Proficiency Testing scheme interlinkage and international laboratory anchorage

O. Leray

France Génétique Elevage, 149 rue de Bercy, F-75595 Paris cedex 12, France

Abstract

The quality of milk recording analytical data is tightly related to the quality of the reference values used by laboratories to calibrate routine testing methods. However the reference methods used by routine laboratories allow differences occur between laboratories within reproducibility ranges. Laboratories's biases as shown in interlaboratory proficiency testing (PT) studies can appears not negligible. Evidence is given that the reference values of a national laboratory network is valid for the group itself but can differ significantly from the reference of other networks depending on analytical methods and PT study conditions. For the sake of the analytical harmonisation within ICAR a general model for PT scheme interlinking is proposed. Laboratory comparison between different independent PT schemes are made possible as well as evaluating lab performance against an international reference as for instance given by ICAR reference laboratory network. Preliminary cautions prior to implementation are given and plea for harmonisation in PT organisation made. Examples from national and international PT trials using somatic cell counting data are presented.

Keywords: milk recording, laboratories, harmonisation, quality assurance, reference, anchorage, network

Introduction

Several hundred millions animal milk samples are analysed every year in milk recording laboratories in the ICAR world. This is rendered possible only by using automated rapid methods mostly based on mid infrared spectroscopy for milk composition and fluoro-opto-electronic methods for somatic cell content. To obtain reliable results these methods so-called routine methods must be regularly calibrated against reference methods and be submitted to regular quality control according to standards and guidelines at the level of the laboratory.

These two last decades have seen a huge internationalization of animal genetic trade (semens, embryos, animals) and genetic evaluation for which the quality of animal performance measurement is of utmost importance so as to allow fair comparison between countries or organisations.

This has justified to ICAR to implement an analytical quality assurance (AQA) system to assure equivalence within and between member organisations thus provide confidence to stakeholders. This system was described in 1994, launched in 1996 then developed progressively till today. It relies on harmonisation of laboratory practices and methods used by laboratories and providing the mean of evaluating laboratory performances within an international reference laboratory network.

The key points of the system are that, through regular international proficiency testing schemes, it allows the reference laboratories

- to evaluate own individual precision and accuracy (bias) for reference methods used against an absolute reference defined by consensus as the average of the results of the group of participating reference laboratories,
- to provide the routine laboratories they monitor with the accuracy traceability to the ultimate truth and the possibility to estimate the real overall uncertainty in the global ICAR system.

Nevertheless if national or regional systems can function separately there is no indication on how effective is the analytical data harmonisation within ICAR. Hence this is the role and the responsibility of ICAR to develop and implement a method that permits to detect and possibly quantify discrepancies between national / regional systems.

Relativeness of reference results and choice of an international reference

A key issue of the ICAR AQA system relates to the so-called reference values used to assess both labs performance and routine methods calibration. Requirement is that they be obtained by internationally standardized reference methods or be standardized methods accurately anchored to the latter.

The exactness of an analytical method is relative since different results lying within standard limits can be obtained for a same milk sample using the same method. Indeed so-called precision figures – repeatability and reproducibility – have been statistically estimated through interlaboratory evaluation method studies according to ISO 5725 and constitute the range of normal observable performances and consequently fix limits for quality control :
- repeatability, r, the largest range not to be exceeded in 95% of cases between duplicates in identical analytical conditions (same laboratory, method, technician, within the closest time),
- reproducibility, R, the largest range not to be exceeded in 95% of cases between duplicates in different conditions (laboratory, device, technician, time) using the same method.

These values enable to establish limits in laboratory performance evaluation by PT and internal laboratory quality control.

From these values can be derived the respective 95% confidence intervals $\pm 2s_L$ and maximum observable ranges $L = 2.8\ s_L$ between two laboratory means thanks to the relationship $s_L^2 = s_R^2 - s_r^2$ (Table 1). Those limits indicate possible not negligible difference between laboratories that are then spread over the whole of routine labs through calibration.

Table 1. Range L and limits +/-2 sL *for laboratory means derived from standard r and R according to ISO 5725.*

Component	ISO	sR	sr	sL	R	r	L	+/-2 sL
Fat	1211	0,02	0,014	0,014	0,056	0,039	0,040	0,029
Protein	8968	0,018	0,014	0,011	0,050	0,039	0,032	0,023
SCC rel (750.10^3 c/ml)	13366-2	6%	3%	5%	17%	8%	15%	10%

Such L or +/-2 sL ranges of possible (accepted) occurrence cannot be found acceptable for every use with regard to the trade value of components which makes critical and questionable the traditional in-house calibration and would plea for more collaboratively obtained reference through centralized calibration (O Leray 2008).

Indeed numerous figures observed in interlaboratory proficiency studies illustrate that still larger biases can occur (Figure1). The same example shows also that in the alternative of centralized calibration the calculated reference (averages) of different groups of laboratories, may differ significantly (e.g. national vs international labs labs larger than 0.01 % protein) whereas in that example ICAR reference provides appropriately the more suitable central position for the reference. Similarly large discrepancies (relative mean biases up to 5-8%) are regularly observed between such groups in somatic cell counting proficiency studies.

Moreover beside reference methods, so-called "secondary reference methods" are permitted by ICAR provided tight anchorage to the international reference methods. For instance case is for fat by the butyrometric method (Gerber) for which national standards exist but no international standards. Ways to relate to ISO 1211 may vary significantly and induce different trueness depending on countries or region (pipette volume, butyrometer calibration, reagents).

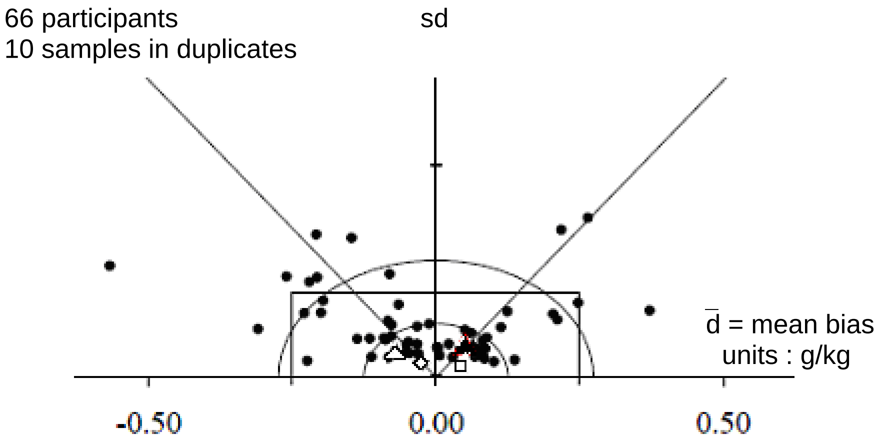

Figure 1. Example of lab score distribution in proficiency study for protein by Kjeldahl method (square □ = country labs ; triangle △= foreign labs ; diamond ◇ = ICAR ref labs)

The here above example and many others illustrated by PT studies, supported by the fact that PT scheme conditions vary and may bias the proper estimation of the reference (unequal and irregular participants numbers, different instruments, calibration material, reagent suppliers), plea for the choice of a unique reference defined at the international level by consensus.

Comparing and assessing laboratories on a same unique scale

The ICAR AQA system requires routine laboratories to participate in local proficiency testing (PT) schemes in their countries or regions so as to evaluate and improve performance if needed. In such studies the reference are calculated as the mean of participating laboratories after excluding abnormal (outlier) results. As a result routine laboratories of difference region (different PT scheme) cannot compare to same reference and it happens that two good performing labs of distinct regions do not agree on results when analysing same samples. This can stem from possible differences (see above) in the reference calculated in the respective scheme but it is unknown how similar or close they are if there is no liaison between scheme and no physical link implemented to measure reference similarity.

O Leray (2008) presented a method of PT scheme linkage and anchorage through the common participation of a laboratory in the both trials between which comparisons are wished. The methods addressed single level comparison and applicable by extension to reference methods where there is theoretically no level effect on the bias (lab-reference) which is supposed constant throughout the concentration range (e.g. ISO 1211 for fat). For other methods where a relationship can exist between the bias and the level (e.g. ISO 13366-2 for SCC) need is for another method. Here is described a general method that it is aimed to apply in the on-going joint IDF-ICAR project of Somatic Cell Counting Reference System (Baumgartner & Bijgaart, 2010).

The general principle consists in a way of calibration of the PT scheme(s) to that one chosen as the ultimate reference. To achieve that, a laboratory participates in each of the trials with analysing the samples of each scheme and it is understood that there is no change happened in the method setting it uses in both sample tests performing. Then a relationship F_A allows to predict the reference of scheme A using the results X of the liaison laboratory with the sample set of scheme A and similarly another relationship F_B to predict the reference of scheme B using the results Z of the liaison laboratory with the sample set of scheme B.
Combining the two relationships as $F_B \circ F_A^{-1}(x) = F_B[F_A^{-1}(x)]$ permits to align the reference of scheme A on reference of scheme B then to used the so-calibrated reference as a new reference for a virtual performance assessment. Where the international ICAR scheme is used for scheme B one can speak of a virtual international evaluation.

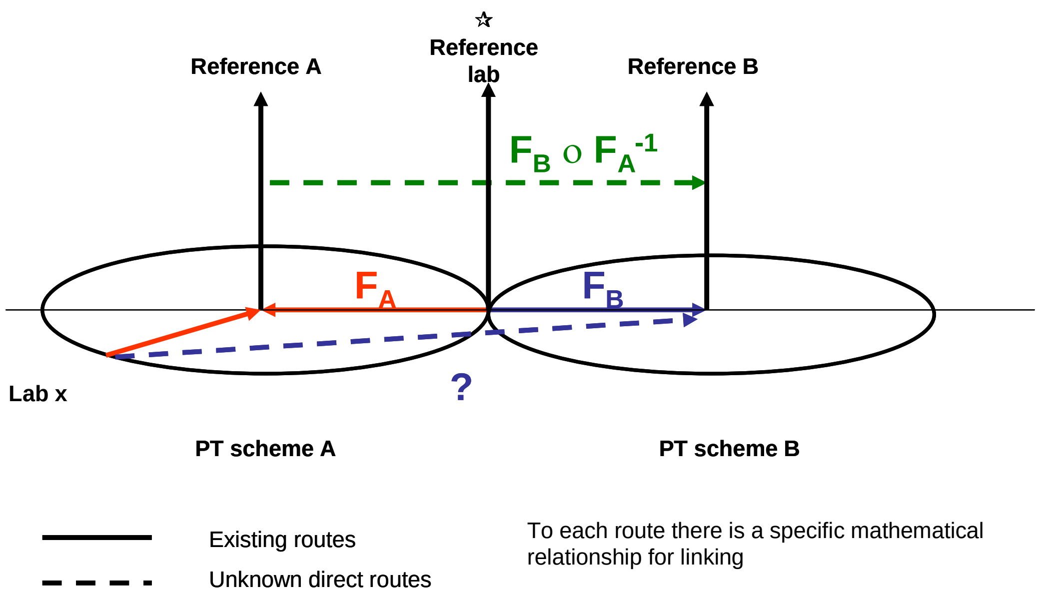

Figure 2. Bridging between two PT schemes by a liaison laboratory (★) and reference alignment of Scheme A to Scheme B.

The nature of the relationships F_A and F_B can be defined by prior polynomial regression optimization through minimizing the regression residual standard deviation before calculating the resulting combination $F_B \circ F_A^{-1}$. In most cases - where there are sufficient participants and but few limited linearity defects in lab data - a simple linear relationship can suffice.

Example of model using linear equation :

F_A : $y = x \cdot b_A + a_A$ $\Leftrightarrow$ F_A^{-1}: $x = y/b_A - a_A/b_A$

F_B : $z = x \cdot b_B + a_B$

$F_B \circ F_A^{-1}$: $z = (y/b_A - a_A/b_A) \cdot b_B + a_B$

$$z = y \cdot (b_B/b_A) + (a_B - a_A.b_B/b_A) \qquad (1)$$

with x the values of the reference lab and y the assigned reference values in Scheme A and z the assigned reference values in Scheme B

Once established equation 1 is used to predict the virtual reference in Scheme B for the samples used in Scheme A (different from those of Scheme B) and the virtual reference values can serve to calculate the virtual scores of laboratories of Scheme A in Scheme B (see Figures 4 and 5).

Where there is no level effect on the bias – i.e. slopes b_A and b_B equal to 1 – the relation (1) simplifies in an addition of a constant term as $z = y + (a_B - a_A)$ as described by O Leray (2008).

Example of application

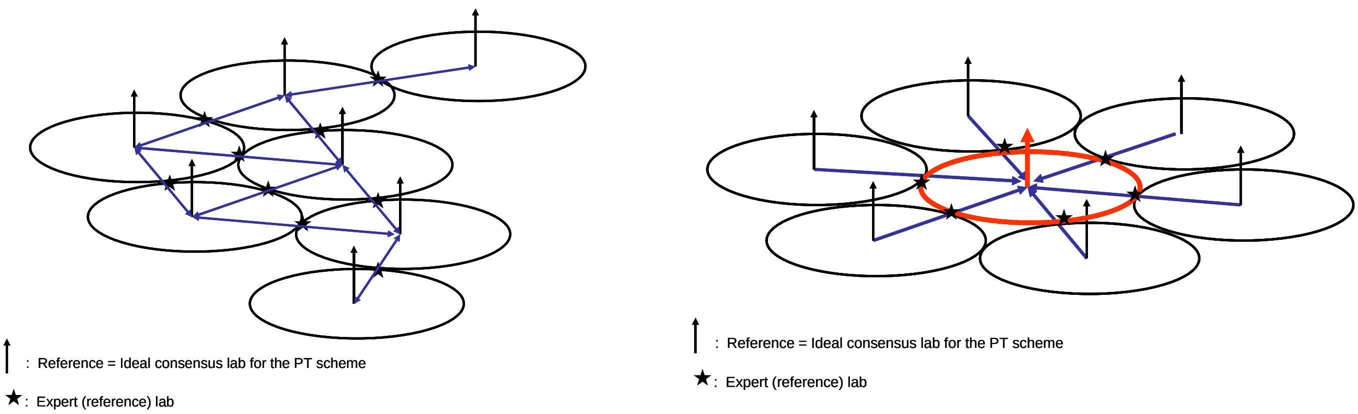

Figure 3. Two-by-two interlinking of PT schemes in a chain or a network (left) ; PT scheme anchorage to a central PT scheme as a model for an international anchorage and traceability (right).

The model of bridging of Figure 2 can be implemented at the level of national / regional schemes in the frame of multi- lateral comparison for instance to resolve local disputes (Figure 3 on left), or as the way to anchor the national / regional PT schemes to a single central scheme that can provide the commonly accepted truth (Figure 3 on right). This is the latter model proposed and developed by the ICAR through its reference laboratory network. However both models can be used complementarily in dedicated reference systems such as the on-going IDF-ICAR project of Reference System for Somatic Cell Counting.

First test application within ICAR

Virtual scoring was made using equation 1 for somatic cell counting in the frame of the ICAR reference laboratory network and PT trial of 2009 (Figure 4). Scores are represented in abscissa as the mean laboratory differences to the reference whereas standard deviation of differences are reported in ordinate (IDF 1999). The reference laboratory of a national PT scheme showed its score (N) reduced in a virtual score close to zero which was then better in line with its score obtained in the international PT trial organised by ICAR. The larger standard deviation of (I) relates to the concentration range of the international significantly higher than in the national PT scheme.

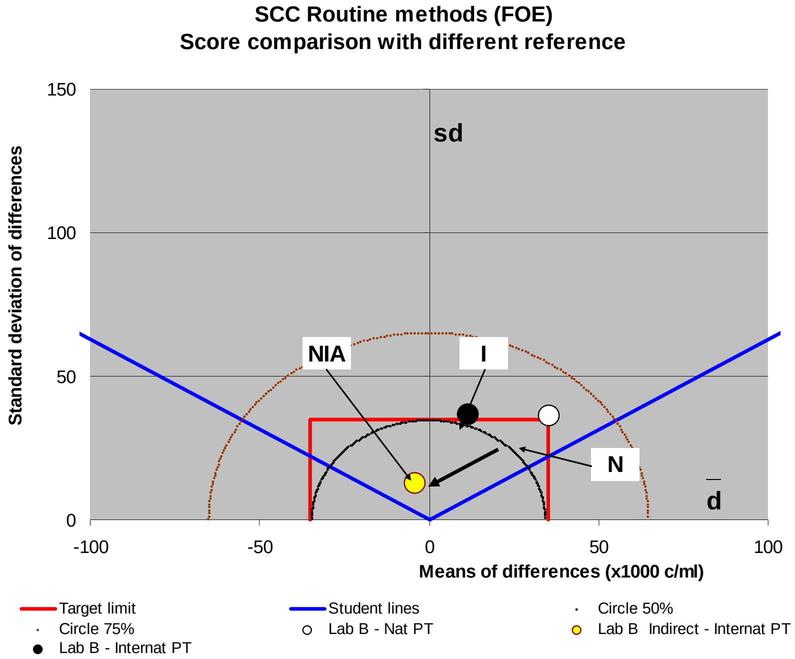

Figure 4. Scores diagram of a reference laboratory in national PT scheme (N), international ICAR PT scheme (I) and national PT scheme internationally aligned (NIA).

Applying the virtual reference to all the laboratories (Figure 5, right) resulted in a shift of the lab population to the left (underestimation) and a reduction of standard deviations in the median part of the diagram. This can be interpreted as a overall slope modification due to either a biased reference or a difference of the reference laboratory performance between national and international schemes. If confidence is given to the results of the reference laboratory both pictures can comfort each other to indicate adequate diagnosis such as possible troubles with PT samples quality or improper calibration materials reflected in a general negative trend. Other statistics such as the residual standard deviation of regressions and quality control records should then confirm.

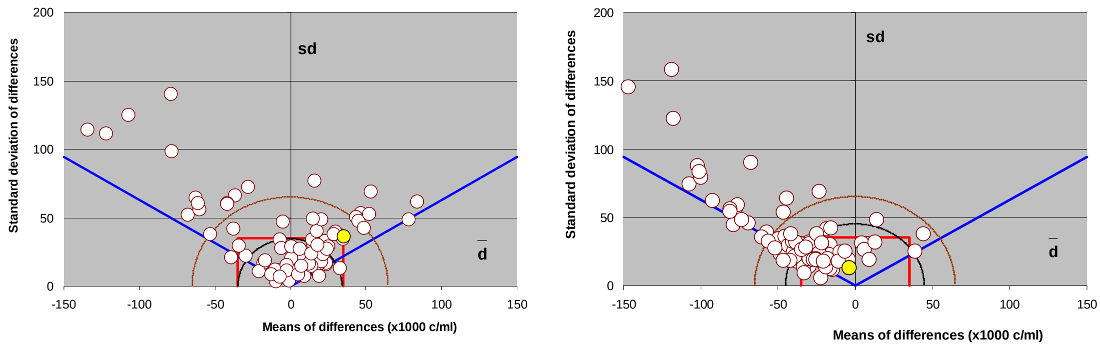

Figure 5. Scores diagrams of participants in national PT scheme : actual as compared to the national reference (left), virtual as compared to the international ICAR reference.

The virtual reference calculation was applied similarly to different PT schemes anchored to ICAR trials in order evaluate the possibility to proceed to a global lab evaluation. Theory was verified on that optimal population centring is obtained onto the average and any national reference not conforming to the international reference would result in larger lab score scattering in the virtual evaluation. Figure 6 illustrates the positions of sub-groups of 20 milk recording laboratories of each scheme before and after national reference alignment onto the international reference of ICAR.

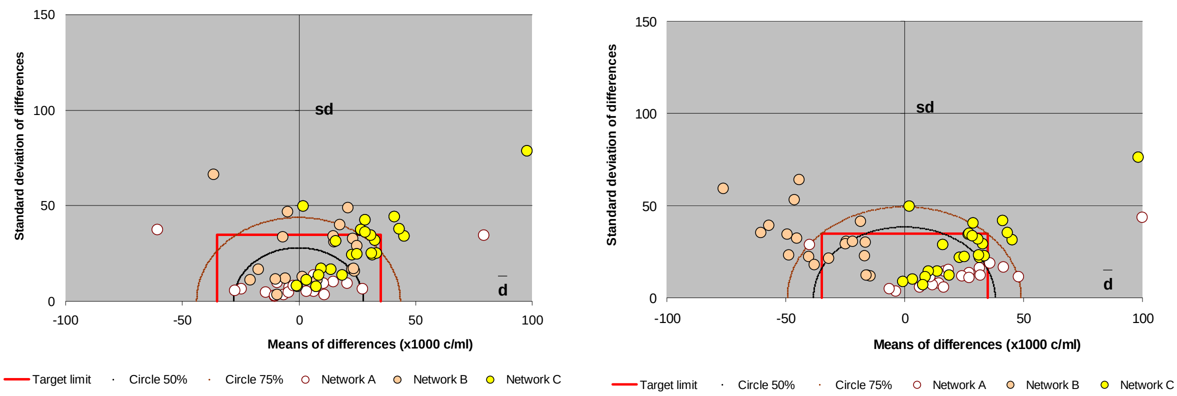

Figure 6. Example of evaluation of laboratories from different national PT scheme : actual as compared to the national reference (left), virtual as compared to the international ICAR reference.

In that example experimental designs (sample types, sample and replicate numbers, concentration ranges) were significantly different between PT schemes thus showing different population pictures with regard to limits stated.

Constraints and requirements

Applying such a system relies on assuring

1- absolute stability of the analytical method used by the reference laboratory during the time gap between the two sample sets testing. This is obtained - either through simultaneous testing in a same test series, subsequently or better by alternating samples of each series, - or linking test run of each sample sets through a significant quality control net with adequate reference materials.

2- the lowest uncertainty of virtual reference estimates through the equation chaining. To achieve this the sample number and the concentration range of the reference PT scheme (Scheme B) should be larger (at least equivalent) to those of the tested PT scheme (Scheme A). The highest correlation of the reference lab with the assigned values of both schemes should be achieved.

Implementing such a system within ICAR would require the former constraints be laid down in appropriate guidelines and a harmonised PT organisation protocol for ICAR be described in line with the ISO 8196 for calibration and ICAR guidelines so as to make PT scheme comparables with regard to performance evaluation.

Conclusion

The interconnection of PT schemes and international anchorage are technically possible for every type of methods thanks to the principle here above developed. The PT scheme interlinking system enables comparing different PT schemes and laboratories of different PT schemes and move forward to harmonisation. It enables to assess national PT schemes through international anchorage to the reference PT schemes organised for the ICAR reference laboratory networks.

Users' awareness is needed on particular cautions necessary to be taken with regard to proper virtual reference estimation and that, all in all, fair comparisons between independent PT schemes can only result from the harmonisation of PT organisation protocols within ICAR.

List of References

Baumgartner, C., van den Bijgaart, H. 2010. International reference system for somatic cell counting in milk - A world wide challenge. Proceedings of the 37th ICAR Biennial Session, Riga (Latvia), 31 May-4 June 2010. ICAR technical Series No.14. 271-276.

International Dairy Federation, (1999). Quality Assurance and Proficiency Testing. Report of Group E29. Bulletin of IDF 342/1999, 23-30.

Leray, O., 2008. ICAR AQA strategy – International anchorage and harmonisation. Fourth ICAR Reference Laboratory Network Meeting. Proceedings of the 36th ICAR Biennial Session, Niagara Falls (USA), 16-20 June 2008. ICAR technical Series No.13. 295-300.

Leray, O., 2008. Interlaboratory reference systems and centralised calibration – Prerequisites and standard procedures. Fourth ICAR Reference Laboratory Network Meeting. Proceedings of the 36th ICAR Biennial Session, Niagara Falls (USA), 16-20 June 2008. ICAR technical Series No.13. 301-305.