Review

Gerard Siest* and Joseph Henny, with the collaboration of Ralph Gräsbeck, Peter Wilding, Claude Petitclerc, Josep M. Queraltó and Peter Hyltoft Petersen

The theory of reference values: an unfinished symphony

Abstract

The history of the theory of reference values can be written as an unfinished symphony. The first movement, allegro con fuoco, played from 1960 to 1980: a mix of themes devoted to the study of biological variability (intra-, interindividual, short- and long-term), preanalytical conditions, standardization of analytical methods, quality control, statistical tools for deriving reference limits, all of them complex variations developed on a central melody: the new concept of reference values that would replace the notion of normality whose definition was unclear. Additional contributions (multivariate reference values, use of reference limits from broad sets of patient data, drug interferences) conclude the movement on the variability of laboratory tests. The second movement, adagio, from 1980 to 2000, slowly develops and implements initial works. International and national recommendations were published by the IFCC-LM (International Federation of Clinical Chemistry and Laboratory Medicine) and scientific societies [French (SFBC), Spanish (SEQC), Scandinavian societies...]. Reference values are now topics of many textbooks and of several congresses, workshops, and round tables that are organized all over the world. Nowadays, reference values are part of current practice in all clinical laboratories, but not without difficulties, particularly for some laboratories to produce their own reference values and the unsuitability of the concept with respect to new technologies such as HPLC, GCMS, and PCR assays. Clinicians through consensus groups and practice guidelines have introduced their own tools, the decision limits, likelihood ratios and Reference Change Value (RCV), creating confusion among laboratorians and clinicians in substituting reference values and decision limits in laboratory reports. The rapid development of personalized medicine will eventually call for the use of individual reference values. The beginning of the second millennium is played allegro ma non-troppo from 2000 to 2012: the theory of reference values is back into fashion. The need to revise the concept is emerging. The manufacturers make a friendly pressure to facilitate the integration of Reference Intervals (RIs) in their technical documentation. Laboratorians are anxiously awaiting the solutions for what to do. The IFCC-LM creates Reference Intervals and Decision Limits Committee (C-RIDL) in 2005. Simultaneously, a joint working group IFCC-CLSI is created on the same topic. In 2008 the initial recommendations of IFCC-LM are revised and new guidelines are published by the Clinical and Laboratory Standards Institute (CLSI C28-A3). Fundamentals of the theory of reference values are not changed, but new avenues are explored: RIs transference, multicenter reference intervals, and a robust method for deriving RIs from small number of subjects. Concomitantly, other statistical methods are published such as bootstraps calculation and partitioning procedures. An alternative to recruiting healthy subjects proposes the use of biobanks conditional to the availability of controlled preanalytical conditions and of bioclinical data. The scope is also widening to include veterinary biology! During the early 2000s, several groups proposed the concept of ‘Universal RIs’ or ‘Global RIs’. Still controversial, their applications await further investigations. The fourth movement, finale: beyond the methodological issues (statistical and analytical essentially), important questions remain unanswered. Do RIs intervene appropriately in medical decision-making? Are RIs really useful to the clinicians? Are evidence-based decision limits more appropriate? It should be appreciated that many laboratory tests represent a continuum that weakens the relevance of RIs. In addition, the boundaries between healthy and pathological states are shady areas influenced by many biological factors. In such a case the use of a single threshold is questionable. Wherever it will apply, individual reference values and reference change values have their place. A variation on an old theme! It is strange that in the period of personalized medicine (that is more stratified medicine), the concept of reference values which is based on stratification of homogeneous subgroups of healthy people could not be discussed and developed in conjunction with the stratification of sick patients. That is our message for the celebration of the 50th anniversary of Clinical
Introduction

Clinical Chemistry and Laboratory Medicine (CCLM) is a journal reporting the majority of the important papers on reference values. The actual editor-in-chief, in his recent editorial [1] is in good agreement with the previous editor-in-chief who created a specific subheading on reference values in the journal.

We would like to present the genesis and the evolution of the concept through the activity of different groups at the international level starting with our own contribution, particularly in the Center of Preventive Medicine (CMP). We have asked friends involved with us in the development of the concept of reference values for comments and historical milestones.

First movement: allegro con fuoco (1965–1980)

This is the time for a triumphant automation. Laboratories became able to produce a large amount of data with a high level of quality. A series of articles consider the issues of biological and analytical variations in long-term studies in order to assess their possible physiological and medical implications [2–6]. Another group of scientists studied the factors contributing to intra-individual variability [7–12] and the effects of preanalytical factors on the intra-individual variation of some serum constituents [13].

In 1968, when we were asked to create and direct the laboratory of the CMP, we were in a good position to understand the importance of biological variability and genetic influences among healthy families. The medical director of the CMP was himself influenced by the work of Roger J. Williams [14] with his book on biochemical individuality.

In this context, it became evident that for the studies of markers of health deviation, it will be important to master and to better understand biological variability of all plasma constituents in healthy people, in healthy families and in chronic diseases. It was also time for a triumphant automation. Laboratories became able to produce a large amount of data with a high level of quality. The first articles on biological variation are published. Our co-workers started a series of PhD studies on nutritional influences, exercises, drug effects and also creating a biobanking organization for retrospective studies. Some other PhDs focused on biological variability of group constituents: enzymes, proteins and some prospective studies on the use of white blood cells. During more than 40 years, the group has produced 10 books and around 300 publications [15].

At the final round table of the first Pont-à-Mousson Biologie Prospective Conference in 1970 [16], M.C. Sanz and the discussants, particularly Pierre Metais, emphasized the importance for preventive medicine of a panel of laboratory tests carried out on a large number of subjects with automatic and standardized methods. The discussion went immediately to normal values and individual variability.

At the same time, IFCC-LM created a new expert panel on the Theory of reference values’ under the leadership of Tom Whitehead and soon after, of Ralph Gräsbeck.

The beginning of the reference values

Ralph Gräsbeck

The Minerva Foundation Institute for Medical Research, Biomedicum U2, 00290 Helsinki, Finland

My school participated in an exchange of teachers between Sweden and Finland. We had an excellent Swede who taught biology, including genetics. I especially remember that he taught that biometric data, e.g., the size of the leaves of a tree, were distributed in the Gaussian manner. Later, as a medical doctor, I absorbed the prevailing belief that laboratory results from normal people were distributed in this way and that the usual normal values were calculated by taking the mean ± 2 SD. Then, starting my scientific career, I participated in a population study involving 1345 persons whose serum vitamin B₁₂ concentrations were assayed [17]. Our
team used an excellent statistician who presented our data as histograms and curves fitted to them. The ordinary Gauss curve did not fit, so he took the logarithm of the observed values, and *hocus-pocus*, the curves fitted excellently. I asked the statistician why, but only got the response that this was the customary thing to do. I was not satisfied and began to dig into the literature and found quite interesting articles (quoted in [18]) casting doubt on the usual Gaussian distribution.

In the late 1960s I became editor of Scandinavian Journal of Clinical and Laboratory Medicine. The editors were supposed to write editorials, so I wrote one together with a statistician Fellman [18]. It criticized the use of the term normal and calculation of normal ranges assuming Gaussian distribution. At about the same time, I was chosen President of the next Scandinavian Congress of Clinical Chemistry and Clinical Physiology outside Helsinki in 1969. The organizers of the Congress had to devise programs for the sessions, and being sensitized to the topic I suggested ‘normal values’ as a theme which was accepted.

In planning the session and my own presentation, I realized the analogous character of the normal values and the controls in the experimental sciences. I also thought that a less ambiguous term than normal values ought to be used (actually [18] already mentions reference population). My schoolmate Nils-Erik Saris was also one of the organizers of the Congress and experienced in producing IFCC recommendations, especially on quantities and units. I contacted him, and we decided to spend an October weekend in my island in the Baltic planning the session and especially clarifying the hazy concept of normal values and trying to produce a paper with suggested solutions of the problems. (A humorous incident occurred during our two-man sauna symposium in the naked state and has been reported) [19].

During the winter season we produced a short paper [20] suggesting the introduction of the term reference values, outlining a number of branches of the field (nomenclature, reference populations, preanalytical factors, etc.) and calling for international collaboration.

Our suggestions clearly interested the audience, which at the end of the Congress formed the General Assembly of the Scandinavian Society sponsoring the Congress. The Assembly decided to activate the rather dormant function of the Society by nominating committees to produce recommendations in several fields of current interest: quantities and units, enzyme assays and normal values. I became chairman of the last-mentioned committee, which soon began to function and changed its name to the Committee on Reference Values. The other members of the Committee were R. Dybkaer from Denmark, S. Skandzen from Norway and M. Hjelm from Sweden. The early history of the Scandinavian Committee is found in [21].

At about the same time an international congress took place in Geneva and Evian. Afterwards I learned that I had become member of an analogous international Committee on Reference Values of IFCC. When it met, T.P. Whitehead was the chairman and the rest of us members of ‘The expert panel on the theory of reference values’. Its other members, except myself, were G. Siest (F), G. Z. Williams (USA) and P. Wilding (UK); Siest and Wilding entered the Panel somewhat later than the others.

The work of the Panel was slow, nothing was published, which evoked criticism. The cause was mainly disagreement on many points, e.g., the exact meanings of the terms reference value and health, but also on practical details such as specimen collection, some favoring strict rules, while others are lax. To provide some guidelines, a fairly long series of educational articles on reference values appeared in the IFCC Newsletter No. 16, February 1977 [22]. A constructive international symposium on Reference Values in Laboratory Medicine was held outside Helsinki in May 1980 [23]. IFCC at that time had rules preventing a member of a panel to serve for more than a fixed number of years. Whitehead therefore resigned in 1976, and I was appointed chairman of the Panel. The panel now produced its first recommendation, albeit only a preliminary one ‘Provisional recommendation on the theory of reference values’ dated 1978 [24]. Then, I was in turn to resign, but continued to be an unofficial adviser and a link to the Scandinavian Committee, which decided to stick to the practical rather than the theoretical aspects of reference values.

In the work of the Scandinavian Committee, Solberg proved himself very competent, so I succeeded in getting him appointed as my successor. Under his guidance the Expert Panel completed an impressive number of recommendations. He and the members of the Panel (including myself) also spread the gospel of reference values in personal articles, one of the most noted ones being a humorous ‘Miss Manhattan’ study correcting the view that the requirements for collecting reference values were almost impossible to satisfy [25].

This is what I experienced as the beginnings of the IFCC activities in the reference value field. Accounts have previously appeared in [21, 23, 26].

At the second Pont-à-Mousson meeting in 1972, we described the first results on our family population and we published the first book on reference values [27]. We were in contact with all the scientists involved in the field and in IFCC but also with Tom Whitehead and Peter Wilding who participated in the Pont-à-Mousson meeting. They were working in a comparable health screening organization BUPA in the UK. Production of reference values required large healthy populations, military recruits, blood donors, medical students, and laboratory workers etc... populations are poor representatives of healthy populations. The very comprehensive studies have been published by health screening centers (Table 1).

The recruitment of healthy subjects is crucial for defining better laboratory test biological variability.

We were in a good position with Peter Wilding to develop the first diagrams on the definition which were at the origin of the IFCC recommendations.

<table>
<thead>
<tr>
<th>CENTRE DE MEDECINE</th>
<th>BUPA</th>
<th>CALAB</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREVENTIVE Nancy</td>
<td>London</td>
<td>Stockholm</td>
</tr>
<tr>
<td>G. Siest</td>
<td>T. Whitehead</td>
<td>I. Jungner</td>
</tr>
<tr>
<td>J. Henny</td>
<td>P. Wilding</td>
<td>G. Walldius</td>
</tr>
<tr>
<td>F. Schiele</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1 Health screening centers contributions.
Memories of IFCC Expert Panel activities in 1974

Peter Wilding, PhD, F.R.C.Path., FACB
Professor Emeritus, University of Pennsylvania, Philadelphia, PA, USA

In 1974, I was invited by Prof. Tom Whitehead to join the IFCC Expert Panel on the Theory of Reference Values. The panel at that time included Prof. Ralph Gräsbeck, Prof. George Z. Williams, Prof. Gerard Siest and myself. It is only in recent years that I have come to truly appreciate the privilege I enjoyed being part of a group that would establish the principles and terminology that governs the topic of reference value use, and assignment, today and in future years. Moreover, I gained so much from the interaction with four outstanding clinical chemists who worked energetically to improve our profession.

For several years the Expert Panel had strived to define its purpose, and determine a strategy, for proposing a clear and definitive nomenclature for the topic. Much of the problem lay in finding a way to present the subject in a simple and unambiguous manner that would facilitate understanding of why we needed reference values, how to create them and how to interpret them.

In 1976 Gerard Siest and I were asked to find a structure for presentation. We met one evening at a London hotel and after a lengthy, and animated, discussion formulated a diagram that we believed would meet the Panel’s needs. The diagram was presented to the Panel, where it was accepted unanimously, and was presented to the profession in several journals in 1978. The diagram (see below) has stood the test of time and laid the basis of numerous derivative articles from 1978 to 1991 and Approved Recommendations from IFCC. None of the five individuals who formulated, and presented, this time-tested proposal are included in the authorship of the Approved Recommendations, but whenever I see, or read, articles using terms such as ‘reference individuals’, ‘reference values’ and ‘reference intervals’, my memories of a long evening in a London Hotel conference room arguing, discussing and debating with my friend Gerard Siest are revived with a tremendous sense of accomplishment over what we created.

The ‘original’ Scheme (Siest and Wilding):

REFERENCE INDIVIDUALS

Comprise a
REFERENCE POPULATION

From which is selected a
REFERENCE SAMPLE GROUP

On which are determined
REFERENCE VALUES

On which is observed a
REFERENCE DISTRIBUTION

From which are calculated
REFERENCE LIMITS

That may define
REFERENCE INTERVALS

This first document corresponds to the first IFCC paper published 13 years later [28].

Outside this expert panel, different groups of scientists were involved in studies on reference values between the years 1970 and 2000. We could classify them into four international groups:

1. **SFBC (Société Française de Biologie Clinique) and French speaking group:**

2. **Scandinavian group:**
   (i) R. Gräsbeck, R. Dybkaer, S. Skandsen, M. Hjelm, H. Solberg, P. Hyltoft-Petersen, N. Lahti, M. Jungner, A. Kallner

3. **American and British group:**
   (i) G.Z. Williams, E.K. Harris, D. Young, J. Boyd, P. Wilding, J. Lott, C. Fraser

4. **Spanish group:**

Due to the activity of the same people in IFCC and the French speaking group, the recommendations have been prepared simultaneously by both groups perhaps with more details in the SFBC one where the meetings were more frequent. Some countries, e.g., Spain also had a reference values committee with good input from our Spanish colleagues.

We can compare the papers published by the four main groups, they are not equivalent. Often papers in French were published before and used as reference documents later for IFCC recommendations that came much too long after. IFCC Expert Panel on theory of reference values was created in 1972 and the last approved recommendations came 19 years later! (Table 2).

The SFBC Reference Values Working group was created in 1973 and recommendations were published in the early 1980s; a last recommendation on the use of reference values [48] was not addressed by the IFCC expert panel. The Spanish group was also very active, publishing six articles of recommendations. Two of them were additional documents to those of the IFCC and SFBC: one on the transferability of reference values, the second one on the intra-individual reference values.

Simultaneously, the Scandinavian group, published in 1993, on behalf of the Committee on Reference Values of the Scandinavian Society for Clinical Chemistry the recommendation for the preanalytical step including management of the biological samples [55].

The works of the different committees were presented in IFCC conferences or other specific meetings. They were
occasions to meet our American and Canadian colleagues. That was the case in 1975 in Toronto, where we presented the concept of reference state and met Claude Petitclerc and Louis Munan. They have developed an ‘atlas of blood data’ and it was the start of a fruitful collaboration.

Claude Petitclerc, MD, PhD, F.R.C.Path.
Department of Biochemistry, CHUM-Hôpital-Notre-Dame, Montreal, QC, Canada

In the 1970s, I was involved in a long and fruitful collaboration with epidemiologists Louis Munan and Anthea Kelly at Sherbrooke University. The Sherbrooke data obtained from a probability sample of families of a large freestanding population was the subject of many publications on reference data and correlates [56–59]. In 1978 I took sabbatical leave to join Gerard Siest’s team.

CMP was the mecca. Not only was the staff remarkable in number and quality but the study population was much larger, family-based and very well-documented. Reference values were systematically produced and published.

Similarities of both sampled populations and laboratory technologies prompted a comparative study that revealed not only population differences but also analytical biases [60]. The problem of transferability of laboratory results and hence reference values had to be addressed. Reference data produced with the technology of the 1970s could not be used decades later. However, the enormous contribution of Siest et al. at CMP was the systematic study of factors of biological variations and their stratification for each substance. This information, independent of technology, was and remains transferable.

More complex was the issue of analytical variability imbedded in the massive amount of population data over the years, especially at the time of data mining. It was then a concern and is still one. A challenge for study of long-term longitudinal variations. A long and fruitful collaboration of 35 years was the outcome of this sabbatical leave.

During my stay, I had the honor to serve on Commission des valeurs de référence de la Société Française de Biologie Clinique (SFBC). Inspired by Gerard Siest’s leadership SFBC was very proactive in publishing guidelines for the production of reference values; being part of such a dynamic group opened the way to relay Gerard Siest on the Expert Panel on Theory of Reference Values of IFCC. I keep a vivid memory of the members of the panel with whom I enjoyed to exchange with for some 10 years under chairman Erik Solberg. Document 2 and 3 were spinoffs of both SFBC and Scandinavian Society for Clinical Chemistry recommendations.

CMP is no longer what it once was, but the people who made it what it was are still around, with us or in our memory. I cherish them all.

Second movement: adagio
(1980–2000)

Between 1980 and 1990, it was a period of maturation but no efforts were made for developing the practical use of reference values by clinicians. We organized in 1983 a meeting in Nancy with Gene Harris to try to develop a new strategy based on his statistical competences and in collaboration with the other statisticians in our group or in Europe.

The importance of Eugene K. Harris contribution
Josep M. Queraltó
Hospital de la Santa Creu i Sant Pau, Servei de Bioquimica, Av. Sant Antoni M. Claret, 167, 08025 Barcelona, Spain

I was very happy to work with Eugene K. Harris during a sabbatical period. I am indebted to him sharing his great scientific culture with me. Although reference values theory clearly is a part of the so-called post-analytical phase, the development of reference values theory is without a doubt a decided drive to refine preanalytical and analytical concepts. The central idea to create a space where a presumptive patient can be tested against non-diseased individuals raised the need for rigorous, practical tools to manage physiological pathological knowledge. Uncertainty, variability, sampling, confidence, tolerance, robustness, expectancy... are concepts with an accurate meaning in mathematical statistics which soon were part of the reference values theory. Accurate use of these terms deserved the special attention of all the groups which conceptualized reference values. Among the statisticians who played a key role to translate statistical science into laboratory medicine was Eugene K. Harris (1927–1997). Gene Harris was a wise man with a broad culture and scientific interest. Gene had an exceptional ability in grasping what those professionals which consult him really needed. At the end of the 1960s, Gene begun to collaborate with Prof. George Z. Williams, who had already published some articles about laboratory data process [61]. Their insights on the components of medical laboratory data variation, biological and analytic, were published in a series of influential articles [2–6]. He realized the relevance of within-person variability in the assessment of biological variation of a number of analytes [62–65]. Eugene Harris contributed significantly in virtually all the aspects of the reference values theory where statistics are required: gaussianity requirements [66], analytical goals [67, 68], analytical variation in monitoring reference individual over time [69, 70], the relative influence of analytical to biological variance [71], the relevance of intra-individual component of biological variance [72, 73], the interpretation of the practical significance of a change in laboratory results [74–78], multivariate regions of reference [79] and the requirement of partitioning reference populations [80, 81], as well as comprehensive reviews [82] and books on statistical aspects of reference values production and use [83, 84]. Moreover, he was persuaded of the need to convey to both clinical chemists and doctors the clearest and sharper statistical concepts together with the straightforward and practical procedures and methods available. At that time, when personal computers did not enjoy today’s facilities, he provided also useful and simple code to help in statistical calculations [84]. The vision and personality of Eugene Harris inspired the work from national and international committees and scientists [32, 52] and today still remains as a reference.

During this period, we started the recruitment of 1000 families of the Stanislas cohort [85] on a more specific research group in CMP focused on individual specific reference values in subject under medication, e.g., contraceptive
pills. We tried also to propose new presentation of reference limits particularly for a better use in prevention on health maintenance including for and by the patients themselves.

The data produced were not limited to biochemical components but also to weight, blood pressure etc. The material collected in the Stanislas cohort project comprised a lot of interesting longitudinal familial data and of biological variability and many publications came out from this tool [85] and particularly the effects of genetics [86], e.g., Apolipoprotein E.

We tried many times particularly as IFCC president to recreate an IFCC reference values committee but the IFCC scientific committee opposed propositions made by a group of scientists. The objectives of this committee would comprise the following:
1. Update IFCC recommendations;
2. Produce practical guidelines;
3. Increase the transferability;
4. Improve the proper and extended use of reference values by clinicians, laboratory scientists and patients.

<table>
<thead>
<tr>
<th>IFCC J Clin Chem Clin Biochem</th>
<th>SFBC (Annales de Biologie Clinique)</th>
<th>Spanish (Bioquimica Clinica)</th>
</tr>
</thead>
</table>
| Selection of individuals for the production of reference values [29] | Sachs C. Objectifs [37]  
Drosdowsky M, Ramon-Bauza F. Age, sexe, puberté, ménopause [38]  
Sachs C, Buret J. Poids corporel et surcharge pondérale [39]  
Aellig A, Sachs C. Facteurs d’environnement: alcool, tabac et mode de vie [40] | |
Buret J, Vernet-Nyssen M. Annexe du Document E:  
| Control of analytical variation in the production, transfer and application of reference values [31] | | |
| Statistical treatment of collected reference values | | |
| Determination of reference limits [32] | | |

Table 2 Comparison between IFCC, French and Spanish documents.
We proposed, without success, a joint CCLM IFCC committee for a better and quicker diffusion of the work and the recommendations.

The founding texts of various scientific societies over the past 20 years are true reference knowledge. They had the merit to disseminate the concept of reference values. Thanks to them, the RIs have become a basic tool for the interpretation of the quantitative results of laboratory tests. However, the evolution of medical practice led to question the current relevance of this concept.


It is obvious that in the late 1990s the concept of reference values is progressively used by all health professionals, including clinical chemists, clinicians, and simultaneously by all official bodies in charge of the establishment of the legislation. Its use by the In Vitro Diagnostic (IVD) industry and laboratories is now recommended by the European Directive 98/79 EC [87] and the ISO 15189 standard [88, 89].

In contrast, the application of the concept remains difficult in clinical and laboratory practice: procedures for estimating reference limits need to be improved. They are too long and too expensive and are not feasible for all laboratories. In 2000, a group of scientists and professionals focused on areas for improvement by proposing that societies and organizations implement the current recommendations [90].

Fundamental works: evolution of the concept and proposals for new procedures

In the early 2000s The Catalan Association of Clinical Laboratory Sciences organized two successive symposia for opening dialogue between professionals and industry representatives. It approached the legal and normative alternatives to the production of specific reference values and the transferability of reference values and the methodological aspects (how define a homogenous starting population, how to define acceptable health status for reference individuals, methodological and statistical aspects, diffusion and tracking of the information) [91].

A year later, in 2004, after the second Catalan Symposium, a special issue of CCLM was published in order to share the basic concepts and new ideas on the field reference values [92]. The expressed opinion were very diverse and do not reflect systematically the official practices. It was the choice of editors to open debates and initiate discussions. Among the 25 contributions it is impossible to say which are the most determining and most helpful. At the risk of being wrong, here are the themes that seem, 10 years later, the most significant: is there a concept of health and normality?, statistical description of reference values, interpretation of reference limits, the concept of common and multicenter reference intervals, and analytical aspects in relation to reference intervals.

Per Hyloft Petersen

NOKLUS, Norwegian Quality Improvement of Primary Care Laboratories, Division for General Practice, University of Bergen, Bergen, Norway

I have been involved in application of the concept in creation of common reference intervals for plasma proteins in the Nordic countries in the 80th and 90th, mostly published as NORDKEM reports and in Upsala Journal of Medical Sciences [93, 94]. NORIP, the other Nordic project on common reference intervals, was chaired by Pål Rustad in Oslo, Norway.

The development of the concept on reference values was a great intellectual conquest which solved many problems in relation to interpretation of clinical laboratory data. The concept, however, is developed for single laboratories, so there may still be differences in measurement results and in the interpretation for the same samples analyzed in different laboratories. Consequently, the next aim was to harmonize measurement results and interpretation of data among laboratories, so that results were interchangeable and clinical explanation would be independent of which laboratory performed the analysis.

Analytical quality: The question on analytical quality of measurements is technical analytical and relates to many elements in the traceability chain from reference method values or international certified reference preparations to the single measurement result. Even these efforts are essential; the final analytical quality can only be proved by control with genuine human materials with target values measured by the reference method. The documentation procedure is considerable and rules for acceptable analytical quality specifications are outlined.

With the analytical quality assured and documented, the next question is whether the reference intervals are dependent on gender, genetic and other biological characteristics.

Models for treatment of the biological information from reference values

A. Even the idea of sampling of reference individuals for estimation of reference intervals is clear in the sense of defining healthy individuals through rule-in and rule-out criteria, the concept of health is more complicated than supposed for varying criteria, and a common question is whether the reference interval should mirror the current apparently healthy population (without any known diagnosis) or be an ideal for all, e.g., based on teetotallers.

B. Further, differences in rule-in and rule-out criteria may be needed for varying components. A reference interval for, e.g., serum-thyroid stimulating hormone (TSH) will need
criteria for absence of thyroid antibodies in all reference individuals, and when, e.g., reference values are to be collected for plasma-glucose and HbA1c in relation to diabetes diagnostic, the criteria will be very strict regarding, e.g., first-generation diabetes relatives. Moreover, the biological concept has been overruled by the international clinical guidelines which recommend diagnostic decision limits for the diabetes diagnosis, neglecting biological variation and other individualities.

C. When two or more subgroups of reference individuals are disclosing deviating reference intervals, it is necessary to decide whether the groups can be combined or should be separated. Varying partitioning criteria have been proposed, but the dominating model is to separate if the result of using the common reference interval results in percentages of reference individuals for lower as well as upper reference limits for both groups are beyond 0.9%–4.1%.

The idea of common reference intervals is extremely attractive and several attempts have been made. For serum electrolytes and some metabolites it seems possible to establish common reference intervals within ethnic groups and selected geographical areas, but for most other components it is only possible within smaller well-defined groups, and it is necessary to test whether common reference intervals can be established.

Further, I have been involved in evaluation of analytical quality specifications, together with Elizabeth Gowans, Callum Fraser, and others, and in proposing criteria for partitioning together with Ari Lahti.

Together with Callum Fraser, Anders Kallner and Desmond Kenny, I arranged the Stockholm conference on analytical quality specifications, published in Scand J Clin Lab Invest 1999;59:475–585, and together with Joseph Henny I edited the CCLM issue on reference values [92]. Moreover, I wrote a paper together with Esther Jensen and Ivan Brandslund: Analytical performance, reference values and decision limits. There is a need to differentiate between reference intervals and decision limits and to define analytical quality specifications which shows many of the restrictions of use of reference intervals and of decision limits [95].

In 2012, a 2nd Special Issue of CCLM devoted to reference values was published. It highlights, once again, the interest in this topic by successive editors of this journal. It brings together a variety of more than 40 articles most of which are devoted to the determination of RIs for particular analytes [96].

Following and/or in parallel with the publication of CCLM’s Specials Issues, numerous articles were published mainly in CCLM and in some other scientific journals (mainly in Clinical Biochemistry and in the Scandinavian Journal of Clinical Laboratory Investigation). Several themes are of particular interest:

The concept of common and/or multicenter reference intervals has been developed in order to avoid laboratories having to produce their own reference values. Two different approaches have been proposed:

1. The first was established in the 1980s by the Nordkem Project [93, 94] followed up some years after on a larger scale by the Nordic Reference Interval Project 2000 (NORIP) to determine common biological reference intervals in wide geographical areas for biochemical quantities and for large populations reference [97–100]. It is also clearly stated that several conditions must be fulfilled: 1) reference population from the different countries must have the same genetic profile and have a similar lifestyle; and 2) certain quality requirements must be met (including traceability of measurement systems, each laboratory using routine analytical systems). This project involved the participation of 102 laboratories in five different countries (Denmark, Norway, Sweden, Finland and Iceland).

Ichihara et al. have been inspired by this protocol as several studies were conducted in South-East Asia some years after [101–103].

2. The second approach was proposed by the Catalan Association for Clinical Laboratory Science [104]. The concept is a little different: several clinical laboratories in the same geographical area, using the same analytical system determine reference intervals in the context of a collaborative project with the IVD industry. Recruitment of individual’s reference is under the responsibility of each clinical laboratory. Each step preanalytical, analytical, statistical treatment follows a well-defined protocol. Metrological traceability is ensured. Several examples of application to specific analytical systems were published in CCLM [105–110].

The approach proposed by the Catalan Association provides only partial response: the RIs proposed are valid only for the considered analytical system. The Scandinavian approach is appealing: the calculated RIs apply to routine methods of each clinical laboratory; potential biases of each routine method are eliminated by the use of common reference materials measured in each participating laboratory. However, this is a very demanding procedure applicable to laboratories having an outstanding experience of standardization as shown by Hyltoft Petersen and Rustad [111]. However, it is necessary to observe a number of rules specified by Ceriotti [112] and Boyd [113].

Improvement of statistical procedures has mobilized the energies, none of those proposed in the multiple recommendations are, to this day, perfect.

1. Data mining: Although the use of data mining procedures are extremely controversial, due to the
lack of information on reference individuals which may lead to selection bias, the Haeckel’s staff, in Germany defends the use of broad data sets from laboratory information systems to derive intra-laboratory reference limits. The RIs are derived by sophisticated statistical procedures, requiring the intervention of highly skilled biostatisticians [114, 115]. The authors note, however, that if the differences between laboratories RIs are observed, they are mainly due to biases between the different analytical procedures [116, 117]. However, Kallner showed that the central 95 percentile derived from Primary Health Care population was wider than those usually assigned, whereas the median almost coincided [118]. He and his coworkers claimed that remaining non-healthy individuals enlarge artefactually the reference intervals.

A similar approach was developed by Grossi et al. [119] in the project REALAB. In a primary health care center, RIs are derived by retrospective analysis of a large set of data collected during three years after a careful selection of non-diseased individuals (based on clinical criteria) and with adapted statistical procedures.

2. **Partitioning**: the most common partition criteria are sex and age. If the reference set of data come from different reference populations (geographical regions, ethnical groups), the origin of each population can be considered as a partition criterion. There is currently no ‘gold standard’ for partitioning a reference sample. Several methods have been proposed by different authors and IFCC [120–122]. Lahti has done an outstanding critical comparison of the main methods proposed [123].

3. **The concept of multivariate Reference Region** was proposed as alternative to univariate reference intervals in the 1970s in order to reduce the number of ‘false-positive’. Even, if this concept has not been really currently used in clinical laboratories practice until now, it remains promising for the future [124].

4. **The appropriate statistical procedures** are a key to derive the RIs. It is difficult to make a choice among the proposed methods. Ichihara and Boyd published in 2010 a review of possible methods [125]. Among them the authors point the multiple regression analysis as a powerful tool to identify the most important factors of biological variation and the analysis of variance (ANOVA) as an effective way to assess the need for partitioning reference values.

5. **Robust approach**: The IFCC recommends that a minimum number of subjects (n=120) is recruited to derive RIs. In some cases it is not possible to reach the suggested number. Horn and Pesce [126, 127] have proposed an alternative method to estimate reference intervals for data sets with a small number of observed values. Robust method to calculate the upper reference limit gives results comparable to traditional methods, but with a confidence interval higher.

6. **Outliers and non-healthy effects individuals**: Faced with the impossibility of recruiting a perfectly homogeneous reference population, statistical techniques are used to mitigate the effects of ‘abnormal’ individuals (in the sample of reference) on the width of the reference range. Horn believes this increase is approximately 10%. This reinforces the importance of proper selection, a priori, of the reference individuals. Usually the Tukey method is most commonly used. Horn proposed a more sophisticated method based in a first stage on a Box-Cox transformation, then in a second stage of labeling extreme values by the method of Tukey [128].

**Analytical aspects**

The analytical bias affects the measurement results and can change RIs, then this may change the interpretation of results. Determining RIs with analytical procedures traceable to a reference system allows transferring results from one laboratory to another [129]. The use of this strategy should be promoted and favored by IDV manufacturers to provide more effective information to clinicians. However, it has no universal value for many years; it concerns only the analytes traceable to a reference system (about 60 to this day). For others the issue of transferability remains. C-RIDL Committee is working hard on this issue and should propose solutions soon. Alternatives could also be considered, such as the use of a common calibrator, but also with other limitations [130].

**The role of the IFCC-LM: creation of C-RIDL (Committee for Reference Intervals and Decision Limits)**

Until 2004 the Scientific Division of the IFCC-LM was no longer interested in the reference values issues. It must be recognized that the guidelines proposed by the IFCC and the NCCLS (now CLSI) were less and less adapted to the needs of the professionals: IVD’s manufacturers and laboratories. Directive 98/79/EC of the European Community requires manufacturers to provide reference limits in the
package inserts of laboratory reagents kits [87]. Other international organizations, such as ISO also recommends the use of RIs in the laboratory reports [88]. To meet the growing demands of international regulations and, largely on the insistence of IVD manufacturers, IFCC-LM decided in 2005 to create the Committee on Reference Intervals and Limits decision (C-RIDL). It was composed of Ceriotti F. (Chairman), Boyd J., Henny J., Kairisto V., Klein G., and Queralto J.M. Almost simultaneously, a working group of the CLSI was created with similar objectives. A joint Working-Group IFCC-LM and CLSI was created in 2005 during the XIX International Congress of the IFCC-LM in Orlando. The purpose of this joint Working-Group was to review the document CLSI (C 28–A) published in its original version in 1995 [131] and revised in 2000 [132] in order:

1. To provide clinical laboratories and diagnostic test manufacturers with updated guidelines for determining reference intervals for quantitative analytes;
2. To provide recommendations regarding procedures that can be used to verify reliable reference intervals for use in laboratory medicine; and
3. To publish a document common to IFCC and CLSI.

Updated document was published in 2008 [133]. The major guidelines of the original recommendations were preserved in the updated document. Three main innovations were introduced: the introduction of a robust method for analyzing references values (especially if the number of data is <100), the transference of RIs and RIs verification, and the reference intervals multicenter studies [134, 135].

Independently, in parallel, C-RIDL published from 2005 to 2010 a series of works. The first seeks to identify universal reference intervals for creatinine. A careful study of the literature showed that very few articles had taken into account the requirements of metrological traceability, thus few data from the literature are transferable from one laboratory to another. The C-RIDL established that for laboratory using a method traceable to IDMS the proposed RIs can be used [136].

The second job of the C-RIDL was devoted to the opportunity to determine Common Reference Intervals. In a preliminary article Ceriotti recalled the position of the IFCC-LM and prerequisites to follow [137]. These are related to the design of the study and the analytical conditions (traceability). Ceriotti also recalls that the populations (that of laboratory and that of the study) must be comparable. An application’s work was made for the measurement of three enzymes (AST, ALT and GGT) measured with commercial analytical systems, but according to the standard methods recommended by the IFCC. Patient’s sera from four regions (Italy, China, Turkey and the Nordic countries) were analyzed. For AST and ALT the use of common RIs appears possible. In contrast, GGT shows significant differences between populations, worldwide RIs do not seem applicable [138].

Finally, two members of the Committee attempted to clarify the position of Reference Intervals vs. Decision Limits. In an article in the eJICC, Ceriotti and Henny recalled that the reference limits should not be confused with the decision limits. They recalled their definitions and characteristics (Conditions Influencing them, Information gathered, Statistical methodologies ...). This is one of the few articles on the role of reference intervals in the process of the interpretation of laboratory results [139]. Incidentally, this topic was raised again a few years later by Haeckel [140] and Hyltoft Petersen [95].

In 2011, Ichihara K. became the new chairman of the C-RIDL renewed (members: D. Armbruster, Barth J., Klee G., Ozarda Y., Pekelharing M.). Now, it aims to derive reliable country specific RIs through multicenter studies at a global level. A protocol was developed by the Committee. It will be applied to markers traceable to a reference system and non-standardized biochemical markers. The ongoing project will demonstrate the feasibility of the protocol and deriving RIs applicable to a global scale if there are no regional and/or ethnic groups’ differences. In 2012 the study is in progress and the results should be published in 2013.

**Deriving works of reference limits**

Besides these fundamental works extensively devoted to the derivation of reference intervals, numerous articles have been published in various scientific journals:

The influence of ethnicity as a partition factor has never been really studied systematically. Horn [141] and Johnson [142] observed for some analytes differences between populations that require separate reference intervals. In this way a large number of works from emerging countries have been published most often in scientific regional journals (they are too numerous to be mentioned here). Most of them follow (more or less) the traditional protocol of the IFCC, applied according to local opportunities and constraints. Each article covers a limited number of analytes and is applied to a group of well-defined populations.

Another group of works relates more specifically to certain periods of life: pediatric age, elderly and pregnancy.

1. Pediatrics reference intervals determination is one of the most difficult tasks to carry out, primarily because of ethical limitations related to blood...
drawing in very young children (or in neonates). Apart from a few specific derivation RIs studies for highly specialized analytes [143–146], the Canadian multicenter study CALIPER, aimed to establish a data base of age- and gender-specific pediatric intervals for 40 serum biochemical markers [147, 148]. As emphasized, Ceriotti [149] the authors have followed the recommendations of the IFCC and CLSI including selection of young children in good health (on the basis of clinical data). However, traceability of analytical data is not truly established. Also transferability of produced RIs may not be always transferable to other analytical systems. The Canadian group, as a service to major clinical laboratories, published a series of articles showing RIs related to various specific analytical systems [150–153].

2. Elderly RIs determination is even more challenging than for young children. The main difficulty is to select healthy individuals: most seniors do not meet the criteria of the IFCC/CLSI. Taking medication(s), unrecognized subclinical diseases can alter the width of the reference range. Hence it is very difficult to differentiate the effect of age, aging or a pathological condition. During this period 2000–2012 some authors are interested in this issue [154–156].

3. Pregnancy is a factor of biological variation due to physiological changes during pregnancy and postpartum. Although changes during pregnancy are well-known some articles are still published [157–159].

4. Genetic is now a well-recognized determinant for many biochemical markers. Use of genetic information for partitioning reference intervals could reduce the misidentification of unusual test results caused by non-disease associated genetic variation, as pointed by Shirts [160]. Siest and co-workers started to be more precise in using genetic information for subgroups stratification, e.g., for ApoE [86] and more recently for haptoglobin [161]. However, the part of biological variability induced by genetic variants is often low and the knowledge of the genetic status of the reference individuals is often lacking. In these conditions the partitioning of RIs by genetic information remain for the moment (and for many long years) only a possible way for the future.

During these years, the theory of reference values was unexpectedly applied to the veterinary laboratory medicine! A French team from the Veterinary School of Toulouse published several theoretical and applied articles, noticeably on reference intervals derived for several animals [162–166].

Fourth movement: finale

What is the situation so far from the CCLM 2000 article which identified gaps and desires of professionals [90]?

On the methodological level:

(i) The importance of careful selection of reference individuals was reaffirmed by the IFCC-LM and confirmed by many authors. Selection from the database without having clinical information to ensure that people are healthy is prohibited. However, there is still no clear guidelines for the issue of RIs for unstable periods of life (e.g., aging or elderly).

(ii) One possible source of reference samples may come from large biobanks that are around the world, especially for studies of large populations. These samples come from individuals generally in good health. They are associated with clinical data often sufficiently comprehensive to select healthy individuals. Unfortunately, the only caveat remains the legal and regulatory restrictions on the use of biological samples in countries other than the original one. We must mention here the first experiment dedicated to the determination of RIs launched by the Scandinavian group which initiates the NORIP Project [167].

(iii) The issue of ethnicity as a factor of variability is still not resolved. We know that there are differences between ethnic groups for some analytes. Information is scarce and far from exhaustive. This is a major obstacle for the transferability of RIs without precautions. How to address this issue in a world where the notion of ethnicity is more and more difficult to define because of population movements? Finally, are the observed variations clinically significant? Genetic profile will help to define better the subgroup.

(iv) Following the initial concept of multicenter reference intervals we can recruit a larger number of reference subjects by dividing the efforts between several laboratories, but analytical bias between different analytical systems remain.

(v) The concept of common reference intervals similar to the previous, emphasizing metrological
traceability, allows the transfer of RIs to all laboratories using methods traceable to the same reference system, but to them only.

(vi) Traceability to a reference system is limited to a small number of analytes, other methods of transference and verification of available RIs are proposed by the IFCC-LM and CLSI, of course, with some limitations.

(vii) New statistical methods have been introduced and validated by the IFCC:
1. The robust method allows deriving RIs on small data sets of values. The accuracy of the so determined limits is less good than with the number of reference samples recommended, but it is an indispensable tool for certain populations (e.g., neonates, elderly, rare analytes, ...).
2. Older forgotten statistical methods: multiple regression analysis, ANOVA are very useful for estimating the factors of biological variation.
3. The statistical methods of partitioning are effective, but they are limited in the presence of several sub-classes.

On the clinical use of RIs:

(i) The concept of RIs is often misunderstood by medical doctors. It has been clearly reaffirmed that RIs should not be confused with the decision limits. The RIs are descriptive of a specific population and are derived from a reference distribution (usually 95% interval). Instead, the decision limits are thresholds above or below which a specific medical decision is recommended.

(ii) What is the use of RIs in clinical practice? This is a (small) help to medical interpretation when there is no medical decision threshold established, when the methods are not standardized (this is the case for many methods in immunochemistry). It is therefore particularly important that RIs are calculated as accurately as possible, in other words, they reflect a 'true' healthy population: sensitivity and/or specificity of the medical diagnosis depend on.

(iii) It would be incomplete if we do not take into account that, for many analytes, the variation is a continuum. Continuous variation of a biomarker may have diagnostic significance more important than a result beyond the upper RL. With this in mind the concept of Reference Change Value could be really useful: unfortunately it is not really used in everyday practice [168].

In terms of access to RIs:
It is important to recognize that comprehensive information is not really available: some textbooks were published a few years ago, but they are not always available. In some articles which were published in various scientific journals, information is sparse, incomplete for only some markers, for well-defined populations and do not always meet the methodological requirements needed. Information from the literature is not always transferable and is not easily accessible to the professional community. This is why the major source of information is that of the IVD’s manufacturers, reported in the package inserts. However, is this information still reliable? Are they sufficiently informative? Are they still applicable to the population of each laboratory? The quality of the information provided in the technical documentation is very heterogeneous. Some manufacturers give very accurate and comprehensive information on how were selected reference individuals, ethnicity, number of individuals, and the statistical methods used. However, this is not always the case.... We can only encourage IVDs manufacturers to follow recommendations proposed by the IFCC and CLSI...

Training of medical doctors and laboratorians:

(i) The misuse of the concept of RIs was noted above. Major educational efforts should be made towards medical doctors so that they understand the purpose, scope and limitations. Publication of educational materials would be helpful. In this respect, even very simplified documents produced by the AACC and published on the website Lab Tests online is a first step, quite remarkable, but ... towards patients [169].

(ii) Improving the presentation of the results has not been a major focus of interest at all for many years. The inclusion of more information on the biological variation factors (and others) in the analysis reports should be possible by new information technologies.

Real progress has been made in recent years, particularly in terms of methodology. CCLM contribution was instrumental in disseminating the work of the IFCC and many independent teams. A non-exhaustive survey of the literature shows that CCLM published in recent years the largest number of articles on the concept of reference values and its applications. There are still many fields to improve and complete: CCLM publishers have the duty to hold one’s position to this topic!

The scope of reference values and/or reference limits will remain attractive for many researchers for many
years to come. However, we believe that future works should not be directed almost exclusively to methodological issues (statistical, analytical, ...), but focus more specifically on the real contribution of the reference limits to the interpretation of laboratory results and to better define their place in medical decision-making. It must not be forgotten that these are based primarily on evidence-based clinical decision limits. This development is absolutely necessary to allow optimal use of laboratory tests and avoid misdiagnosis. This is a very exciting issue for the young scientists which needs, however, a close cooperation between laboratorians and medical doctors. It is on this condition that laboratory medicine can contribute to the improvement of Public Health for the benefit of patients.

Conflict of interest statement

Authors’ conflict of interest disclosure: The authors stated that there are no conflicts of interest regarding the publication of this article.

Research funding: None declared.

Employment or leadership: None declared.

Honorarium: None declared.

Received October 9, 2012; accepted October 18, 2012; previously published online November 24, 2012

References

83. Harris EK, Yasaka T, editors. Maintaining a healthy state within the individual. Ifip-Imia Working Conference on Maintaining a Healthy State Within the Individual; 1986; Kobe, Japan: North Holland.


Prof. Gérard Siest received his pharmacy diploma, his specialization in laboratory medicine and his PhD from the Universities of Strasbourg and Nancy. He was Professor of Biochemistry, Molecular Biology and Molecular Pharmacology and in charge of the direction of the Center for Preventive Medicine laboratory and some years later all the research of this health screening organization. He developed a research team linked to CNRS and INSERM dealing with drug metabolism, more particularly the UDPGTs and on the genetic influence on laboratory tests and on reference values. Two proteins were studied more deeply: apolipoprotein E and gamma-glutamyltransferase. He created and was in charge of the regional postgraduate course on drug metabolism and biochemical pharmacology. He was President of the European Society of Biochemical Pharmacology and of the International Federation of Clinical Chemistry (IFCC) for 6 years. Gérard Siest was Editor-in-Chief of Clinical Chemistry and Laboratory Medicine (CCLM) from 1998 until 2008. In 2010, he founded a new European Society of Pharmacogenetics and Theranostics (ESPT) for which he is now President and starting to be Editor-in-Chief of the journal Drug Metabolism and Drug Interactions (DMDI) that is also the official journal of ESPT. In addition, he is involved in many boards of pharmacogenomics and laboratory medicine journals. He has published over 600 peer review publications. The scientific conferences organized every 2 years, in Santorini (Greece) under his responsibility are very successful. The last one was held from 30 September to 2 October 2012.

Joseph Henny pursued his studies at the University of Lorraine (formerly Nancy). Then he served as director of the clinical laboratory at the Centre for Preventive Medicine in Nancy. Since 2009, he is responsible for the constitution of a Biobank devoted to large population studies related to the ‘Constances Cohort Project’ launched by UMR INSERM 1018 (Centre for Research in Epidemiology and Population Health) – Versailles Saint-Quentin University. He is also in charge of the scientific and technical coordination of the Medical Laboratories Network of the French Health Periodic Examination Centres. He is participating in several regional and national programs in the domains of Public Health in collaboration with regional and/or national health operators (such as French Institute for Public Health Surveillance). He serves as reviewer for various journals. His scientific interest is primarily in the field of the study of the effects of biological variations on laboratory tests, he is working on development of reference values and related concepts. Further, he has been an active participant in numerous studies seeking to better understand biologic and pre-analytical variability, including standardization of laboratory tests. Of note, since the 1980s Joseph Henny has been actively involved in the Société Française de Biologie Clinique (SFBC). He has also occupied the post of treasurer and general secretary of the SFBC. He is an active member of several national scientific societies and member of several working groups and committee at national and international level. He was a full member of the Reference Intervals and Decision Limits Committee (C-RIDL) of the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC-LM).