This preprint version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/ Restricted use of this manuscript is permitted provided the original work is properly cited. The authors assert their moral rights, including the right to be identified as an author.

this pre-print manuscript downloaded from http://brendongill.com/unsecure/wpg_manuscript2015a.html

Evaluation Protocol for Review of Method Validation Data by the AOAC Stakeholder Panel on Infant Formula and Adult Nutritionals Expert Review Panel

Brendon D. Gill[1]*, Harvey E. Indyk[1], Christopher J. Blake[2], Erik J.M. Konings[2], Wesley A. Jacobs[3], and Darryl M. Sullivan[4]

[1] Fonterra Cooperative Group Ltd, PO Box 7, Waitoa, New Zealand

[2] Nestlé Research Center, Nestec Ltd, PO Box 44, 1000 Lausanne, Switzerland

[3] Abbott Nutrition, 3300 Stelzer Rd, Columbus, OH 43219

[4] Covance Laboratories, 3301 Kinsman Blvd, Madison, WI 53704

* Corresponding author

Abstract

Methods under consideration as part of the AOAC Stakeholder Panel on Infant Formula and Adult Nutritionals process are to be evaluated against a set of Standard Method Performance Requirements (SMPRs) via peer review by an expert review panel (ERP). A validation protocol and a checklist have been developed to assist the ERP to evaluate experimental data and to compare multiple candidate methods for each nutrient. Method performance against validation parameters mandated in the SMPRs as well as additional criteria are to be scored, with the method selected by the ERP proceeding to multi-laboratory study prior to Final Action approval. These methods are intended to be used by the infant formula industry for the purposes of dispute resolution.

Gill, B.D.; Indyk, H.E.; Blake, C.J.; Konings, E.J.M.; Jacobs, W.A.; Sullivan, D. (2015) Evaluation protocol for the review of method validation data by the AOAC Stakeholder Panel on Infant Formula and Adult Nutritionals Expert Review Panel. *Journal of AOAC International* 98, 112–115.
doi.org/10.5740/jaoacint.14-158

Introduction

In April 2010, an agreement between AOAC International and the International Formula Council (IFC) was reached on a 3-year project to develop standards for the analysis of micronutrients in infant formulas and adult nutritionals. The infant formula industry required updated methods that were proven to be accurate, precise, and robust using contemporary techniques for dispute resolution purposes. A set of nutrients was identified by infant formula manufacturers as high priority because of either the lack of a reference method or the existence of a reference method that was considered to be outdated or not validated for a broad range of infant formulas. The Stakeholder Panel on Infant Formula and Adult Nutritionals (SPIFAN) was formed under the direction of AOAC to ensure independence, freedom from conflicts of interest, and to facilitate a wide representation from industry, academia, and regulatory authorities. The expert review panel (ERP) was appointed by AOAC in a similar manner, whereby experts were selected based on demonstrated expertise and the requirement for a balanced analytical representation (1). AOAC and IFC signed another 3-year agreement in June 2013 to continue the SPIFAN process to include additional nutrients for the development of Standard Method Performance Requirements (SMPRs) and method evaluation.

SMPRs are developed by working groups, and candidate methods are selected for evaluation by a process of peer review. SMPRs document the criteria for the performance of an analytical method that stakeholders establish for a dispute resolution method. These SMPRs are developed by voluntary consensus of participating experts representing a range of different industry stakeholders, including infant formula manufacturers, regulatory authorities, contract laboratories, academic institutions, raw material and standards suppliers, and instrument manufacturers. The establishment of SMPRs occurs with endorsement by a stakeholder panel vote through a transparent, open, and balanced process (1, 2).

A set of infant formula and adult nutritional products (SPIFAN kit) was manufactured as representative of the wide range of commercially available products for use during both the single laboratory validation (SLV) phase and the multi-laboratory testing (MLT) phase of method evaluation. The SPIFAN kit is a valuable tool for determining the scope of matrixes for which candidate methods are applicable.

During the evaluation of candidate methods, ERP meetings are open to all interested parties, and any participant is able to contribute to the discussion. The author of a method, if present, provides immediate and valuable answers to key questions during the discussion. Decisions are made by a voting panel that consists of a vetted group of approximately 12–15 subject matter experts. This gives all involved the opportunity to provide input and provides for an open and transparent method review process.

A validation protocol and an evaluation spreadsheet were developed to assist reviewers to compare and evaluate the SLV data of multiple candidate methods for each nutrient and to select a single method for

the assessment of method reproducibility by MLT, which may be used by the infant formula industry for dispute resolution purposes.

Validation Parameters

All candidate methods for each nutrient are subjected to a common SLV protocol using the SPIFAN matrixes (3).

Applicability

The applicability statement of the SMPRs defines the form(s) of analyte that must be included within the scope of a method under consideration. Typically, the SMPRs dictate that the method must also demonstrate its applicability to all forms of infant, adult, and/or pediatric formulas (powders, ready-to-feed liquids, and liquid concentrates). To meet this requirement, methods are subjected to SLV utilizing the SPIFAN kit to cover the broad range of pediatric and adult formula product types.

Analyte Forms

A critical component in the assessment of a method is whether the correct analyte forms are measured as specified in the SMPRs. As a minimum, all analyte forms identified in the SMPRs should be included as part of the method validation. Other forms of the analyte may be desirable but are not mandatory.

Range

The analytical range defined in the SMPRs is generally chosen for a particular analyte to be from 20% of the lower limit to 200% of the upper limit required by global regulatory authorities. For nutrients for which no regulatory limits exist, 20% of the lowest concentration or 200% of the highest concentration to which products are formulated is used.

The establishment of linearity dose-response over the required range is demonstrated by the analysis of a minimum of six levels that span the desired working range. The relative error of back-calculated concentrations determined against a least-squares line of best fit is calculated. Although no criterion for linearity is specified in the SMPRs, it is recommended that calibration errors be < 5% over the entire calibration range.

LOD and LOQ

The LOD is defined as the lowest concentration of analyte that can be detected, and the LOQ is defined as the lowest concentration that can be reliably quantitated under the stated conditions of the test (4). The

standard protocol is to perform 10 independent analyses of a blank material or, if there is no detectable signal, a blank material spiked at a low level. The LOD and the LOQ are then calculated as described in Equations 1 and 2 (3).

Precision

The precision of an analytical method is the variation in the measured results from multiple analyses. To ensure the veracity of an estimate of method precision, confirmation of analyte homogeneity within the product is necessary. Precision is typically subdivided into three types: repeatability, intermediate precision (or intra-laboratory reproducibility), and reproducibility.

(a) *Repeatability.*—The degree to which measured results agree when evaluated between successive measurements of the same analyte determined under the same conditions is termed repeatability. Replicate test portions are analyzed with the mean ($\bar{x}$) and RSD (S_r) of pairs (Equations 3 and 4), and the RSD_r (Equation 5) calculated (5).

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ x } 100 \qquad (3)$$

where

x_i = individual measured value *i* and n = number of individual values

$$S_r = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{2n}} \qquad (4)$$

where

d_i = difference between duplicate results for test sample *i* and n = number of pairs of duplicates

$$RSD_r = \frac{S_r}{\bar{x}} \text{ x } 100 \qquad (5)$$

where

S_r = RSD for test sample and $\bar{x}$ = mean of replicate analyses

(b) *Intermediate precision*.—The degree to which measured results agree when evaluated a number of times on different days, within the same laboratory, is termed intermediate precision or intra-laboratory reproducibility. A large number of parameters contribute to intermediate precision: variations in days, analysts, equipment, etc., resulting in numerous permutations by which intermediate precision may be assessed. Hence, intermediate precision is not a required parameter in the SPIFAN SMPRs. However, in the absence of an MLT study, the inclusion of intermediate precision values is a suitable substitute for reproducibility, and hence is a requested component of precision in the ERP checklist. Replicate test portions can be analyzed with the intermediate precision SD (S_{iR}) and the RSD_{iR} as calculated in Equations 6 and 7 (6).

$$S_{iR} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \quad (6)$$

$$RSD_{iR} = \frac{S_{iR}}{\bar{x}} x\ 100 \quad (7)$$

(c) *Reproducibility*. —The degree to which measured results agree when evaluated a number of times between different laboratories is termed reproducibility. Reproducibility is a required parameter in the SPIFAN SMPRs and is assessed using a multi-laboratory trial as described by the joint International Union of Pure and Applied Chemistry, International Organization for Standardization, and AOAC harmonized protocol (7).

The Horwitz ratio (HorRat) is a simple parameter that indicates the suitability of a method by comparing its precision against an expected value. The HorRat is expressed as the ratio of the measured reproducibility to the predicted value obtained from the Horwitz curve (Equations 8 and 9). HorRat values of less than 1 demonstrate better than expected reproducibility whereas HorRat values of greater than 1 demonstrate poorer reproducibility, with acceptable HorRat values in the range 0.5–2.0. The HorRat may be applied to repeatability, although with less reliability. Typically, repeatability is estimated as one-half to two-thirds of the predicted reproducibility ($pRSD_R$), with estimated HorRat limits of 0.3–1.3 (2).

$$pRSD_R = 2\ x\ C^{-0.1505} \quad (8)$$

where

C = concentration expressed as mass/mass or mass/volume fraction

$$HorRat = \frac{RSD}{pRSD_R} \quad (9)$$

The HorRat is a useful reportable indicator as part of any validation study under consideration by the SPIFAN ERP. However, concerns over the applicability of this metric, particularly to analytes at low concentration, limit its usefulness to that of a guideline rather than a strict criterion (8). Industry needs and regulatory requirements form the basis of the precision criteria stated in the SMPR, despite the possible variance from expected precision as estimated by the HorRat (9). For this reason, the HorRat is neither a requirement as part of the ERP checklist nor a performance requirement enumerated in the SMPRs.

Accuracy

Bias is a measurement of systematic error representing the difference between the measured value and the true value. However, as the true value is unknown, an accepted reference value is substituted for the true value; thus bias can be determined by the analysis of a certified reference material (CRM) and by

comparing the results of the candidate method with those of a reference analytical method (10). Spike recovery is an additional practical means to assess accuracy and is a required metric for ERP approval.

(a) *Analysis of CRM*.—A CRM contains values that are traceable to an accurate concentration accompanied by an uncertainty and a stated level of confidence. The concentration of target analyte in a CRM measured by the candidate method should agree with the assigned value within the defined uncertainties; however, if the results show a statistically significant difference ($p < 0.05$), the method is biased (11).

Bias is evaluated by replicate analyses (nine replicates are recommended) of an appropriate CRM, e.g., National Institute of Standards and Technology 1849a. Differences between the measured value and the certified value are determined via the mean and SD of the differences, and the *T*-statistic (Tstat) is calculated (Equation 10). The null hypothesis that there is no difference between the measured results and the certified value (H_0: $d = 0$) is applied. The probability (*P*-value) of the observation under the null hypothesis is assessed at the $\alpha = 0.05$ level of confidence. The degrees of freedom (DF) can be estimated using the Satterthwaite approximation (Equation 11; 12).

$$T_{stat} = \frac{\bar{d}}{\sqrt{\frac{S_m^2}{n_m} + \left(\frac{U_{CRV}}{k}\right)^2}} \qquad (10)$$

$$DF = \frac{\left(\frac{S_m^2}{n_m} + \left(\frac{U_{CRV}}{k}\right)^2\right)^2}{\frac{1}{n_m - 1}\left(\frac{S_m^2}{n_m}\right)^2 + \frac{1}{n_{CRV} - 1}\left(\left(\frac{U_{CRV}}{k}\right)^2\right)^2} \qquad (11)$$

where

$\bar{d}$ = mean of the differences between the measured results and the certified value;

S_m = SD of differences between the measured results and the certified value;

n_m = number of replicate analyses of the CRM;

U_{CRV} = uncertainty of the certified reference value; k = coverage factor used for calculating the expanded uncertainty; and

$n_{CRV} - 1 = DF_{CRV}$, where this is calculated for k, using a two-sided t distribution at $\alpha = 0.05$. An empirical model for doing this calculation was obtained by fitting 1/DF against log $t_{0.05}$ to a fourth order polynomial for integral values of DF from 2 to 100. The maximum relative error over the fitted range was 0.06%.

(b) *Comparison with other methods*.—Comparison of quantitative data obtained from a candidate method with that from a reference method is not required because of the additional effort needed in setting up a second analytical method and because of the lack of appropriate reference methods for some nutrients (3).

(c) *Spike recovery*.—The determination of analyte recovery is an important means to evaluate the accuracy of an analytical method through multistage sample preparation procedures, since components of the matrix may interfere with the separation, detection, or accurate quantitation of the analyte. Analyte recovery experiments are performed by the analysis of a well-characterized analyte standard in the presence of the sample matrix. One limitation of this technique to evaluate accuracy is that procedural effects on the spiked analyte may be different from that for the innate analyte.

The SPIFAN kit contains a number of products that were not supplemented with micronutrients during manufacture, and are therefore ideal materials for spike recovery experiments. The recovery tests are performed by preparing test portions spiked at two different concentrations of analyte: either at 50 and 150% of typical label claim, or as over-spikes added at 50 and 100% of endogenous concentrations. Duplicate unspiked samples and spiked materials at each concentration should be analyzed on each of 3 days, calculated independently, and reported as a range of recoveries. These samples are then processed through the test method and the recovery percent (R%) is calculated (Equation 12).

$$\text{Recovery}\,(\%) = \frac{C_{FS} - C_{US}}{C_{SS}} \text{ x } 100 \qquad (12)$$

where

C_{FS} = measured concentration of the fortified (spiked) test sample;

C_{US} = measured concentration of the unfortified (unspiked) test sample; and

C_{SS} = concentration of standard spiked into the fortified test sample.

To evaluate how the spike recovery was performed, additional information is requested of the Study Director (Appendix 1).

Suitability Factor and Weighting

Because multiple candidate methods may match the target SMPRs, an evaluation system has been developed to facilitate discrimination between methods delivering equivalent performance. A Student's *t*-test may be used to evaluate whether two similar methods give comparable results before further evaluation. The evaluation system involves a ranking of similar methods based on suitability scores and weighting factors of the various performance characteristics.

Validation parameters are assessed by reviewers and graded by a suitability factor score (1, 3, or 5). The score for each validation parameter is evaluated against the SMPRs where applicable, with methods that demonstrate compliance with the SMPRs given a value of 5.

The weighting factors (1, 2, or 3) are empirical values indicating the relative importance of each parameter. In the evaluation of method performance, they are part of the calculation used to determine an overall

performance rating by prioritizing key parameters during the SLV. The highest weighting factors in the ERP evaluation are given to accuracy and precision (Table 1).

Additional Evaluation Parameters

In instances in which more than one method per analyte is under review by the ERP, performance against the SMPRs may be comparable, and qualitative aspects of the analytical methods can be an important tool for differentiating between methods. During the evaluation of the method under consideration, the ERP is asked to consider additional information (Appendix 2).

Conclusions

As part of the SPIFAN process, candidate methods undergo validation using SPIFAN matrixes, and the data are reviewed by the ERP. A checklist was developed for use by ERP reviewers to evaluate and compare performance data of candidate methods, with the intention that a single method per nutrient would be selected for use by the SPIFAN community for dispute resolution purposes.

References

(1) Sullivan, D. (2012) *J. AOAC Int.* 95, 287–297. http://dx.doi.org/10.5740/jaoacint.Sullivan_Intro

(2) Official Methods of Analysis (2012) 19th Ed., AOAC INTERNATIONAL, Gaithersburg, MD, Appendix F

(3) Official Methods of Analysis (2012) 19th Ed., AOAC INTERNATIONAL, Gaithersburg, MD, Appendix L

(4) Eurachem (1998) *The Fitness for Purpose of Analytical Methods: A Laboratory Guide to Method Validation and Related Topics*, LGC, Teddington, UK

(5) Bruce, P., Minkkinen, P., & Riekkola, M.L. (1998) *Microchim. Acta* 128, 93–106. http://dx.doi.org/10.1007/BF01242196

(6) Lynch, J.M. (1998) *J. AOAC Int.* 81, 679–684

(7) Official Methods of Analysis (2012) 19th Ed., AOAC INTERNATIONAL, Gaithersburg, MD, Appendix D

(8) Thompson, M. (2000) *Analyst* 125, 385–386. http://dx.doi.org/10.1039/b000282h

(9) Lynch, J.M., Barbano, D.M., Fleming, J.R., & Nicholson, D. (2004) *Inside Laboratory Management* 8, 24–28

(10) Thompson, M., Ellison, S.L.R., & Wood, R. (2002) *Pure Appl. Chem.* 74, 835–855. http://dx.doi.org/10.1351/pac200274050835

(11) Sharpless, K.E., & Duewer, D.L. (2008) *J. AOAC Int*. 91, 1298–1302

(12) Ryan, T.P. (2000) *Statistical Methods for Quality Improvement*, 2nd Ed., John Wiley & Sons, New York, NY

Table 1. Weighting factors of validation parameters

Weighting	Parameter[a]
1	Analyte forms, analytical range, range of matrixes, intermediate precision, LOD, reproducibility
2	LOQ, spike recovery
3	Bias against Certified Reference Material, repeatability

[a] Parameters evaluated against values defined in the SMPRs.

Appendix 1. Spike Recovery Information

Was a portion of the product spiked?

Were all spike recovery samples taken from the same spiked product?

Were multiple portions of the product spiked and was each spiked product analyzed?

Were products weighed into extraction vessels before spiking?

Were products spiked before they were weighed into extraction vessels?

Were all spike recovery samples prepared on the same day?

Were all spike recovery samples prepared on multiple days?

Were spike recoveries calculated from theoretical additions?

Was the spiking solution analyzed and the results used to calculate spike recoveries?

What form of the nutrient was spiked into the product?

What kind of solution or solvent was used to deliver the spike to the product?

What percentage of the spiked sample was the spike blank?

Appendix 2. Additional Method Information

Is there adequate proof of performance via system suitability?

Was feedback from users of the method since being awarded First Action Official Methods Status positive?

Did the method author consider the ERP's specific recommendations?

Was bias evaluated against an established method? Is there a bias?

Is the analytical equipment commonly available in most laboratories?

Are unique or proprietary equipment/accessories required in the method?

Does the method require any special safety precautions?

Any other considerations?