BRIEFING

Appendix XVIII: Guidance On Developing and Validating Non-Targeted Methods For Adulteration Detection. The USP Expert Panel on Non-Targeted Methods for Milk Ingredients (an Expert Panel to the Food Ingredients Expert Committee) proposes this new Appendix to the *Food Chemicals Codex* to provide guidance that reflects current scientific thinking on how to develop and validate non-targeted analytical methods. Non-targeted testing for potential adulterants in foods and food ingredients is becoming a more common approach to identifying products and determining whether or not more specific analytical testing for adulteration might be advised.

This guidance document is one in a series of documents that USP intends to offer to provide general information and assistance to the food industry. The document is intended to assist users in supply chain management by providing information that can generally be applied to testing and authentication of raw materials with a variety of analytical techniques. This document is not overly prescriptive by design to allow for differences in parameters such as testing techniques, data analysis, and ingredient variability that we would expect to exist within the food industry.

The proposal is targeted for publication in the *Third Supplement* to *FCC 10*.

This Appendix to the *Food Chemicals Codex* is intended to elaborate guidance frameworks and tools to assist users in the development and validation of non-targeted analytical methods to counter food fraud.

Add the following:

▲APPENDIX XVIII: GUIDANCE ON DEVELOPING AND VALIDATING NON-TARGETED METHODS FOR ADULTERATION DETECTION

CONTENTS

PURPOSE

Detecting food fraud (economically-motivated adulteration, or EMA) is a challenging analytical task because for any food or food ingredient at risk of adulteration there may be numerous potential adulterants, many of which are unknown. Even for the subset of known potential adulterants, the time and cost associated with traditional targeted and quantitative analyses may preclude their effective use in screening applications. A non-targeted method consists of an analytical measurement that is sensitive to multiple potential adulterants coupled with a statistical model that recognizes deviations from the signal associated with the nominal mate-rial: it is not calibrated for any specific adulterants. Such methods have significant practical benefits but due diligence is required in their development, validation and implementation to ensure sensitivity and specificity.

This document provides guidance on how to develop and implement one-class, non-targeted classification methods for the detection of EMA-related adulterants in food, independent of the analytical technology used. It is not intended to cover the use of multi-class classification methods that that are often represented as non-targeted methods (e.g. NMR fingerprinting coupled with PLS-DA models to classify the geographic origin of a sample[1]).

This guidance is intended for use specifically for food fraud/EMA. This is due to the fact that EMA-related adulterants are typically present in concentrations consistent with the sensitivity limits of non-targeted methods (typically above 0.1% concentration). Detecting ideologically motivated intentional adulteration of the food supply (such as terrorism), while possible with non-targeted methods, is not the intended application of this guidance. Agents used for intentional adulteration would typically be expected at concentrations below the sensitivity limits of most non-targeted methods (e.g., below 1 ppm).

OVERVIEW

A non-targeted method for detecting adulteration is one which models the properties of the authentic material rather than the properties of the adulterants or any of the adulterant's characteristics. This type of non-targeted method is typically one of several potential tools used in a raw materials authentication scheme alongside other screening and confirmatory tools. As shown in *Figure 1,* these methods are usually carried out by comparing the measured sample (U) to a set of reference standards (Sn) representative of "Typical" samples using a preselected analytical procedure. Classification criteria are pre-established in order to classify the sample tested as "Typical" or "Atypical",[2] for example criteria based on statistical significance of a distance from the centroid and/or hyperplane of the model. These methods are generally (but not exclusively) multivariate and, if so, may include chemometric data preprocessing and analysis.

A non-targeted method for detecting food fraud/EMA asks the question: "Is the test sample (U) Typical or Atypical compared to a reference set (Sn) of Typical samples?" A "Typical" outcome suggests that, within the known performance of the method and the applied statistical conditions, the test sample (U) exhibits similar properties to the reference set (Sn). This outcome does not disprove the presence of adulterants, as the adulteration level could be below the limit of detection of the method, or the test sample may be adulterated with a material that the analytical method is not capable of detecting. An "Atypical" outcome suggests that the unknown sample is not consistent with the reference set and therefore possibly adulterated. A sample with an Atypical outcome could be a truly adulterated sample, or it could be an authentic, unadulterated sample with compositional or matrix parameters outside that represented in the reference set. A single Atypical result does not generally provide a sufficient degree of evidence to deem a material as adulterated, but rather should be a trigger for additional analyses to verify the nature of the material.

Non-targeted methods can be sensitive enough to detect anomalies from adulteration provided the unknown sample's properties are different enough from those captured in the model of the reference set. Everything else being equal, the more the derived signal of an adulterant (as produced by the method) differs from the profile of a Typical sample, the more sensitive the method will be to it. The higher the concentration of the adulterant, the farther the result will be from the centroid of the model, and the stronger the derived signal. In many cases, adulterants that are chemically similar to the authentic material (such as corn syrup used to adulterate honey, whey used to adulterate milk, or industrial grade food additives used in place of food-grade equivalent additive) may be less easily detected by non-targeted models than adulterants that are more distinct chemically from the Typical samples (such as high-nitrogen industrial chemicals).

Since it is impossible to validate such a method against all possible adulterants, a pragmatic approach is required. This guidance recognizes that non-targeted methods can in reality be non- or partially- targeted in terms of their creation, but will require some form of targeted validation to be of practical benefit.

1 Scampicchio, M., D. Eisenstecken, et al. (2016). "Multi-method Approach to Trace the Geographical Origin of Alpine Milk: a Case Study of Tyrol Region." *Food Analytical Methods* 9(5): 1262-1273.

2 This guide uses the terms Typical and Atypical in a pure binary sense; other authors have used terms including Positive and Negative, or Inclusivity Panel/Exclusivity Panel to describe fractions of probabilities falling wholly within one side or the other. This guide has used the descriptors "correct" and "incorrect" to describe the veracity of the result, while this has been referred to elsewhere as True/False. Intermediate fractions, which straddle the uncertainty region, are also given specific titles such as Specific Superior and Specific Inferior Test Materials (SSTM and SITM) [LaBudde & Harnly: *Journal of AOAC International* Vol. 95, no. 1, 2012]

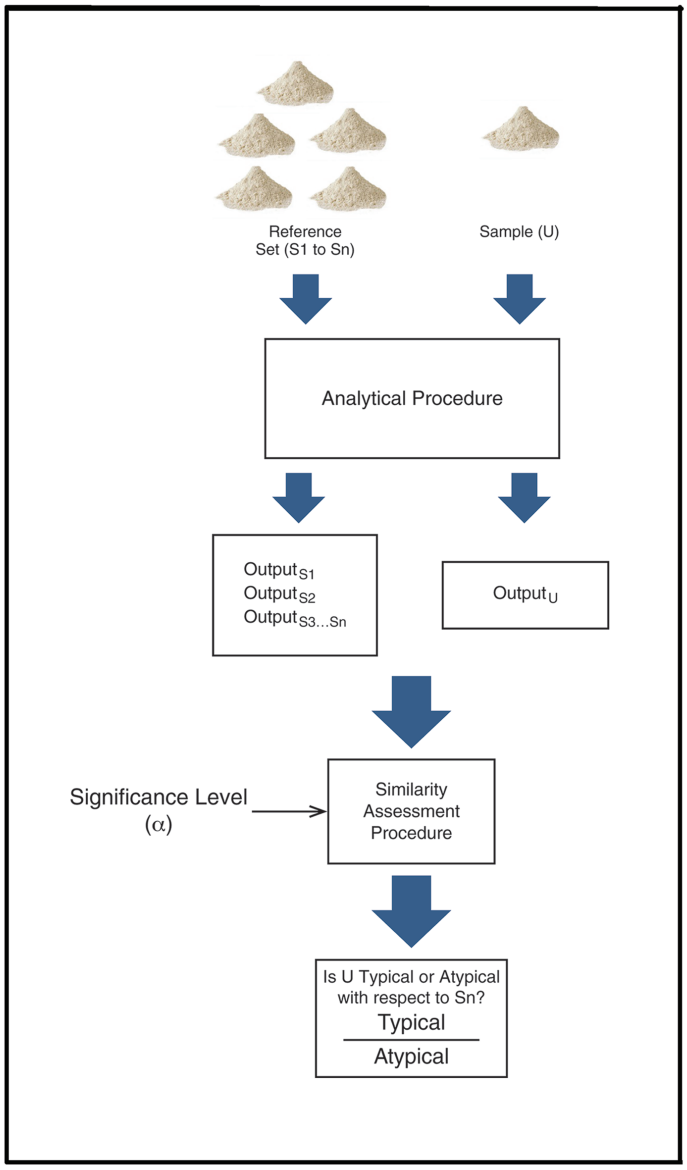

Figure 1: Essential elements of a non-targeted adulterants detection method used in authentication.

OUTLINE AND SCOPE

Glossary of Terms

Adulterant: Any undeclared biological or chemical agent, foreign matter, or other substance in food that may (though not necessarily) compromise food safety or suitability.[3]

Economically-Motivated Adulteration (EMA): More generally referred to as "food fraud," EMA is the fraudulent addition of non-authentic substances or removal or replacement of authentic substances without the purchaser's knowledge for the economic gain of the seller. An EMA-related adulterant (which is the focus of this guide) is an adulterant added to food by a supplier for economic gain.

Food Defense: Food defense is the protection of food products from intentional contamination or adulteration where there is an intent to cause public health harm and/or economic disruption.[4]

Incorrect Atypical Result: An incorrect Atypical result is a result for a Typical sample that the method has incorrectly identified as Atypical.[5]

Incorrect Typical Result: An incorrect Typical result is a result for an Atypical sample that the method has incorrectly identified as Typical.[6]

Limit of Detection (LoD): The lowest concentration of an adulterant that can be detected within the stated confidence interval of the method's result.

Model: A mathematical expression used to relate the response from an analytical instrument to the properties of samples, or to capture the underlying structure of a calibration data set.[7]

Model Boundaries: Multivariate or univariate boundaries that define whether an output is Atypical or Typical (with respect to Sn) in the similarity assessment procedure.

Non-targeted Method: A method that determines the similarity of a sample (U) to a reference standard or set (Sn). It has a binary output — the sample is Atypical or Typical with respect to the known sample set. The concept of non-targeted methods covers a spectrum from truly non-targeted (largely theoretical) to semi-targeted (most practical applications), but for the purposes of this paper any broadly nonspecific adulterant detection method is treated as non-targeted, as the same principles are applicable.

Sensitivity Rate: The ability to correctly recognize unacceptable samples/material as Atypical (i.e., possibly adulterated).[8]

Significance Level: For a non-targeted method, the significance level chosen to set model boundaries controls the rate of false declarations of nonconformity with Sn. The significance level is often expressed as a Confidence Interval (CI), typically 95% CI or 99% CI. As the significance level increases, fewer samples are rejected by chance, but fewer adulterated samples will be correctly rejected because the limit of detection (LOD) threshold increases.

Specificity Rate: The ability to correctly recognize acceptable samples/material as Typical (i.e. unlikely to be adulterated).[9]

Validation Set: A set of samples that are independent of the Reference set that are used to validate the method as a whole — this is composed of both Typical and Atypical samples.

EXAMPLES OF IN-SCOPE AND OUT-OF-SCOPE METHODS

In scope

1. Detection of adulterants in skim milk powder as an ingredient using near-infrared (NIR) spectroscopy. Unknown samples purported to be "skim milk powder" are measured using an NIR diffuse reflectance procedure with predefined conditions for sampling, scanning and preprocessing. The resulting spectra are statistically compared to the reference set of spectra, which is a broadly representative group of samples known to be genuine samples of "skim milk powders" of interest to the user. These reference samples were measured and preprocessed in the same pre-defined manner. The outcome would be either Typical, implying a lower probability of adulteration, or Atypical, implying an increased probability of adulteration.
2. Analysis of paprika by paper chromatographic fingerprinting. In this example ground paprika spice procured from a supplier is screened to ensure that it is not adulterated with unknown colors using a non-targeted analysis. Colors, in the case of food fraud/EMA, would be added at concentrations sufficient to enhance the color of the spice. The method prepares a color extract from the paprika sample which is spotted onto reverse phase chromatographic paper and developed. The chromatogram is visually compared to a reference set of chromatograms generated from samples of paprika thought to be genuine by the user using a predefined set of qualitative characteristics, e.g., the size and R_F of the spots in the chromatograms of the sample correspond to those of the reference set. The outcome would be either "Typical" for samples conforming to the predefined criteria and imply a lower probability of adulteration, or "Atypical" for samples not conforming to the criteria and indicating an increased probability of adulteration. A sample with an Atypical outcome could be a sample truly adulterated with a color, e.g. sudan IV, or it could be a genuine, unadulterated sample with processing or compositional parameters outside that represented in the reference set.

3 Adapted from the definition of "Food contaminant" from the *Codex Alimentarius* glossary at http://www.fao.org/3/a-y8705e/y8705e07.htm)
4 http://www.fda.gov/downloads/Food/FoodDefense/FoodDefensePrograms/UCM478509.pdf
5 An incorrect Atypical result could lead to increased costs through unnecessary inventory management activities and reference testing.
6 An incorrect Typical result would lead to a food safety concern should the true adulterant levels in the sample exceed food safety limits.
7 *PF* Online 41(6), 2015, USP general chapter <1039> *Chemometrics,* available at www. usppf.com
8 For non-targeted methods this is the number of correctly identified Atypical samples divided by the total number of (verified) Atypical samples.
9 For non-targeted methods this is the number of correctly identified Typical samples divided by the total number of (verified) Typical samples.

3. Analysis of raw milk by mid-infrared (MIR) spectroscopy. The method uses routine MIR analysis of liquid milk to build a database of typical samples. A chemometric approach is taken to define a model that suitably describes unadulterated milk that is routinely encountered. When a test sample is introduced to the analyzer, its spectrum is evaluated by the model and an assessment is made whether the sample is Typical or Atypical. An Atypical result is then followed up by a subsequent targeted (possibly quantitative) model to assist narrowing down the most likely adulterants present, before appropriate reference analytical tests are applied to determine the nature and concentration of the adulterant.
4. Analysis of non-protein contents of skim milk powder for detection of N-rich compound adulteration. This is a single analyte method, however it is nonspecific and for the purpose of this guide is considered to be a non-targeted method. The aim is to determine whether some unknown type of nitrogen compound has been added to the sample. The method uses a wet chemistry technique to segregate protein from non-protein nitrogen (NPN), and Kjeldahl is run on the latter to quantify its nitrogen contents. In this test, users are comparing the resulting non-protein nitrogen content of the unknown sample (U) to the statistically determined acceptance criteria range predefined by authentic skim milk powders; samples falling outside the range are deemed suspicious. A sample with an Atypical outcome could be a sample truly adulterated with a compound (e.g., melamine), or it could be a genuine, unadulterated sample with an inherent NPN content outside that represented in the reference set.
5. Analysis of glycerin using MIR spectroscopy for authentication (identification) purposes. The aim is to determine whether unknown EMA-related adulterants, e.g., diethylene glycol, have been added to the material. A reference standard glycerol spectrum is subtracted from an unknown sample absorbance spectrum so as to maximize residual IR peaks. The residual spectrum is analyzed via a human expert or a hit quality index algorithm to assess similarity to pure glycerin. Samples falling outside the set criteria (subjective or numerical) are deemed to be Atypical and therefore suspicious.

All the examples listed above are one-class classifiers—see *Appendix 1* for comparison to other one-class approaches. In each example, an Atypical result would require further testing with complimentary reference methods to provide verification. Refer to the section entitled *Interpretation and Next Steps for a "hit" (an Atypical result indicating that U is Atypical)* for guidance on interpretation of results.

Out of scope

1. Multi-class classifiers.
   - Multi-class classifiers are based on the characteristics of multiple reference standards and are outside the scope of this guidance.
     - Example: Authentication of olive oil geographic origin by NMR spectroscopy. A method that assigns the sample to one of four possible geographic origins (e.g., Italy, Spain, Morocco, or Portugal) is a multi-class classifier and outside the scope of this guidance. A method that assesses whether or not the sample is consistent with the expected geographic origin (e.g., Italy) could be a non-targeted method and thus within the scope of this guidance.
2. Quantitative methods for specific adulterants.
   - Methods that return a concentration of a known adulterant are considered targeted methods and thus outside the scope of this guidance. Non-targeted methods may have a quantitative component: Example 4 above, for instance, involves a quantitative measure of non-protein nitrogen. That method is in-scope because the measurement is nonspecific in nature.
     - Example: Determination of melamine, dicyandiamide, and cyromazine in milk powder by LC-MS/MS. While the method detects multiple adulterants, they are named species and thus it is a targeted method.
3. Food defense.
   - The deliberate contamination of food with the intent to cause harm is outside the scope of this document, as the adulterant quantities involved are typically orders of magnitude lower than EMA levels, and a substantially different approach is required.
     - Example: Following a hoax leveled against the New Zealand dairy industry, a routine screening program was developed and implemented for the pesticide Sodium Monofluoroacetate (referred to colloquially as 1080[10]) in raw milk and infant formula by LC-MS/MS. The method employed targets a single analyte and has an extremely low limit of quantification at 0.5 ng/mL. It is targeted, implemented for food defense reasons, and not economically motivated and therefore is out of scope.

10 Pronounced "Ten Eighty"; 1080 is routinely used by the New Zealand Department of Conservation for the control of various pests.

STEPS FOR DEVELOPMENT AND VALIDATION—THE GENERIC THOUGHT PROCESS

The process of generating a non-targeted method includes setting out the requirements through an Applicability statement, understanding the specific threats to be covered (if any), choosing an appropriate analytical technology and mathematical processing technique, creating and testing the reference set, generating the model and establishing the boundary, then testing and finally validating the model. For non-targeted methods, the validation process does not limit the scope of applicability of the method, but does give a degree of calibration for those adulterants validated.

Establish an Applicability Statement[11]

The applicability statement is a general statement about the intended purpose of the method—what must it do to be useful in your specific application? Key points to cover are the intended matrix, the purpose, and an indication of sensitivity, specificity, and significance.

When setting model boundaries there is a trade-off between the sensitivity rate and the specificity rate of the method, where the risk the organization is willing to accept must be balanced against possible increased inventory management and test costs. Increased risk can result from setting too low a sensitivity rate, which will result in a higher rate of incorrect Typical results, while increased costs can be incurred from too low a specificity rate, which may result in a higher rate of incorrect Atypical results. Identifying the most likely adulterants may help guide the choice of an appropriate acceptable risk level, and this may in turn guide appropriate sensitivity and specificity rates. Where possible, consult relevant risk assessment information as to levels of adulterant that make adulteration economically viable.

The risk level for both sensitivity and specificity rates should be explicit in the Applicability Statement as they will be critical in the Validation stage, a key decision point for the acceptance of the method.

- Example 1: "A rapid non-targeted method for detecting the adulteration of milk powder with nitrogen-rich compounds added at economically motivating levels (e.g., risk threshold = 0.1% for melamine, which is a food safety risk) with a sensitivity rate of 99% and a specificity rate of 95%, both with a Confidence Interval of 95%".
- Example 2: "A rapid non-targeted method for detecting the adulteration of milk powder with any foreign material at economically motivating levels (e.g., risk threshold = 5% for maltodextrin, which is a non-food safety risk) with a sensitivity rate of 90% and a specificity rate of 95%, both with a significance level of $p = 0.01$".[12]

Assess How to Determine Range and Levels of Adulterants to Validate the Model

Based on a thorough risk assessment, determine what types of adulterants are likely and at what concentration range(s) they could be present, taking into account the concentration level at which use of the adulterant becomes economically viable. In some cases, the intent may truly be to detect any possible type of adulterant, but in other cases a more narrowly defined scope may be sufficient.

Key aspects to consider in this assessment of adulterants:

- Cost of the adulterant;
- Ease of obtaining the adulterant in viable quantities (to meet the economically viable levels determined above);
- Compatibility with the matrix (e.g., solubility, color);
- Viable methods to adulterate the matrix;
- Economic benefit provided;
- Significance of the food safety impact on the food manufacturing process, if any, taking into account any dilution effect from other ingredient addition;
- Potential to violate market restrictions or agreements.

As appropriate, adulterants should be divided into classes based on chemical similarity (see *Figure 2,* e.g., small molecule high nitrogen containing compounds, vegetable protein isolates) as this will indicate both an appropriate analytical approach (next section) and possible other similar chemical compounds that could indicate other potential adulterants. It is essential this assessment is reviewed and updated regularly.

11 In López MI, Colomer N, Ruisánchez I, Callao MP. 2014. Validation of multivariate screening methodology. Case study: Detection of food fraud. *Analytica Chimica* Acta 827:28-33, López et al. use a multivariate screening methodology applied to IR ATR spectral data for detection of hazelnut adulteration by addition of almond and chickpea, with an adulteration range of 7%.
In Tengstrand E, Rosén J, Hellenäs K-E, Åberg KM. 2013. A concept study on non-targeted screening for chemical contaminants in food using liquid chromatography–mass spectrometry in combination with a metabolomics approach. *Anal Bioanal Chem* 405:1237–1243, Tengstrand et al. use a metabolomics approach on data obtained from UPLC-TOF-MS to detect very low levels of chemical contaminants in orange juice.

12 A significance of p=0.01 equates to a Confidence Interval of 99%

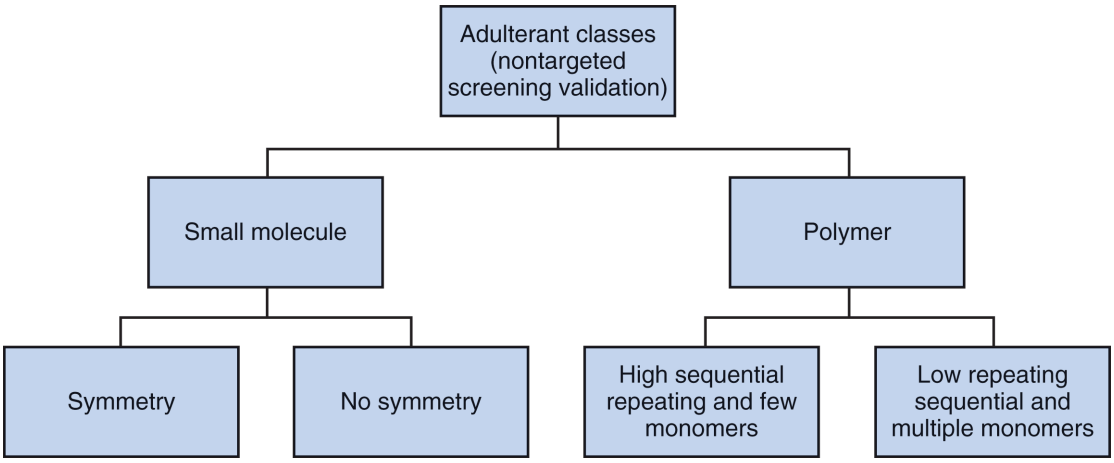

Figure 2: Adulterant class assessment based on chemical similarity.

Select an Appropriate Analytical Approach

Assess a range of viable analytical approaches that can rapidly test the matrices in question and have the potential to detect differences due to adulteration at economically motivating levels, and choose the most appropriate one.

- Example: Kjeldahl analysis of the non-protein nitrogen (NPN) fraction of milk powder isolated by tannic acid precipitation and comparison of the resulting NPN value to an established specification range that encompasses the NPN contents of authentic milk powders.[13]

Carry out small-scale experiments to characterize the performance potential of the method. This can be done by performing small-scale prevalidation studies, similar to those below in *Validation*, to determine the variance of authentic samples, the robustness of the chosen method, and the sensitivity of the method towards adulterants. A latter part of the analytical method selection process will include adoption of an appropriate statistical technique by which to analyze the results. The specific method used will depend upon the analytical technique and the applicability statement. These are comprehensively reviewed by Reidel et al[1411] and the USP general chapter *Chemometrics* <1039>.[15]

Method Development and Optimization[16]

Two sets of samples are required initially, a reference set, used in the creation of the models, and a test set, used to challenge the models for optimization. A validation set will also be required later and is detailed under that section.

It is recognized that heterogeneous materials will likely lead to non-normal distributions which can be a significant issue if extreme, therefore best practice sampling and modelling techniques must be used.

Method development and optimization for non-targeted methods are based on the reference set, with the boundary between Typical and Atypical being determined by the variability present in the reference set. Optimization of the model occurs through a process of prediction of the Test set using the model, comparing to the criteria set out in the Applicability Statement, and if necessary reducing the variability within the Reference set (by categorical splitting, perhaps into seasonality, production line etc.), and this cycle could be repeated as necessary to be aligned with the requirements of the Applicability Statement.

13 It is noted that this is a single analyte method, but the possible nitrogen containing compounds it will detect is nonspecific, hence considering it to be a non-targeted method.

14 Riedl J, Essenger S, Fauhl-Hassek C. 2015. Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. *Analytica Chimica* Acta 885:17-32

15 Guidance on multivariate analysis and chemometrics can be found in many published articles including USP <1039> *Chemometrics, PF* Online 41(6), 2015, general chapter <1039> *Chemometrics,* available at www. usppf.com.

16 López et al. refer to Typical samples as Compliant samples. Their reference sample set is comprised of 28 hazelnut pastes.
Tengstrand et al. refer to Typical samples as blank samples. Their reference set was 13 samples, all in duplicate.

REFERENCE SET

Define a population of representative authentic Typical samples and data acquisition conditions (different authentic sample variants, and instruments, operators, days of analysis, of relevance to the Applicability statement) and analyze them as the Reference set (Sn). There are two key factors to consider when selecting representative Typical samples:

- The selection of representative unadulterated samples or standards is critical to the success of any non-targeted method. This can be best assured by careful control and documentation of the samples' provenance, however secondary or reference methods can also be employed to verify the absence of likely contaminants.
- The samples must be representative of Typical. They must be sampled and handled in such a way to be representative of the bulk and also capture all variation that may exist in the normal sample population. Variation is typically assessed by review of sample quality parameters, and can arise from seasonal, aging, processing and sampling effects. All sources of potential variation must be considered and covered to form a truly representative sample set.

TEST SET

The test set is an approximately equal mix of Typical samples (which must not form part of the Reference set), and Atypical samples. As the name suggests, Test sets are used to test the adequacy of models created using the reference set, and also to optimize models that have been selected. These test set samples are one or a combination of:

- Known provenance (i.e., authentic samples with fully traceable history)
- Deliberately spiked
- Laboratory tested using reference test methods (to establish true adulteration levels, if any)

Ideally, at least three samples for each adulterant should be used, at varying levels of concentration—one at levels around the risk threshold that was identified in the Applicability Statement, the others around half and twice that level. The choice of adulterants should cover the classes previously identified and be randomly selected within those classes; additional specifically targeted adulterants may be included at this point where necessary.

If the decision is made to spike Typical samples, care should be taken in selecting which samples to spike. To avoid an optimistic assessment of the method sensitivity, select samples that are near the model centroid (and thus furthest from the model boundary), as these are likely to require higher spiking levels to trigger Atypical results. Alternatively, for a more heterogeneous material, a random selection of samples could be chosen, accepting that the overall variance will likely increase.

The adulterants of interest that were selected should not be mixed if possible, as only one adulterant type per sample should be used; care must be taken to ensure that deliberate spiking does not compromise the sample matrix unduly (the intention is to make the sample as "genuinely adulterated" as possible). The actual adulterant levels of each sample must be known, preferably through appropriate reference testing.

The test set should be analyzed using the chosen methodology in one or more laboratories, and changes to the method protocol and/or reference set composition may be necessary to further optimize it for non-targeted testing. Consider that any analytical method will have a range of inherent variables that when combined contribute to the overall uncertainty. It is important to distinguish method variables from sample variation that is covered above.

- Method variation includes such factors as variation between instruments, both on a one-off basis as well as how such variation may change over time.
- Methods with steps involving sample handling will be subject to sample homogeneity issues, and variation between different technicians.
- Methods involving the use of different chemicals or consumables (such as filters/chromatography columns) may be subject to variation between batches.
- Reference testing variation between different laboratories providing the adulterant analyses may require attention (improved reporting and QC limits).

In-silico methods (simulating an analysis computationally) may be a valuable addition to the validation of a non-targeted method. In a case where the way in which the adulterant contributes to the analytical signal is well understood, the signal may be reproduced synthetically. For example, with spectroscopic methods it is often (though not always) the case that the spectrum of a mixture is approximately equal to a linear combination of the spectra of the pure components, weighted in proportion to their concentrations.

An example workflow of in-silico method (assuming linear additivity of respective signals):

1. Develop and validate the method using carefully chosen adulterants and levels as described in this guidance.
2. Measure the analytical response of further (potential) adulterant species, either as pure materials or mixed in known proportion into samples of authentic material.
3. Generate potentially a very large number of synthetic analytical responses by combining the responses of the adulterants, at potentially numerous concentration levels, with the responses of one or more authentic samples.
4. Submit these synthetic data to analysis by the procedure.
5. Inspect the results for unexpected outcomes. For example, an adulterant that had been assigned to a general class of adul-

terants (expected to have similar responses) during the risk assessment may turn out to be significantly harder or easier to detect than other adulterants in the same group.

6. At this point the risk assessment can be revisited with new information about the sensitivity of the method towards various adulterants, and it may be deemed necessary to create additional spiked samples to extend the validation.

It is important to be aware of potential limitations with in-silico methods. The computational procedure for combining the adulterant response with that of the authentic material may not be equivalent to actually measuring a physical mixture. For example, the absorbance [or log(1/R)] intensity of near-infrared spectra when measured in diffuse reflectance is dependent on the light-scattering properties of the sample. If the physical form of the adulterant deviates significantly from that of the authentic material, simply scaling its spectrum by a proportional concentration may lead to a significant over- or under-estimation of the sensitivity. Additionally, if there is any chemical interaction or reaction between the adulterant and the matrix of authentic material, this may change its response either qualitatively or quantitatively. For these reasons, in-silico methods, while useful, cannot be used as a substitute for validation with realistically prepared samples.

Establish Performance Criteria

To meet the Applicability statement requirements, determine the sensitivity rate and the specificity rate:

Sensitivity rate is the number of **correct Atypical predictions** from the method divided by the **total** number of **true Atypical samples**.

$$\text{Specificity Rate} = \frac{\text{Correct Typicals}}{\text{Total Typicals}}$$

Specificity rate is the number of **correct Typical predictions** from the method divided by the **total** number of **true Typical samples**.

$$\text{Specificity Rate} = \frac{\text{Correct Typica}}{\text{Total Typical}}$$

From any known set of samples, each replicate will be reported as one of two conditions, Atypical or Typical, and this reported condition will either be Correct or Incorrect, as illustrated below:

Table 1: Possible Method Result States

		Actual Sample State	
		Typical	Atypical
Method Prediction[a]	Typical	Correct Typical	Incorrect Typical
	Atypical	Incorrect Atypical	Correct Atypical

[a] The López et al. paper refers to correct Typical as True Positive (TP), incorrect Typical as False Positive (FP), correct Atypical as True Negative (TN) and incorrect Atypical as False Negative (FN). This is the opposite to the definitions proposed in this guide. As a result of their chosen reporting convention, they have defined sensitivity as the ability of the model to recognize its own samples and specificity as the ability of the model to distinguish external samples; again, opposite to this guide.

Tengstrand et al. do not refer to either Typical or Atypical samples as such, but do refer to peaks generated by suspected contaminants or impurities as "positive", and identify a number of false-positive peaks in their work. These would be interpreted as Atypical and incorrect Atypical respectively in this guide.

Using the samples from the Test set, each sample is tested multiple times across all facilities that will be implementing the method to build a robust picture of the method reproducibility. The replicate results are pooled for each sample to allow calculation of the sensitivity or specificity rates. The rates are calculated for each sample, allowing indication of the sensitivity of the method towards each adulterant tested, at each concentration level that was tested. This will afford a range of results that will provide a good approximation of the expected:

- Correct Atypical result rate,[17]
- Incorrect Atypical result rate,
- Correct Typical result rate, and
- Incorrect Typical result rate,[18]

17 With the possible exception of spiked samples, which are manufactured and potentially not representative and therefore may lead to an optimistically high sensitivity rate

18 With the possible exception of spiked samples, which may lead to an optimistically high specificity rate as described in the text.

These results can be plotted as a receiver operating characteristic (ROC)[19] curve that illustrates the performance (probability of being deemed Atypical correctly vs probability of being deemed Atypical incorrectly) of any selected adulterant that has been tested at varying concentrations.[20]

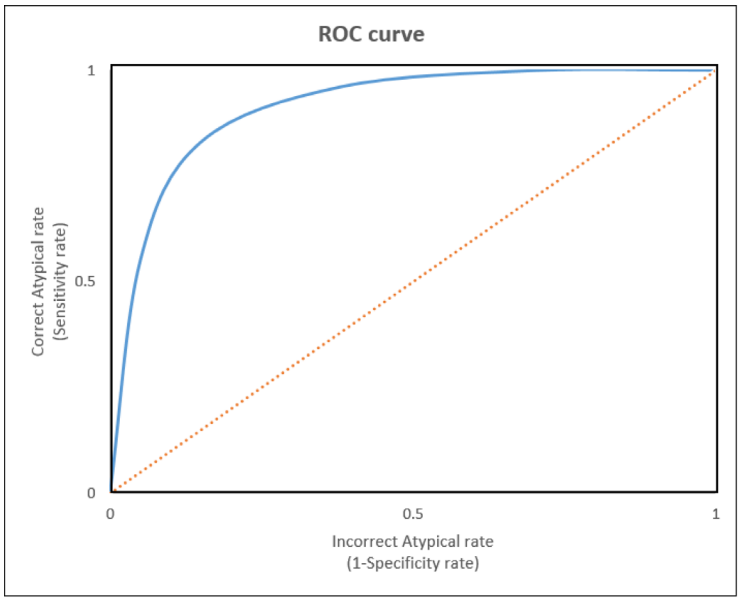

Figure 3: Example receiver operating characteristic (ROC) charts used to characterize the relationship between Correct Atypical rate (sensitivity rate) and Incorrect Typical rate (1-specificity rate) at various discrimination thresholds

The sensitivity and specificity rates are determined by the chosen discrimination threshold (the model boundaries)—see *Figure 3*. The area under the curve (AUC) is a good indication of the method performance. A random decision method is represented by the diagonal line with an AUC of 0.5, whereas a good non-targeted method should have an AUC closer to 1. This chart will provide visual assistance that will help determine whether the method is likely to be suitable for use.

The acceptable rates are a business and technical decision that will depend on the action required when an alert is triggered, as well as results from a business risk analysis. A relatively low specificity rate may be acceptable if the method is used as a low-level screen and several sequential Atypical results are required to trigger an alert. A higher specificity rate might be used if a single result is to be used to trigger action, such as refusing to accept a raw material. Increased focus on the Incorrect Atypical rate over the Incorrect Typical rate is generally appropriate in cases where models are built and/or validated with artificially manipulated samples (that is, samples generated for the purposes of method development) as these may not be completely representative. Charts should be generated to assist choosing the optimal threshold for the most robust and appropriate model, which meets the parameters of the Applicability statement.

Validate the Optimized Method[21]

The validation criteria can be assessed against two sample sets, Atypical and Typical (or there can be one set with both types of samples), of approximately the same size. It is important that the samples used for validation are independent from any samples used in the Reference set or any samples used establishing performance criteria. Number of samples used in this experiment will vary depending on range and number of classes of adulterants.

19 In López et al, the ROC curves are derived similarly, and depict sensitivity against "1-specificity" for each of the thresholds studied. The optimal threshold is the one that shows highest values for both sensitivity and specificity. Once the model boundaries are established by setting the optimal α value, the final quality parameters are calculated.
Tengstrand et al. do not use ROC charts in their paper.

20 "Introduction to Statistical Quality Control" (7th Edition), D.C. Montgomery, ISBN-13: 978-1-118-32416-5.

21 Validation in the López et al. paper was performed using samples spiked with almond paste or chickpea flour at various levels (1%–8%).
Tengstrand et al. use a validation set of three samples (in duplicate) each spiked with either seven mycotoxins at 4 µg/mL, or 18 pesticides at 25 µg/mL, or one pharmaceutical (sulfadoxin) at 1,000 µg/mL.

ATYPICAL SAMPLES

Atypical samples can be:

- Samples from true adulterated product, with identified adulterants and known concentrations from reliable reference methods;
- Deliberately adulterated reference samples or reference standards (e.g. USP's Skim Milk Powder with Melamine reference standards);
- Deliberately spiked samples containing a specific adulterant at a known concentration.

The expected outcome is that these Atypical samples must be reliably identified by the method within the specified acceptable sensitivity criteria from the Applicability Statement. Ideally, at least 3 samples for each adulterant chosen for validation should be used, at varying levels of concentration—one at levels around the risk threshold that was identified in the Applicability Statement, the others around half and twice that level. The choice of adulterants should cover the classes previously identified and be randomly selected within those classes. Ideally, only one adulterant type per sample should be used, and care must be taken to ensure that deliberate spiking does not compromise the sample matrix unduly (the intention is to make the sample as "genuinely adulterated" as possible). These samples are analyzed as Test samples, U, during the experiment.

From these data the sensitivity rate for each sample needs to be calculated. This is only valid for the adulterants and concentrations presented to the method, but will be indicative for the families of those adulterants. The proportion and nature of Incorrect Typical outcomes (1–sensitivity rate) are of principle concern as an indication of the effectiveness of the method in a food safety application. To assist with the selection of adulterants to validate against, start with the Applicability statement. From there select adulterants from three groups:

- Group 1: High-priority adulterants based on historical information that are in the scope of the Applicability statement.
  - Example (Non-protein nitrogen method): melamine, urea
- Group 2: Other plausible and representative adulterants (consider classes based on chemistry) in scope of the Applicability statement based on knowledge (chemical, sensory, and physical properties; economics). An example of how to consider classes for spectroscopic techniques is provided in *Figure 2*.
  - Example (Non-protein nitrogen method): dicyandiamide, IBDU (inexpensive N-rich compounds, widely available)
- Group 3: Randomly select some additional adulterants that represent different in-scope classes to investigate the generalizability of the method performance results (how well would it detect a random N-rich compound that we would not expect). The more you select the test, the more you can infer.
  - Example (Non-protein nitrogen method): L-arginine, aminotriazole

Follow the section on method development and optimization above to determine how to select authentic samples to spike adulterants into. The sensitivity rate for the model needs to meet that which is specified in the Applicability Statement *at the risk threshold for the specific application*, however the concentration ranges chosen should cover both lower and higher levels that must display increasing and decreasing sensitivity rates respectively.

TYPICAL SAMPLES

A set of randomly chosen authentic samples are included in the validation set and analyzed as unknown samples, U, during the experiments to establish the specificity rate. The expected outcome is that these Typical samples must be reliably identified within the specified acceptable specificity criteria. When analyzing the validation data there are a number of questions that the validation process needs to answer:

- For what type of adulterants is the method likely to fail? This information helps form the scope of applicability of the method.
- What is the sensitivity of the method at the predetermined risk threshold?
- How generalizable are the sensitivity results to other adulterants? This question aims to determine how much you can infer about the sensitivity of the method for other adulterants not tested during validation. It depends on the design of the adulterants chosen for the validation study (see previous section *Assess how to determine range of adulterants and levels).*
- Specificity: what is the likelihood of a good sample failing the test? This is the Incorrect Atypical rate, which is independent of the presence of any adulterants, and is determined using the results from the Typical (authentic) samples validation set. It is important to determine the likelihood of a good sample failing the test and whether there are any non-adulteration variations/defects that might cause an unadulterated sample to fail the test.

USING/MAINTAINING/MONITORING/REVALIDATION[22]

All methods are built on historical data, but to remain useful they need to be monitored for performance. Some examples of where this variation can originate from are listed below- they are components of the method and variation in any one area can lead to invalidation of the method over time:

- Instrumentation
  - Mechanical deterioration
  - Detector drift
- Test method
  - Environmental changes
  - Consumable supplier changes
- Analyzed material
  - Matrix effects in the material from processing changes
  - Seasonal or environmental shifts impacting composition

Monitoring

While validation is generally a one-off process to certify a method, ensuring that a method remains effective over time requires regular and ongoing monitoring. Monitoring is practiced by testing known samples at a chosen frequency and analyzing the output for drift or step-changes. The same sample selection criteria as applied for validation also applies to selection of monitoring samples.

Where possible, a number of representative and unadulterated Typical samples should be regularly collected (from the source where possible, or of known provenance otherwise) and split into multiple sub-samples. A proportion of these bulk samples should be adulterated through spiking in the same way and with the same adulterants as used for the original validation, ensuring the samples are homogenous prior to splitting. These spiked samples must be analyzed using reference methods, and will allow ongoing validation of the sensitivity rate.

At each monitoring event, a randomly selected set of these subsamples (representing an appropriate diversity of processing variables and adulterated/unadulterated samples) is then measured at the chosen frequency. The **incorrect** results for both Typical and Atypical samples from all the instruments are collated as soon as possible after each monitoring event. Once enough have been collected (minimum of 10 of each class of sample), the in-practice sensitivity and specificity rates are calculated and compared against the stipulations in the Applicability statement.

The incorrect results should also be plotted over time using the reference results, so that a statistical analysis can be performed to determine the current state of the method and to monitor systematic trends in the method performance. In addition, statistically significant shifts in the measured parameter (e.g. spectral changes for NIR/MIR) of Typical samples are sought; these will trigger an alert that there has been a change in the overall system or sample matrix that must be investigated.

Monitoring can be performed locally at the instrument, or centrally if there are multiple installations:

- Local monitoring, where routine pilot samples are sourced and applied locally to monitor instrument performance; however, this will only be comprised of Typical samples and hence will not monitor performance of the method against Atypical samples;
- A sub set of the validation sample set with both routine and selectively spiked samples/known contaminated samples could be applied and monitored locally;
- Ring trial (Inter-laboratory monitoring)—if the method is applied across an instrument network then an ongoing ring lab system should be established. Bulk samples are prepared as above and sent to each of the test facilities for analysis. Results are collated and the performance of each individual instrument is monitored by a central team.

Good monitoring practice will encompass the following:

- **Frequency:** Routine monitoring will take the form of an experiment similar to validation but on a smaller scale, repeated at regular intervals and documented with appropriate statistics and control charts. The period between monitoring needs to provide confidence in ongoing testing, balanced against budgetary and downtime considerations. For example, daily monitoring may be appropriate in a dedicated process lab within the manufacturing facility, but only monthly in a commercial lab with low testing volumes for the product requiring non-targeted testing.
- **Suitable sample numbers:** The number and variety of samples used for monitoring must be sufficient to allow reasonable assessment of actual performance in a timely manner—too few samples may allow undue influence over the monitoring, as will insufficient variety.
- **Stability:** The samples used must be shown to be stable over time, and with regards to any physical handling they will be subjected to. Consideration should be given to breaking bulk samples into multiple single-use samples that are individually sealed and stored in controlled conditions in order to minimize the need for unnecessary handling.
- **Charting[23]:** There should be appropriate graphing of the monitoring sample results in terms of rate of alerts, daily mean/median results, and variations. Limits need to be clearly defined, as well as actions to be taken upon limit breach.

22 *PF* Online 41(6), 2015, general chapter <1039> *Chemometrics*, available at www. usppf.com

23 "Some Theory of Sampling", W.E. Deming, Dover Publications, ISBN-13: 978-0486646848, ISBN-10: 048664684X.

- **Homogeneity:** If bulk samples are to be split and used repeatedly for monitoring, the samples used need to be as homogenous as possible. This is standard laboratory practice but bears repeating as non-homogeneity can lead to significant loss of confidence in the method.

Internal Control Plan

When monitoring a non-targeted method, one is typically tracking the number of instances and scale of both incorrect Typical results and incorrect Atypical results over a set period of time or events. Results should show the performance of the method on the samples vs. the expected results, for each instrument.

It is important to record trends over time so emerging or ongoing issues with individual instruments or the method as a whole can be detected. Control charting can be used to show any shift from the center point of the data.[24] A statistical approach should be taken when analyzing the data[25] for drift in order to tune the monitoring system's reaction point—too fast and the system may react to noise, but too slow could lead to important influencing factors to be missed for several monitoring cycles. Both the cumulative sum (CUSUM) and the exponentially weighted moving average (EWMA) methods have been shown to be effective tools for monitoring.

The frequency of control charting is a business decision that should be based on factors including the number of analyses performed between monitoring events, and a view of historic variation. Often a new system will be monitored and charted at a high frequency, this can be relaxed over time once some sense of inherent variation over time is established.

Method Updates

Method updates are required when there has been a significant change in samples being analyzed (due to season, climate other changes), or the existing models are performing less effectively in routine monitoring. An updated method will be created by including new input data alongside the existing data and carrying out an internal validation assessment of performance for the method. This will be contrasted with the performance of the existing method using the same validation set and a decision on updating reached. The updated methods can then be installed and validated as described earlier.

INTERPRETATION AND NEXT STEPS FOR A "HIT" (AN ATYPICAL RESULT INDICATING THAT U IS ATYPICAL)

Result Monitoring and Trending

Any qualitative screening system will produce a significant amount of data in the form of many Typical results interspersed with a few Atypical results. Each of these Atypical results will be linked to a specific test on a "lot" of material. That "lot" of material must have accompanying information that uniquely identifies it. Over time it is important to plot the incidence of Atypical results and determine if there are any obvious trends—preferably using statistical control chart techniques that are tuned to the specific system needs.

For example:

- Liquid milk being purchased from a range of farms—an Atypical result on a single day from one farm could be due to random error; the same result from the same farm across multiple successive days should trigger a follow-up action.
- Olive oil being supplied from various suppliers from a specific geographic region under protected geographical indications—if multiple suppliers receive Atypical alerts at the same time, this could be an indicator of method degradation (e.g., calibration model drift), but a single supplier receiving multiple Atypical alerts would need to be further investigated for possible adulteration with oils from other regions.

The method of plotting these data will depend upon the volume and nature of results; the most important factor is to be able to identify trends over time. The types of trends will become more apparent as more data are collected over time.

Follow-up Actions

REFERENCE TESTING

Once a sample is identified as an Atypical result this should be confirmed with a routine reference laboratory test to verify that the sample is unusual. The type of reference test(s) used will depend on the nature of the risk(s) for that area and food type. If water into milk is a common risk, then a freezing point depression test would be used.

CONFIRMATORY ANALYSES

These can be distinguished from reference testing in that they would be carried out in an accredited laboratory and provide a result that could be legally credible.

---

24 "Engineering Statistics" (5th edition), D.C. Montgomery, G.C. Runger, N.F. Hubele, ISBN-13: 978-0470631478, ISBN-10: 0470631473, Wiley (2010).

25 See USP general chapter *Analytical Data—Interpretation and Treatment* <1010>, accessible at http://www.uspnf.com for information on acceptable practices of data analysis of chemical and other analyses.

ACTION TO PREVENT FURTHER ADULTERATION

Depending on the nature of the adulteration, type of food and business model, an action to prevent or reduce the risk of adulteration may be taken. This may range from a low-level acknowledgement of an issue with the raw material being screened to legal action based on confirmatory analyses.

APPENDIX 1

Table 2: Comparison of Non-Targeted Methods to Other Approaches Used in Authentication

Function of Method	Components		Method	General Sensitivity for Detecting Adulterants	Authentic Sample Misclassification Rate
	Primary Material[a]	Secondary Material[b]			
Authentication (called identification at USP)	Known for reference samples (1 class)	Not known	1-class classifier	Low	Very Low (target is 0% misclassification rate of authentic samples)
Non-targeted detection of adulterants	Known for reference samples (1 class)	Not known	1-class classifier	Medium	Low (but case-by-case determination based on performance needs)
Targeted detection of adulterants	Known for ref samples (1 class)	Known for 1 or more materials	Direct analysis	High	NA

[a]Primary material—material to be authenticated, identified, or tested for adulterants.
[b]Secondary material—contaminants, adulterants, or substituted materials.

APPENDIX 2

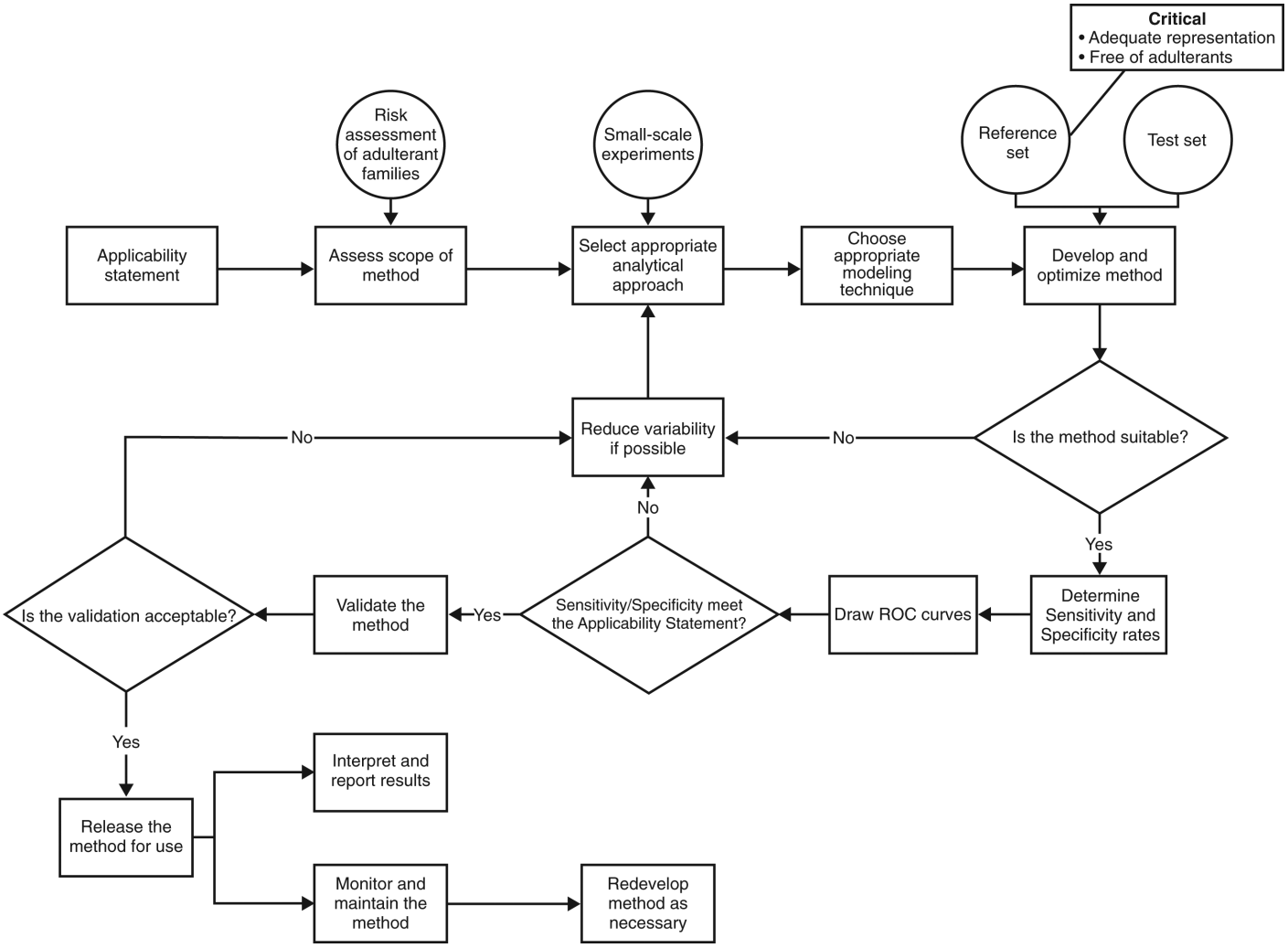

Figure 4: Flowchart of critical steps in non-targeted method development and validation.

▲3S (FCC 10)