Arguably, one of the most important elements of any clinical laboratory test is the reference interval, the values that help physicians interpret their patients’ test results. Although it is frequently recommended that laboratories establish their own reference limits with their own methods and local patient populations (1), few laboratories have the resources to do all the work required to achieve this goal.

In this issue of Clinical Chemistry, Adeli and colleagues provide, in a series of 3 papers (2–4), an exceptional compilation of high-quality reference interval data. Following a protocol they developed for the Canadian Laboratory Initiative on Pediatric Reference Intervals (CALIPER)(2) (5), which was undertaken to address the dearth of pediatric reference interval data, they arranged for roughly 12 000 reference individuals to provide blood and urine samples to be analyzed for a large number of common laboratory tests. These individuals, males and females from multiple age groups, were selected from across Canada in an effort to represent the entire Canadian population. Their health was vetted by questionnaire, personal interview, and brief physical examination. Analyses were done largely by a single laboratory using well-defined, standard laboratory instrumentation; exceptions included complete blood counts (CBCs) which, for specimen integrity reasons, were done at the individual collection sites, using instrumentation from a single manufacturer to ensure consistency. Their data analysis followed the CLSI protocol (1) completely with respect to detection of outliers, need for partitioning, and use of nonparametric statistics.

Laboratories using these methods and located in Canada may now have the high-quality reference interval data they always wished they had but could not arrange to collect themselves. Actually, the data may be applicable to a much wider audience.

With respect to methods, although these studies were generated with specific instrumentation and reagents, the results may be transferable to other instrument/reagent combinations (1). Indeed, much of the original CALIPER data have been adapted to other common platforms (6). A companion editorial in this issue of Clinical Chemistry (7) provides a more detailed description of the procedures involved.

As noted by the authors, their proposed reference intervals are designed to apply across Canada, because they weighted the data to represent Canada. Interestingly, there are profuse data on age and sex but no discussion of race or ethnicity, except in the third article (4), where the authors state that small sample sizes precluded analysis of the effects of these factors. This is unfortunate, because race and ethnicity can exert powerful effects on reference intervals. In a recent study (8), it was noted that, at least for creatine kinase, there were clinically significant differences not only by sex but also by race and ethnicity (>2-fold). Of note, creatine kinase was not among the analytes in the current studies.

Once the issues of method transferability and population comparability have been addressed, it is quite possible that the reference interval data presented in these articles will have very widespread applicability, far beyond these particular methods and Canada. In other words, many of us may have the very data that we have long sought.

Unfortunately, now that we have these data, a new series of questions arises, and answering these questions is more subjective and probably more difficult.

In these articles, Adeli and colleagues chose to use the central 95% of the partitioned data to calculate reference intervals. Although this is the traditional and most commonly used approach, is it the best one to use? If an apparently healthy individual has a value in the 99th percentile [e.g., a 52-year-old man with a serum calcium of 10.3 mg/dL (2.58 mmol/L)], is that really “abnormal”? If so, what disease does he have? What follow-up, if any, is needed? Perhaps none, but that result may get an asterisk (or a similar flag), one that looks just like the flag applied to a serum calcium of 14.5 mg/dL (3.63 mmol/L), a value that almost certainly needs follow-up. What may we need are asterisks of different sizes or colors, or some other way of delineating results that are outside some arbitrary central distribution but not necessarily abnormal, or indicative of disease, or requiring further action.

Things get even more complicated when one adds the effects of imprecision—do we really want a physician to interpret a serum calcium of 10.3 mg/dL (2.58 mmol/L) differently from a serum calcium of...
10.2 mg/dL (2.55 mmol/L)? Not to mention the effects of other tests on the interpretation of a given test—for example, “low” serum calcium values are not uncommon in hospitalized patients, but these same patients may also have low serum albumin values, in which case the low calcium may actually be appropriate. Nonetheless, the low calcium value almost certainly has an asterisk appended to it. As laboratory directors, we should aspire to do a better job than simply adding asterisks to every result that exceeds by any amount the limits associated with an arbitrary percentage of a healthy population.

All of these problems are exacerbated because physicians seem to be ordering ever-increasing numbers of tests on healthy people (screening) and because patients now have direct access to their laboratory test results. As a result, those asterisks are appearing more frequently, and they are being seen by people who have no training in laboratory medicine.

One possible, partial solution to some of these issues is to increase the proportion of the reference population included in the reference interval (for example, from 95% to 99%). This simple maneuver would decrease the number of false positives 5-fold (from 5% to 1%) and, in all likelihood, would not result in obscuring clinically significant abnormalities.

Another aspect of these articles that deserves comment is the authors’ repeated use of the word “required” in connection with partitioning (e.g., “Albumin, total protein, total calcium, total bilirubin, sodium, and chloride all required 4–5 age partitions” (2)). Statistics may indicate that these analytes qualify for partitioning, but, as laboratory directors, we can (and should) exercise judgment as to whether such partitioning is clinically appropriate. In my view, far fewer reference intervals (perhaps even just 1) for each of these analytes may be sufficient. Indeed, the authors themselves seem to suggest as much, finishing the assertion above by adding “although changes in concentration across ages were relatively minor.”

Among the analytes tested were several for which decision limits exist: hemoglobin A1c, cholesterol, LDL cholesterol, HDL cholesterol, and triglycerides. In these cases, the intention is not to substitute the observed reference intervals for the decision limits. Rather, as the authors emphasize, the data they collected can be used to do a better job than simply adding asterisks to every result that exceeds by any amount the limits associated with an arbitrary percentage of a healthy population.

As the authors point out relative to the hemoglobin A1c scatterplot (which includes all the data before application of the exclusion criteria), not only did a surprising number of individuals meet the criteria for diabetes mellitus (who presumably were eliminated once the exclusion criteria were applied), but also a large number of individuals met the criteria for prediabetes. Indeed, the upper limit of the “reference intervals” falls well into the prediabetic range. With increased public health efforts directed toward reducing the prevalence of diabetes, one might expect the proportion of prediabetic individuals to decrease in future surveys.

Serum creatinine represents a particularly interesting case. Although perhaps not conventionally considered a test with decision limits, it really should be. Laboratories still provide reference limits for creatinine, but it is well known that they can be misleading. For example, a 67-year-old white woman with a creatinine of 1.0 mg/dL (88.4 μmol/L), which falls within the proposed reference interval of 0.6–1.0 mg/dL (53–88.4 μmol/L), has an estimated glomerular filtration rate (eGFR) of 55 mL·min⁻¹·(1.73 m²)⁻¹, suggestive of chronic kidney disease (10). In other words, it is the eGFR result that provides useful information, for which an accurate creatinine is vital.

One other aspect of the articles by Adeli and colleagues warrants attention. As noted earlier, to obtain high-quality reference interval data, these studies closely followed CLSI guidelines, including the use of well-defined exclusion criteria. However, it is disheartening that such a high proportion of individuals was excluded; for individuals 60–79 years old, the figure was 79%! Similar figures have been cited in other studies (11). In the future, consideration should be given to changing some of the exclusion criteria to partitioning criteria (e.g., the use of prescription drugs), which would allow comparison of results from these individuals to the others. It is conceivable that their results are comparable for at least some of the analytes, and where different, detailed descriptions of those distributions could be provided.

In summary, Adeli and colleagues have collected and analyzed an immense volume of high-quality data, providing central 95% reference limits, partitioned by age and sex, for many common analytes. Whether individual laboratories can reliably use these data depends first on establishing method and population comparability. Even the large number of laboratories that can make use of the data, however, should keep in mind that they do not have to adopt the partitions indicated or the universal use of central 95% limits. All of us should remember not to impose reference limits, no matter how well they are determined, on assays for which decision limits exist. And finally, for groups such as the elderly, in which the overwhelming majority may not qualify to be included in reference interval studies using conventional exclusion criteria, we need to find alternative strategies to provide comparison data to help physicians interpret their test values. Adeli and colleagues have provided us all with an amazing collection of data. How we use those data—when we apply those asterisks—is up to each of us.
Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

Authors’ Disclosures or Potential Conflicts of Interest: No authors declared any potential conflicts of interest.

References