CONFERENCE CONTRIBUTION

Lars Jorhem

Non-use and misinterpretation of CRMs. Can the situation be improved?

Received: 10 June 1997 / Revised: 7 July 1997 / Accepted: 9 July 1997

Abstract A survey of 82 scientific papers on trace elements in foods found the use of Certified Reference Materials to be less than anticipated. Less than 50% of the papers reported use of CRMs. When used, the evaluation is usually very crude and provides the user with very little information about his analytical performance. This is mainly due to lack of information/communication between the producers of CRMs, the users and those writing guidelines for the use of CRMs. Several specific problems are pointed out and remedies to improve the situation are described.

Introduction

When an analyst purchases a certified reference material (CRM) for use in the laboratory to check the quality of the results, he/she gets one or several bottles of the CRM, plus a certificate and sometimes a report. The certificate is usually very elaborate on the number of participants in the certification procedure, the different techniques used and the statistical evaluation of the certified intervals. There is, however, a curious lack of information on how to *use* the CRM. You might find information on how to store, dry and homogenize the content of the container prior to use. None of the major producers of CRMs, however, have any information in the report/certificate on how the user should interpret the results from the analysis of the material.

Several international organizations have produced guidelines on how to use CRMs. These guidelines are seemingly unknown to most of the users (apparently to the producers as well), since they are never cited in any published papers.

The users of CRMs thus have no help from the producers and very limited help from the guidelines (if they are found at all). The evaluation procedure most commonly used, which has developed into something of an informal standard procedure, is simply to compare the found mean and standard deviation (SD) with the certified interval. Since the certified interval is expressed in different terms by the various producers, it is difficult to know what the comparison represents.

Non-use or use

In order to find out how CRMs are actually used, we made a survey on the use of CRMs in food-related publications on the subject of trace elements for the years 1990 to 1996.

All together 82 papers published in the following journals were checked for use of CRMs. The ratio of non-users/total number is shown within brackets.

J. Food Comp. Analysis	(22/37)
Food Add. Cont.	(4/15)
J. AOAC Int.	(2/4)
J. Sci. Food Agr.	(1/2)
Z. Lebensm. Unters. Forsh.	(13/24)

As can be seen, in 42 papers there was no mention of CRM results and it must therefore be assumed that no CRMs were used. Since the importance of incorporating CRMs in the AQA-activities today is well recognized, it is surprising that firstly; so many laboratories still do not use CRMs and secondly; that scientific journals accept papers describing analytical results without the use of, e.g. reference materials, as part of the verification of the analytical results. Of these 42 papers, 13 came from countries within the EU, 8 from European countries outside the EU and 12 from North America.

In 40 of the papers, however, results of the use of CRMs are described. Of these papers, 29 came from countries within the EU, 1 from a European country outside the EU and 7 from North America.

L. Jorhem
National Food Administration, Box 622,
S-751 26 Uppsala, Sweden

Evaluation

Several observations on how the CRMs were actually used can then be made:

- None of the papers makes reference to any user guide, e.g. ISO-Guide 33 [1], that gives guidance on the use of CRMs.
- In almost all of the 40 papers the result is presented as mean ± SD (in a few cases the 95% confidence interval is calculated), which is then compared with a certified mean and an interval that is anything but SD.
- The comparison between the found results and the certified means and intervals are often presented in rather vague, or “non-statistical”, terms:

  – the results are within range –, – are similar to –, – were close to – , – agreed well with –, – were in good agreement with –, – agreed with certified value of –, – the method/results were verified by –, – indicate good agreement with –, – the result was X ± Y% of the certified value –, – rejection if greater or less than certified value ± 20% –. Only the latter has attempted any kind of statistical evaluation of their results.
- Then the paper sometimes contains a discussion on why certain results are satisfactory although they do *not* agree with the certified interval.

Regardless of what statistical procedure that lies behind the certified intervals, they are the result of a very elaborate procedure that defines the certified interval with a high degree of probability. With this in mind, it seems a little bit odd to compare the results found with the certified intervals in such imprecise terms.

Certificates

Considering the way certified intervals are described in the manufacturers certificates, the confusion in interpretation is somewhat understandable. The information provided by the producers on the accompanying certificate is, to say the least, rather limited (the full report from the certification gives more information, but this is not automatically provided by all producers) and typically includes only one of the following:

BCR: 95% Confidence Interval (no SD)

IAEA: Confidence Interval (no SD)

NIES: Estimate based on consideration of 2 times the Standard Deviation of the mean of the acceptable values, and of the 95% Confidence Intervals for the mean of individual method (no SD)

NIST: 1. 95% Confidence Interval (no SD)

2. 95% Confidence Interval plus an additional allowance for systematic error among the methods used (no SD)

3. 95/95% Tolerance Limit (no SD)

4. The uncertainties of the values include allowances for inhomogeneity, method imprecision, and an estimate of possible biases of the analytical methods used (no SD)

NRCC: 95% Tolerance Limit (no SD)

How can the average user, who usually is not a statistician, understand the difference between the various types of intervals? The 95% CI and 95/95% TL are comprehensible for most analysts, but the other intervals are more problematic to understand and must create some confusion (what is a 95% TL?).

Guidelines

Then there are the user guides [e.g. 1, 2, 3, 4], that never seem to be used. As mentioned, they were not referred to in any of the 40 papers in which CRMs were used. There are probably several reasons for this:

- Their existence is not widely known.
- The guides are difficult to find and some are rather expensive.
- The evaluation procedures are not seen as useful.
- All evaluation procedures are based on SD, which is not provided on the certificate by the CRM-producers, as seen above.

There seems to be something missing: The producers of CRMs never refer to user guidelines. The user guidelines do not satisfy the needs of the user. The users place no demands on the producers. The producers apparently do not confer with each other, and they do not confer with the users (see Fig. 1).

Also, the use of CRMs is encouraged in collaborative trials of analytical methods [5]. No guideline on how to evaluate the results from these CRMs is, however, provided. CRMs are also recommended for the establishment

Fig. 1 Are we striving towards the same goal?

of control charts in AQA-procedures [2, 3]. Why CRMs are recommended is unclear, since the certified means and intervals are not being utilized for any evaluation of the results. Any in-house reference material with an acceptable homogeneity would suffice.

Why not just proceed as usual and compare the results with the certified intervals?

This cannot be recommended for several reasons:

- It is not a statistically validated procedure, although it has developed into a *de facto* norm.
- The result of this comparison provides the user with little/no information about the analytical performance.
- When the certified interval is something else than the 95% CI or the 95/95% TL it is difficult for the user to interpret what it represents.
- The 95% CI is good for characterization purposes, but not for evaluation of results.
- It can be paradoxical in the evaluation.

The paradox

The CI (unadjusted) is dependent on the number of results, the more results, the narrower the interval becomes. When based on more than six results the 95% CI is narrower than the SD.

A number of laboratories are participating in the certification procedure. Fig. 2 shows the results from a typical characterization exercise. The participating laboratories all have good analytical records, their results have been scrutinized in detail for analytical errors and have thereafter been subjected to normal outlier elimination procedures. All results in the figure are thus acceptable as being part of the basis for the mean ± 95% CI. The bars show the mean and 95% CI for each laboratory (since $n = 5$ or 6, this is approximately the same as the SD). The bottom bar shows the mean of means and its 95% CI. It can then be noted that seven laboratory means fall outside the 95% CI. It can also be noted that for labs 17, 11 and 12 there is no overlap of the laboratory and the total CIs. This is the normal outcome of a certification procedure based on 95% CI, when this number of laboratories participate.

But, suppose that, e.g., labs 11, 12 or 17 now want to use this CRM in the daily work using the same method. The new results must be assumed to fit the earlier distribution reasonably well, otherwise the new, or the earlier, results must be erroneous. From the user's point of view, the results must now be regarded as unacceptable since, not only is the mean outside the 95% CI, but there is also no overlap of the two intervals (being the definition on agreement between results given in the NIST Handbook for SRM-Users [2]).

How can results that are perfectly good for the characterization of the CRM later be regarded as not acceptable?

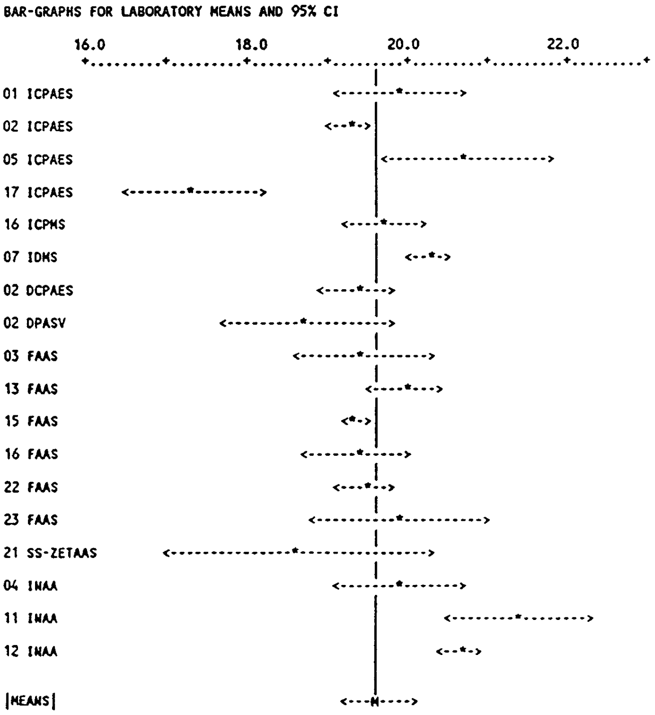

Fig. 2 The normal outcome of an average certification procedure as carried out by the BCR (SM & T). The bottom bar represents the mean of means and 95% CI

Simply because the 95% CI is well suited for *characterization* purposes, but highly unsuitable for *evaluation* of the results.

The above-mentioned shortcomings strongly indicate that there is a need to develop procedures to make CRMs more useful to the customers.

Suggestions for improvement of the current situation

- The producers agree on a common model for presentation of the certified interval. This would greatly simplify the comparison with the results found.
- The producers provide the consumers with the information needed for an acceptable evaluation, i.e. the certificate needs to contain more information.
- The guideline producing bodies cooperate with producers and users to establish evaluation procedures that are relevant and easy to use. Evaluation procedures do not necessarily have to be very complicated. One possible model is the use of Z-scores [5], which can be based on e.g. the SD, the Horwitz equation, or some other type of target value. A similar method has been described in which the random or systematic errors can be evaluated [7]. Another model, which has also been described earlier [8] takes several questions into account:
  – What is the permissible range for laboratory means accounting for both between and within laboratory variation?

  - What is the permissible range for the user’s individual results?
  - What is the permissible range for the user’s means when the RM is used repeatedly, accounting only for within-laboratory variation?

- The producers start a dialogue with the users. It is the needs of the users that should be the foundation of the producers’ existence.
- The formation of a “users’ association” (on Internet?). The users probably have much in common, but no forum for exchange of ideas or problems.
- Scientific journals should be encouraged not to accept papers that lack proper validation of results (e.g. by use of CRMs).

References

1. ISO Guide 33 (1989) Uses of certified reference materials. International Organisation for Standardisation. Case Postale 56, Genève, Switzerland
2. NIST Special Publication 260-100. (1993) Standard Reference Materials. Handbook for SRM Users. Standard Reference Materials Program. National Institute of Standards and Technology, Gaitersburg, MD 20899, USA
3. NMKL Report no: 13 (1993) Guidelines in quality assurance in food laboratories – Reference materials and other calibration tools (in Swedish). Nordic Committee on Food Analysis. C/o VTT, Pb 203, SF-02151 Espoo, Finland
4. Eurachem. Quantifying Uncertainty in Analytical Measurement. First ed. (1995) Prepared by the working group on Uncertainty in Chemical Measurement
5. AOAC Official Methods Program (1995) Associate Referee’s Manual on Development, Study, Review and approval process. AOAC International
6. Thompson M, Wood R (1993) Pure Appl Chem 65:2123–2144
7. Jorhem L, Slorach S, Engman J, Schröder T, Johansson M (1995) SLV-Rapport 4/1995
8. Jorhem L, Schröder T (1995) Z Lebensm Unters Forsch 201: 317–321