Estimating Reference Intervals

Gary L. Horowitz, MD

As indicated in the thought-provoking article in this issue of the Journal by Katayev and colleagues, there are few things more important than the reference intervals we report along with our laboratory measurements. Unfortunately, we laboratory professionals give them far too little attention, adopting manufacturers’ recommended intervals, often without even verifying them ourselves and, rarely, if ever, establishing our own values. Thus, the article “Establishing Reference Intervals for Clinical Laboratory Test Results: Is There a Better Way?” is certain to attract much attention.

Katayev and colleagues make several assertions that bear comment. They claim that there is no clear guideline as to which technique to use, but the recommendations of the International Federation of Clinical Chemistry and Laboratory Medicine and the CLSI are quite clear. If one collects samples carefully from 120 vetted healthy people, then the technique of choice is nonparametric analysis. The reason for this is that the nonparametric technique requires no knowledge of, and makes no assumptions about, the nature of the data distribution. In other words, the reference interval values obtained are valid no matter what the underlying distribution is. If one has fewer samples, again from carefully screened, apparently healthy people, one can use a parametric technique with as few as 40 points, so long as the original data (or some transformed version of the data) exhibit a gaussian distribution. And with even fewer than 40 data points, one can use robust techniques to get an estimate of the reference interval. Notwithstanding the assertions by Katayev and colleagues, the reference intervals obtained from properly collected and analyzed data do not vary depending on the technique used.

Katayev and colleagues also make the point that it becomes prohibitively difficult to collect sufficient data for all the potential partitions for which one might want reference intervals. They mention specifically sex and age (eg, deciles), but one might also include others (eg, fasting and race or ancestry). In this regard, it is particularly interesting that Katayev and colleagues did not specifically mention differences (or lack thereof) for any partitions, with the exception of sex for hemoglobin and creatinine. In his original article, Hoffman demonstrated the use of his technique with “just” 500 points. Clearly, Katayev and colleagues had more than enough data to look at every analyte by sex and by age. One wonders, for example, whether there was any effect of sex and age on calcium or on thyroid-stimulating hormone (TSH).

The fact that something is difficult to do does not negate the importance, or usefulness, of doing it. As a particularly good example, a group in the Netherlands recently did a superb reference interval study. By collecting data for 1,444 people and using the recommended nonparametric method of analysis, they determined that creatine kinase reference intervals varied tremendously not only by sex but also by race/ancestry as well. Specifically, the 97.5th percentile for women varied from 201 to 313 to 414 IU/L for “white Europeans, South Asians, and blacks,” respectively; for men, the corresponding values were 322, 641, and 801 IU/L. By using the manufacturer’s reference intervals (defined as a single partition), the authors showed that the proportion of healthy women whose values were “abnormal” was 8% for white Europeans, 16% for South Asians, and 42% for blacks; for men, the corresponding values were 17%, 32%, and 62%.
The truth, though, is that Katayev et al\textsuperscript{1} are absolutely correct in pointing out that few laboratories are capable of undertaking a reference interval study of this magnitude. What can we do? As recommended in the CLSI document,\textsuperscript{3} we can verify, rather than establish from scratch, a reference interval established elsewhere by collecting samples from just 20 carefully vetted, healthy people. If no more than 2 of these 20 samples have values outside the proposed interval, then it is statistically valid to adopt the proposed interval. Any laboratories that attempted to do this for creatine kinase would likely have discovered that there was a major problem with the manufacturer’s proposed interval.

Of course, if you find that you cannot adopt the manufacturer’s reference interval, what do you do next? What if you could find other laboratories that used the same method, each of which collected data from its own set of 20 reference subjects? If you pooled the data from 6 laboratories, you would have 120 different values, enough to establish a reference interval using the recommended technique (at least for 1 partition). In the same way, if you could pool data from 20 or even 50 laboratories, all using the same method, you would have 400 or even 1,000 different values, which might allow you to establish reference intervals for many different partitions. As it turns out, the College of American Pathologists offers, among its proficiency testing program services, a service that does just this for its participants.\textsuperscript{7} It is, at least in my opinion, an excellent, although underused, service.

It is important, too, that laboratory professionals realize that for many analytes where national or international guidelines apply, they should not attempt to establish, or verify, their own traditional reference intervals; they should use the indicated “decision limits.” Examples include cholesterol, high-density lipoprotein cholesterol, triglycerides, glucose, glycated hemoglobin (hemoglobin A\textsubscript{1c}), and neonatal bilirubin. In these cases, the laboratory’s job is to ensure accuracy—that its results on patient specimens match those that were used to establish the guidelines. In many laboratories, it is assumed the its methods are accurate, which is not entirely unreasonable if homogeneous systems are used in which the reagents, calibrators, and instruments have been validated by the manufacturers, who have done remarkably good jobs with many, if not all, of these analytes. Some laboratories may go further by subscribing to proficiency survey programs that use commutable materials and establish values by reference methods.\textsuperscript{8,9} The main point, though, is that efforts that would have been directed to establishing, or verifying, reference intervals should, for these analytes, be directed instead toward verifying accuracy of the method.

In this connection, it is interesting that Katayev et al\textsuperscript{1} included creatinine among the analytes in their report. Because of differences in sex, age, and race, establishing a reference interval for creatinine is fraught with problems. Rather, laboratories should report the estimated glomerular filtration rate (GFR), using the Modification of Diet in Renal Disease Study equation, whenever they report serum creatinine values. As noted on the National Kidney Disease Education Program Web site, estimated GFR “provides a more clinically useful measure of kidney function than serum creatinine alone.”\textsuperscript{11} For example, a 55-year-old non-African American woman with a creatinine level of 1.0 mg/dL, a value at the upper end of the reference interval, has an estimated GFR of 58 mL/min/1.73 m\textsuperscript{2}, indicative of chronic kidney disease.\textsuperscript{12} The “normal” creatinine value is clinically misleading.

The inclusion of TSH in the article by Katayev et al\textsuperscript{1} is interesting as well. Because the incidence of so-called subclinical hypothyroidism is higher among women than among men, why did they not evaluate values for women and men separately? And, to the extent that subclinical hypothyroidism also increases with age, why did they not look at their data as a function of age as well as sex? Assuming their data had confirmed these phenomena, they could have questioned the validity of the current reference intervals, as others have. That is, if a TSH value in an asymptomatic older woman is outside the central 95% of values in younger people of both sexes, what should a concerned clinician do? Many would argue that it may be wise to do nothing.\textsuperscript{13-15} If that is the case, is the value genuinely “abnormal”?\textsuperscript{16}

And this begs the question of why we typically use 97.5% as the upper limit of the reference interval. To the extent that physicians are screening (ie, ordering tests for people without symptoms or signs), then, by definition, 2.5% of “normal” people will have “abnormal” results and, therefore, possibly be subjected to additional, unnecessary testing. Is there any real medical benefit in following up on a calcium level in the 98th percentile in the absence of symptoms or signs? This may be part of what drives up health care costs—follow-up on tests that are outside the central 95% of results but that, by themselves, warrant no further action. As laboratory professionals, we need to think beyond simply providing central 95% reference intervals and more about what we want clinicians to do with the information we provide. Perhaps the reference interval for calcium should include the central 99.0% of calcium values.

The issue of what constitutes abnormal becomes even more controversial when we turn to prostate-specific antigen (PSA) screening. A PSA value of 5.0 ng/mL is typically flagged as abnormal, but what does it mean in a man with no symptoms and no signs? Without doubt, as a result of screening, we detect prostate cancer at earlier stages, but it...
is not clear that, overall, patients live longer or benefit in any other ways. Are too many men undergoing biopsies and even therapy, the benefits of which are less certain than the costs and adverse effects?16,17

Even if one grants the attractiveness of the proposal by Katayev et al,1 it does nothing to help individual laboratories that may use different methods or that may serve populations with different backgrounds or that may not have the computing power to replicate their analyses. What are they to do? I would submit that they can still use the Hoffman technique as a powerful quality assurance tool.

As Katayev and colleagues1 point out, the huge advantage of the Hoffman technique is that it does not require that samples be obtained from healthy people. Indeed, in his article, Hoffman2 demonstrated the technique using sample sets with admixtures of 20% abnormal values. Furthermore, the Hoffman technique does not require sophisticated statistics and computer analyses. In the original description, Hoffman plotted, on gaussian probability paper, the cumulative frequency of test values vs concentration. (It should be noted that this is very different from Figure 1 in the article by Katayev and colleagues; presumably, the computer techniques used made Figure 1 equivalent to the original description.) The effect of gaussian probability paper is to give much more weight to the central part of the distribution, lessening the contributions of values as they deviate from the center, values that are progressively less likely to be from healthy people.

If the reference intervals generated in this way are strikingly different from the reference intervals in use (for example, if 30% of samples are labeled as abnormal), the laboratory needs to do some troubleshooting. Maybe the original reference interval study was flawed, or maybe the method was not implemented correctly, or maybe the local population is different, or maybe the method has drifted over time, or maybe the “reference interval” is really a “decision limit.” Whatever the cause, it is incumbent on laboratory professionals to make sure they understand why the apparent reference interval by the Hoffman technique is so different from the reference interval in use. As noted at the outset, I believe that this exercise is at least as important as reviewing quality control and proficiency testing.

My assessment is that the proposal by Katayev and colleagues1 is not a “better way” to establish reference intervals, but I am indebted to them for making me think hard about the whole issue of reference intervals, for encouraging me to read in detail Hoffman’s2 classic article, and for enabling me to see that there are other, equally important applications of their ideas from which we can all benefit.

References
15. Surks MI, Ortiz E, Daniels GH, et al. Subclinical thyroid disease: scientific review and guidelines for diagnosis and management. JAMA. 2004;291:228-238.