REFERENCE INTERVALS: PRACTICAL ASPECTS

Gary L. Horowitz, MD
Beth Israel Deaconess Medical Center
Boston, MA 02215

Arguably, one of the most important elements of a laboratory test is the reference interval, the values that help clinicians interpret their patients' test results. Interestingly, though, laboratorians spend surprisingly little time formally addressing the issue of reference intervals. Typically, they adopt the intervals given them by manufacturers, neither establishing the intervals themselves nor even verifying the applicability of those intervals to their patients. In addition, for many analytes, conventional reference intervals, values typically representing the central 95% of healthy individuals, have been superceded by decision limits. But, once again, individual laboratories rarely verify that their methods provide accurate values, a requirement if one is to use these decision limits.

In this communication, I would like to review these three topics in some detail:

1) how laboratories can (and should) establish the accuracy of their values for those tests that have decision limits (e.g., cholesterol, glycated hemoglobin, neonatal bilirubin)

2) how laboratories can (and should) **verify** the applicability of reference intervals they adopt from other sources

3) how laboratories can **establish** reference intervals.

Decision Limits

If one were to perform a conventional reference interval study on 120 apparently healthy American adult men, one would find that the central 95% of values would range from roughly 139 mg/dL to 273 mg/dL.[1] At one time, that was indeed the "reference interval" for total cholesterol.

However, based on many studies, culminating in the publication of the NCEP ATP guidelines, [2] we learned that the "typical" cholesterol was not necessarily a healthy cholesterol. Indeed, we learned that apparently healthy individuals whose cholesterols were above 200 mg/dL were at increased risk of coronary artery disease and that, by lowering their cholesterol levels, their risk could be lowered as well.

Thus, the current reference interval, or better yet, decision limit, for cholesterol is 200 mg/dL. This is the value most laboratories use as their "upper limit of the reference interval". But how does a laboratory know that its cholesterol values are accurate (that is, that they match the values that a certified laboratory would get by the reference method on this same sample)? And that is what they must want – if their values are low by 10 mg/dL or high by 10 mg/dL (just 5%), a large percentage of their clinical samples will be miscategorized, since many samples have concentrations near the decision limit (50^{th} percentile is 202 mg/dL).[1]

Typically, laboratories assess their performance by daily internal quality control samples (for day-to-day precision) and by periodic external quality control (proficiency testing) samples (for comparability to other laboratories). The latter represent samples measured as patient samples but reported to a central agency and compared, typically, to other laboratories using the same methods. Ideally, these results would be compared to "truth", but, unfortunately, external quality control samples are rarely native human serum and, as a result, exhibit "matrix effects". In brief, they act differently from human serum, and different methods may yield different values on these external quality control samples even when they yield comparable results on native human serum samples.[3] Thus, agencies grading external quality control samples typically resort to peer group grading to overcome this problem.

As a result, for analytes like cholesterol, individual laboratories may know that their results agree with their peers, but that does not, in and of itself, insure that their results are accurate. The only way to achieve that level of confidence is to participate in surveys where "matrix-neutral" material (a material that acts exactly like human serum, as, for example, fresh frozen human serum) is used and where results are compared to the reference method.

Fortunately, there are several such surveys in existence, and more will probably follow. [4,5] Analytes where such surveys are particularly important are listed in Table 1. What these analytes have in common is that clinicians use national (or even international) guidelines rather than conventional reference intervals (a defined central percentage (e.g., 95%) of an apparently healthy population). In these cases, one cannot compensate for a biased method by having a comparably biased reference interval.

Clinicians assume that all methods are accurate. But the data shows otherwise, as exemplified from data from two recent proficiency surveys from the College of American Pathologists (CAP). The CAP GH2 Survey uses fresh whole blood collected from diabetic patients to assess glycated hemoglobin (A1c) performance. Each sample is tested by the reference method to establish the true value. As shown in Figure 1, data from a recent glycated hemoglobin survey showed that for one sample, whose true value was 8.40%, more than 50% of reported values were less than 7.9% for one widely-used peer group (method) and more than 50% of reported values were greater than 8.7% from a second widely-used peer group.[6] Clinicians using these laboratories, if not the laboratories themselves, would be very surprised indeed to know that differences of more than 1% on the same sample are "acceptable" for accreditation purposes.[7]

Similarly, to the extent that clinicians are using the recently published clinical practice guidelines for neonatal hyperbilirubinemia,[8] accuracy for this analyte is equally important. The CAP NB Survey includes some samples that are "matrix-neutral" and on which the true value is established by the reference method. On the NB-A 2008 Survey, of 1959 laboratories participating, four methods accounted for 77% of the values on the matrix-neutral sample; with a true value of 21.7 mg/dL, the means of these four methods ranged from 22.2 to 24.1 mg/dL Thus, for the commonly used platform whose mean value was 24.1 (11% above the true value), roughly 50% of the reported values were more than 10% above the true value.[9] Aware of these problems, laboratories can (and should) put pressure on manufacturers to address the issues or change to better (more accurate) methods.

Obtaining Samples from Reference Individuals

Whether one wants to verify or establish reference intervals, one needs to collect samples from reference individuals. Ideally, the individuals should be selected from a reference population using specific criteria, including exclusion criteria (which might include recent surgery, tobacco use, over-the-counter medications, etc.) and partitioning criteria (which might include age, gender, race, etc.). [10] Note that a condition such as pregnancy might be an exclusion criterion for one study but a partitioning criterion for another study.

It is important that laboratories obtain written informed consent from each reference individual, not only to collect samples but also to inform the participants the data will be used to calculate reference intervals.

Once reference individuals have been identified, it is important that pre-analytical factors be addressed carefully before samples are collected and analyzed. Subject preparation (e.g., fasting, physical activity, medication), sample collection (e.g., time of day, tourniquet time, tube type), and sample handling (e.g., clotting time, centrifugation, storage) should all be standardized. [11]

With high-quality samples from well-qualified reference individuals in hand, one can proceed with the analytical measurements, which themselves need to be standardized, especially when data from more than one site may be aggregated in order to achieve adequate numbers of samples. Once the measurements have been completed, the data analysis can begin.

Verifying Reference Intervals

When laboratories adopt new methods, they typically validate them extensively in terms of their analytic performance, including imprecision, dynamic range, analytic sensitivity, analytic specificity, and correlation with their current method. When it comes to defining a reference interval, most laboratories, intimidated by the formidable task of doing a formal reference interval study (see next section), simply adopt the reference interval suggested by the manufacturer of their reagents. Although they may perform a cursory review of the intervals, it rarely involves much in the way of formal statistics or actual patient materials.

It turns out that, by collecting samples from just 20 reference individuals, a laboratory can ***verify*** that a manufacturer's reference intervals can safely be adopted for its population.[12] This is in marked contrast to the 120 reference individuals typically recommended to ***establish*** reference intervals (see next section).

According to current guidelines, [12] if no more than 2 of 20 samples fall outside the suggested reference interval, one can infer that the proposed reference interval can be adopted. If 3 or more fall outside the limits, then one may have a problem, and one would be required to collect more data.

Most laboratories should be capable of collecting samples from 20 reference individuals, so it is surprising how few laboratories actually even attempt to verify their reference intervals.

If more laboratories did so, an interesting opportunity presents itself. If several laboratories, using the same methods, each collected 20 samples, one might be able to pool that data to obtain the 120 samples needed to ***establish*** reference intervals. Indeed, one might even have sufficient numbers of samples to evaluate gender differences, age differences, racial differences, etc.

This is the idea behind multi-center trials to establish reference intervals. With careful attention to exclusion and partitioning criteria, pre-analytic factors, and analytic methods, one should be able to pool data from many individual sites, thereby greatly reducing the number of samples each site needs to collect. Ideally, quality control samples would be embedded in the trials to insure that each participating laboratory was performing the measurements comparably.

Among its Proficiency Test offerings, the CAP has a "Reference Range Service",[13] which provides something very much along these lines. Table 2 summarizes some of the data we received on two analytes as a result of our participation, which involved submitting observations from just 20 reference individuals. For TSH in particular, we had two reference individuals whose values were outside the manufacturer's reference interval (which, as noted earlier, does not invalidate the manufacturer's reference interval for our use); when our data was pooled with data from other laboratories using the same method, these two individuals' values were, in fact, identified as "not normal".

Even without pooling data, though, it may be possible to generate reliable reference intervals from fewer than 120 points by virtue of using such modern statistical techniques as the robust method. By using robust measures of location (center) and scale (spread), the method does not make any assumptions about the underlying distribution of the data. The method involves somewhat more expertise in computing, but the iterations required can be done in typical spreadsheet programs like Microsoft Excel. The exact number of observations required varies. A more detailed discussion of the method is beyond the scope of this article, but details are available elsewhere. [14]

Establishing Reference Intervals

For a variety of reasons, the method most often recommended for establishing reference intervals is the non-parametric approach.[15] First, the nature of the underlying distribution of the data does not matter. Second, no statistical expertise is required; one simply puts the values obtained from reference individuals in rank order by concentration (rank 1 is the lowest, rank 2 is the next lowest, etc.), and the central n% becomes the reference interval. In addition, the confidence limits of the endpoints of the interval can similarly just be taken from the data points themselves.

It turns out that 120 observations provide enough data to determine both the central 95% of the distribution and the 90% confidence limits on both endpoints. That is, with 120 observations, rank 3 is the 2.5^{th} percentile; rank 118 is the 97.5^{th} percentile; ranks 1 and 7 define the 90%

confidence interval of the 2.5th percentile; and ranks 114 and 120 define the 90% confidence interval of the 97.5th percentile.[15]

For each individual partition (e.g., for gender, for age brackets, for race), one needs 120 observations. (Of course, to prove that one does not need separate reference intervals for different partitions, one must first collect the data for each partition and show that there any differences are not significant.)

Collecting such data is no mean feat, but the absence of such data may have serious consequences, as was reflected in a recent study involving CK. [16]. The authors did an exemplary job establishing reference intervals by the non-parametric technique. All together, they collected data on 1444 adult reference individuals. Exclusion criteria included cholesterol-lowering drug therapy and strenuous exercise within three days of sample collection. The authors were able to partition their data by gender and by ancestry as they had more than 120 observations in each category. As shown in Table 3, the authors concluded that the upper limit of the reference interval (97.5th percentile) ranged from 201 to 841 for their six partitions, roughly 1.6-fold to 4.6-fold greater than the manufacturer's suggested upper limit. Put differently, anywhere from 8% to 62% of their reference individuals would be categorized as abnormally high (>97.5th percentile) using the manufacturer's reference interval.

Admittedly, few laboratories could do this study on their own, but how many laboratories could have realized there was a problem simply by measuring CK levels on just 20 reference individuals to verify the manufacturer's claim? In my view, we really have no excuse for not verifying, even on a recurring basis, the adequacy of our reference intervals. As noted by the authors of the CK paper, it is humbling to note how many individuals may have been deprived of cholesterol-lowering medications because of poorly-established reference intervals.

Conclusions

Many, if not most, of the methods we use in clinical laboratories are more than adequate from an analytical perspective. In contrast, the reference intervals that accompany the test results on laboratory reports deserve more scrutiny. For those tests where accuracy is important, laboratories should participate in surveys that go beyond peer-group assessment and establish adequate accuracy. For other tests, laboratories should at least verify, with 20 reference individuals, the appropriateness of their current reference intervals. Establishing, as opposed to verifying, reference intervals is clearly more difficult because of the daunting numbers of reference individuals required. But the ability to pool data from several laboratories using the same method and the availability of new statistical techniques may ease the burden considerably.

Acknowledgement

The author thanks all the members of the Working Group on Reference Intervals from CLSI, with whom he has been collaborating for the past two years on updating C28 and from whom he has learned much of the material appearing in this article. In particular, the author calls attention to Drs. James Boyd, Ferruccio Ceriotti, and Paul Horn, whose statistical knowledge and patience in sharing it are truly noteworthy.

REFERENCES

1. National Heart Lung and Blood Institute, Distribution of cholesterol values. (last accessed 7/10/2008)
http://www.nhlbi.nih.gov/guidelines/cholesterol/atp3full.pdf

2. Report of the National Cholesterol Education Program Expert Panel on detection, evaluation, and treatment of high blood cholesterol in adults. Arch Intern Med 1988;148:36-69.

3. Miller WG, Myers GL, Rej R. Why commutability matters. Clin Chem 2006;52:553-554.

4. Canadian External Quality Assessment Laboratory (CEQAL), Certification--Total Cholesterol (last accessed 7/10/2008)
http://www.ceqal.com/services.php

5. College of American Pathologists, Accuracy-Based Lipid Survey (last accessed 7/10/2008)
http://www.cap.org/apps/docs/proficiency_testing/surveys_catalog/chapter_10.pdf

6. College of American Pathologists, A1c Grading 2006 (last accessed 7/10/2008)
http://www.cap.org/apps/docs/committees/chemistry/2006GH2BDiscussion.pdf

7. Holmes EW, Erbahin C, Augustine GJ, Charnogursky GA, et al. Analytic bias among certified methods for the measurement of hemoglobin A1c: a cause for concern? Am J Clin Pathol 2008;128:540-547.

8. American Academy of Pediatrics, Subcommittee on Hyperbilirubinemia. Clinical practice guideline: management of hyperbilirubinemia in the newborn infant 35 or more weeks of gestation. Pediatrics 2004;114:297-316

9. College of American Pathologists, Neonatal Bilirubin Accuracy (last accessed 7/10/2008)
http://www.cap.org/apps/docs/committees/chemistry/Educational_Grading_Discussion.pdf

10. PetitClerk C, Solberg HE. Approved recommendation on the theory of reference values. Part 2. Selection of individuals for the production of reference values. J Clin Chem Clin Biochem 1987;25:639-644.

11. Solberg HE, Petitclerc C. Approved recommendation (1988) on the theory of reference values. Part 3. Preparation of individuals and collection of specimens for the production of reference values. Clin Chem Acta. 1988;177:S1-S12.

12. Clinical Laboratory and Standards Institute. "How to Define and Determine Reference Intervals in the Clinical Laboratory; Approved Guideline—Second Edition" CLSI document C28-A2. Wayne, PA: 2000;24.

13. College of American Pathologists, Reference Range Service (last accessed 7/10/2008)

http://www.cap.org/apps/docs/proficiency_testing/surveys_catalog/chapter_11.pdf

14. Horn PS, Lesce AJ. Reference Intervals: A User's Guide. Washington, DC: 2005;47-57.

15. Clinical Laboratory and Standards Institute. "How to Define and Determine Reference Intervals in the Clinical Laboratory; Approved Guideline—Second Edition" CLSI document C28-A2. Wayne, PA: 2000;13-22.

16. Brewster LM, Miaruhu G, Sturk A, van Montfrans GA. Distribution of creatine kinase in the general population: implications for statin therapy. Am Heart J 2007;154:655-61.

Table 1. Analytes for Which Accuracy is Particularly Important

Cholesterol (and its Fractions)	Cardiac Risk Assessment
Creatinine	Calculation of Estimated GFR
Glucose	Diagnosis of Diabetes
Glycated Hemoglobin (A1c)	Management of Diabetes
Neonatal Bilirubin	Management of Neonatal Hyperbilirubinema

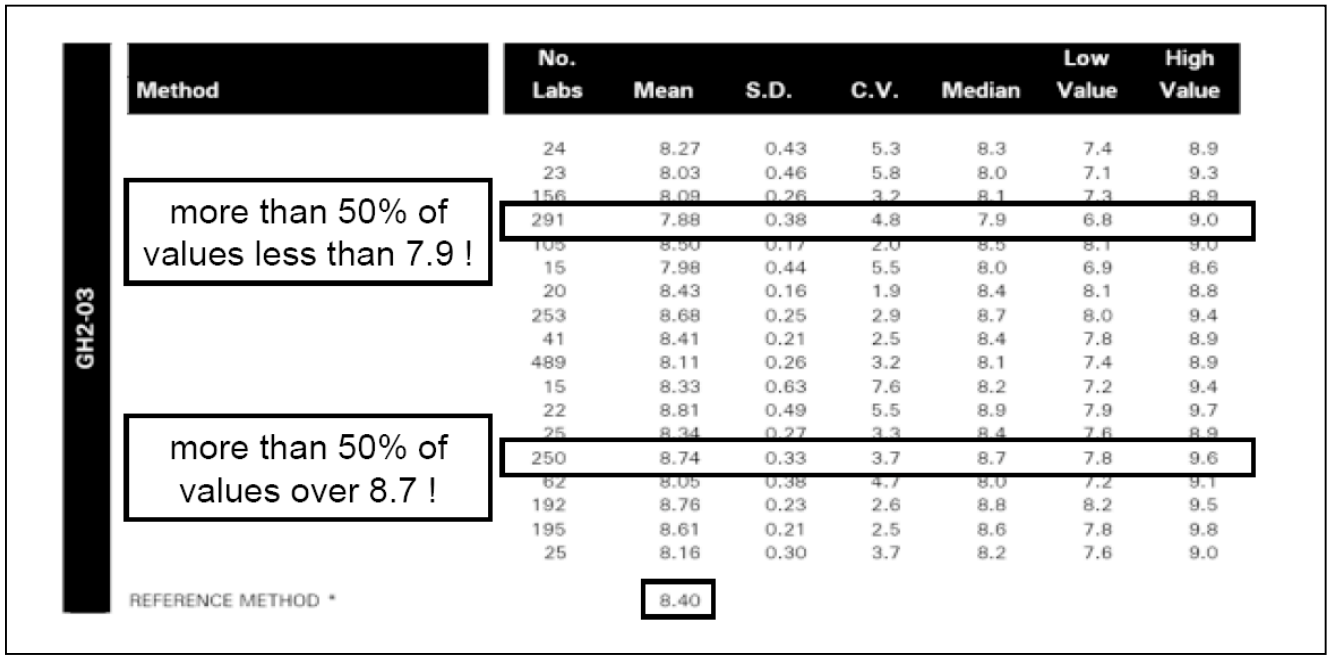
Table 2. CAP Reference Range Service Data

Analyte	Source of Reference Individuals	Number of Reference Individuals	Minimum Value	Mean Value	Maximum Value	2.5th percentile	97.5th percentile	Manufacturer's Suggested Reference Interval
Calcium	Author's Laboratory	20	8.60	9.35	10.10	__	__	8.4 – 10.2 mg/dL
Calcium	All Roche MODULAR	335	8.00	9.44	10.90	8.45	10.29	8.4 – 10.2 mg/dL
TSH	Author's Laboratory	18	0.940	2.169	4.370	__	__	0.27 – 4.2 uU/mL
TSH	All Roche MODULAR	267	0.090	2.032	5.360	0.59	4.38	0.27 – 4.2 uU/mL

Table 3. CK Reference Interval Data [16]

Gender	Ancestry	Number of Reference Individuals	Calculated 97.5th Percentile (IU/L)	Manufacturer's Suggested 97.5th Percentile (IU/L)	Ratio Observed/Suggested Upper Limit	Reference Individuals Whose Values Exceeded Manufacturer's 97.5th Percentile
Women	White	252	201	140	1.4	8%
Women	South Asian	147	313	140	2.2	16%
Women	Black	387	414	140	3.0	42%
Men	White	251	322	174	1.9	17%
Men	South Asian	123	641	174	3.7	32%
Men	Black	183	801	174	4.6	62%

Figure 1. Section of report from the author's laboratory detailing performance of various methods on CAP Survey Sample GH2-03 from the A-mailing in 2006. Each row corresponds to a specific method. For each method, the number of laboratories submitting data, as well as the mean value, standard deviation, minimum and maximum values are indicated. The true value, 8.40, is indicated at the bottom of the figure.



Method	No. Labs	Mean	S.D.	C.V.	Median	Low Value	High Value
	24	8.27	0.43	5.3	8.3	7.4	8.9
	23	8.03	0.46	5.8	8.0	7.1	9.3
	156	8.09	0.26	3.2	8.1	7.3	8.9
	291	7.88	0.38	4.8	7.9	6.8	9.0
	105	8.50	0.17	2.0	8.5	8.1	9.0
	15	7.98	0.44	5.5	8.0	6.9	8.6
	20	8.43	0.16	1.9	8.4	8.1	8.8
	253	8.68	0.25	2.9	8.7	8.0	9.4
	41	8.41	0.21	2.5	8.4	7.8	8.9
	489	8.11	0.26	3.2	8.1	7.4	8.9
	15	8.33	0.63	7.6	8.2	7.2	9.4
	22	8.81	0.49	5.5	8.9	7.9	9.7
	25	8.34	0.27	3.3	8.4	7.6	8.9
	250	8.74	0.33	3.7	8.7	7.8	9.6
	62	8.05	0.38	4.7	8.0	7.2	9.1
	192	8.76	0.23	2.6	8.8	8.2	9.5
	195	8.61	0.21	2.5	8.6	7.8	9.8
	25	8.16	0.30	3.7	8.2	7.6	9.0
REFERENCE METHOD *		8.40					