VERIFICATION OF REFERENCE INTERVALS

GIAN LUCA SALVAGNO, MD, PhD
University of Verona, Italy

THE LABORATORY IN DIAGNOSIS

The use of laboratory methods in diagnosis is, for the most part, a growth of recent years. While it is true that the laboratory examination of the urine is a procedure which has been practiced by several generations of physicians, the more complicated hematologic, bacteriologic and serologic examinations have come into general use only recently.

In the average medical school instruction in clinical microscopy is seldom given in a hospital, and the subject is often taught by the professor of pathology, who, most likely, has excellent laboratory training, but little clinical experience. In comparatively few schools has the student the opportunity of applying the laboratory methods at the bedside, for in the majority of American schools bedside instruction worthy of the name does not exist.

JAMA 100 YEARS AGO

FEBRUARY 9, 1907
THE LABORATORY IN DIAGNOSIS

The effect on the student of the conditions of instruction now existing is often to give him a false idea of the value of laboratory work, and this idea may or may not be dispelled during his final years. If he happens to come under the guidance of a clinical teacher who has passed through the laboratory as part of his training, and who thoroughly realizes the value as well as the limitations of laboratory work, he may graduate with a just conception of the subject. Very often, however, his teacher is a man who, at the time laboratory methods came into vogue, was too busy to take the time to master laboratory procedures, or too old to grasp their significance. In such a case the student often leaves the walls of his alma mater with a false conception of the use of the laboratory in diagnosis.

JAMA 100 YEARS AGO
FEBRUARY 9, 1907

THE LABORATORY IN DIAGNOSIS

A large proportion of the profession in America to-day consists of men who graduated before laboratory procedures, other than the ordinary routine urinary examination, were taught in the schools. Many of them are progressive men, anxious to learn new methods, enthusiastic, hard working, but too often, of necessity, only half trained in laboratory work. It has been impossible from lack of time and opportunity for many of these men to become expert themselves, and they frequently employ recent graduates, or even students, as laboratory assistants who, depending on their training, have good or erroneous ideas of the value of laboratory methods.
THE LABORATORY IN DIAGNOSIS

There is danger that occasionally laboratory findings may be allowed to take the place of the keen thinking and the educated senses which our professional ancestors used to such good purpose. It lies in the hands of the clinical teachers of this country to overcome the obvious tendencies in this direction. Every practitioner should preferably do his own laboratory work, and, if this is impossible, he should at least know how to properly interpret and correlate the findings of others. Too many errors are made when the physician, lacking in knowledge of the significance of laboratory findings, relies on the reports of the pathologist who lacks in clinical experience, and who bases his returns on the examination of abstract specimens from a concrete patient of whose history he is usually almost completely ignorant.

JAMA 100 YEARS AGO

FEBRUARY 9, 1907
The aim of a Clinical Laboratory Test is to determine the concentration or activity of a diagnostically relevant analyte in a body fluid in order to provide information on the clinical situation of a patient.

Perhaps most notable was a Pulitzer-prize winning series that was broadcasted by a Washington, D.C., television station and followed by a compelling article in the Feb. 2, 1987, edition of The Wall Street Journal that highlighted scandals involving several commercial laboratories which inaccurately analyzed Pap smears. Several women died from undetected cervical cancer because of these inaccurately analyzed tests. Of the ten of thousands of laboratories run by physicians at that time, only a handful performed cytology testing, but the public demanded that Congress take action in response to these preventable deaths. Congress reacted by holding oversight hearings which ultimately guided the drafting of CLIA ‘88, which later was signed into public law (PL 100-275) on Oct. 31, 1988. The law is far-reaching and regulates POL testing on a national scale.

In Schneider’s 1960s paper entitled “Some thoughts on normal, or standard, values in clinical medicine”, he states: “... practical medicine is basically founded on comparison. If medicine is to be scientific, we must not only understand the structural, functional and chemical relations operating in individuals, but we must also understand the basis of our comparisons...”

According to Horn and Pesce: “the reference interval is the most widely used medical decision-making tool”, even if its practical usefulness is lower than its theoretical power. This is due to the fact that obtaining a “good” reference interval is a very demanding activity, in terms of time, money and knowledge.
It is hard to underestimate the importance of clinical laboratory test results. Nearly 80% of physicians’ medical decisions are based on information provided by laboratory reports. A test result by itself is of little value unless it is reported with the appropriate information for its interpretation. Typically, this information is provided in the form of a reference interval (RI) or medical decision limit. An RI as defined by Ceriotti is an interval that, when applied to the population serviced by the laboratory, correctly includes most of the subjects with characteristics similar to the reference group and excludes the others. No RI is completely “right” or “wrong.” The majority of RIs in use today refer to the central 95% of the reference population of subjects. By definition, 5% of all results from “healthy” people will fall outside of the reported RI and, as such, will be flagged as being “abnormal.”

There are many problems associated with the calculation of RIs. The latest edition of the Clinical and Laboratory Standards Institute–approved guideline, “Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory,” recognizes the difficulties and controversies surrounding the establishment of RIs and the verification process. "...the working group recognizes the reality that, in practice, very few laboratories perform their own reference interval studies." "...Instead of performing a new reference interval study, laboratories and manufacturers refer to studies done many decades ago, when both the methods and the population were very different."
Recruiting a valid group of reference subjects and obtaining informed consent in today’s environment is costly, time-intensive, and virtually an impossible task for most laboratories. The challenge is further magnified in establishing RIs for different age groups (eg, pediatric patients and geriatric patients), uncommon sample types (eg, cerebrospinal fluid and aspirations), timed collections, challenge tests, and serial measurements.

A laboratory can elect to “transfer” the RIs that were in use with an older method (or from another laboratory) to a new method. To do this, the laboratory must first demonstrate that the 2 methods produce comparable results. It is well known that analytic systems drift over time, and there is no guarantee that the method of today is producing results that are comparable to those that were produced at the time of the original RI study. This technique is the main reason why many laboratories today are using RIs that were established decades ago and are out-of-date.

An alternative approach for establishing RIs is to do an indirect so-called a posteriori study of the patient data already collected and stored in the laboratory database. This is appealing because the data are readily available and will result in time and cost savings. A number of publications discuss this approach. Most of these studies were able to report clinically relevant and meaningful RIs. All of them used various sophisticated filters to exclude results from “unhealthy” subjects, and some used data from hospital laboratories and some from outpatient care settings or noninstitutionalized population study databases. Most of these studies used complex statistical algorithms to derive the final interval. However, current guidelines do not endorse these methods as a primary approach for establishing RIs, mainly out of concern for the fact that most of the data may not come from reference or healthy subjects. This position may be justified for test results collected from hospitalized patients but is questionable when considering a very large number of results that have been collected in outpatient settings.
But what is meant by a “good” reference interval? It is an interval that, when applied to the population serviced by the laboratory, correctly includes most of the subjects with characteristics similar to the reference group and excludes the others. Usually, we consider “health-related” reference intervals to mean that the subjects with values within the interval have a lower probability of being affected by a specific disease, while those outside the interval have a higher statistical probability of having the disease or, at least, that the observed value is not normal for a healthy person. The percentage of unhealthy people included in the reference interval or, vice versa, the percentage of healthy subjects outside the interval, defines the “goodness” of the interval.

The factors responsible for this misclassification were already recognised by Schneider who identified the three contributing causes namely, intraindividual, interindividual and analytical variabilidad. Intra- and interindividual variabilidad are inextricably bound. Nevertheless, the relative sizes of these two sources of variation can substantially affect the usefulness of the reference interval as a guide to the status of an individual.

One way to improve the usefulness of reference intervals is to reduce the interindividual variabilidad by partitioning the intervals as much as possible. Stratification by age and gender is the minimum pre-requisite. But other ways include by race, ethnic group, body mass index or nutritional habits (e.g. vegetarians). Herein is the problem of the selection of an appropriate number of reference subjects, properly screened to exclude relevant pathologies, and subdivided by gender, age, race, ethnicity and lifestyle.
Particularly, the source of reference intervals may be quite different and heterogeneous (manufacturer’s recommendation, textbooks and literature, internal studies by healthy volunteers testing, other laboratories, undefined) (11, 12). To try to minimize these differences among laboratories, results are often expressed as multiples of the upper reference limit (URL). This approach has been used in international guidelines (13) and some authors in the past welcomed it as “physician liberation units” (14). Recently, it was proposed a more sophisticated version of that approach called “quantity quotient reporting” (15). Studies have demonstrated, however, that the expression of results as URL multiples paradoxically increases the scatter of results due to both inter-method differences and the different laboratory reference limits (16).

Cautions needed in the adoption of traceable reference intervals

Boyd has cautioned against the uncritical use of “common” reference intervals, stressing the need to fully document the possible presence of differences in tests results across populations due to biological or environmental factors (22). Until it can be reasonably shown that no differences exist for a given test between the population served by individual laboratories and that used in defining traceable reference intervals, their adoption should indeed be discouraged.

The selection of individuals for the production of reference values needs much attention. Selection of “normal reference” subjects is from many years a key point in all recommendations for the development of reference intervals (23).
Review Article

Reference intervals: the way forward

<table>
<thead>
<tr>
<th>Table 3 Differences between reference intervals and decision limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference intervals</td>
</tr>
<tr>
<td>Conditions influencing them</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Information gathered</td>
</tr>
<tr>
<td>Being or not being part of the reference population</td>
</tr>
<tr>
<td>Statistics</td>
</tr>
</tbody>
</table>

ROC, receiver-operating characteristic
Summary of analytical and biological requirements in the adoption of traceable reference intervals

<table>
<thead>
<tr>
<th>Responsibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>IFCC, JCTLM, National Metrology Institutes</td>
</tr>
<tr>
<td>IVD manufacturers</td>
</tr>
<tr>
<td>Clinical laboratories</td>
</tr>
<tr>
<td>Clinical laboratories, EQAS organizers</td>
</tr>
</tbody>
</table>

Analytical requisite:
- Existence of a reference system
- Availability of commercial assays producing traceable results
- Correct implementation of traceable assays in the clinical setting
- Control of the performance of commercial assays within stated limits of uncertainty
- Compatile pre-analytical conditions

Biological requisite:
- Accurate definition of traceable reference intervals, providing information on the influence of biological and environmental factors
- Validation of the applicability of the traceable reference intervals to the laboratory's own population

EQAS, external quality assessment scheme; IVD, in vitro diagnostic.

Reference intervals: the way forward

<table>
<thead>
<tr>
<th>Category</th>
<th>Pre-requisite</th>
<th>Responsibility for implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytical</td>
<td>Existence of a reference measurement system (traceable-based)</td>
<td>IFCC, JCTLM, national metrological institutes</td>
</tr>
<tr>
<td>Analytical</td>
<td>Existence of traceable routine methods</td>
<td>Manufacturers</td>
</tr>
<tr>
<td>Analytical</td>
<td>Correct implementation of methods in clinical laboratories</td>
<td>Clinical laboratories</td>
</tr>
<tr>
<td>Analytical</td>
<td>Control of the performance of routine methods to keep them within stated limits for uncertainty</td>
<td>Clinical laboratories, EQAS organizers</td>
</tr>
<tr>
<td>Clinical</td>
<td>Accurate definition of reference intervals, providing information on the influence of biological and environmental factors</td>
<td>Joint effort between IFCC, manufacturers and clinical laboratories</td>
</tr>
<tr>
<td>Clinical</td>
<td>Comparable preanalytical phase</td>
<td>Clinical laboratories</td>
</tr>
<tr>
<td>Clinical</td>
<td>Validation of applicability of common reference intervals to the laboratory's own population</td>
<td>Clinical laboratories</td>
</tr>
<tr>
<td>Clinical</td>
<td>Adoption of common reference intervals</td>
<td>Clinical laboratories</td>
</tr>
</tbody>
</table>
15th EFLM Continuous Postgraduate Course in Clinical Chemistry and Laboratory Medicine
October 24-25, 2015 – Zagreb, Croatia

Reference intervals as a tool for total quality management

Special Issue: Quality in laboratory diagnostics: from theory to practice

Clinical Chemistry Section, Department of Life and Reproduction Sciences, University of Verona, Italy
*Corresponding author: guiluca.salvagno@univr.it

Neonatal period (1–6 months)
Infancy (6 months–3 years)
Childhood (3–5 years)
Pre-pubertal (6–11 years)
Puberty (11–18 years)
Adulthood (18–45 years)
Pre-menopausal
Post-menopausal
Maturity (45–65 years)
Old age (> 65 years)

Figure 4. General forms of kurtosis.

Figure 5. Interval of reference: Gaussian curve.
**Reference intervals as a tool for total quality management**

Special Issue: Quality in laboratory diagnostics: from theory to practice


**Figure 2. Non-Gaussian distribution (non-normal).**

**Figure 3. Tendency to asymmetry (skewness).**

**Figure 6. Interpercentile intervals: nonparametric distribution.**
Non-parametric statistical techniques are used to analyze the data not having a specific type of probability distribution. In general, when observing non-Gaussian distributions (non-normal) (Figure 2a-b), their description is assigned to other indices such as median, percentiles classes, and more others. Moreover, in this second category of data distribution, other methods become more useful, including the so-called and important ones “bootstrap methods”. Sometimes non-Gaussian distributions can be normalized via appropriate processing techniques (12). This is the general case of distributions obtained from experimental data, for which the assumption of normality is always verified. In constructing a reference range from individual data, often the difficulty of achieving a perfect Gaussian distribution is apparent. Even after sampling the data from a population which is presumed to be normally distributed, it is often necessary to take some approximations of the data to comply with the assumption of normality. In this regard, several statistical tests have been put in place, which comprises the distribution of experimental data with a hypothetical Gaussian distribution (13–15). These methods are called mathematical/statistical goodness-of-fit test tests. Among them, the most known and used is the Kolmogorov-Smirnov, although its real discriminant power is questioned by some researchers, especially when the parameters of the distribution are estimated based on data rather than being specified a priori. Accordingly, other tests have been proposed that are best suited for this purpose, among them the test of Shapiro-Wilk (for distribution of samples greater than 2,000 subjects it should be replaced by the test for normality of Stephen) and the test of D’Agostino-Pearson. None of these tests can however indicate the type of non-normality observed in the case where the distribution is showing tendency to asymmetry (skewness) and kurtosis at both (Figure 3).
"Which is the best choice?"

Or......
# 15th EFLM Continuous Postgraduate Course in Clinical Chemistry and Laboratory Medicine

October 24-25, 2015 – Zagreb, Croatia

## Table of Contents

1. Scope
2. Introduction
3. Standard Procedures
4. Terminology
   4.1 A Note on Terminology
   4.2 Definitions
   4.3 Abbreviations/Acronyms
   4.4 Clarifications
5. Use of System International d’Unités (SI Units)
6. Protocol Outline for Establishing Reference Values
   6.1 New Analyte or Analytical Method
   6.2 Multicenter Reference Interval Studies
   6.3 Previously Measured Analyte
7. Selection of Reference Individuals
   7.1 Introduction
   7.2 Exclusion and Partitioning
   7.3 Sample Questionnaire
   7.4 Selection of Reference Individuals
8. Preanalytical and Analytical Considerations
   8.1 Subject Preparation
   8.2 Specimen Type, Collection, Handling, and Storage
   8.3 Analytical Method Characteristics
9. Analysis of Reference Values
   9.1 Minimum Number of Reference Values
   9.2 Treatment of Outlying Observations
   9.3 Partitioning of Reference Values
   9.4 Examples
9.5 Confidence Intervals for Reference Limits
10. Transference
    10.1 Transference: Comparability of the Analytical System
    10.2 Transference: Comparability of the Test Subject Population
11. Validation
    11.1 Validation: Subjective
    11.2 Validation: Using Small Numbers of Reference Individuals
    11.3 Validation: Using Larger Numbers of Reference Individuals
12. Presentation of Reference Values
    12.1 Introduction
    12.2 Laboratory Presentation
    12.3 Mantlefied Presentation
13. Other Issues
    13.1 Qualitative Analysis
    13.2 Therapeutic Drug Levels
    13.3 Time-Dependent Challenge Tests
    13.4 Individual Variation
    13.5 "Critical Values"
14. Summary

---

**EP28-A3C**
### EXCLUSION CRITERIA

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol consumption</td>
<td>Illness, recent</td>
</tr>
<tr>
<td>Blood donor</td>
<td>Lactation</td>
</tr>
<tr>
<td>Blood pressure, abnormal</td>
<td>Obesity</td>
</tr>
<tr>
<td>Drug abuse</td>
<td>Occupation</td>
</tr>
<tr>
<td>Drugs, prescription</td>
<td>Oral contraceptives</td>
</tr>
<tr>
<td>Drugs, over the counter</td>
<td>Pregnancy</td>
</tr>
<tr>
<td>Environment</td>
<td>Surgery, recent</td>
</tr>
<tr>
<td>Fasting or nonfasting</td>
<td>Tobacco use</td>
</tr>
<tr>
<td>Genetic factors</td>
<td>Transfusion, recent</td>
</tr>
<tr>
<td>Hospitalization, current/recent</td>
<td>Vitamin abuse</td>
</tr>
</tbody>
</table>

### EXEMPLE OF PARTITIONING FACTORS

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Geographic location</td>
</tr>
<tr>
<td>Blood group</td>
<td>Posture when sampled</td>
</tr>
<tr>
<td>Circadian variation</td>
<td>Race</td>
</tr>
<tr>
<td>Diet</td>
<td>Sex</td>
</tr>
<tr>
<td>Ethnic background</td>
<td>Stage of menstrual cycle</td>
</tr>
<tr>
<td>Exercise</td>
<td>Stage of pregnancy</td>
</tr>
<tr>
<td>Fasting or nonfasting</td>
<td>Tobacco use</td>
</tr>
</tbody>
</table>
### Pre-analytical Factors for Considerations

<table>
<thead>
<tr>
<th>Subject Preparation</th>
<th>Specimen Collection</th>
<th>Specimen Handling</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Prior diet</td>
<td>• Environmental conditions during collection</td>
<td></td>
</tr>
<tr>
<td>• Fasting vs nonfasting</td>
<td>• Time</td>
<td></td>
</tr>
<tr>
<td>• Abstinence from pharmacologic agents</td>
<td>• Body posture</td>
<td></td>
</tr>
<tr>
<td>• Drug regimen</td>
<td>• Specimen type</td>
<td></td>
</tr>
<tr>
<td>• Sampling time in relation to biological rhythms</td>
<td>• Collection site</td>
<td></td>
</tr>
<tr>
<td>• Physical activity</td>
<td>• Site preparation</td>
<td></td>
</tr>
<tr>
<td>• Rest period before collection</td>
<td>• Blood flow</td>
<td></td>
</tr>
<tr>
<td>• Stress</td>
<td>• Equipment</td>
<td></td>
</tr>
</tbody>
</table>

**What’s next?**
An appraisal of statistical procedures used in derivation of reference intervals
Kiyoshi Ichihara* and James C. Boyd† on behalf of the IFCC Committee on Reference Intervals and Decision Limits (C-RIDL)

Common reference intervals for aspartate aminotransferase (AST), alanine aminotransferase (ALT) and γ-glutamyl transferase (GGT) in serum: results from an IFCC multicenter study
Ferruccio Cirrieti*†, Joseph Hunter*, Josep Queralt*, Nan Zou+, Yosio Urauchi*, Rauerng Chen*, James C. Boyd† and Munre Puntisguth†† on behalf of the IFCC Committee on Reference Intervals and Decision Limits (C-RIDL) and the Committee on Reference Systems for Enzymes

The evolution of the reference value concept
Ralph Gräsbeck*
Minerva Foundation Institute for Medical Research, Biomedical Helsinki, Helsinki, Finland

Editorial
Reference values and the journal: why the past is now present
Mario Plebani and Giuseppe Lippi
The IFCC recommendation on estimation of reference intervals, The RefVal Program
Helge Erik Solberg*
Department of Clinical Chemistry, Rikshospitalet, Oslo, Norway
Helge Erik Solberg*
Department of Clinical Chemistry, Rikshospitalet, Oslo, Norway
Take a home message

- Reference intervals “is an interval that, when applied to the population serviced by the laboratory correctly includes most of the subjects with characteristics similar to the reference group and excludes the others” (Ceriotti F. Prerequisites for use of common reference intervals. Clin Biochem Rev. 2007;28:115-121).

- It has been recommended that an Reference Intervals be established by selecting a statistically enough group of healthy people.

- Clinical laboratory professionals, who want to provide a high quality report, should comply with the requests of quality standards like ISO 15189:2012, which specifically asks for a periodical review of biological reference intervals.