Supplementary information

Contents

Number of pages: 11

Number of figures: 6

Number of tables: 1

S1. Organisation

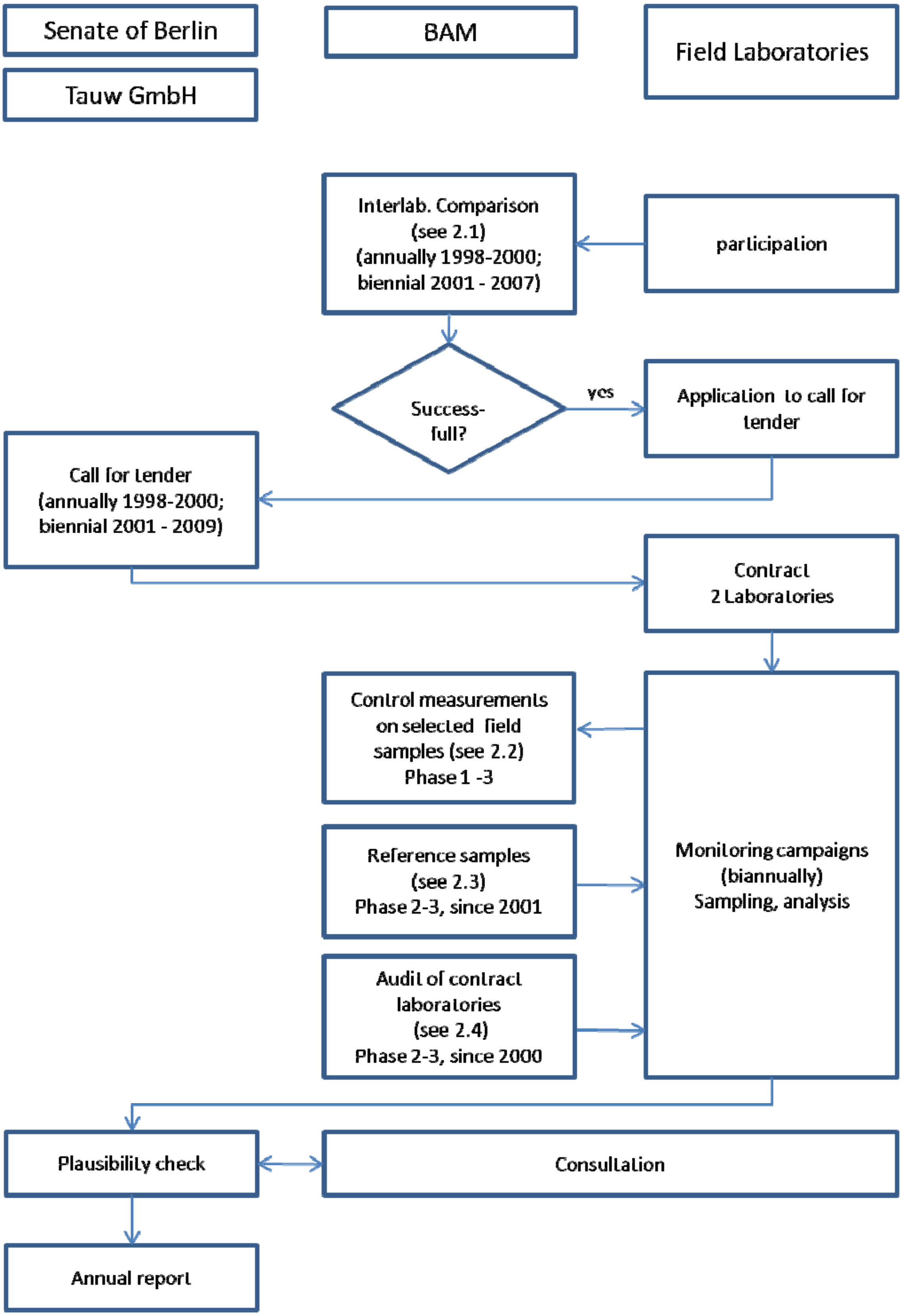
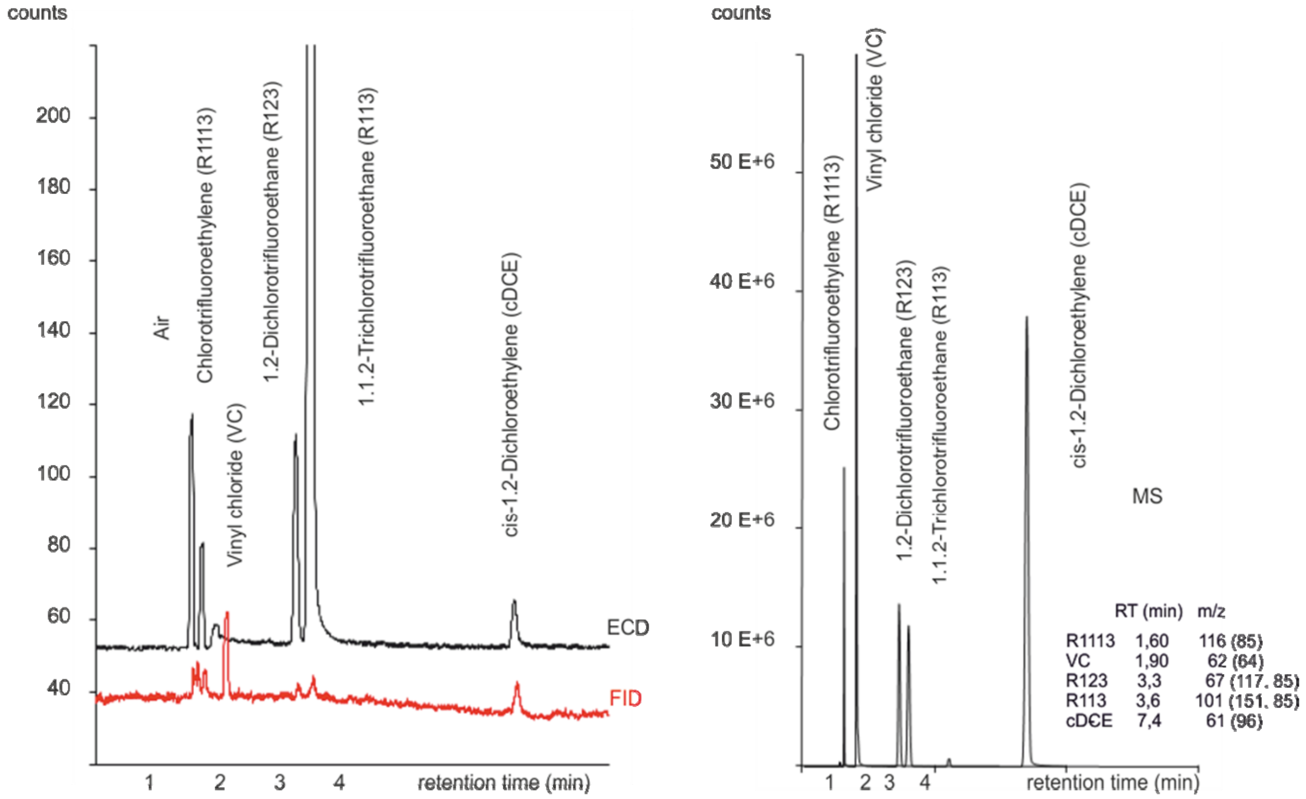

Senate of Berlin
Tauw GmbH
BAM
Field Laboratories
Interlab. Comparison (see 2.1) (annually 1998-2000; biennial 2001 - 2007)
participation
Success-full?
yes
Application to call for tender
Call for tender (annually 1998-2000; biennial 2001 - 2009)
Contract 2 Laboratories
Control measurements on selected field samples (see 2.2) Phase 1 -3
Reference samples (see 2.3) Phase 2-3, since 2001
Audit of contract laboratories (see 2.4) Phase 2-3, since 2000
Monitoring campaigns (biannually) Sampling, analysis
Plausibility check
Consultation
Annual report

Figure S1: Organisation of the project and quality assurance measures

S2. Chromatograms (ECD, FID, MS)



	RT (min)	m/z
R1113	1,60	116 (85)
VC	1,90	62 (64)
R123	3,3	67 (117, 85)
R113	3,6	101 (151, 85)
cDCE	7,4	61 (96)

Figure S-2: Chromatographic detection of the halogenated organics mentioned in the main text. FID and ECD should be applied in parallel to enable the quantification of R1113 and vinyl chloride

S3. Statistical Analysis

S3.1. Aims and Tasks

This supplement provides information on the data assessment carried out to demonstrate the constant improvement of laboratory performance due to the different quality assurance measures applied. The analysis is based upon the ordinal data obtained by grouping laboratory results into categories describing the relative deviation of the result obtained by the laboratory from the reference value. Categories range from relative deviations of 10% to 100%. Occurrences in the different categories form a pattern describing the average laboratory performance over the three phases considered.

Questions to be answered by the assessment of the data are

i. Do the patterns differ between the various phases?

ii. Is the most recent phase in time the best one?

iii. Do differences between analytes exist?

Auxiliary is the question related with item point ii), namely what should be considered "best". Theoretically, the ideal situation would be if all laboratories always found the reference value, at least within a stated measurement uncertainty for the latter. In practice, laboratories will always have a certain bias at the time of measurement which changes (hopefully reduces) over time. Given this, the best one would be able to attain through quality assurance measures is that this bias is random and follows a normal distribution with zero mean (no persistent systematic bias between reference and field measurements, i.e. at least on average, filed laboratories match the reference value) and an appropriately small standard deviation. In the following, these assumptions have been made to define the best laboratory performance attainable in practice, and used for answering the item raised under point ii) question.

S3.2. Data, parameters and tools

Original laboratory results were converted into normalised (absolute) deviations with respect to the corresponding reference value(s) according to

$$d_{imk} = \frac{\left|x_{imk} - x_{imk,ref}\right|}{x_{imk,ref}} = \begin{array}{lll} \dfrac{x_{imk}}{x_{imk,ref}} - 1 & if & x_{imk} > x_{imk,ref} \\ 1 - \dfrac{x_{imk}}{x_{imk,ref}} & if & x_{imk} \le x_{imk,ref} \end{array} \qquad \text{(ES-1)}$$

were the indices refer to the phases (i = 1… 3), the kind-of-analyte (m = 1… 7), and the range (k = 1… 3). Note that, for the ideal case as described below, the expected value for the $\frac{x_{imk}}{x_{imk,ref}}$ ratio is unity, while the expected value for d_{imk} is zero due to the shift by unity.

Reference values were dependent on the range (low, medium and high) and the kind-of-analyte, differing in the low and medium range by an order of magnitude according to the attainable sensitivity for the analyte under consideration. Therefore, the normalised (absolute) deviations were categorised in 10 classes making them, within the framework of evaluating laboratory performance, comparable over all ranges, the kind-of-analyte, and between the time s. The classes were 0 - 10%, 11 – 20%, 21 – 30% .... 91 – 100% and are denoted in the formulae for the sake of simplicity as 0, 1, 2 .... 9. This yields occurrences (number of observations in a class as ordinal data) of normalised (absolute) deviations within the specified classes according to

$$o_{imk}(n) = count\{d_{imk} \in [n \cdot 0.1, (n+1) \cdot 0.1)\} \qquad \text{(ES-2)}$$

were n refers to the class (n = 0… 9), and the indices to the same metadata as in eq. ES-1. The n = 0 class reflects the class closest to the expected value of d_{imk}. The very few observations with $d_{imk} > 1$ (deviation above 100%) were disregarded and not classified. The $o_{imk}(n)$ provide, in dependence on n, the "pattern" of laboratory performances in the ranges, for the kind of analyte under consideration, and for the three phases. Since the total number of available data points is rather small for some ranges, occurrences were normalised (i.e. referred to the total number of observations within the range) and summed up over the three ranges yielding distribution densities

$$p_{im}(n) = \sum_{k=1}^{3} \frac{o_{imk}(n)}{\sum_{n=0}^{9} o_{imk}(n)} \qquad \text{(ES-3)}$$

for all analytes and the three phases. For the seven analytes considered, and the phases 1997 – 2000, 2001 – 2004, and 2005 – 2010 one obtains the distribution densities collected in Figure S-3.

Improvements for all analytes over time can clearly be seen from the graphs: The maximum distribution density shifts considerably towards zero (the reference value), and the "spread" of data over the ten classes becomes smaller. There are, however, differences between the analytes (in particular for R113, R123, and R 1113) which may mainly be attributed to the fact that QA measures for the latter three analytes started later than those for the former.

Without any limitations to the general purpose of the study (assessment of laboratory performance over all analytes covered), and for the reasons of further enhancing the number of observations in a class (for phase three, classes 0 to 4 for all analytes contain 374 observations out of 402), the $p_{im}(n)$ in eq. ES-3 may be summed up over the seven analytes to yield the overall performance of all laboratories involved, for all ranges and analytes, according to

$$p_i(n) = \frac{\sum_{m=1}^{7} p_{im}(n)}{7} \qquad \text{(ES-4)}$$

with the exception that the denominator in eq. ES-4 is 5 (instead of 7) due to the lack of data for R123 and R1113 in phase one (1997 – 2000). This provides the overall performance pattern depicted in Figure 5 in the main text.

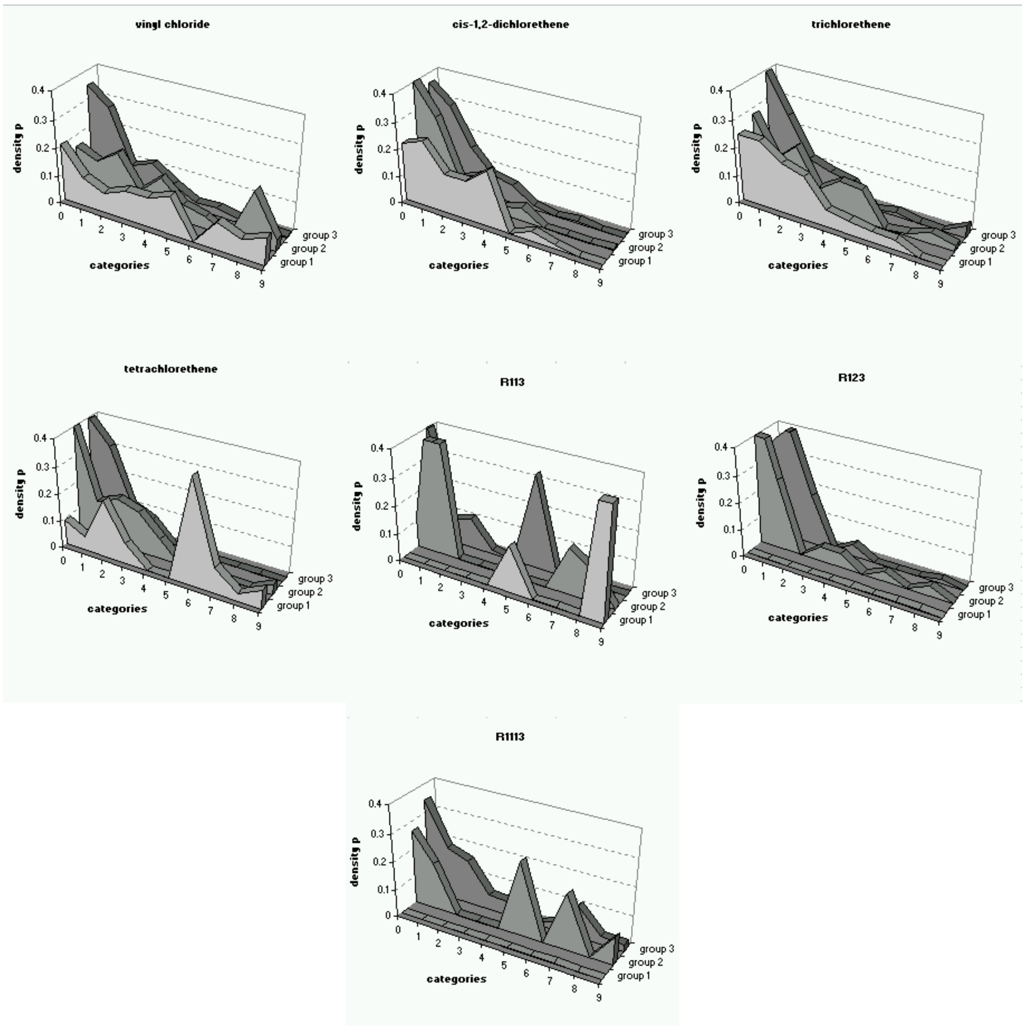


Figure S-3: Distribution densities of observations within the defined deviation classes 0 to 9 for the seven analytes considered, and the three phases (=groups).

It is obvious from Fig. 5 in the main text that the overall performance of the laboratories improved significantly from phase one to three: While in the beginning, the distribution is flat a rather similar to a rectangular (in other words, noise-like), it becomes much more pronounced at the later stages, with more and more observations within a limited, and allocated around the reference value, deviation interval not much larger than 20%.

Comparison between the patterns may be carried out using the Kolmogorov-Smirnov test for distributions which is distribution-independent and thus can cope even with experimental distributions of unknown origin. This would provide chi-square values for the analyte comparisons one-by-one and thus answer the question i) of clause S1, but in a rather ineffective way. Since one still deals with the rest of the questions raised, one would rather try to adjust an "ideal" situation to the patterns obtained for the three phases, again using the Kolmogorov-Smirnov test criterion. This test uses the statistic

$$\chi_i^2 = \sum_{n=0}^{9} \frac{\left[o_i(n) - k \cdot p_n(N(1+\Delta_i,\sigma_i))\right]^2}{k \cdot p_n(N(1+\Delta_i,\sigma_i))} \qquad \text{(ES-5.1)}$$

or

$$\frac{\chi_i^2}{k} = \sum_{n=0}^{9} \frac{[p_i(n) - p_n(N(1+\Delta_i,\sigma_i))]^2}{p_n(N(1+\Delta_i,\sigma_i))} \qquad \text{(ES-5.2)}$$

where the left-hand side of eq. ES-5.2 is the chi-square value per observation with k being the total number of observations (large in all cases considered here), and $p_n(N(1+\Delta_i,\sigma_i))$ the cumulative density of observations within a certain interval of a normal distribution $N(1,\sigma_i)$ with mean unity and standard deviation σ_i, the index i referring to the corresponding phase. Note that although the ideal model distribution has an expectation of unity, one would allow here a certain shift of the experimental distribution with respect to the model, namely Δ_i, to be interpreted as a persistent overall bias. The cumulative density of observations within a defined class n, $p_n(N(1+\Delta_i,\sigma_i))$, is

$$p_n(N(1+\Delta_i,\sigma_i)) = \begin{matrix} \int_{0.9}^{1.1} N(1+\Delta_i,\sigma_i,x)\cdot dx \quad for \quad n=0 \\ \int_{1-(n+1)\cdot 0.1}^{1+(n+1)\cdot 0.1} N(1+\Delta_i,\sigma_i,x)\cdot dx - \int_{1-n\cdot 0.1}^{1+n\cdot 0.1} N(1+\Delta_i,\sigma_i,x)\cdot dx \quad for \quad n>0 \end{matrix} \qquad \text{(ES-6)}$$

The deviations Δ_i in eq. ES-5.1 or eq. ES-5.2 (if not zero) reflect an (additional) shift of the best fit to the experimental data from a distribution centred at the reference value(s). Zero bias estimates would indicate that laboratory performance follows the ideal situation (with a certain spread) for the corresponding phase, and the smallest standard deviation would indicate that the performance of this time frame was the best attainable.

All other distributions (be this cumulative for phases two and one, or the analyte-per-analyte distributions in the various phases) can now be assessed against the maximum (or best) performance attained which is the one in phase three, i.e. the one after a quite considerably long period of QA measures applied. This refers to both the combined for all analytes distributions in the other phases, and the distributions attained analyte-per-analyte.

Assessment may easily be done by looking at the corresponding $\frac{\chi_i^2}{k}$ value representing the deviation of the actual distribution from the best attainable. It should be noted that these are indistinguishable if the $\frac{\chi_i^2}{k}$ value remains below a value of roughly two. This is, in a statistical sense, indeed a rough estimate since the chi-square values at a significance level of 0.05 and considerable large numbers of observations are smaller than $2\cdot k$ and even smaller than twice the number of degrees of freedom of the data set (which depends on how many parameters are assessed using the same data set and may vary from k-1 to k-3). However, as will be seen from the graphs depicted under clause S3, deviations of distributions (in particular for the early phases) assume values well above 100 or even 1000 such that a discussion of whether 1.69 or 2 is critical to make the deviations significant remains largely obsolete.

These approaches provide the answers to questions ii) and iii) as given above. Question i) is best answered by hierarchical cluster analysis and a discriminant analysis. In both cases, the pattern $p_{im}(n)$ for the different analytes are considered objects belonging to group i, i.e. the corresponding phase i. Cluster analysis will reveal similarities between the objects, and discriminant analysis will show whether the manifold of objects in a group (phase) are significantly different (by pattern) from the objects in another group.

S3.3. Results and Discussion

Results are ordered according to the questions raised under clause S1.

S3.3.1. Do the pattern differ within the various phases?

The result of a hierarchical cluster analysis of the object in groups 1 to 3 (phases one to three) is shown in Figure S-4. The distance measure is Euclidean distance, and the linkage method used is Ward.

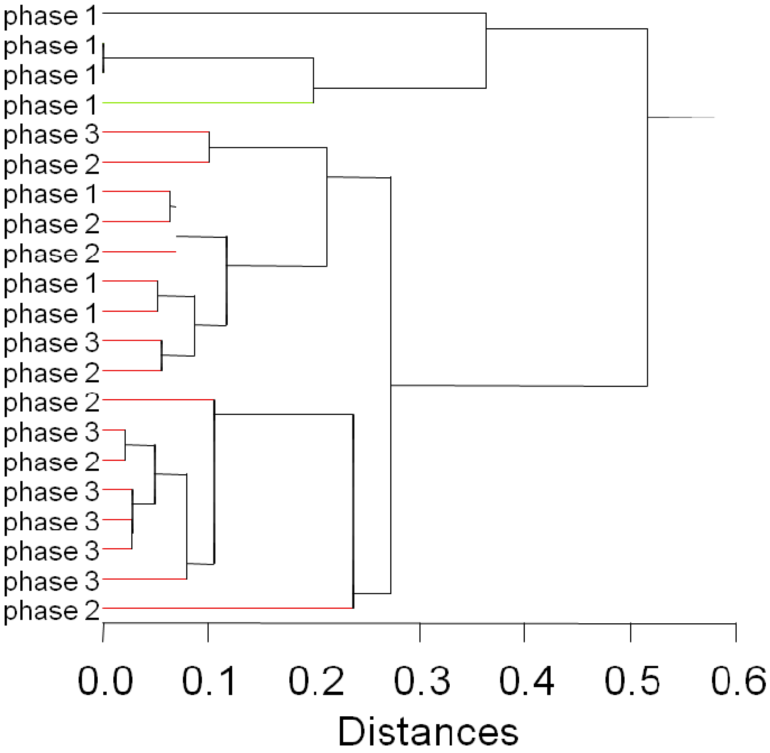


Figure S-4: Hierarchical cluster analysis of the pattern for the different analytes in phases one to three.

Most of the objects (analyte pattern) from phase 1 are distant from the rest and form a separate group at the top of the graph. The next agglomeration (downwards) is mainly formed by objects from group 2, occasionally stained by objects from group 1 and 3, while the major part of group 3 objects form a separate cluster at the bottom of the graph, again mixed with some objects from group 2. This leads to the conclusion that the objects within every group are closer to each other than any objects from other groups, all this with a certain overlap.

The above is confirmed by a discriminant analysis of the objects in the three phases as shown in Figure S-5.

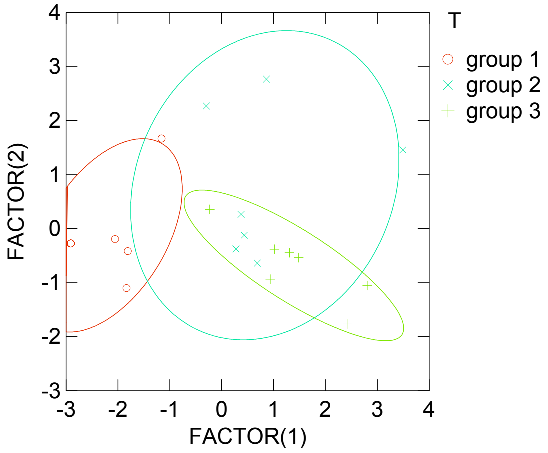


Figure S-5: Discriminant analysis of the pattern for the different analytes in phases one to three.

The patterns differ between the various phases, with a significant difference of all patterns in phase one with respect to the rest, but certain overlap exists. Pattern improvement will be shown under clauses S3.2 and S3.3.

The main features revealed in the discriminant analysis could already be seen from the plot of patterns in Fig. S-3 and coincide with the results of the cluster analysis (Fig. S-4).

S3.3.2. Is the most recent period in time the best one?

Eq. ES-5.2 (optimization criterion) and eq. ES-6 (cumulative class densities of the theoretical distribution) are used to optimise a normal distribution against the experimental pattern within the three phases, using the minimum $\frac{\chi_i^2}{k}$ value as the fit criterion. This gives, for the two parameters means and standard deviation fitted, the results summarised in table TS-1:

Table TS-1: Best-fit parameters mean and standard deviation for the regression of a normal distribution to the experimental class densities

	bias Δ_i (%)	s_i (%)
1997 - 2000	5.4	60.0
2001 - 2004	0.2	23.3
2005 - 2010	0.25	20.2

It is obvious that the best fit to the experimental class densities for time frames 3 and 2 has a zero bias, demonstrating that the laboratory values are centred at the references and deviate from the reference value randomly, but with different standard deviations. Phase three (2005 to 2010) has the smallest standard deviation and therefore reveals the best laboratory performance. For phase one, the best fit to the experimental class densities is still a normal distribution but with a persistent, overall bias against the reference value of around 5.4 %, and a very large standard deviation, It was mentioned earlier that the class density distributions for phase one remind a rectangular distribution rather than something pronounced with respect to the reference(s). Finally, phase three reveals not only the smallest standard deviation, but a value of this standard deviation (20%).

S3.3.3. Do differences between analytes exist?

This question is answered by assessing the analyte-per-analyte distributions in the various phases against the maximum (or best) performance attained which is the one in phase three, i.e. the one after a quite considerably long period of QA measures applied. Assessment may easily be done by looking at the corresponding $\frac{\chi_i^2}{k}$ value representing the deviation of the actual distribution from the best attainable. It should be noted that these are indistinguishable if the $\frac{\chi_i^2}{k}$ value remains below a value of roughly two. This is, in a statistical sense, indeed a rough estimate since the chi-square values at a significance level of 0.05 and considerable large numbers of observations are smaller than 2·k and even smaller than twice the number of degrees of freedom of the data set (which depends on how many parameters are assessed using the same data set and may vary from k-1 to k-3). However, as can be seen from graphs S-6 (same as Fig.6 in the main body of the text), deviations of distributions (in particular for the early phases) assume values well above 100 or even 1000 such that a discussion of whether 1.69 or 2 is critical to make the deviations significant remains largely obsolete.

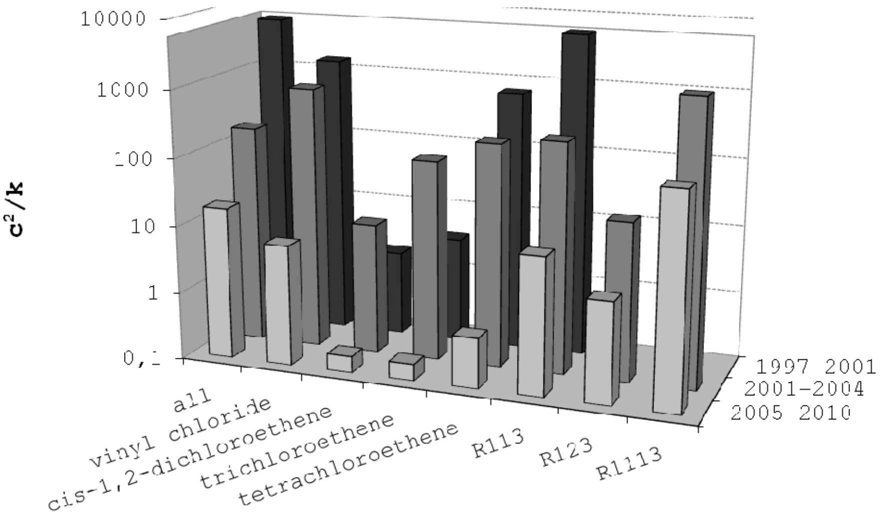


Figure S-6: Per-analyte chi-square values for the deviation of observations within the defined deviation classes 0 to 9 for the three phases from the best laboratory performance attained in the third phase (2005 to 2010).

Figure S-6 clearly shows that

- the deviations of the class distributions are huge in phase one (1997 – 2000) and reduce step-by-in the later phases

- phase three doubtlessly represents the best laboratory performance not only on average but analyte-by-analyte
- actually obtained distributions for certain analytes still differ from the ideal situation to various extents: While all pattern in phase one are, beyond any doubt, far from being acceptable for stating acceptable laboratory performance, all analyte pattern in phase three show acceptably small deviations from the ideal situation, in some cases even compatibility.

To summarise, laboratory performance pattern for the various analytes are (still) different even in the last phase (2005 to 2010), a fact caused by technical reasons and the peculiarities of the analytical technique(s). However, the huge disagreement between laboratories, and with respect to the reference values could be substantially diminished over time.

S3.4. Concluding Remarks

All of the above data assessment and statistical tests, be this multivariate, analyte-by-analyte, or average performance tests clearly indicate a significant improvement over time. Phase three (2005 to 2010) nearly resembles an "ideal" model where the deviations of determinations by field laboratories from the reference value follow a normal distribution with a standard deviation of 20% which is regarded as satisfactory.