Long-Term Variability in Immunofluorescence Titer of Antibodies to Nuclear Antigens Observed in Clinical Laboratory Proficiency Testing Surveys

Mark H. Wener, MD; Susan Fink, MD, PhD; Christine Bashleben, BS, MT(ASCP); Stephanie Sindelar, BSc, MLS(ASCP); Michael A. Linden, MD, PhD

Tests for antibodies to nuclear antigens (ANAs) are important for clinical diagnosis of several autoimmune rheumatic diseases, are central to epidemiologic classification of systemic lupus erythematosus and subsets of systemic sclerosis, and have been used to screen for eligibility in some clinical trials for lupus therapies. The ANA test performed on HEp-2 cells has been considered the American College of Rheumatology gold standard for ANA testing and has been considered the reference ANA method by an international initiative. The recently revised classification criteria for systemic lupus erythematosus set an ANA titer of 1:80 or equivalent on HEp-2 cells as the “permissive” entry criterion for lupus classification. Reliable and consistent ANA titer results are therefore necessary for clinical, research, and regulatory purposes. Despite that need, clinical experience and small studies have described that ANA results differ when performed in different laboratories. Reports from external quality assessment proficiency testing surveys, which involve sending the same specimen to multiple laboratories participating in compliance and quality assessment programs, have demonstrated that inconsistency between results from different laboratories is common. Furthermore, these analytical differences can lead to statistically significant differences in the operating characteristics of ANA tests, including the calculations of cutoffs for positive and negative results.

Many factors may account for the differences between results from different laboratories, including the choice of HEp-2 cell kits used by different laboratories. For example, previous studies performed within individual laboratories have demonstrated that ANA results vary depending on the kit used, even when the testing is performed by the same staff. However, previous studies have not examined the
long-term contribution of the kits to the variation in a large population of laboratories.

To systematically investigate the variability in ANA results reported by laboratories in the United States, data from 11 years of laboratory proficiency testing conducted by the College of American Pathologists were analyzed. These data were used to quantify the variability of ANA titer reporting and the potential influence of the HEp-2 kit source on the ANA titer results. In addition, these results were used to examine the stability of ANA titer rankings of various kits during the 11 years studied, and to determine how the titer rankings were influenced by the observed ANA patterns.

METHODS

Data Source

Data were extracted from the HEp-2 ANA test participant summary reports (PSRs) of 165 test events sent between 2008 and 2018 from the College of American Pathologists (CAP; Northfield, Illinois) to laboratories subscribing to the external quality assessment Diagnostic Immunology S proficiency survey series. (More than 95% of participant laboratories are in the United States.) Individual laboratories were instructed to treat the proficiency testing survey specimens as if they were routine clinical specimens and to report whether each specimen was negative or positive, together with the method and kit used by that laboratory. For the ANA test, the criterion for assessing individual laboratory performance was based on a requirement that the laboratory report the consensus-positive or consensus-negative result if there was agreement among at least 80% of reporting laboratories. Because proficiency survey specimens are designed in part to evaluate the agreement with consensus, these specimens may underrepresent borderline titer samples for which consensus would not be reached. Results reported for consensus-positive and consensus-negative specimens were considered true positives and true negatives, respectively, whereas the opposite results were considered false negatives and false positives, respectively. When ANA results from a laboratory were reported as positive, the pattern and titer results were also sent to the CAP. The analysis in this manuscript is based on the PSRs that summarized the number of laboratories reporting positive or negative results separately for each kit (reported in the CAP PSR as “Method/Manufacturer”) used by at least 10 laboratories. For specimens reported as positive, the ANA patterns were reported. The ANA titers reported in the PSRs were partitioned into the following bins: less than or equal to 1:10; 1:16 to 1:20; 1:32 to 1:40; 1:64 or 1:80; 1:128 or 1:160; 1:256 or 1:320; 1:512 or 1:640; greater than 1:640; 1:1024 or 1:1280; 1:2048 or 1:2560; greater than or equal to 1:5120. For the purposes of this study, the titers divisible by 10 (eg, 1:10, 1:20, 1:40) were used for calculating titer results. Any ANA titers reported to the CAP as greater than 1:640 without an endpoint titer were assigned a titer of 1:1280, and titer results reported as greater than or equal to 1:5120 were assigned a titer of 1:5120. The screening titers of ANAs (the cutoff above which a result would be considered positive) was not reported to the CAP by individual laboratories. Because one of the goals of this study was to investigate the consistency of kits relative to each other, HEp-2 slide immunofluorescence or immunoperoxidase kits reported in the 2018 CAP PSR with less than 4 preceding years of data were excluded in this evaluation. In different years of the study, 8 to 11 different kits were available for analysis using the PSRs; therefore, when ranking of kits was performed, ranks were normalized to 8.

For the 11-year study period, 91,614 individual ANA results from 165 samples with 1602 kit-specimen combinations were used for determining agreement of reporting positive and negative results. Of the 165 samples, 88 (53.3%) were reported as positive by more than 80% of laboratories, and therefore were considered consensus-positive specimens. For these positive specimens, 54,403 results, with 824 specimen-kit combinations, were available to analyze variability in ANA titer reporting. On average, 65.3 (median, 44) laboratories reported titer results for each sample-kit combination.

The consensus patterns reported included homogeneous, speckled, nucleolar, centromere, and mixed patterns, and those pattern designations were used to determine the influence of patterns on the titer rankings of different kits.

Statistical Methods

Data reported by the CAP to participating laboratories in PSRs were copied into an Excel (Microsoft, Redmond, Washington) spreadsheet. The medians and the geometric mean titers were calculated for each specimen overall as well as for each specimen-kit combination. The data from Excel spreadsheets were exported into SAS University Edition (SAS, Cary, North Carolina). One-way and multiple analyses of variance, nonparametric tests, and ranking of titers were performed using SAS procedures. For multiple analysis of variance, each specimen’s identifier, its ANA pattern, the kit, and the testing date were included as independent variables, and the log of the geometric mean ANA titer for each specimen/kit combination was considered the dependent variable. The rank order of the geometric mean titer of each individual kit was determined for each specimen, to evaluate the relative position of each kit’s results relative to the other kits despite the absence of some of the kits during earlier test episodes. The rankings at all time points were converted into quantiles of 8 (because as few as 8 kits were evaluable at some time points), with the cut points for the 8 quantiles determined using the SAS default. The statistical significance of differences in kit results and the titer ranks of kits were evaluated by the nonparametric Kruskal-Wallis test. The yearly mean rankings were also calculated and were used to determine the consistency of rankings relative to each other throughout the 11 years of data. The rankings were also used to compare the overall relative performance with the ranking for the common patterns reported, to determine if different patterns were associated with different titer ranks of each kit. Statistical significance was considered P < 0.01, but other P values are shown.

The rates at which individual kits were discordant with consensus qualitative results were also calculated. The false-positive rate of each kit-specimen combination was determined as the proportion of positive results reported from laboratories using a particular kit for a specimen reported as consensus positive. Conversely, the false-negative rate was determined as the proportion of negative results for specimens that were positive by consensus.

RESULTS

Consensus regarding whether a specimen was positive or negative was achieved for all 165 specimens in these surveys. The 88 positive samples covered a range of titers, with individual samples having a median titer as low as 1:80, and other samples with a median titer of 1:1280. A mean of 34.2% (range, 19.5%–54.7% in different specimens) of the laboratories reported the same titer as the median individual titer for a given specimen. The median titer plus or minus 1 titer (ie, a span of three 2-fold titers) was reported by 81.2% of laboratories (ranging from 60.6% to 90.0% for different specimens). Reported titers were within the range of plus or minus 2 titers (span of five 2-fold titers) for a mean of 96.8% of laboratories, with a range of 90.2% to 99.2% for different specimens.

Different individual laboratories frequently reported very divergent results. For 54 of the 88 positive specimens, there were 1 or more laboratories that reported a titer of 1:40 whereas 1 or more laboratories reported a titer 1:5120 for the same specimen—that is, a span of 8 serial 2-fold titers. None of the specimens had a span of less than 7 serial 2-fold titers. Figure 1 illustrates the range and distribution of
results reported for a representative single proficiency specimen.

One contribution to this variability in reported titers is the use of kits from different reagent manufacturers. For individual survey samples, the geometric mean titers of individual kits differed from the overall geometric mean titers by as much as 1.8-fold, and the difference between the geometric means of the highest kit and the lowest kit averaged 3.2 (about one and a half 2-fold titers), with a maximum of 8.3 (ie, more than three 2-fold dilution titers).

The differences between kits demonstrated substantial consistency over time. For each specimen, individual kits were ranked according to their geometric mean titer. Summary information about the titer rank of each kit is graphed in Figure 2, which demonstrates the relatively consistent ANA titer ranking of individual kits across all 88 positive specimens. Identity of kit manufacturers corresponding to kit letter codes are contained in Supplemental Table 1 (see supplemental digital content file at https://meridian.allenpress.com/aplm in the August 2021 table of contents). Differences between kits were highly significant
Correlations between mean titer rank of each kit and the rank (as false-positive). The kit false-positive rate is correlated with the titer reported as positive in consensus-negative specimens are designated positive specimens are designated as false-negative, and results nuclear antigens (ANAs). (Results reported as negative in consensus-negative specimens but 1 were reported as negative by more than 98% of laboratories. Despite this general concordance, more than 1% of laboratories reported positive results in 14 of the 77 consensus-negative specimens (18.2%). Kits with a higher ANA titer rank demonstrated a higher rate of positive results among the consensus-negative specimens (P = .008 for correlation; Figure 5). Conversely, there was a tendency for kits with a higher ANA titer rank to have a lower rate of negative results in consensus-positive specimens (P = .17; Figure 5).

**DISCUSSION**

It has long been recognized that ANA tests performed in different laboratories can yield different results. Tan et al.\(^1\) investigated the frequency of ANA in normal individuals and noted wide variation in ANA titers reported worldwide from experienced laboratories. In a study\(^2\) of 18 Italian laboratories testing shared specimens, qualitative (positive/negative) agreement in ANA results was generally acceptable, but quantitative/titer results were variable. A recent small study\(^3\) observed substantial variation in ANA results ordered clinically from different US commercial laboratories.

The observations and analysis presented here confirm and extend previous observations that HEp-2 ANA titers tend to vary depending on the kits/reagents used. The strengths of this study include the large number of laboratory results (91,614), tested in many laboratories (typically more than 500 for each specimen, with the vast majority based in the United States), and the long duration (11 years) of the compiled observations that form the basis of the analyses. We have used a novel titer rank approach to facilitate comparing a variable number of kits with a variety of ANA patterns and titers during many years. We have documented that the differences between kits are relatively consistent during 11 years, and that they are similar for common ANA patterns. Furthermore, we have quantified the differences in titers provided by using different kits.

Many factors contribute to differences in HEp-2 ANA titer results. Among the factors controlled by kit manufacturers include the potential genetic variations that may have arisen during 11 years, and that they are similar for common ANA patterns, the titer rank of each kit was quite consistent (Figure 4). The high correlation demonstrates that kit rankings were similar for all patterns evaluated. Nevertheless, for a few individual specimen samples, there were some observable differences in the relative ranking of kits. The HEp-2000 cell variant substrate overexpresses recombinant SSA/Ro antigen; as expected, specimens positive for anti-SSA display a distinct cellular staining pattern on those cells and demonstrate relatively higher reactivity.

There was a high level of consensus among laboratories for reporting proficiency testing specimens as positive, and good concordance for consensus-negative specimens. Of the 88 consensus-positive specimens, 65 (74%) were reported as positive by all laboratories and only 1 was reported as positive by less than 90% of laboratories. Among the 77 consensus-negative specimens, 8 specimens (10.4%) were reported as negative by all laboratories, and all specimens but 1 were reported as negative by more than 98% of laboratories. Despite this general concordance, more than 1% of laboratories reported positive results in 14 of the 77 consensus-negative specimens (18.2%). Kits with a higher ANA titer rank demonstrated a higher rate of positive results among the consensus-negative specimens (P = .008 for correlation; Figure 5). Conversely, there was a tendency for kits with a higher ANA titer rank to have a lower rate of negative results in consensus-positive specimens (P = .17; Figure 5).
buffers used in serum and reagent diluents; and calibrators or controls supplied for the kits. The data presented in this study demonstrate that the systematic differences between kits are statistically and likely clinically significant, but they account for only some of the variation observed. Differences between individual laboratories using the same kits also lead to substantial variation in ANA titer results (see, for example, Figure 1). Thus, knowledge of the HEP-2 kit or reagent used by an individual laboratory does not provide a clinician sufficient information for definitive interpretation of ANA titer rank from that laboratory.

Individual laboratories are ultimately responsible for the results produced, and have substantial control over other factors, including choice of initial and serial serum dilution concentrations tested to determine the endpoint titer; laboratory conditions during specimen and antibody incubations; microscope manufacturer and lenses; microscope illumination type, intensity, and optical filters; criteria for considering a given microscopic appearance to be at endpoint; ambient lighting conditions used during microscopy; choice and use of controls and calibrators; choice and validation of reference normal range; and training and consistency of technical staff. For nonautomated fluorescence microscopy, the ultimate performance of ANA testing depends most on an individual laboratory establishing performance characteristics to meet its expectations. For example, individual laboratories routinely establish acceptable parameters of analytical precision. To improve overall performance, they also have the opportunity to correlate the ANA screening and titer results with results from specific ANA subset antibody testing; to determine clinical specificity with respect to reference/normal populations; potentially to determine clinical sensitivity in working with clinical colleagues; and to evaluate other factors. Some of these evaluations can be difficult, expensive, and/or time-consuming, but individual laboratories, depending on their focus and resources, do have the opportunity to address these issues. Clinicians using the same clinical laboratory often learn the performance expected from a given laboratory, and can provide helpful feedback to laboratory colleagues. Conversely, experienced clinicians often expect greater precision and consistency from the laboratory they are accustomed to use, and therefore interpret ANA results with that background knowledge about individual laboratory performance.

Although proficiency testing programs are designed primarily as a quality measure to allow laboratories to compare their individual performance with those of other laboratories, they have additional value to assess laboratory performance in populations. Proficiency testing results have been used to describe reagent performance for various tests, including tests important in autoimmune rheumatic diseases. There is also recognition that results of external quality assessment, such as proficiency testing, can play an important role in assessing the need for kit improvement, quality assessment, such as proficiency testing, can play an important role in assessing the need for kit improvement. Nevertheless, there are limitations in using proficiency testing specimens such as those in our study to compare kit performance. For these specimens, the consensus ANA patterns were known, but neither the specific antigens recognized by the specimens nor the diagnosis of the patients providing the specimens were available. Another limitation was that rare ANA patterns and uncommon autoantibody specificities were not tested in these specimens. A potential limitation of the long-term studies described here concerns the possibility that the same serum source was distributed for testing more than once, as separate proficiency testing specimens. If that occurred, then the variability present in different sera (with different binding avidities, specificities, and other properties) could be underrepresented by these studies. (It is not known how often the same sera were sent labeled as distinct specimens for testing in this series, but repeat sampling may have occurred in a minority of testing events.) A common limitation of external quality assessment programs is that specimens distributed to laboratories may have been prepared with preservatives or stabilizers not found in clinical specimens, and kits may be differentially influenced by these extraneous materials. These factors limit the conclusions that can be drawn when differences between kits are observed only with survey material. Thus, findings from proficiency testing surveys benefit from confirmation of their commutability (ie, their ability to provide information valid in clinical specimens) by alternative approaches, including verification that the survey specimens give results that are equivalent to clinical specimens. Further studies to address the commutability are under way (S.F., unpublished data, 2020).

Although our data demonstrate significant and sometimes substantial variations in ANA quantitation, these observations also suggest an approach for improvement. The rankings between kits were remarkably consistent during 11 years, which indicates consistency among the kit manufacturers that supply HEP-2 cells and reagents. Because differences between kits are relatively constant and systematic, rather than random and sporadic, these observations suggest that there is the potential to harmonize the kits relative to each other in order to improve the consistency of ANA quantitation. If manufacturers can be encouraged by professional organizations such as rheumatology and pathology societies, and permitted by regulatory agencies to adjust their kits to achieve harmonization, the quality measures they currently use to achieve consistency might be directed also toward achieving harmonization. Variability within individual laboratories will still need to be addressed, because even laboratories using the same kit vary substantially, as shown in Figure 1. However, a significant initial improvement could be achieved by appropriately aligning, harmonizing, or normalizing kits. For example, there could be a method for normalized scoring across all kit platforms, analogous to the international normalized ratio used for adjusting prothrombin time results. Another approach to begin to address ANA titer variation would be for the CAP and other organizations conducting proficiency surveys to provide more feedback to laboratories that lie outside 95th and 99th percentile distributions for ANA titers of survey specimens, using all methods to establish the population distributions.

Automation of HEP-2 immunofluorescence ANA has been introduced and its use is expanding, although information about use of automation by laboratories participating in these surveys is not available. Automation might be anticipated to lead to greater reproducibility and consistency for each automated kit and instrument. However, if different automated platforms are not harmonized with respect to each other, then the potential for overall improvement in precision and consistency may not be
achieved.\textsuperscript{17,18} Recently, progress has been made by professional organizations to improve the accuracy, consistency, and standardization in reporting the patterns of nuclear and cytoplasmic staining observed in HEP-2 cells.\textsuperscript{19}

The additional problem of achieving consistency in ANA quantitation or titration also warrants efforts for improvement. Because distinguishing between negative and positive ANA and low-level and significant-level ANA results has a key role in evaluation and diagnostic classification of patients with several autoimmune rheumatic diseases, improved quantitation offers the possibility of more efficient and more accurate diagnosis of these conditions. For the ANA test performed on HEP-2 cells to continue as a putative gold standard, ANA testing on HEP-2 cells should be linked to improved consistency of quantitation.

In summary, the remarkable finding in this large study is that the differences between kits are systematic, are consistent over time, and are similar for the common ANA patterns. Attempts at harmonization between kits therefore have the potential to be successful. Achieving the changes necessary for harmonization will likely benefit from collaboration among vendors, laboratories, and regulatory agencies, with substantial input and influence from interested clinical groups such as clinical laboratory professionals and rheumatologists.

The authors thank Thomas Long, MPH (College of American Pathologists), for his helpful data analysis guidance.

References