic kits intended to provide an accurate analysis of serum autoantibodies in systemic rheumatic diseases. These kits have incorporated various technologies such as indirect immunofluorescence (IIF), immunodiffusion (ID), immunoblotting (IB), ELISA, and more recently, antigen array assays.

The use of ELISA kits to detect autoantibodies relevant to systemic rheumatic diseases has become commonplace because they offer sensitivity, high performance, and relatively low cost. Unfortunately, there has been little done to standardize these kits, and postmarketing surveillance and quality assurance are largely left to the manufacturers. A number of studies have compared the autoantibody kits provided by different manufacturers⁶⁻¹³. However, these studies were evaluated in a single laboratory and some were limited by the spectrum of kits under evaluation.

We undertook to evaluate a number of commercial antinuclear antibody (ANA) ELISA kits by having at least 2 academic laboratories evaluate the kits by utilizing highly characterized sera that were provided by the Centers for Disease Control and Prevention (CDC) in Atlanta, USA. In addition, these sera were provided to each participating manufacturer and the performance of the kits was analyzed by the respective manufacturer. Our study found significant interlaboratory variation of test results when kits from the same manufacturer were used.

MATERIALS AND METHODS

Twenty commercial purveyors of enzyme immunoassay (EIA) kits were approached in order to determine their interest and willingness to participate in the original study^{14,15}. These manufacturers were advised that they would be sent coded sera containing mixtures of the Arthritis Foundation/Centers for Disease Control and Prevention (AF/CDC) reference reagents, and that they were to use their own test kits to analyze the antibody content of these sera and report the data, preferably in optical density (OD) units (or in their own arbitrary units). They were informed that our study was designed to critically evaluate the performance of EIA-based methods for detection of autoantibodies and that the data would be published as a comprehensive evaluation of this methodology without divulgence of the specific performance of any individual manufacturer. The 9 participating manufacturers were (in alphabetical order): Cambridge Life Sciences (Cambridge, UK), Elias (Freiburg, Germany), Helix Diagnostics (Sacramento, CA, USA), Immunoconcepts (Sacramento), Imtec Immunodiagnostika (Zepernick, Germany), Incstar (Stillwater, MN, USA), Inova Diagnostics (San Diego, CA, USA), MBL (Nagoya, Japan), and Shield Diagnostics (Dundee, Scotland). The 9 participating manufacturers are randomly designated I through IX as established¹⁴.

Independently, 12 investigators in academic institutions who have done research in this field agreed to participate in the parallel study. Many of these laboratories are also certified clinical diagnostic laboratories in their respective geographic locations. Investigators included M.J. Fritzler, Canada; T. Gordon, Australia; J.A. Hardin, New York; J.R. Kalden, Germany; R.G. Lahita, USA; R.N. Maini, UK; J.S. McDougal; N.F. Rothfield, USA; J.S. Smolen, Austria; Y. Takasaki, Japan; E.M. Tan, USA; and A. Wiik, Copenhagen. The 12 academic laboratories are designated "a" through "l" in the ordering established in Figure 1.

Design of test samples. Serum samples (Table 1) were prepared by M. Byrd, CDC. Briefly, the AF/CDC standard reagents are designated CDC1, CDC2, etc. Samples containing a single undiluted serum are designated 7x. In samples that contain a mixture of 3 sera, the relative volumes of each standard in any sample are shown as multiples of 4x, 2x, and 1x, with a total unit volume of 7x. For example, the vertical column of CDC1 (Table 1) shows that this reference reagent was used in different relative volumes of 4x, 2x, and 1x, and the horizontal line for sample A shows that this sample contained 4x unit volumes of CDC1, 2x unit volumes of CDC2, and 1x unit volume of CDC4. With this system, the sensitivity, specificity, and dose-response of different test kits could be evaluated, and it could be ascertained whether antibodies of other specificities would interfere with each other in the EIA system.

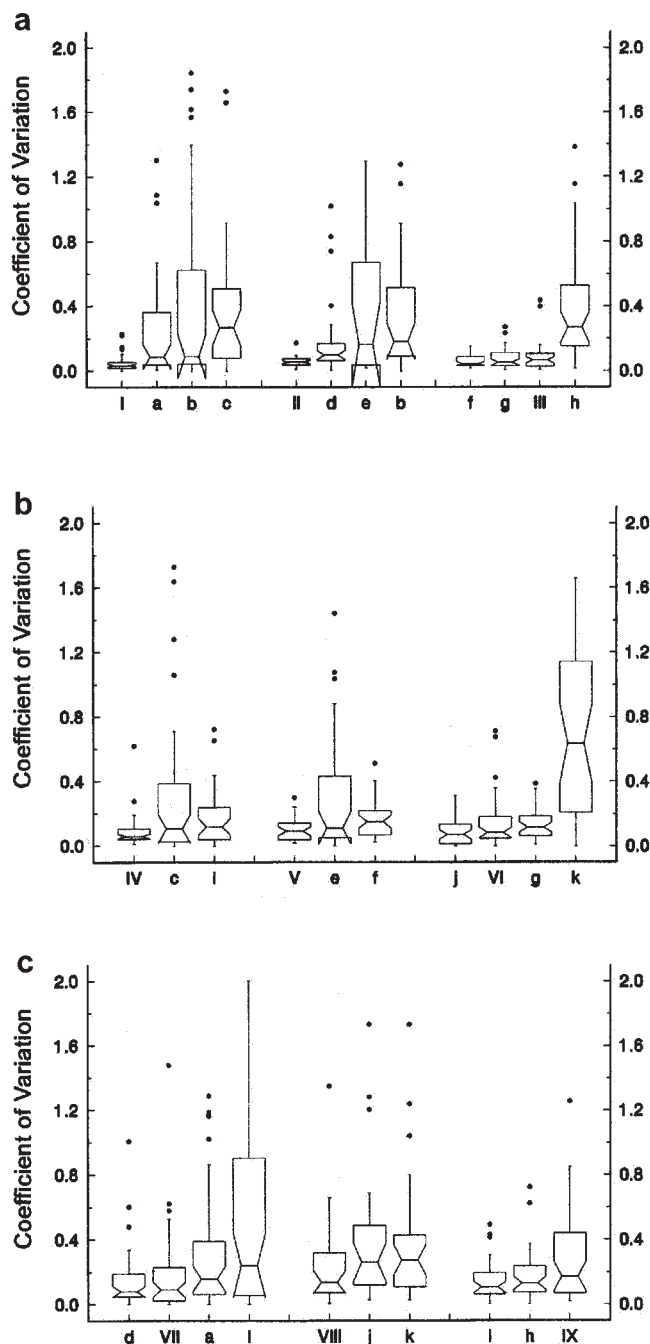
The study was designed so that mixtures of antibodies would contain different proportions of anti-dsDNA, anti-SSB/La, anti-U1 RNP, anti-Sm, and anti-SSA/Ro (see first 5 columns of Table 1), because such combinations of antibodies might be expected in diseases such as systemic lupus erythematosus (SLE). Similarly, combinations of anti-Scl-70 (CDC9), antinucleolar (fibrillarin) antigens (CDC6), and anti-centromere (CDC8) might be seen in scleroderma. In addition, use of antibody of a defined specificity [e.g., anti-SSB/La (CDC2) at 4x, 2x, and 1x relative volumes] made it possible to examine whether EIA could be used for quantitation of antibody content.

As seen in Table 1, multiple myeloma sera, which were used undiluted (7x), were included in samples R, S, and T. These multiple myeloma sera, a gift from Dr. H. Spiegelberg (University of California, San Diego, USA), contained an average of 35 mg/ml of IgG. Sera Fz and Ba were of the IgG1 subclass, and Fr was of the IgG3 subclass. Serum Ba was known to contain cryoprecipitates, but the other 2 did not. These multiple myeloma sera were included in the study to determine whether high concentrations of immunoglobulin or cryoprecipitates might lead to false-positive results. Serum CDC10, containing anti-Jo-1 (histidyl tRNA synthetase), and normal human serum were also used undiluted.

Each participating manufacturer received a set of 26 serum samples for analysis. Each set was prepared by the CDC and consisted of one aliquot from each of the 21 samples A to U (Table 1), 2 replicate samples randomly selected from samples A–J, 2 replicate samples randomly selected from samples K–P, and one replicate sample randomly selected from samples Q–U. A coding scheme was adopted to ensure that the participating personnel remained blinded to the identity of the serum samples. Only the biostatistician (JAK) and Martha Byrd at the CDC knew details of the randomization and coding scheme.

Report forms were prepared for use by the participants. For each test sample, manufacturers' laboratory personnel were requested to determine (in duplicate) optical densities (OD) at serum dilutions of 1:100, 1:400, 1:1600, and at the manufacturer's recommended dilution (if different). In addition, manufacturers were asked to indicate whether their kits gave positive or negative results for each antibody at the recommended dilution that their kit was designed to detect. For example, no test kit was designed to detect antibodies to fibrillarin (contained in CDC6), and only 2 manufacturers had test kits designed to detect antibodies to centromere antigens (CDC8).

Independently, 2 manufacturers' kits were randomly assigned to each of the participating academic laboratories, along with one set of the coded 26 serum samples prepared as described above. The 5 duplicate sera varied among the participants, with no 2 laboratories having the same 5 duplicate sera. For example, academic laboratory "a" received kits from manufacturers designated VII and X, and was requested to evaluate the 21 distinct serum samples and the 5 duplicates with each kit. As with the commercial laboratories, a coding scheme was adopted to ensure that the academic laboratory personnel remained blinded to the identity of the serum samples. As detailed above, the directions to the academic laboratories were identical to those of the manufacturers in terms of evaluation and reporting of results. Due to technical difficulties, one academic laboratory (l) reported results from only one kit; the other academic laboratories reported results



from 2 kits and the coded serum samples. In summary, each manufacturer's kit was tested by 2 or 3 different academic laboratories, in addition to testing by the manufacturer's in-house laboratory, and each serum sample was tested 3 to 4 times by a particular kit.

Individual characteristics of the CDC-based ANA reference sera were established in Tan, *et al*¹⁴. Operating characteristics (sensitivity and specificity) of the various kits were determined relative to these standards.

Data analyses. Intralaboratory variability was assessed with coefficients of variation (CV), as in Tan, *et al*¹⁴. For each antibody kit evaluated by any particular laboratory, CV were calculated from the OD or units/ml values at the manufacturer's recommended serum dilution level, using the replicate results with the duplicate test samples of CDC sera. Box plots were prepared from the CV calculated within each laboratory, separately for each

Figure 1. Grouped box plots of test reproducibility. Each grouping represents results from a particular manufacturer's set of test kits. The manufacturers are denoted "I" through "IX," and the academic laboratories are denoted "a" through "l." Ordering of the manufacturers is consecutive, from high manufacturer reproducibility to low manufacturer reproducibility. Ordering of the academic laboratories from a to l is related to the assignment of the manufacturer kits. Within each grouping, reproducibility was determined from coefficients of variation from the optical densities (or units/ml) at that manufacturer's recommended serum dilution, using all the replicate results with the 5 duplicate test samples. In each box, the median of the CV is depicted by a horizontal line segment within the rectangle, and the upper and lower quartiles of the data are depicted by the top and bottom of the rectangle. The notches in the rectangle approximate a 95% confidence region about the median. The vertical lines at each end of the box approximate a central 95% confidence region for the entire sample values. Closed circles denote observed values outside these limits. Within each group, the ordering of academic laboratories and that manufacturer is in terms of median CV, from smallest median CV to largest median CV.

set of manufacturer's antibody kits evaluated by that laboratory. In addition, separate 2-way analyses of variance (ANOVA) were performed with all of the CV data obtained from each manufacturer's antibody kits, so as to compare CV computed from the different laboratories with the various antibody kits provided by that manufacturer (i.e., the 2 factors evaluated in this statistical model were laboratory and antibody kit).

We also computed 3-way ANOVA to assess interlaboratory reliability, that is, the degree of concordance among laboratories in their OD determinations. The 3 factors evaluated in this statistical model were laboratory, antibody kit, and serum sample. We summarized these analyses by means of interlaboratory correlation coefficients: the interlaboratory correlation coefficient is the correlation between 2 laboratories' OD determinations of the same serum sample with the same antibody kit. Further details relating to this particular analysis are given in Appendix 1.

For every set of manufacturer's kits evaluated by each laboratory, operating characteristics (i.e., sensitivity and specificity for each ANA) were determined using the pooled data of all test samples at the recommended dilution levels. The manufacturers' interpretations were adopted for designation of a positive assay result. Some laboratories occasionally reported results that were called "borderline" or "weakly positive." In contrast to a previous study¹⁴, we here decided to include as positive those results that were called "borderline" or "weakly positive" in addition to those called "definitely positive." Overall sensitivity and specificity for each set of manufacturer's kits were calculated separately from the manufacturer's results (modified slightly from the statistics reported in Tan, *et al*¹⁴ because of the aforementioned change in classification of borderline or weakly positive results) and from each laboratory assigned that manufacturer's kits. In these calculations, individual sensitivity and specificity values were pooled across the ANA tested by at least 8 of the manufacturers. Global assessments of sensitivity and specificity for each manufacturer's kits were computed by combining the overall determinations per laboratory, using a random effects model under which we explicitly allow for interlaboratory variability¹⁶.

RESULTS

Reproducibility

We computed CV as in Tan, *et al*¹⁴ for each of the academic laboratories participating in this study. Results are depicted in Figure 1, where we present box plots of the summary CV by laboratory (commercial laboratory values are from Tan, *et al*¹⁴).

More formally, we submitted all CV obtained from each manufacturer's kits in turn to 2-way ANOVA, to compare

Table 1. Key to serum samples: relative volumes of reagents in each sample*.

Sample	CDC1 (dsDNA)	CDC2 (SSB/La)	CDC4 (U1 RNP)	CDC5 (Sm)	CDC7 (SSA/Ro)	CDC9 (Scl-70)	CDC3 (Speckled)	CDC6 (Nucleolar)	CDC8 (Centromere)	CDC10 (Jo-1)	MM (Fz)	MM (Ba)	MM (Fr)	NHS
A	4x	2x	1x											
B	2x			1x	4x									
C	1x		4x		2x									
D		4x	2x	1x										
E		2x		4x	1x									
F	2x	1x			4x									
G		1x	4x	2x										
H			1x	4x	2x									
I	1x		2x	4x										
J		2x		1x	4x									
K						4x	1x	2x						
L						2x	1x	4x						
M						1x	4x	2x						
N							4x	1x	2x					
O						1x		2x	4x					
P						2x	1x		4x					
Q										7x				
R											7x			
S												7x		
T													7x	
U														7x

* CDC: Centers for Disease Control and Prevention; MM: multiple myeloma; NHS: normal human serum.

the variability of the CV within each laboratory with variability due to different antibody kits. That is, the 2 factors evaluated in this statistical model were the laboratory (manufacturer, along with all the academic laboratories that had been assigned kits from that manufacturer) and antibody kit. We summarize our findings:

1. With each of the 9 sets of manufacturers' kits, there were statistically significant differences between laboratories in terms of CV. Generally, the smallest CV were found with the manufacturers, although there were a few exceptions, as can be seen in Figure 1.
2. With the exception of kits from manufacturer III, there were no significant differences between the various antibody kits provided by each manufacturer in terms of reproducibility (CV).
3. With the exception of kits from manufacturers VIII and IX (the "worst" manufacturers in terms of reproducibility, from Tan, *et al*¹⁴), at least one laboratory had average CV of 10% or less for each set of manufacturers' kits.

Interlaboratory reliability

We next assessed interlaboratory reliability, by means of 3-way ANOVA, as outlined in the Appendix. We summarize reliability by interlaboratory correlation coefficients (ICC), i.e., the estimated correlation between OD measurements made by 2 different laboratories, using the same antibody kit (from the same manufacturer) and the same serum sample. The ICC are given for each manufacturer in Table 2. There is a rather clear separation of manufacturers into 2 subsets:

the ICC of kits from manufacturers III, VI, I, IX, and VIII were all impressively high, in the neighborhood of 0.90, whereas the ICC of manufacturers VII, IV, V, and II were lower, ranging from 0.70 to 0.81. (We remark that only with manufacturer VI were the components of variance attributable to laboratories not significantly different from 0.)

Operating characteristics

We computed overall sensitivity and specificity for each set of manufacturer's kits, separately for each laboratory evaluating that kit. Findings are presented in Figure 2. We also assessed heterogeneity of sensitivities and specificities via chi-squared statistics: with regard to sensitivities, nominally significant differences were found with laboratories IV, c, i (chi-square = 16.13, degrees of freedom = 2, $p = 0.0013$); with regard to specificities, nominally significant differences were found with laboratories IV, c, i (chi-square = 11.13, $DF = 2$, $p = 0.004$), and laboratories VIII, j, k (chi-square = 9.21, $DF = 2$, $p = 0.01$). Lastly, we combined individual laboratory estimates of sensitivity and specificity into overall measures by means of random effects models; these results are shown in Figure 3. We have essentially the same orderings of manufacturers' kits in terms of sensitivity and specificity as found by Tan, *et al*¹⁴; on the other hand, standard errors of the estimates do vary, because of our allowance here for interlaboratory variability.

DISCUSSION

This comparative study was undertaken to address funda-

mental questions relating to routine use of EIA kits for measuring ANA: How reproducible are the measurements? What are the intrinsic sources of uncertainty in such measurements? Do the kits produce valid results?

In previous studies^{14,15}, we have investigated operating characteristics of 9 EIA kits (reproducibilities, sensitivities, specificities) and reliabilities of quantitative measurements, when these kits were used by their commercial purveyors to assess coded mixtures of reference sera from the CDC. We have here expanded the scope of our previous studies, by placing the kits in the hands of academic laboratories and asking the academic laboratories to evaluate the same coded mixtures of reference sera as had been sent to the commercial laboratories. Our prior belief was that by using the academic laboratories we would obtain a more practical picture of routine performance of the kits as compared to the commercial laboratories, as the academics have no vested interest in the kits, and laboratory practices might differ dramatically between the commercial and the academic laboratories.

To our surprise, we found a disconcertingly large range of performance characteristics in the various laboratories. From our examination of reproducibility as summarized by CV, we suggest that average CV of 10% or less (from replicate results of duplicate samples, as herein) are a realistic and achievable goal for laboratories (perhaps with the exception of kits from manufacturers VIII and IX). It should be understood that some of the participating academic laboratories are primarily involved in basic and clinical research and they do not routinely provide an autoantibody testing service. Other participating laboratories that were certified in their respective jurisdictions would be required to employ competent and certified technologists. Nevertheless, a key point of this study is that attention to manufacturers' instructions, and in particular the use of manufacturers' recommended dilutions, is critical to test performance and reliability.

A common technique of assessing laboratory performance is to send replicate sera samples to different laboratories, and to compare the interlaboratory results, much as in the present study. However, the utility of this procedure as a method of improving a particular laboratory's performance is not altogether obvious. We found that OD determinations by different laboratories, using identical sera samples and the same antibody kits, were highly correlated (ICC about 0.90) for 5 of the 9 manufacturers (III, VI, I, IX, VIII). In the other 4 cases, ICC ranged from 0.70 to 0.81. In general, the inclusion of laboratories with widely disparate evaluations (that is, poor reproducibility) in summary calculations of interlaboratory correlation coefficients will tend to lower the summary estimates of the ICC. Moreover, in every instance but one (kits from manufacturer VI), the components of variance attributable to laboratories in our ANOVA analyses were found to differ significantly from zero. That is, there

were statistically significant differences between the various laboratory determinations of OD. Hence, laboratory differences in OD determinations appear to be the norm rather than the exception. Thus, we hesitate to pronounce a particular laboratory as providing the gold standard of absolute OD determinations with any set of antibody kits against which other laboratories' determinations can be rated. The variables that can be attributed to this variation are likely related to different equipment used to perform the OD analyses. Our study emphasizes that it is imperative that equipment used for clinical assays should be calibrated to a reference sample. Once again, the practice of research and clinical service laboratories may differ. For example, many research laboratories rely on intratest and intralaboratory calibration and standardization, but this may not relate directly to ANA kits that require both intra- and interlaboratory calibration to a reference standard such as those provided in ANA kits.

It is clear from Figure 2 that there can be substantial differences between laboratories in terms of their assessments of positive assay results. The most discrepant findings were the differing sensitivities of laboratories I, a, b, and c, and those of laboratories IV, c, and i. There are, of course anomalies with our findings: for example, it is curious, and perhaps unexpected, that laboratory c tended to read relatively low OD values for positive sera with kits from manufacturer I, and high OD values with kits from manufacturer IV.

Previous studies in individual laboratories that compared EIA kits from different manufacturers to conventional assays such as indirect immunofluorescence (IIF) and double immunodiffusion (DID) concluded that there was significant discordance between conventional assays and EIA^{7,9} and between kits from different manufacturers⁸. In one study EIA were found to be more sensitive than DID⁶; another study that used a cross section of serum referred to a rheumatology laboratory found moderate to good agreement between ANA-IIF and anti-DNA results with 2 commercial EIA kits¹⁰. Some studies utilized selected sera^{6,8} and others used unselected sera^{7,10}. Analysis of the design of some studies suggests that lack of agreement between EIA and conventional assays may depend on the diagnosis of the patients under study. For example, one study found high concordance between assays performed on SLE and primary Sjögren's syndrome sera¹⁰, whereas a study that found high discordance studied sera from children with juvenile rheumatoid arthritis⁹.

Tan, *et al*¹⁴ focused on the EIA kits themselves, and in particular highlighted deficiencies in intrinsic properties of the kits (sensitivities and specificities). Here, we wish to shift focus to the clinical laboratories. Many laboratories are required to participate in a number of laboratory improvement and quality assurance programs such as the one administered by the College of American Pathologists (<http://www.cap.org>). The Clinical Laboratory Improvement

Table 2. Interlaboratory correlation coefficients (ICC) for each set of manufacturer's kits.

Manufacturer	I	II	III	IV	V	VI	VII	VIII	IX
ICC	0.90	0.70	0.93	0.79	0.76	0.91	0.81	0.86	0.87

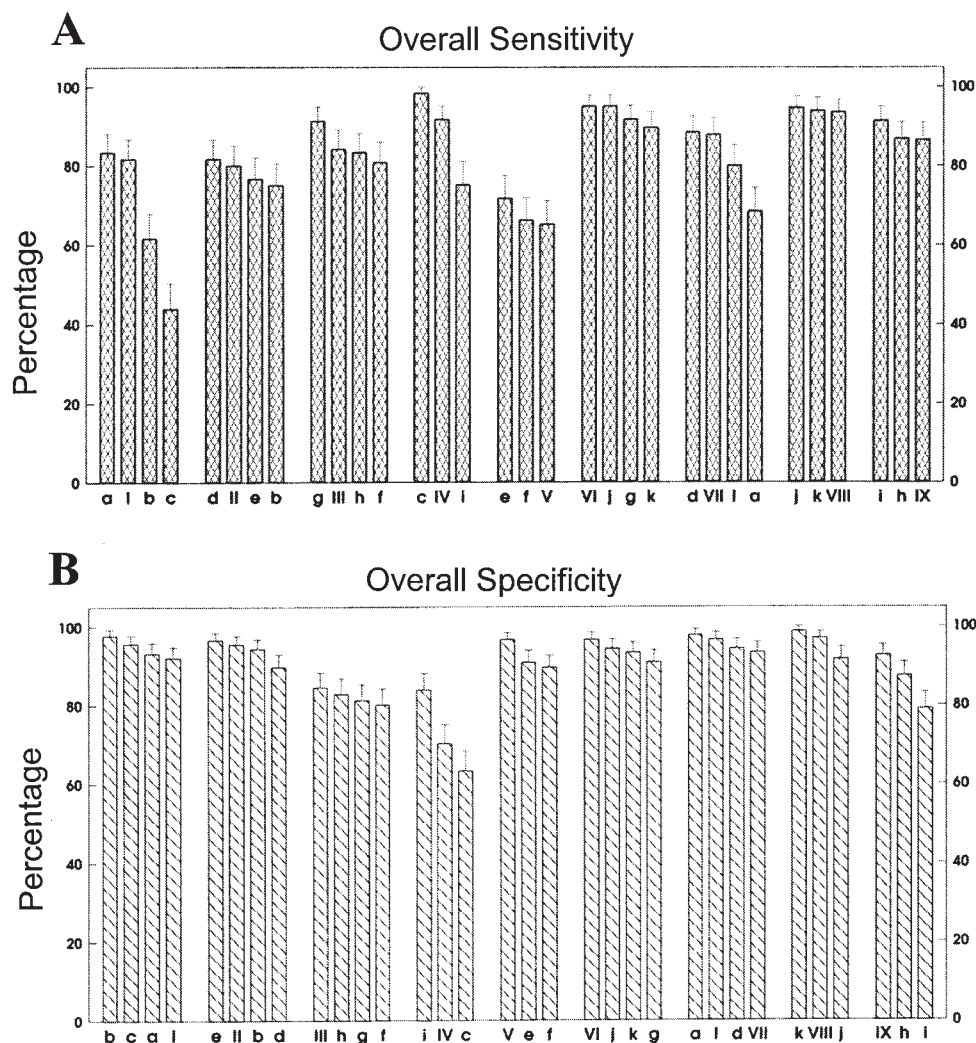


Figure 2. Overall sensitivity (A) and specificity (B) for the 9 manufacturers and the academic laboratories. Overall values were calculated by pooling each manufacturer's or academic laboratory's results for assay positivity or negativity, for the antinuclear antibodies tested by at least 8 of the manufacturers (see Tan, *et al*¹⁴). The groupings reflect the consecutive ordering of the manufacturers, from I to IX. Within each group, the ordering of academic laboratories and manufacturer is from highest to lowest sensitivity (A) or specificity (B). Each manufacturer evaluated solely his own antibody kits, whereas academic laboratories generally evaluated more than one set of kits. The lines extending above each of the individual bars represent 1 SD of the estimated sensitivity or specificity.

Amendments of 1988 set standards for all laboratories engaged in clinical testing (Fed Reg, 1992). These standards include requirements for trained supervisory and testing personnel, record keeping and instrument maintenance, daily quality control practices, result reporting, and laboratory inspection and maintenance. Whether these standards are being met in routine practice is questionable: although

one might expect the academic laboratories to be rather proficient in the implementation of the EIA kits, we found numerous instances of gross errors that should have been flagged by the laboratories themselves. Indeed, many of these errors could have been precluded had the personnel read the manufacturer's instructions. We suggest that quality control procedures for daily performance of tests in the clin-

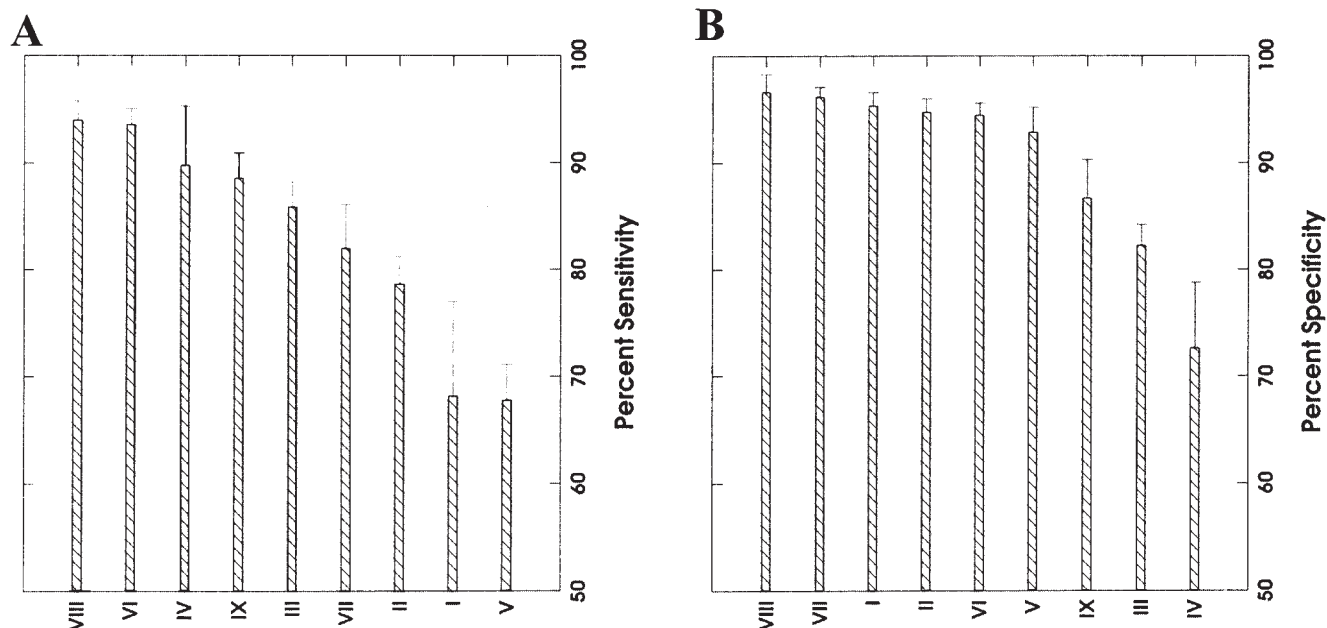


Figure 3. Pooled overall sensitivity and specificity of each manufacturer's kits. The pooled values represent weighted averages of the sensitivities and specificities from the academic laboratories and manufacturers in Figure 2, using a random effects model (see text for details). Ordering of manufacturers is in terms of decreasing sensitivity or specificity (from top to bottom). The horizontal lines extending to the right of each bar depict 1 SD of the pooled estimate of sensitivity or specificity.

ical laboratory setting not be ignored, and that a minimal performance target of CV in EIA assays of 10% or less with trained technicians be established. Clinicians should be aware of the performance issues raised in our study, and be involved by asking their laboratory director how that laboratory assesses its own performance and the performance of the commercially available testing systems.

Finally, this study assessed the performance of only one set of kits provided by manufacturers and did not assess variation of lot-to-lot performance. Although it is anticipated that each new lot is equal to or better in performance, thorough testing of each new lot in comparison to previous ones is highly recommended.

REFERENCES

- Hargraves MM, Richmond H, Morton R. Presentation of two bone marrow elements: the "tart" cells and the "L.E." cell. *Mayo Clin Proc* 1948;27:25-8.
- Robinson WH, DiGennaro C, Hueber W, et al. Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nature Med* 2002;8:295-301.
- Tan EM. Autoantibodies in pathology and cell biology. *Cell* 1991;67:841-2.
- Tan EM. Autoantibodies in diagnosis and identifying autoantigens. *The Immunologist* 1999;7:85-92.
- von Muhlen CA, Tan EM. Autoantibodies in the diagnosis of systemic rheumatic disease. *Semin Arthritis Rheum* 1995;24:323-58.
- Jaskowski TD, Schroder C, Martins TB, et al. Comparison of three commercially available enzyme immunoassays for the screening of autoantibodies to extractable nuclear antigens. *J Clin Lab Anal* 1995;9:166-72.
- Jaskowski TD, Schroder C, Martins TB, et al. Screening for antinuclear antibodies by enzyme immunoassay. *Am J Clin Pathol* 1996;105:468-73.
- Emlen W, O'Neill L. Clinical significance of antinuclear antibodies: comparison of detection with immunofluorescence and ELISA. *Arthritis Rheum* 1997;40:1612-8.
- Fawcett PT, Rose CD, Gibney KM, et al. Use of ELISA to measure antinuclear antibodies in children with juvenile rheumatoid arthritis. *J Rheumatol* 1999;26:1822-6.
- Ulvestad E, Kansenstrom A, Madland TM, et al. Evaluation of diagnostic tests for antinuclear antibodies in rheumatological practice. *Scand J Immunol* 2000;52:309-15.
- Bossuyt X. Evaluation of two automated enzyme immunoassays for detection of antinuclear antibodies. *Clin Chim Lab Med* 2000;38:1033-7.
- Jansen EM, Deng J, Beutner EH, Kumar V, Chorzelski TP. Comparison of commercial kits for the detection of anti-nDNA antibodies using Crithidia luciliae. *Am J Clin Pathol* 1987;87:461-9.
- Avina-Zubieta JA, Galindo-Rodriguez G, Kwan-Yeung L, et al. Clinical evaluation of various selected ELISA kits for the detection of anti-DNA antibodies. *Lupus* 1995;6:370-4.
- Tan EM, Smolen J, McDougal JS, et al. A critical evaluation of enzyme immunoassays for the detection of antinuclear antibodies of defined specificities. I. Precision, sensitivity and specificity. *Arthritis Rheum* 1999;42:455-64.
- Tan EM, Smolen JS, McDougal JS, et al. A critical evaluation of enzyme immunoassay kits for the detection of antinuclear antibodies of defined specificities. II. Potential for quantitation of antibody content. *J Rheumatol* 2002;29:68-74.
- Fleiss FL. The statistical basis of meta-analysis. *Stat Methods Med Res* 1993;2:121-45.
- Eliaszewicz M, Young SL, Woodbury MG, et al. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. *Phys Ther* 1994;74:777-88.

APPENDIX

Interlaboratory Reliability

For any particular manufacturer, let Y_{ijk} denote the mean OD measurement from the laboratory $i = 1, 2, \dots, I$, using antibody kit $j = 1, 2, \dots, J$, and serum sample $k = 1, 2, \dots, K$. From our experimental design, I is generally 3 or 4, J can vary from 6 to 9 (depending on manufacturer), but K is always 21 (since every laboratory evaluated each of the 21 different sera).

The statistical model may be written: $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \varepsilon_{ijk}$.

Here, the underlying mean μ is a constant;

the α_i are independent normal variates, each with mean 0 and variance σ_A^2 ;

the β_j are constants satisfying $\sum_{j=1}^J \beta_j = 0$;

the γ_k are independent normal variates, each with mean 0 and variance σ_C^2 ;

the interaction terms $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, $(\beta\gamma)_{jk}$ are independent normal variates, all with means = 0 and respective variances σ_{AB}^2 , σ_{AC}^2 , σ_{BC}^2 , subject to the constraints that $\sum_j (\alpha\beta)_{ij} = 0$ for any i ,

and $\sum_j (\beta\gamma)_{jk} = 0$ for any k ;

the error terms ε_{ijk} are independent normal variates, each with mean 0 and variance σ_ε^2 .

This is a standard mixed models analysis of variance, laboratories and serum samples representing random effects, and antibody kits fixed effects. We define the interlaboratory correlation coefficient ICC as the correlation between independent determinations by

laboratories i and i' of the same serum sample with the same antibody kit. That is, $ICC =$

$$\text{corr}(Y_{ijk}, Y_{i'jk}) = \frac{\text{covariance}(Y_{ijk}, Y_{i'jk})}{\sqrt{\text{var}(Y_{ijk}) \text{var}(Y_{i'jk})}}.$$

From the statistical model posited above,

$$\text{cov}(Y_{ijk}, Y_{i'jk}) = \sigma_C^2 + \sigma_{BC}^2$$

and

$$\text{var}(Y_{ijk}) = \sigma_A^2 + \sigma_C^2 + \sigma_{AB}^2 + \sigma_{AC}^2 + \sigma_{BC}^2 + \sigma_\varepsilon^2,$$

$$\text{hence } ICC = \frac{\sigma_C^2 + \sigma_{BC}^2}{\sigma_A^2 + \sigma_C^2 + \sigma_{AB}^2 + \sigma_{AC}^2 + \sigma_{BC}^2 + \sigma_\varepsilon^2}$$

Each of the components of variance is estimable from the ANOVA decomposition, which was performed in SYSTAT. See Eliasziw et al.¹⁷ for detailed discussion of the underlying model and analysis.

We remark that we might also have considered a model in which the laboratory effects are fixed and not random. In general, this would produce ICC's slightly higher (closer to 1) than the random laboratory decomposition. For purposes of generalizing to other laboratories, the random effects assumption for the laboratory component seems more appropriate than the fixed effects assumption.