Accred Qual Assur (2016) 21:175–183
DOI 10.1007/s00769-016-1207-y

GENERAL PAPER

Probabilistic comparison and assessment of proficiency testing schemes and laboratories in the somatic cell count of raw milk

Thomas F. H. Berger[1] · **Werner Luginbühl**[2]

Received: 11 December 2015 / Accepted: 31 March 2016 / Published online: 29 April 2016

Abstract The somatic cell count (SCC) of milk is one of the main indicators of the udder health status of lactating mammals and is a hygiene criterion of raw milk used to manufacture dairy products. An increase in SCC is regarded as one of the primary indicators of inflammation of the mammary gland. Therefore, SCC is relevant in food legislation as well as in the payment of ex-farm raw milk and it has a major impact on farm management and breeding programs. Its determination is one of the most frequently performed analytical tests worldwide. Routine measurements of SCC are almost exclusively done using automated fluoro-opto-electronic counting. However, certified reference materials for SCC are lacking, and the microscopic reference method is not reliable because of serious inherent weaknesses. A reference system approach may help to largely overcome these deficiencies and help to assure equivalence in SCC worldwide. The approach is characterised as a positioning system fed by different types of information from various sources. A statistical approach for comparing proficiency tests (PTs) by assessing them using a quality index P_Q and assessing participating laboratories using a quality index P_L, both deriving from probabilities, is proposed. The basic assumption is that PT schemes are conducted according to recognised guidelines in order to compute performance characteristics, such as z-scores, repeatability and reproducibility standard deviations. Standard deviations are compared with the method validation data from the ISO method. Input quantities close to or smaller than the reference data of the method validation or the assigned value of the PT result in values for P_Q and P_L close to the maximum value. Evaluation examples of well-known PTs show the practicability of the proposed approach.

Keywords Reference system · Somatic cell count · Proficiency testing · Statistical approach · Quality index

Introduction

The somatic cell count (SCC) of milk is one of the main indicators of the udder health status of lactating mammals and one of the hygiene criteria of raw milk used to manufacture dairy products. Somatic cells excreted through milk include various types of white blood cells and some epithelial cells. Its composition and concentration change dramatically during periods of inflammation. An increase in SCC is therefore regarded as one of the primary indicators of inflammation of the mammary gland [1]. Therefore, SCC is relevant in food legislation [2–4], in the payment of ex-farm raw milk serving as a price setting quality parameter; when measured in individual animals, it also has a major impact on farm management and breeding programs. Consequently, somatic cell count determination is one of the most frequently performed analytical tests in dairy laboratories worldwide, with an estimated more than 500 000 000 tests per year [5].

Electronic supplementary material The online version of this article (doi:10.1007/s00769-016-1207-y) contains supplementary material, which is available to authorized users.

✉ Thomas F. H. Berger
thomas.berger@agroscope.admin.ch

Werner Luginbühl
info@chemstat.ch

[1] Agroscope, Institute for Food Sciences (IFS), Schwarzenburgstr. 161, 3003 Bern-Liebefeld, Switzerland

[2] ChemStat, Aarstrasse 98, 3005 Bern, Switzerland

SCC data for routine measurements are nowadays almost exclusively obtained through the application of automated fluoro-opto-electronic counting. Guidance on this application is available through ISO 13366-2 | IDF 148-2 [6]. Part of the guidelines focus on calibration and calibration control; however, certified reference materials (CRM) for SCC are lacking. Laboratories therefore calibrate with 'secondary' reference materials, which are types of milk, more or less well defined in its properties, using assigned 'reference values' for counting. These reference values may derive from the application of the reference method, which is a direct microscopic SCC, according to ISO 13366-1 | IDF 148-1 [7], often in combination with the results of automated counting. Routine testing laboratories usually rely on these secondary reference materials and their assigned values. Others base their calibration on the performance in proficiency tests (PTs), and some rely on the standard settings of the instrument manufacturer. The reasons for lack of full reliance on the microscopic reference method are an insufficient definition of the measurand and a poor precision [5]. To overcome the large uncertainty of the microscopic reference method, reference material providers can additionally rely on a set of routine measurement data, often coming from a selected group of laboratories. However, such reliance bears the risk of circular calibration [8, 9]. If at least a part of the participating laboratories do not also rely on other PTs, they may start correcting their instruments to the assigned value, and an undefined drift within the large uncertainty of the reference method begins. The existing PTs therefore need to be interlinked based on a quantitative scale. At this juncture, there is no 'true' value to assess the competence of a laboratory.

A reference system approach may help to largely overcome these deficiencies and help to assure equivalence in somatic cell counting worldwide. A reference system is characterised as a positioning system fed by different types of information from various sources—that is, from reference materials, reference method analysis, routine method results and PT results of laboratories operating in a laboratory network structure [10].

The purpose of this work is to propose a statistical approach for comparing PTs by assessing them using a quality index P_Q and assessing participating laboratories using a quality index P_L, both deriving from probabilities. The approach was developed in the framework of the SCC Reference System Working Group (International Dairy Federation [IDF] and the International Committee on Animal Recording [ICAR] [5, 10]) by the participating organisations. The basic assumption is that the PT schemes are conducted according to recognised guidelines such as the Harmonized Protocol [11] and ISO 13528 [12] or ISO 5725 [13] in order to compute performance characteristics such as z-scores, repeatability and reproducibility standard deviations. The existence of a CRM (as an estimate of a 'true value') is not required in the following considerations. The situation is comparable to the summarising assessment of medical and similar studies, where meta-analysis is a well-proved tool using variances and frequencies for weighting and as objective criteria. However, given the fact that reliable estimates of the population variances are available (see below), we preferred to develop a probabilistic approach.

Method

Assessing PTs by a quality index P_Q derived from probabilities

This approach makes use of the precision parameters repeatability standard deviation σ_r and reproducibility standard deviation σ_R of automated fluoro-optic SCC measurement as reported in the international standard ISO 13366-2 | IDF 148-2 [6].

Assume that in a given PT the estimates s_r and s_R (or the standard deviation between laboratories, s_L) of the repeatability and reproducibility standard deviations, σ_r and σ_R, respectively, are computed (for one level) using the results from p laboratories. Each laboratory measures the test material n times. Then, a quality index P_Q based on the probabilities derived from Chi-square distributions can be constructed.

From standard statistical results, the following equation relating the estimated and the population repeatability variances with the Chi-square distribution with v degrees of freedom holds for normally distributed measurements (see also ISO 5725-4 [13]):

$$\hat{\chi}^2_{(r)} = \frac{v s_r^2}{\sigma_r^2} \sim \chi^2_v \qquad v = p(n-1), \tag{1}$$

and similarly

$$\hat{\chi}^2_{(R,r)} = \frac{v\left(s_R^2 - \left(1-\frac{1}{n}\right)s_r^2\right)}{\sigma_R^2 - \left(1-\frac{1}{n}\right)\sigma_r^2} \sim \chi^2_v \qquad v = p-1, \tag{2}$$

which by $s_L^2 = s_R^2 - s_r^2$ is the same as

$$\hat{\chi}^2_{(L,r)} = \frac{v\left(s_L^2 + \frac{s_r^2}{n}\right)}{\sigma_L^2 + \frac{\sigma_r^2}{n}} \sim \chi^2_v \qquad v = p-1. \tag{3}$$

Therefore, we can estimate the probabilities $P_{(r)}$ and $P_{(L,r)}$:

$$P_{(r)} = P\left(\chi^2_v > \hat{\chi}^2_{(r)}\right) = 1 - P\left(\hat{\chi}^2_{(r)}\right) = 1 - P\left(\frac{v s_r^2}{\sigma_r^2}\right) \tag{4}$$

$$P_{(\mathrm{L,r})} = P\left(\chi_v^2 > \hat{\chi}^2_{(\mathrm{L,r})}\right) = 1 - P\left(\hat{\chi}^2_{(\mathrm{L,r})}\right) = 1 - P\left(\frac{v\left(s_\mathrm{L}^2 + \frac{s_\mathrm{r}^2}{n}\right)}{\sigma_\mathrm{L}^2 + \frac{\sigma_\mathrm{r}^2}{n}}\right). \quad (5)$$

The known variances σ_r^2 and σ_L^2 are derived from the values of σ_r and σ_R, as published in standard ISO 13366-2 | IDF 148-2 [6].

$P_{(\mathrm{r})}$ and $P_{(\mathrm{L,r})}$ may then be combined to define the PT quality index P_Q as the product of these probabilities:

$$P_\mathrm{Q} = P_{(\mathrm{r})} P_{(\mathrm{L,r})}. \quad (6)$$

P_Q can be (approximately) interpreted as an estimate of the probability that the set of p laboratories within the PT can achieve a repeatability standard deviation as small as σ_r and simultaneously a standard deviation between laboratories as small as σ_L.

If the reference value θ of the test material is known, or the assigned value θ is accepted as reliable, then the z-scores (based on an accepted standard deviation for proficiency assessment, σ_p [11]) of the p laboratories can be combined. To reduce the influence of extreme z-score values, a robust mean estimator $\bar{z}_{(\mathrm{rob})}$ according to Huber is necessary, known as A15 (without an iterative update of the robust estimation of the standard deviation) or as 'Huber proposal 2', or H15 (with an iterative update of the robust estimation of the standard deviation) (Algorithm A, described in Annex C [12]), [14, 15]. The robust sum of z-scores is therefore

$$Z_p = p \cdot \bar{z}_{(\mathrm{rob})}, \quad (7)$$

and a probability $P(Z_p)$ for $Z \cdot \sqrt{p}$ larger than $|Z_p|$ may be derived on the basis of the realisation $\hat{Z}$ of the standard normal random variable Z, i.e. $\hat{Z} = Z_p / \sqrt{p} \sim N(0,1)$:

$$P(Z_P) = 2P\left(Z > \left|\hat{Z}\right|\right) = 2P\left(Z > \frac{|Z_p|}{\sqrt{p}}\right) = 2\left(1 - \Phi\left(\frac{|Z_p|}{\sqrt{p}}\right)\right), \quad (8)$$

where $P(\cdot)$ stands for probability and $\Phi(\cdot)$ indicates the distribution function of the standard normal distribution.

An alternative combination of z-scores is possible because the sum S_p of the squared z-scores is Chi-square distributed with p degrees of freedom [11]: $S_p = \sum_{i=1}^{p} z_i^2 \sim \chi_p^2$.

The quality index P_Q has three components in this case: two are related to precision measures and one is related to the trueness of the p mean values.

$$P_\mathrm{Q} = P_{(\mathrm{r})} P_{(\mathrm{L,r})} P(Z_p) \quad (9)$$

It is still possible to modify this quality measure by multiplication with a further expression (factor) $q = f(q_1, q_2, q_3, \ldots, q_m)$ made up of the PT-specific quality indices $q_1, q_2, q_3, \ldots, q_m$ to obtain

$$P_\mathrm{Q} = P_{(\mathrm{r})} P_{(\mathrm{L,r})} P(Z_p) q. \quad (10)$$

The m quality indices $q_{i1}, q_{i2}, q_{i3}, \ldots, q_{im}$ may be used to model m PT_i characterising criteria. The components of $q_i = f(q_{i1}, q_{i2}, q_{i3}, \ldots, q_{im})$ could be defined in such a way that higher values in the resulting q_i indicate higher quality.

To compare up to k PTs in such a way, it may be better to compute normalised values, especially if the P_Q values were calculated according to Eq. (10):

$$\tilde{P}_{\mathrm{Q},i} = \frac{P_{\mathrm{Q},i}}{\sum_{j=1}^{k} P_{\mathrm{Q},j}}. \quad (11)$$

Comparing PT schemes over time based on the quality index P_Q or its elements

There are various possibilities to construct quality control charts for a given PT scheme.

The following quality or performance characteristics may be plotted versus the number of rounds, 1, 2, …, t:

- s_r or s_r^2 or $\hat{\chi}^2_{(\mathrm{r})}$ or $P_{(\mathrm{r})}$
- s_L or s_L^2 (or s_R or s_R^2) or $\hat{\chi}^2_{(\mathrm{L,r})}$ or $P_{(\mathrm{L,r})}$
- Z_p or $P(Z_p)$
- P_Q
- the fraction of 'satisfactory' z-scores, i.e. $|z| \leq 2$, as proposed by Gaunt and Whetton [16].

The sums or cumulative averages of these characteristics over t rounds may be used as numerical indices to compare PT schemes quantitatively over time.

Assessing laboratories by a quality index P_L derived from probabilities

Again, this approach makes use of the precision parameters repeatability standard deviation σ_r and reproducibility standard deviation σ_R of automated SCC measurements, as reported in the international standard ISO 13366-2 | IDF 148-2 [6].

Assume that the values of σ_r and σ_R, as published in standard ISO 13366-2 | IDF 148-2 [6], are known and that an accepted reference value θ has been established.

A single laboratory within a PT can be rated similar to the rating shown above if it provides a repeatability standard deviation s_r and a mean value $\bar{y}$ of n replicates at a given level (estimates of s_r and $\bar{y}$ for σ_r and θ, respectively).

With

$$\hat{\chi}^2_{(\mathrm{r})} = \frac{v\, s_\mathrm{r}^2}{\sigma_\mathrm{r}^2} \sim \chi_v^2, \quad v = n - 1 \quad (12)$$

we can estimate the probability $P_{(r)}$

$$P_{(r)} = P\left(\chi_v^2 > \hat{\chi}_{(r)}^2\right) = 1 - P\left(\hat{\chi}_{(r)}^2\right) = 1 - P\left(\frac{vs_r^2}{\sigma_r^2}\right). \quad (13)$$

The difference $\bar{y} - \theta$, standardised by $\left[\sigma_R^2 - \left(1 - \frac{1}{n}\right)\sigma_r^2\right]^{\frac{1}{2}}$, is a standard normal variate:

$$\tilde{z}_n = \frac{\bar{y} - \theta}{\left[\sigma_R^2 - \left(1 - \frac{1}{n}\right)\sigma_r^2\right]^{\frac{1}{2}}} \sim N(0,1), \quad (14)$$

which is used to compute the probability

$$P(\tilde{z}_n) = 2P(Z > |\tilde{z}_n|) = 2(1 - \Phi(|\tilde{z}_n|)). \quad (15)$$

$P_{(r)}$ and $P(\tilde{z}_n)$ may be combined to define the laboratory quality index P_L as the product of these probabilities:

$$P_L = P_{(r)}P(\tilde{z}_n). \quad (16)$$

P_L can be (approximately) interpreted as an estimate of the probability that a certain laboratory having participated in a PT can achieve a repeatability standard deviation as small as σ_r and simultaneously a difference between the assigned value of the PT θ and its own mean value $\bar{y}$ as small as the standard deviation between laboratories σ_L.

Again, it is possible to modify this quality measure by multiplication with a further expression (factor) $q = f(q_1, q_2, q_3, \ldots, q_m)$ made up of the laboratory-specific quality indices $q_1, q_2, q_3, \ldots, q_m$ to obtain

$$P_L = P_{(r)}P(\tilde{z}_n)q. \quad (17)$$

The components $q_{i1}, q_{i2}, q_{i3}, \ldots, q_{im}$ of q_i should be defined in such a way that higher values in the resulting q_i indicate higher quality.

A normalised quality index $\tilde{P}_{L,i}$ may be preferred to compare a set of p laboratories, especially if the P_Ls were calculated according to Eq. (17):

$$\tilde{P}_{L,i} = \frac{P_{L,i}}{\sum_{j=1}^{p} P_{L,j}}. \quad (18)$$

Comparing laboratories over time based on the quality index P_L or its elements

There are various possibilities to construct quality control charts for a given laboratory (see also ISO 13528 [12]). The following quality or performance characteristics may be plotted versus the number of rounds, 1, 2, …, t:

- s_r or s_r^2 or $\hat{\chi}_{(r)}^2$ or $P_{(r)}$
- $\tilde{z}_n$ or $P(\tilde{z}_n)$ (or z-scores as reported by the PT provider)
- P_L
- the fraction of 'satisfactory' z-scores, i.e. $|z| \leq 2$, as proposed by Gaunt and Whetton [16].

The sums or cumulative averages of these characteristics over t rounds may be used as numerical indices to compare laboratories quantitatively.

Data

For the testing of the assessment schemes for PTs and laboratories using the probabilistic approach, the data from five national and international PTs were chosen (see Table 1). The PTs took place between September 2010 and October 2011. The data sets were well known, meaning that the evaluation had been finished and feedback had been received.

Each level of a PT was handled as an individual comparison. PTs and laboratories were anonymised, and, where known, the multiple participations of a certain laboratory were each handled as an individual participant.

An Excel® spreadsheet was used for the evaluation. Firstly, the data of the different PTs and levels were arranged according to the necessary information, which included laboratory labels/codes (and the instrument type, if known), number of replicates n, mean values $\bar{y}$ as reported by the laboratories, repeatability and reproducibility standard deviations s_r and s_R of the laboratories and reference values (consensus or 'true' values) θ as well as the s_r of the PT or PT level. Additionally, the robust sum of the z-scores was calculated according to Eq. (7).

Secondly, the quality indices P_Q (assessing PTs) were calculated by inserting the data into the specific Excel® spreadsheets. Additionally, the population repeatability standard deviations σ_r and the population reproducibility standard deviations σ_R from ISO 13366-2 | IDF 148-2:2006 [6] had to be implemented. As the reference values θ are mostly between the published values in the ISO IDF standard, an interpolation table was used to calculate the relevant σ_r and σ_R. ISO 13366-2 | IDF 148-2:2006 [6] mentions, e.g. for the levels of 150 000 SCC/mL and 300 000 SCC/mL repeatability values of 6 % and 5 % and reproducibility values of 9 % and 8 %, respectively. For a reference value of 162 000 SCC/mL a s_r of 5.92 % or 9 590 SCC/mL and a s_R of 8.92 % or 14 450 SCC/mL were interpolated. Quality indices $q_1 \ldots q_m$, as proposed in Eq. (10), were not used because thus far no considerations of the characters and values of the factors have taken place. Therefore, the weight w for the difference $1 - q$ is of no meaning. The upper part of Fig. 1 shows a calculation example (with p being the number of laboratories participating in the PT).

Thirdly, the quality indices P_L (assessing the laboratories) were calculated by inserting the data in the specific Excel® spreadsheets. Additionally, the population repeatability standard deviations σ_r and the population reproducibility standard deviations σ_R from ISO 13366-2 | IDF 148-2:2006 [6] had to be implemented. As mentioned above, for the calculation of the quality indices P_Q for the PTs, an interpolation table is needed to calculate the relevant σ_r and σ_R. Again, a weight of $w \in [0,1]$ for the

Table 1 PTs used for the calculation of the quality indices P_Q and P_L

Name	Organiser	Date	No. of levels	No. of participating laboratories
AIA Isl	Associazione italiana allevatori (AIA), Laboratorio Standard Latte (http://www.aia.it/lsl)	March 2011	6	27
Characterisation of Agroscope SCC Standard	Agroscope, Institute for Food Sciences (http://www.agroscope.ch)	September 2010	2	21
Characterisation of Agroscope SCC Standard	Agroscope, Institute for Food Sciences (http://www.agroscope.ch)	March 2011	2	21
Cornell	Cornell University, Department of Food Science (http://foodscience.cals.cornell.edu/extensior/dairy-milk-products)	October 2011	8	8
ICAR	Actalia-Cecalait (http://www.cecalait.fr: http://www.icar.org/pages/Sub_Committees/sc_milk_laboratories.htm)	September 2011	10	15

difference $1 - q$ could be chosen, but, as mentioned above, thus far no considerations of the characters and values of the factors have taken place. Figures 2 and 3 show graphical evaluation and calculation examples. In addition to the evaluation of the participating laboratories in a specific PT by calculating the individual quality indices P_L, it is also possible to calculate, for example, the median quality indices from different PTs in order to have an indicator regarding the comparability of a certain laboratory or instrument over time and in different PTs (see Fig. 4).

Discussion

P_Q and P_L are influenced by their input variables. The three variables and performance characteristics z-score, repeatability and reproducibility standard deviations are calculated according to recognised standards, and they are compared with the specific method validation data from the ISO standard. It follows that input quantities close to or smaller than the reference data of the method validation or the assigned value of the PT result in values for P_Q and P_L close to the maximum value of 1.

The outcome of a PT is influenced by the competence of the participating laboratories. If the laboratories perform well and the overall repeatability s_r of p laboratories is close to or even smaller than σ_r of the standard, then the probability $P_{(r)}$ and the quality index P_Q of the concerned PT or PT level become larger or close to the maximum value of 1 (solid circle in Fig. 1, PT no. 6). Otherwise, if a larger part or most of the laboratories show a poor performance and s_r therefore is larger than σ_r, the probability $P_{(r)}$ and the index P_Q become smaller (dashed circle, PT no. 16). The same is true for P_Q and the probability related to the inter-laboratory standard deviation $P_{(L,r)}$, calculated from the PT's reproducibility s_R (solid and dashed circles in Fig. 1, PTs nos. 4 and 28). If the mean values of the laboratories in the PT are close to the assigned value, then the robust absolute sum of p z-scores $|Z_p|$ according to Eq. (7) becomes small, and the related probability $P(Z_p)$ and the index P_Q become large or close to the maximum value of 1 (solid circle, PT no. 26). For large values of $|Z_p|$, the probability $P(Z_p)$ and the index P_Q become small (dashed circle, PT no. 1). The summarising quality index P_Q is almost equally influenced by the probabilities $P_{(r)}$, $P_{(L,r)}$ and $P(Z_p)$ and therefore allows no conclusion on the PT's performance concerning the repeatability, inter-laboratory standard deviation and z-scores achieved by the participating laboratories.

Regarding the assessment of a laboratory, the influence of its repeatability s_r and the mean value of a laboratory $\bar{y}$ is shown in Figs. 2 and 3. If s_r is larger than σ_r, the probability related to the repeatability standard deviation $P_{(r)}$ becomes small as well as the corresponding quality index P_L. In cases where s_r is close to or smaller than σ_r, the opposite is true, and the probability $P_{(r)}$ as well as the quality index P_L become larger or close to the maximum value of 1. If the mean value $\bar{y}$ is larger or smaller than the reference value (consensus value, 'true' value) θ, then the absolute z-score $|\tilde{z}_n|$ becomes larger, and the related probability $P(\tilde{z}_n)$ as well as the corresponding quality index P_L become small. In cases where the mean value $\bar{y}$ is close or equal to the reference value θ, the absolute z-score $|\tilde{z}_n|$ becomes small, and the related probability $P(\tilde{z}_n)$ as well as the corresponding quality index P_L become large or close to the maximum value of 1. The summarising quality index P_L is almost equally influenced by the probabilities $P_{(r)}$ and $P(\tilde{z}_n)$ and therefore allows no conclusions on the laboratory's performance concerning repeatability and comparability to the assigned value (this differentiation is provided by the results of the PTs reported to the participants).

Quality indices P_L of laboratories or even of different instruments of a laboratory may be evaluated using, for example, control charts (value vs time) or statistical measures such as mean or median. In applying the test data, a

PT No	s_r	s_R	s_L	σ_r	σ_R	σ_L	p	n	Z_p (rob)	$\hat{\chi}^2_{(r)}$	$\hat{\chi}^2_{(L,r)}$	$P_{(r)}$	$P_{(L,r)}$	P_{Zp}	P_Q	$\tilde{P}_Q$
17	36.45	39.00	13.87	13.73	21.56	16.62	8	2	0.623	56.394	16.184	0.0000	0.023	0.826	0.000	0.0000
16	22.48	33.39	24.69	12.65	19.61	14.98	8	2	-0.289	25.275	19.821	0.0014	0.006	0.919	0.000	0.0000
20	3.00	7.00	6.32	3.33	4.83	3.50	15	2	0.740	12.174	35.031	0.6658	0.001	0.849	0.001	0.0001
14	16.87	29.47	24.17	15.38	24.78	19.43	8	2	0.643	9.620	10.254	0.2927	0.175	0.820	0.042	0.0051
19	12.00	73.00	72.01	29.52	59.04	51.13	15	2	0.037	2.479	24.130	0.9999	0.044	0.992	0.044	0.0053
1	8.60	10.26	5.60	12.62	19.56	14.94	22	15	9.403	143.031	3.253	1.0000	1.000	0.045	0.045	0.0054
15	13.55	34.52	31.74	16.86	28.09	22.47	8	2	0.906	5.170	11.897	0.7393	0.104	0.749	0.058	0.0069
12	12.62	39.70	37.64	18.67	33.01	27.22	8	2	1.178	3.658	11.446	0.8866	0.120	0.677	0.072	0.0087
25	14.00	52.00	50.08	22.35	44.10	38.02	15	2	-0.698	5.886	21.524	0.9816	0.089	0.857	0.075	0.0090
24	23.00	98.00	95.26	41.88	83.76	72.54	15	2	-0.051	4.524	21.300	0.9954	0.094	0.989	0.093	0.0112
3	10.28	19.80	16.92	14.67	23.35	18.17	21	15	6.265	144.369	17.041	1.0000	0.650	0.172	0.112	0.0135
2	12.82	15.99	9.56	18.40	32.39	26.66	22	15	6.280	149.504	2.929	1.0000	1.000	0.181	0.181	0.0218
28	3.28	9.13	8.52	5.99	8.81	6.46	14	2	-0.388	4.208	16.992	0.9941	0.200	0.917	0.182	0.0220
11	11.50	31.70	29.54	17.24	29.07	23.41	8	2	0.216	3.560	9.435	0.8945	0.223	0.939	0.187	0.0226
23	4.00	16.00	15.49	10.53	15.99	12.03	15	2	0.990	2.164	17.339	1.0000	0.239	0.798	0.190	0.0230
18	9.22	28.33	26.78	16.24	26.61	21.08	8	2	0.541	2.578	9.231	0.9580	0.237	0.848	0.192	0.0232
26	14.00	39.00	36.40	19.98	36.18	30.16	15	2	0.030	7.365	17.958	0.9467	0.209	0.994	0.196	0.0237
27	13.55	71.79	70.50	35.66	71.33	61.78	14	2	0.062	2.021	14.783	0.9999	0.321	0.987	0.317	0.0382
22	11.00	28.00	25.75	17.51	29.84	24.16	15	2	-0.147	5.920	13.741	0.9811	0.469	0.970	0.446	0.0539
8	5.36	12.06	10.80	9.39	14.13	10.56	27	2	-1.767	8.794	21.887	0.9996	0.695	0.734	0.510	0.0615
4	14.56	19.48	12.94	20.85	38.55	32.42	21	15	2.918	143.370	3.362	1.0000	1.000	0.524	0.524	0.0633
21	6.00	19.00	18.03	14.04	22.14	17.12	15	2	0.202	2.739	12.262	0.9998	0.585	0.958	0.561	0.0677
13	6.58	19.02	17.85	14.96	23.93	18.68	8	2	0.144	1.548	5.170	0.9919	0.639	0.959	0.608	0.0734
7	15.07	35.89	32.57	23.37	46.75	40.49	27	2	2.132	11.223	15.965	0.9967	0.937	0.682	0.637	0.0768
9	12.75	33.24	30.70	29.62	59.23	51.30	27	2	2.383	5.006	8.671	1.0000	0.999	0.646	0.646	0.0780
5	16.48	27.09	21.51	19.10	34.02	28.15	27	2	0.993	20.089	15.958	0.8270	0.937	0.848	0.658	0.0794
6	7.71	16.44	14.51	15.96	26.01	20.53	27	2	-1.097	6.304	11.382	1.0000	0.994	0.833	0.828	0.1000
10	16.62	49.59	46.72	35.48	70.95	61.45	27	2	0.649	5.929	13.697	1.0000	0.977	0.901	0.880	0.1062
															8.284	1.000

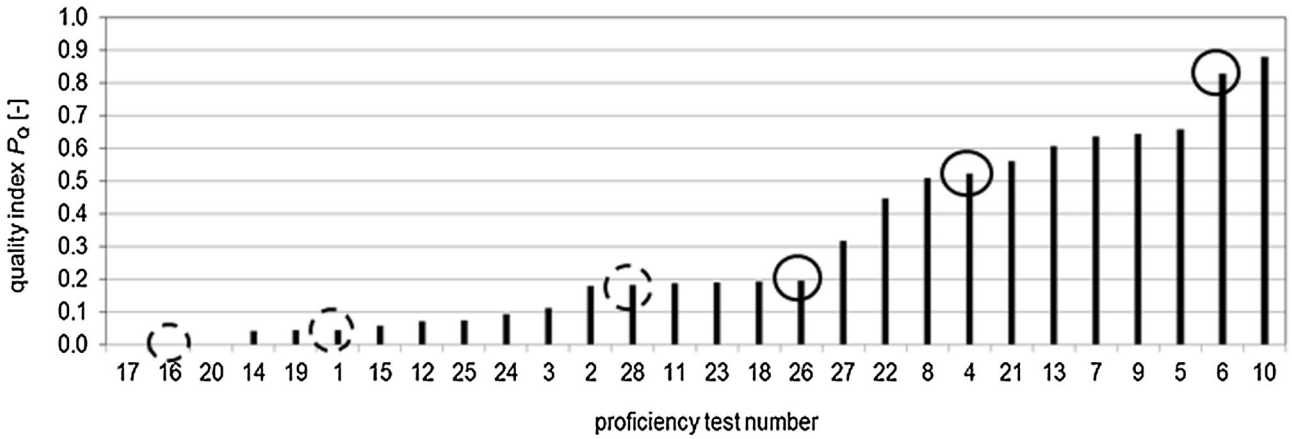

Fig. 1 Calculation example of quality indices P_Q (assessing PTs) and parameters influencing it. Values for the parameters s_r, s_R, s_L, σ_r, σ_R in somatic cells/µl. For explanations, refer to the text. Calculation is accessible in the Electronic Supplementary Material ESM

good discrimination of the laboratories and their median values are revealed (Fig. 4). The reasons for the discrimination may be different but are also a result of a differing analytical performance. Figure 5 shows the quality indices P_L (median) and their corresponding standard deviations of the laboratories having participated two or more times in a PT or PT level. The data show that some laboratories performed consistently at the same level and that others had greatly varying quality indices. However, frequency of participation seems not to be a determining factor [17]. As stated above, the outcome of a PT is influenced by the competence of all of the participating laboratories. It follows, also, that the outcome of each laboratory in a PT is influenced by the others, and a situation is conceivable where only one laboratory measured the correct value while all others show a bias. However, the well-performing

Lab No	s_r	$\bar{y}$	θ	σ_r	σ_R	n	$\tilde{Z}_n$	$\hat{\chi}^2_{(r)}$	$P_{(r)}$	$P(\tilde{z}_n)$	P_L	$\tilde{P}_L$
1	56.57	283.00	261.00	13.73	21.56	2	1.143	16.976	0.000	0.253	0.000	0.0000
3	6.36	261.50	261.00	13.73	21.56	2	0.026	0.215	0.643	0.979	0.630	0.2191
4	0.00	269.00	261.00	13.73	21.56	2	0.416	0.000	1.000	0.678	0.678	0.2357
5	4.95	236.50	261.00	13.73	21.56	2	-1.273	0.130	0.718	0.203	0.146	0.0508
6	2.12	263.50	261.00	13.73	21.56	2	0.130	0.024	0.877	0.897	0.787	0.2736
7	6.36	261.50	261.00	13.73	21.56	2	0.026	0.215	0.643	0.979	0.630	0.2191
8	85.56	305.50	261.00	13.73	21.56	2	2.312	38.833	0.000	0.021	0.000	0.0000
9	0.71	207.50	261.00	13.73	21.56	2	-2.779	0.003	0.959	0.005	0.005	0.0018
sum											2.875	1.000
mean							0.623					

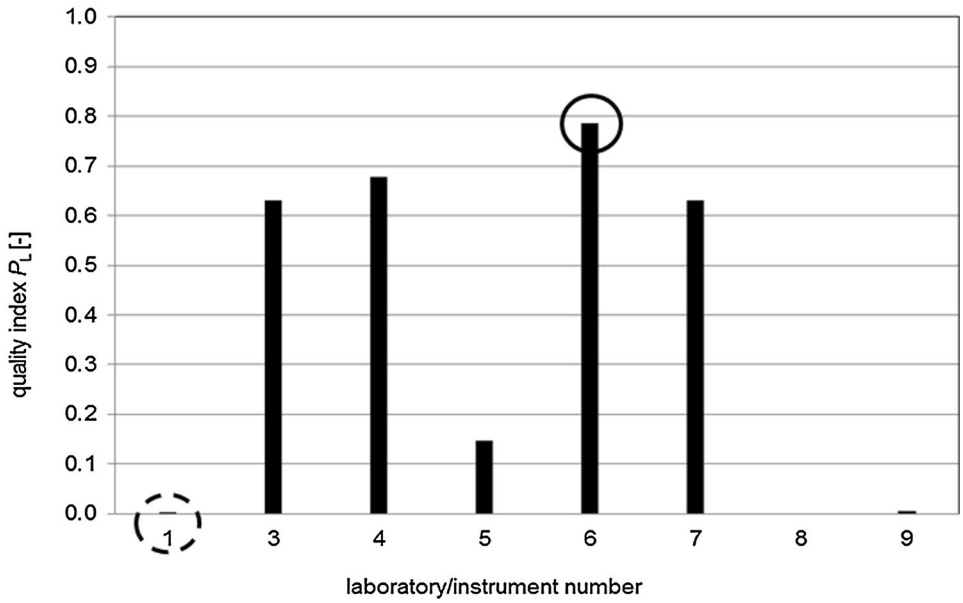

Fig. 2 Graphical evaluation and calculation example of quality indices P_L (assessing laboratories) and parameters influencing it from PT 197 (Cornell, October 2011). Values for parameters s_r, $\bar{y}$, θ, σ_r, σ_R in somatic cells/µl. The mean of $\tilde{z}_n$ was calculated using the robust estimator A15. If s_r is larger than σ_r, the probability related to the repeatability standard deviation $P_{(r)}$ and the probability related to the inter-laboratory standard deviation $P_{(L,r)}$ become small as well as the as the corresponding quality index P_L (*dashed circles*, laboratory no. 1). In cases where s_r is close to or smaller than σ_r, the opposite is true, and the quality index P_L becomes larger or even close to the maximum value of 1 (*solid circles*, laboratory no. 6). Calculation is accessible in the ESM

laboratory or instrument will show a mean value $\bar{y}$ larger or smaller than the 'biased' reference value θ and the related probability $P(\tilde{z}_n)$, and the corresponding quality index P_L will become small and influence the laboratory's median. Such influences are difficult to control. With some experience, a laboratory will participate preferably in well-known and broadly supported PTs. If the PT also disposes an acceptable quality index P_Q, as proposed in this paper, it could be a driver for a laboratory to participate in such a PT. But as mentioned above the quality indices P_Q, and P_L are influenced by different factors and therefore do not allow detailed conclusions on performance details of PTs and laboratories. The approach described in this paper allows an easy general and long-term comparison of PTs and laboratories participating in PTs. It is limited to this and for a detailed assessment of an individual PT or PT scheme or laboratory further information will be necessary, e.g. such as used to calculate the indices mentioned in this paper or by the analysis of the individual results.

In Eqs. (10) and (17), the possibility to modify the quality measure by multiplication with further expressions is mentioned. Such expressions (factors) $q = f(q_1, q_2, q_3, \ldots, q_m)$ made up of m PT and laboratory-specific quality indices q_{i1}, q_{i2}, q_{i3}, …, q_{im} may be used to model m PT$_i$ characterising criteria (e.g. frequency of the PT, number of participants, number of test levels, inter-linkage to other PTs, [summarised] competence index of participating laboratories and of the PT provider, frequency of laboratories' PT participation, competence of the laboratory and laboratory bias [by considering the z-score, e.g. $q_i(z_i) = 2(1 - \Phi|z_i|)$]). Further criteria are mentioned by Golze [18]. The components of q_i need to be defined in such a way that higher values in the resulting q_i indicate higher quality. As yet, no experts in the field of automated somatic cell counting have established such indices and experience in this regard is lacking. The need for using such indices might appear as soon as a system like that described in this paper is set up, and more data than are presented here are integrated. The brackets in the graphical evaluation of the median quality indices in Fig. 5 mark groups of laboratories and instruments and their numbers of times of participation. The median quality indices show a

Lab No	s_r	$\bar{y}$	θ	σ_r	σ_R	n	$\tilde{z}_n$	$\hat{\chi}^2_{(r)}$	$P_{(r)}$	$P(\tilde{z}_n)$	P_L	$\tilde{P}_L$
1	0.00	97.00	94.00	5.99	8.81	2	0.388	0.000	1.000	0.698	0.698	0.1373
2	2.12	101.00	94.00	5.99	8.81	2	0.906	0.125	0.723	0.365	0.264	0.0519
3	1.41	80.00	94.00	5.99	8.81	2	-1.812	0.056	0.813	0.070	0.057	0.0112
4	4.24	98.00	94.00	5.99	8.81	2	0.518	0.502	0.479	0.605	0.289	0.0570
5	0.71	105.00	94.00	5.99	8.81	2	1.424	0.014	0.906	0.154	0.140	0.0275
6	0.71	90.00	94.00	5.99	8.81	2	-0.518	0.014	0.906	0.605	0.548	0.1078
7	0.71	93.00	94.00	5.99	8.81	2	-0.129	0.014	0.906	0.897	0.813	0.1599
8	4.95	112.00	94.00	5.99	8.81	2	2.330	0.683	0.409	0.020	0.008	0.0016
9	1.41	94.00	94.00	5.99	8.81	2	0.000	0.056	0.813	1.000	0.813	0.1600
10	3.54	80.00	94.00	5.99	8.81	2	-1.812	0.348	0.555	0.070	0.039	0.0076
11	7.78	91.00	94.00	5.99	8.81	2	-0.388	1.686	0.194	0.698	0.135	0.0266
12	4.95	97.00	94.00	5.99	8.81	2	0.388	0.683	0.409	0.698	0.285	0.0561
13	0.71	86.00	94.00	5.99	8.81	2	-1.036	0.014	0.906	0.300	0.272	0.0536
15	0.71	92.00	94.00	5.99	8.81	2	-0.259	0.014	0.906	0.796	0.721	0.1419
sum											5.082	1.000
mean							-0.388					

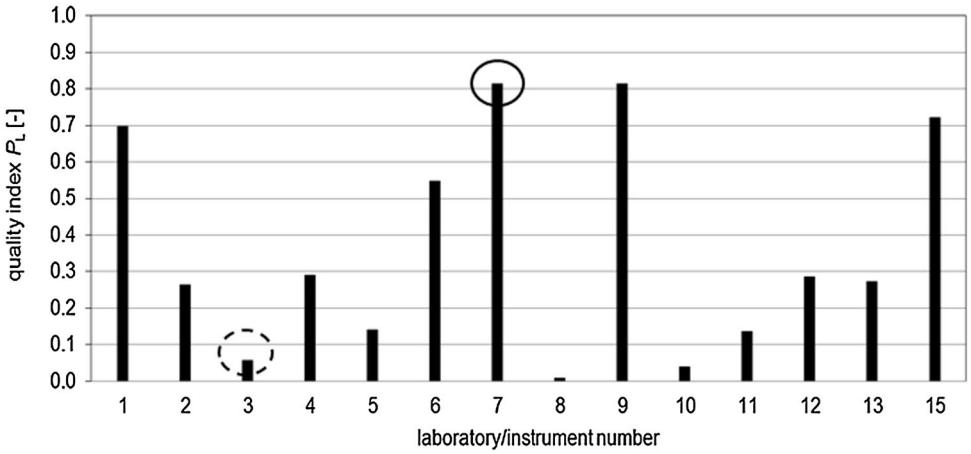

Fig. 3 Graphical evaluation and calculation example of quality indices P_L (assessing laboratories) and parameters influencing the quality index P_L in assessing laboratories from PT 113 (ICAR, September 2011). Values for parameters s_r, $\bar{y}$, θ, σ_r, σ_R in somatic cells/µl. The mean of $\tilde{z}_n$ was calculated using the robust estimator A15. If $\bar{y}$ the mean value of the laboratory, is larger or smaller than the reference value (consensus value, 'true' value) θ, then $|\tilde{z}_n|$ becomes larger, and the related probability $P(\tilde{z}_n)$ as well as the corresponding quality index P_L becomes small (*dashed circles*, laboratory no. 3). In cases where the mean value $\bar{y}$ is close or equal to the reference value θ, $|\tilde{z}_n|$ becomes small, and the related probability $P(\tilde{z}_n)$ as well as the corresponding quality index P_L becomes large or close to the maximum value of 1 (*solid circles*, laboratory no. 7)

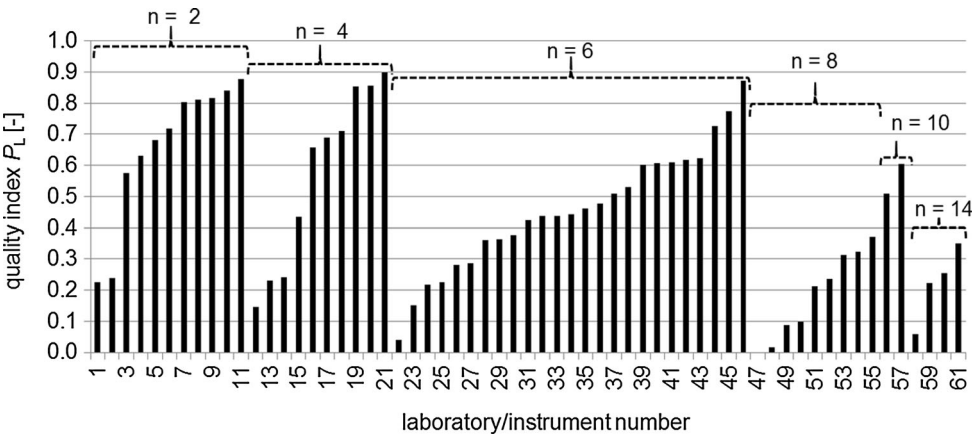

Fig. 4 Graphical representation of the median quality indices P_L of all participating laboratories and instruments in the test data sets (61 laboratories or instruments, 5 PTs and 28 PT levels, none of the laboratories participated in all PTs). *Brackets* mark groups of laboratories and instruments and their number of times of participation

tendency to decline with higher numbers of times of participation. If such a tendency were to become obvious with more data sets, the use of specific quality indices might be necessary.

A model such as that described here can be used for all types of PTs where measurands are quantified. To set up a system as described here, a neutral and trustworthy body is needed to collect the sensitive data from PT trial organisers. Participating laboratories need to give authorisation for the evaluation of their data. Results must be anonymised, and it would be in the responsibility of PT providers and laboratories to communicate

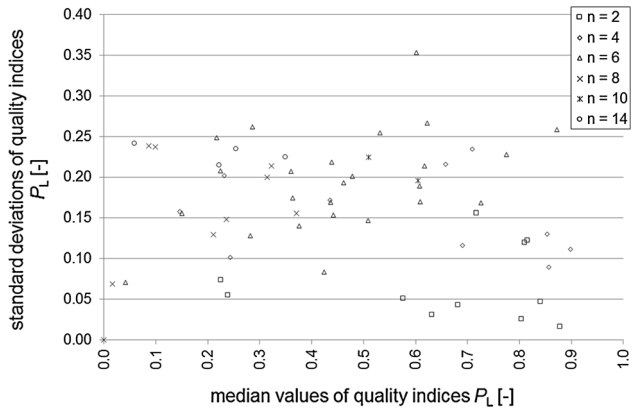

Fig. 5 Standard deviations of the laboratory/instrument quality indices P_L

their codes to their customers in order to demonstrate their competence.

References

1. Pyörälä S (2003) Indicators of inflammation in the diagnosis of mastitis. Vet Res 34:565–578
2. Regulation (EC) (2004) No 853/2004 of the European Parliament and of the Council of 29 April 2004 laying down specific hygiene rules for on the hygiene of foodstuffs. Off J Eur Union. 139/55 Annex II, Section IX, Brussels
3. Grade "A" (2009) Pasteurized Milk Ordinance, 2009 Revision. US Department of Health and Human Services, Public Health Service, Food and Drug Administration, Silver Spring
4. Beal R, Eden M, Gunn I, Hook I, Lacy-Hulbert J, Morris G, Mylrea G, Woolford M (2001) Managing mastitis: a practical guide for New Zealand dairy farmers. Livestock Improvement, Hamilton
5. Baumgartner C (2008) Architecture of reference systems, status quo of somatic cell counting and concept for the implementation of a reference system for somatic cell counting. Bull IDF 427
6. ISO 13366-2 | IDF 148-2:2006. Milk—enumeration of somatic cells, Part 2—guidance on the operation of fluoro-opto-electronic counters. International Organization for Standardization, Geneva and International Dairy Federation, Brussels
7. ISO 13366-1 | IDF 148-1:2008. Milk—enumeration of somatic cells, Part 1—microscope method (Reference method). International Organization for Standardization, Geneva and International Dairy Federation, Brussels
8. Petley BW (1985) Fundamental physical constants and the frontier of measurement. Adam Hilger, Bristol
9. Pendrill LR (2005) Meeting future needs for metrological traceability—a physicist's view. Accred Qual Assur 10:133–139
10. Orlandini S, van den Bijgaart H (2011) Reference system for somatic cell counting in milk. Accred Qual Assur 16:415–420
11. Thompson M, Ellison SLR, Wood R (2006) The international harmonized protocol for the proficiency testing of analytical chemistry laboratories. Pure Appl Chem 78:145–196
12. ISO 13528:2005. Statistical methods for use in proficiency testing by interlaboratory comparisons. International Organization for Standardization, Geneva
13. ISO 5725:1994. Accuracy (trueness and precision) of measurement methods and results. Parts 2, 4, 6. International Organization for Standardization, Geneva
14. Analytical Methods Committee of the Royal Society of Chemistry (1989) Robust statistics: How not to reject outliers. Part 1. Basic concepts. Analyst 114:1693–1697
15. Analytical Methods Committee of the Royal Society of Chemistry (2001) MS EXCEL add-in for robust statistics. (http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/Software/RobustStatistics.asp)
16. Gaunt W, Whetton M (2009) Regular participation in proficiency testing provides long term improvements in laboratory performance: an assessment of data over time. Accred Qual Assur 14:449–454
17. Thompson M, Lowthian PJ (1998) The frequency of rounds in a proficiency test: Does it affect the performance of participants? Analyst 123:2809–2812
18. Golze M (2001) Information system and qualifying criteria for proficiency testing schemes. Accred Qual Assur 6:199–202