BRIEFING

Appendix XVIII: USP Guidance on Developing and Validating Non-Targeted Methods for Adulteration Detection, *FCC 11* page 1510. On the basis of comments and data received, it is proposed to add a commercial example to illustrate the application of the guidance document.

Additionally, minor editorial changes have been made to update this appendix to current *USP* style.

This proposal is targeted for publication in the *Third Supplement* to *FCC 11*.

(FI: K. Xie) C200932

Change to read:

Appendix XVIII: ~~USP~~

▲▲3S (FCC 11)

Guidance on Developing and Validating Non-Targeted Methods for Adulteration Detection

Proposed in: Dec 2018

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.

The content in this Guidance is intended to be used as an informational tool and resource only. It is not intended to and does not constitute legal advice and is not warranted or guaranteed by USP to be accurate, complete, adequate, or current. Your use of the content in the Guidance is at your own risk. USP accepts no responsibility or legal liability for the use and/or accuracy of the Guidance or for decisions based on its content.

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 material: 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,[1] 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 multiclass classification methods 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).[2]

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 (e.g., 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. Furthermore, such 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, and can be generated by any laboratory analysis technique that produces data from which patterns can be mathematically extracted or provides quantitative measurements that characterize authentic samples. 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 samples (Sn) representative of "Typical" samples using a preselected analytical procedure. Classification criteria are pre-established in order to classify the sample tested as "Typical" (implying a lower probability of adulteration) or "Atypical" (implying an increased probability of adulteration).[3] 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, however an authentic, unadulterated sample with compositional or matrix parameters outside that represented in the reference set could also show as Atypical. 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 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; hence this guidance recognizes that non-targeted methods can in reality be non- or partially targeted in terms of their design, but will require some form of targeted validation to be of practical benefit.

Lastly, non-targeted methods are best employed using samples from material as close to the source of adulteration as possible, prior to any downstream processing—this is usually at the inventory receiving point of the material.

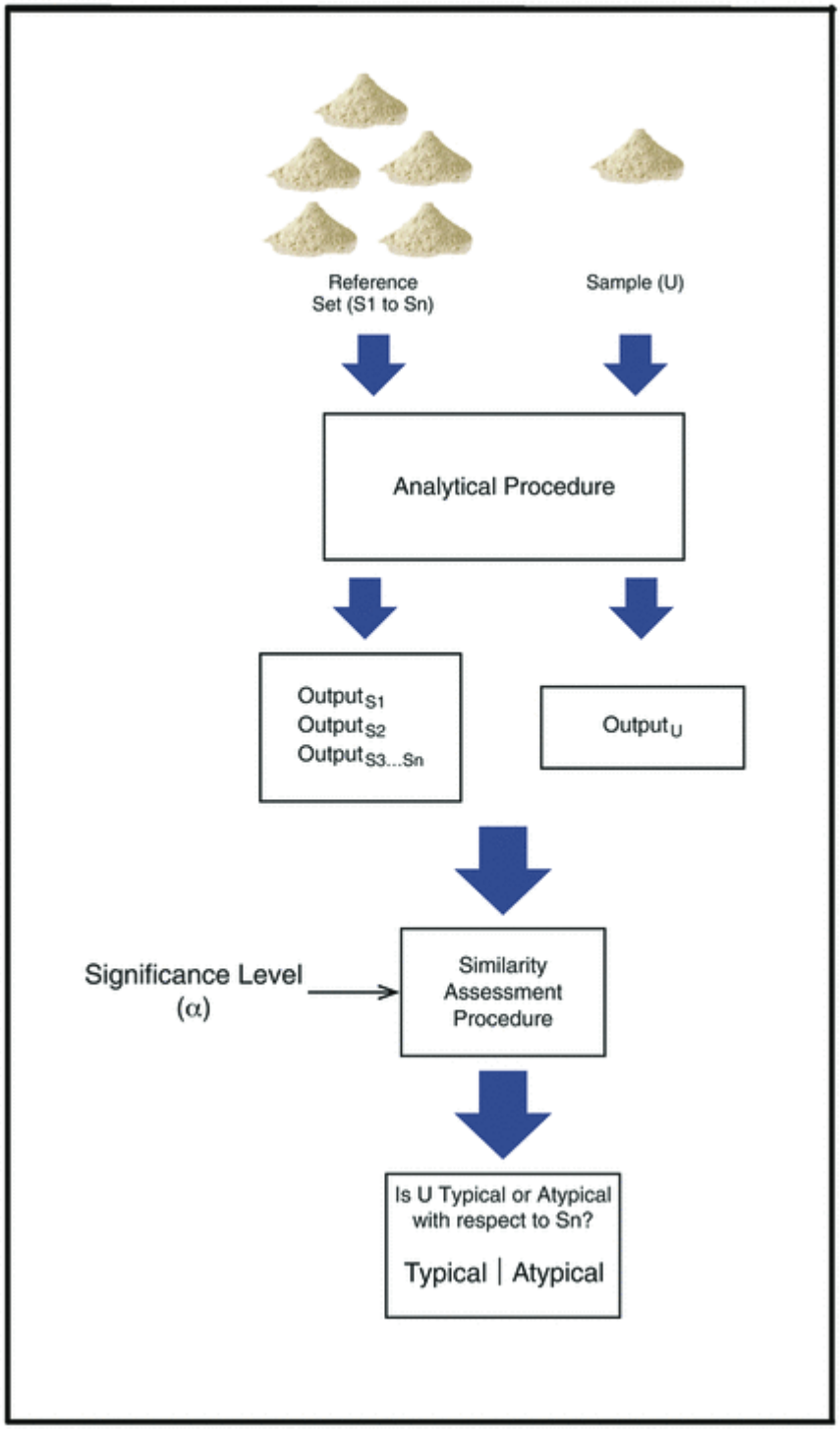

Figure 1. Essential elements of a non-targeted adulterants detection method used in authentication.

Outline and Scope

Glossary of Terms

Adulterant: Any undeclared biological or chemical agent, foreign matter, or other substance in food which may (though not necessarily) compromise food safety or suitability.[4]

Atypical sample: An Atypical sample is a sample that is not representative of one of the variations of the material for which the method is being developed. It is known to be contaminated by at least one adulterant, and has been either selected due to this fact or has been created by spiking a Typical sample with an adulterant.

Correct Atypical result: A Correct Atypical result is a result for an Atypical sample that the method has correctly identified as Atypical.

Correct Typical result: A Correct Typical result is a result for a Typical sample that the method has correctly identified as Typical.

Economically motivated adulteration (EMA): More generally referred to as "food fraud," EMA is the fraudulent addition of nonauthentic 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 guidance) 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.[5]

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

Incorrect Typical result: An incorrect Typical result is a result for an Atypical sample that the method has incorrectly identified as Typical.[7]

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.[8]

Model boundaries: Multivariate or univariate boundaries that define whether an output is Atypical or Typical (with respect to Sn) in the similarity assessment procedure, also referred to as Decision Thresholds.

Multiclass (or Two-class) classification: Samples are assigned to two (or more) classes and a model is constructed based on the combined characteristics of all the classes.[9] This is also known as hard modeling since unknowns must be assigned to one of the classes. Partial least squares-discriminant analysis (PLS-DA) is an example of two-class classification.

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.

One-class classification: Samples are assigned to one or more classes and a model is constructed for each class (one or more) based only on the characteristics of the class.[10] This is known as soft modeling as unknowns may belong to one class, more than one class, or none of the classes. Soft independent modeling of class analogy (SIMCA) is an example of one-class classification. As applied to non-targeted adulterant testing, a model is constructed for only one class, the Typical samples.

Reference set: A set of samples that are used to generate the method model—this is composed of Typical samples only. The samples in the reference set must fully encompass all the variability from all pertinent attributes of the material that will be analyzed using the method once it has been developed.

Sensitivity rate: The ability to correctly recognize unacceptable samples/material as Atypical (i.e., possibly adulterated).[11]

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 detected because the threshold increases.

Specificity rate: The ability to correctly recognize acceptable samples/material as Typical (i.e., unlikely to be adulterated).[12]

Test set: A set of samples that are used in the optimization of the method model—this is composed of both Typical and Atypical samples that must encompass the same degree of variability of the target material as the Reference set.

Typical sample: A Typical sample is a sample that is representative of one of the variations of the material for which the method is being developed. It must be known to be free of any adulterants, and has been selected randomly from the same sampling point and in the same manner as a routinely encountered sample would be.

Validation set: A set of samples that are independent of the Reference and Test sets that are used to validate the method as a whole—this is composed of both Typical and Atypical samples that must encompass the same degree of variability of the target material as the Reference and Test sets.

Change to read:

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 predefined 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 adulterated with a color, e.g., sudan IV, or it could be a genuine, unadulterated sample with processing or compositional parameters outside those represented in the reference set.

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 using the model and an assessment is made whether the sample is Typical or Atypical. An Atypical result is then followed 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 nonprotein 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 guidance 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 nonprotein nitrogen (NPN), and Kjeldahl is run on the latter to quantify its nitrogen contents. In this test, users are comparing the resulting nonprotein 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. Contamination of Black Cohosh root, which is used as a dietary supplement, with plant material from other species. A combination of flow injection mass spectrometry (FIMS) and nuclear magnetic resonance spectrometry (NMR) are used to detect whether Black Cohosh (*Actaea racemosa*)[13] root has been adulterated with plant material from other species. A reference set of Black Cohosh root material is created and initially assessed by DNA sequencing to ensure authenticity; a spectral fingerprint of the reference set is then derived by modeling spectra of the reference set samples from analysis by FIMS and NMR. This fingerprint is then used as a one-class classification method to distinguish samples that are dissimilar to the samples that comprise the reference set. An unknown sample is analyzed using the same techniques under the same conditions and the resultant spectra are compared to the fingerprint to determine if the spectra is statistically similar to the samples in the reference set.
6. 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 an 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 ▲*14*▲3S (FCC 11)—see ~~*Appendix 1*~~▲*Annex 1*▲*3S (FCC 11)* 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)* for guidance on interpretation of results.

Out of Scope

1. Multiclass classifiers.
   - Multiclass 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 multiclass 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 nonprotein 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 threat against the New Zealand dairy industry, a routine screening program was developed and implemented for the pesticide Sodium Monofluoroacetate (known in New Zealand as 1080)[15] 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.

Change to read:

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, validating, and implementing appropriate quality assurance measures for the model. Each step may need revisiting to refine it further, including the Applicability Statement. The validation process does not limit the scope of applicability of the method, but does give a degree of calibration for those adulterants involved in the validation process.

See *Figure 3* in ~~*Appendix 1*~~ ▲*Annex 1* ▲*3S (FCC 11)* for a flowchart summarizing the process.

Establish an Applicability Statement[16]

The applicability statement is a general statement about the intended purpose and scope of the method—it entails key aspects of expected achievements for the specific situation and circumstances. 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 or laboratory customer is willing to accept must be balanced against possible increased inventory management and test costs. This 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 and consequential restocking and production disruptions. 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 or revision of the method.

Any deliberate restrictions of application of the method should also be stated at this stage, otherwise the method should cover all nominated product that will be routinely encountered (i.e., if the method will only be employed for product from a particular supplier or country, or the assumption will be all stipulated product regardless of origin), understanding that a broader application will have more variability, which in turn will require a larger reference and test set to generate a successful 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."[17]

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 may be divided into classes based on chemical similarity (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.

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 nonprotein 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.[18]

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. An important 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 Riedl et al.[19] and the USP general chapter *Chemometrics* ‹1039›. [20]

Method Development and Optimization[21]

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. The number of samples required will depend on the complexity of the model. It is recognized that heterogeneous materials will likely lead to nonnormal distributions, which can be a significant issue if extreme; therefore best practice sampling and modeling techniques must be used, both to ensure full representation of the intended target material, and homogenous samples to allow analytically identical sample splitting.

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 removal of extreme outliers, determined using appropriate statistical techniques) or 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. This repetition must be performed with acknowledgement that each cycle will further narrow the applicable scope accordingly.

Comprehensive capture of sample metadata is imperative to ensure robust non-targeted methods, as the detailed description of sample provenance, handling, and analysis can greatly assist in understanding incorrect predictions the model may generate in the Test set, which in turn can assist in any eventual decision to limit the scope of application of the method to a particular subset of the Reference samples (e.g., to remove all samples from a particular geographic area of origin, etc.). Similarly, during the modeling process, any samples removed as outliers need to be fully justified and documented.

REFERENCE SET

Define a population of fully representative authentic Typical samples (covering all relevant geographic regions of origin, growing seasons, etc.) and sample data acquisition conditions (different authentic sample variants, and instruments, operators, days of analysis, etc.) 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 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, and secondary or reference methods can also be employed to verify the absence of those adulterants defined in the scope.
- The samples must be representative of Typical. They must be sampled and handled in such a way to be representative of routine samples, 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:

- Typical samples of known provenance (i.e., known nonadulterated samples that have been authentically sourced with fully traceable history)
- Genuinely adulterated (i.e., Atypical) samples that have been laboratory tested using reference methods (to establish actual adulterant identities and levels)
- Typical samples that have been deliberately spiked.

Ideally, at least three samples for each selected adulterant should be used, at varying levels of concentration—one at levels around the risk threshold that was specified in the Applicability Statement, the others around half and somewhat exceeding 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. Routinely sourced adulterated samples are obviously preferred in the selection of Atypical samples, as these are truly representative of target material, however these can be difficult to procure, and therefore might be saved for the Validation stage. 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, either use an appropriate sample-selection algorithm, or 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 (or again, an appropriate algorithm may be preferred).

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, and consideration should be given to the purity of the adulterants used for spiking, as the adulterants used in EMA may be of poor quality. 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.[22] 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 adulterants (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 for the chosen model at various significance levels, or decision thresholds:

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

$$Sensitivity\ Rate = \frac{Correct\ Atypicals}{Total\ Atypicals}$$

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

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

From any known set of sample, each replicate will be reported as one of two conditions at each decision threshold, 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 guidance. 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 guidance. 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 guidance.

Using all 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 repeatability and reproducibility for each nominated adulterant at varying concentrations, and at each of the chosen decision thresholds. The replicate results for each sample are aggregated to allow calculation of the performance 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,[23]
- Incorrect Atypical result rate,
- Correct Typical result rate, and
- Incorrect Typical result rate,[24] as functions of adulterant and concentration.

These results for each sample at various decision thresholds can be plotted as a receiver operating characteristic (ROC)[25] 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.[26]

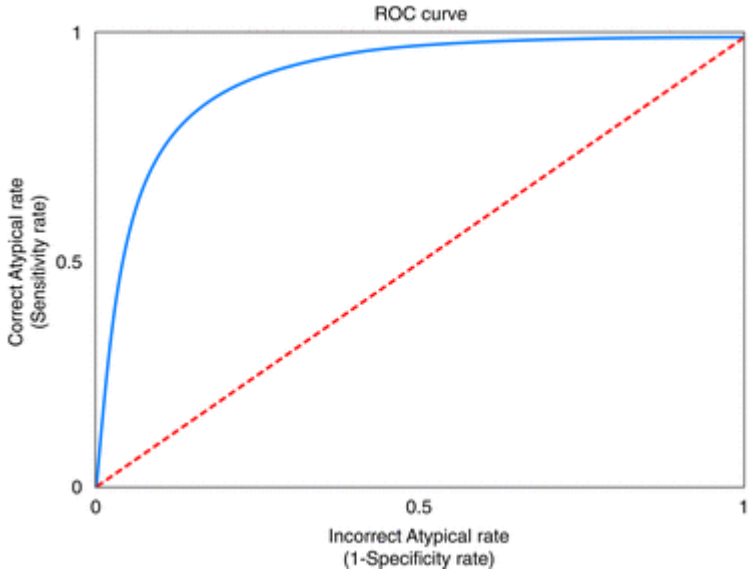

Figure 2: Example receiver operating characteristic (ROC) charts used to characterize the relationship between Correct Atypical rate (sensitivity rate) and Incorrect Typical rate (1-specificity rate) at various

decision thresholds.

The sensitivity and specificity rates are determined by the chosen decision threshold (the model boundaries)—see *Figure 2*. 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, or whether some aspect needs to be revised.

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 before triggering an alert (i.e., a higher tolerance for Incorrect Atypical results). 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 (i.e., a lower tolerance for Incorrect Atypical results). 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 spiked 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[27]

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 in the Test set to establish performance criteria. Number of samples used in this experiment will vary depending on range and number of classes of adulterants.

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.

At least three 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 somewhat exceeding 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 no aspect of deliberate spiking compromises 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: (Nonprotein 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).
  - Example: (Nonprotein nitrogen method): dicyandiamide, isobutylidene diurea (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 to test, the more you can infer.
  - Example: (Nonprotein 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 an authentic 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 an authentic sample failing the test and whether there are any nonadulteration variations/defects that might cause an unadulterated sample to fail the test.

Quality Assurance and Control Measures

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 in laboratory
  - Consumable supplier changes
  - Operator induced variability

- Analyzed material
  - Matrix effects in the material from processing changes
  - Seasonal or environmental shifts impacting composition

Method 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 incorrect predictions each of Typical and Atypical samples), 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 versus 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 which 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 not monitor performance of the method against Atypical samples;
- A subset 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 verification, 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.
- **Control charting:**[28]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.
- **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.[29]

A statistical approach should be taken when analyzing the data[30] 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 (updated) 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)

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 a different type of pasture on the same farm or supplementary feeding; 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., samples from new tree species not covered by calibration model, or 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 ANALYSIS

These can be distinguished from reference testing in that they would be carried out in a qualified laboratory and provide a result that could be scientifically defensible.

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.

Change to read:

~~Appendix 1~~

▲Annex 1▲ 3S (FCC 11)

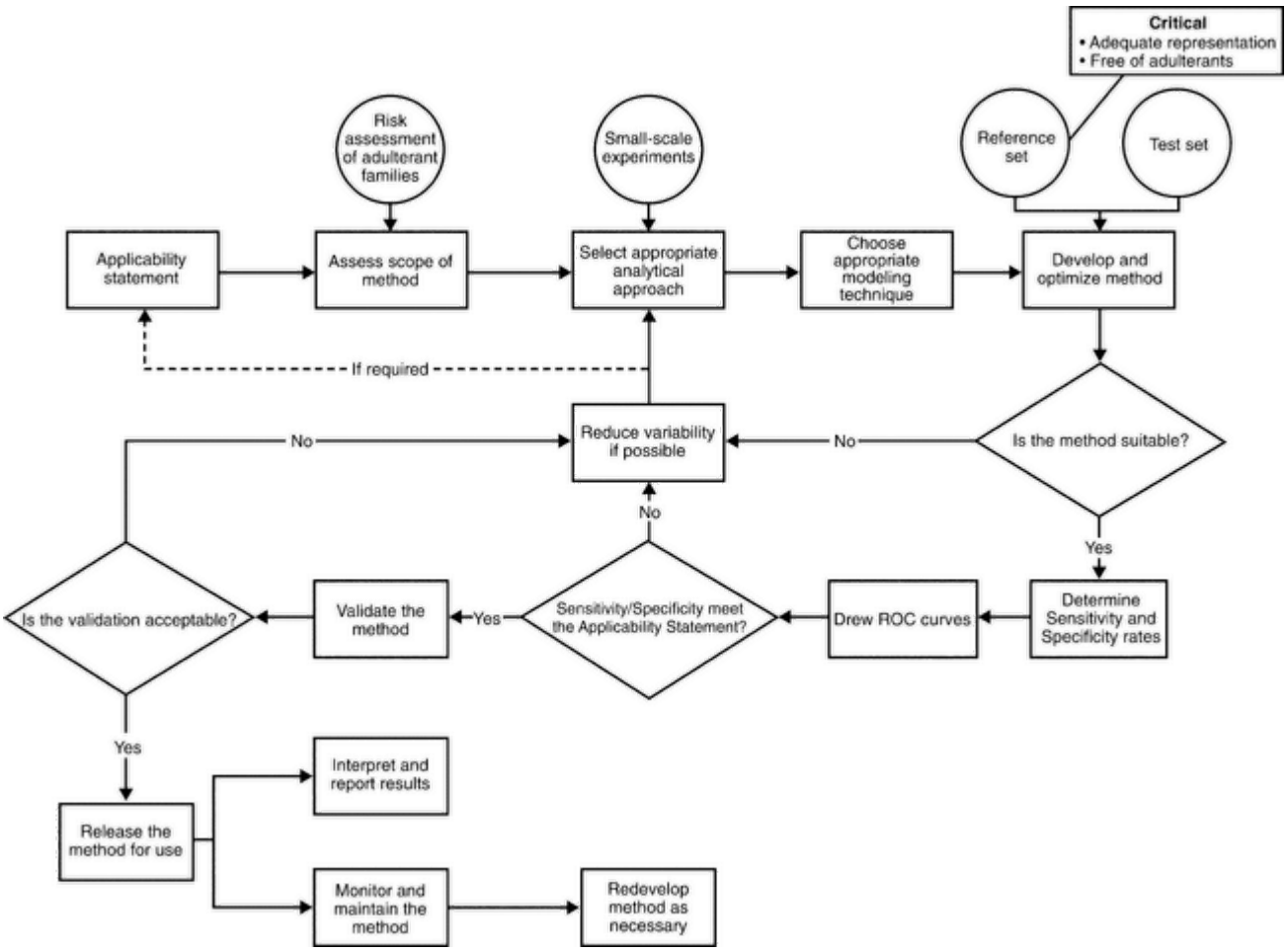

Figure 3: Flowchart of critical steps in non-targeted method development and validation.

Add the following:

▲Annex 2

Example of an Assessment of a Non-Targeted Method

Applying the principles of the guidance document, a method was created for the non-targeted detection of adulterants in liquid unprocessed milk, using high-capacity mid-IR instruments situated in a large milk payment laboratory that tests incoming raw milk to ensure the quality and safety of the materials. This method was then assessed using samples of known levels of contamination with various adulterants.

SUMMARY

A non-targeted method was created using a broadly similar approach to that outlined in the guidance document, and subsequently validated at a central milk testing facility. The laboratory tests thousands of samples daily and multiple analyses are performed simultaneously on these samples by FTIR instruments.

An assessment trial was developed to test the non-targeted method with 12 common dairy adulterants at various concentrations, spiked into milk collected from routine incoming milk samples. The method development and validation represented in this example were performed prior to the publishing of the USP guidance document and used proprietary software, and therefore did not address every aspect of the guidance. Many of the guidance document's principles were broadly used by the method developers, however, and the example serves as an appropriate model for users who may wish to better understand the direct application of the guidance document in this context.

The non-targeted method described in this example is now routinely used in conjunction with traditional targeted methods to screen incoming milk.

Applicability Statement

It was established that the applicability statement would be a rapid non-targeted method for detecting adulteration of unprocessed liquid bovine milk with nitrogen-rich compounds added with a Sensitivity rate of 99%, a Specificity rate of 95% and a confidence interval of 95% (the detection threshold was derived from previous work on similar methods and is adulterant-specific, see *Table 2*).

Method Development

Based on a risk assessment of economic benefits, adulterant availability and health risks, the following adulterants were included in the method development: dicyandiamide, hydroxyproline, melamine, urea, amidourea, ammonium sulfate, cyromazine, biuret, allantoin, thiourea, 4-aminotriazine, and cyanuric acid.

Sets of unprocessed liquid milk samples spiked with five levels of adulterants and a blank sample were tested. Each set was comprised of a blank, two adulterated samples that were below the trigger limit (the boundary), one at the limit, and two samples above the limit. Based on available viable methods and infrastructure, mid-IR milk testing analyzers were selected as an appropriate analytical approach.

A reference set was built using unprocessed liquid milk that had a high degree of variability, including differences in geographical origin, age, composition and breed of dairy cattle. The test set was built using the same milk pool and the samples were deliberately spiked with known concentrations of the chosen adulterants. The notional boundary described in the guidance surrounding the multivariate cloud of Typical samples was created using a combination of spectral variables in the reference set. Using certain algorithm, these variables can produce a numerical score (it is termed the Abnormal Spectra Module or ASM) in the commercially available system tested), and can be loosely described as distance from the mean Typical sample across various attributes.

Any ASM result that was equal to or higher than 3 (the boundary) was reported as a FAIL (i.e., an Atypical sample), while anything lower than 3 was reported as a PASS (a Typical sample).

The linearity of the correlation between adulterant concentration and ASM score was not assessed, but given the binary nature of the output, it was assumed to be linear.

According to the guidance document, the Specificity and Sensitivity for each adulterant should be tested. This example did not fully evaluate these two criteria, and it is recognized that Specificity and Sensitivity will need to be more comprehensively assessed in the development stage of the process.

Validation of the Optimized Method

The validation criteria were assessed against Typical and Atypical sample sets.

The validation set was prepared by adding, individually, 12 common dairy adulterants (identified in the *Applicability Statement*) to routine incoming milk samples at an addition rate that approximated ASM scores of 0, 1.5, 2.25, 3, 4.5, and 6 for each adulterant. This allowed for a blank sample, two adulterated samples that were below the boundary, one at the boundary, and two samples above the boundary for each adulterant.

The samples in the validation set were labelled SB, S1, S2, S3, S4, and S5 in ascending order of adulterant concentration.

The actual concentrations equating to the threshold (ASM score of 3) were formulated as sample S3 for each adulterant, and are listed in below:

Table 2

Adulterant	Concentration (µg/g)
Dicyandiamide	325

Adulterant	Concentration (µg/g)
Hydroxyproline	950
Melamine	325
Urea	550
Amidourea	600
Ammonium sulfate	275
Cyromazine	275
Biuret	600
Allantoin	300
Thiourea	350
4-Aminotriazine	2000
Cyanuric acid	150

The spiked samples each contained only a single adulterant, meaning a total of 72 bulk adulterated samples were created (12 adulterants, six sample concentrations). All samples were measured using four separate FTIR instruments, three separate runs in duplicate or triplicate over two days to eliminate possible interference of instrument, human, or environmental factors. Additional confirmatory reference tests were used as performance checks. The results of analysis provided a total of approximately 2200 individual data points.

Analysis

In this example, the overall trend of the results (see *Figure 4*) was as expected, with no failures (Atypical samples) on the blank samples (SB), and increasing failure rates as the adulterant concentrations increased (with corresponding decreasing pass rates (Typical) as adulteration concentrations increased).

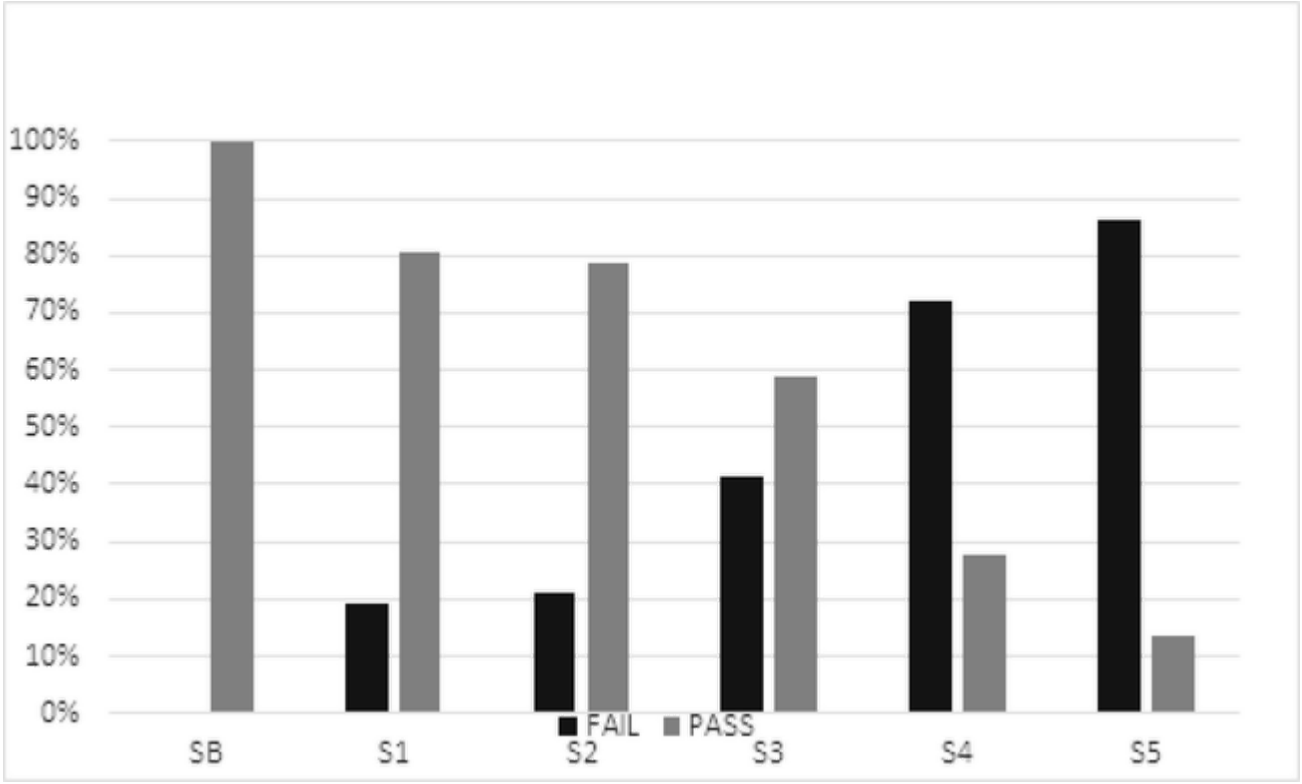

Figure 4: Overall pass and failure rate of all samples.

Once the results were split into individual adulterants, however, it became clear that many of the adulterants were either not detected satisfactorily or not at all. Sensitivity and Specificity were calculated as per the guidance, with Sensitivity being defined as the proportion of correctly identified Atypical samples to total Atypical samples, and Specificity being defined as the proportion of correctly identified Typical samples to total Typical samples.

Sample S3 for all adulterants was excluded from the Sensitivity and Specificity calculations because it was specifically formulated to lie so close to the threshold that it would be subject to spectral noise due to minor compositional discrepancies. The Sensitivity and Specificity for samples with individual adulterants are highlighted in *Table 3*.

Only three of the 12 adulterants tested were detected at the required Sensitivity and Specificity: dicyandiamide, hydroxyproline, and melamine.

A small proportion of the low range Typical samples S1 and S2 were failed (Incorrect Atypical results). Although the Specificity rate did not meet the applicability statement, the laboratory decided the performance of urea as a primary detection method was acceptable in retrospect, as it would be accompanied by a simultaneous spectral targeted analysis.

Table 3

Adulterant	Sensitivity	Specificity
Dicyandiamide	100%	100%
Hydroxyproline	100%	100%
Melamine	100%	100%
Urea	100%	91%

Adulterant	Sensitivity	Specificity
Amidourea	100%	23%
Ammonium sulfate	100%	23%
Cryomazine	100%	0%
Biuret	83%	99%
Allantoin	83%	97%
Thiourea	82%	100%
4-Aminotriazole	15%	100%
Cyanuric acid	9%	100%

Conclusion

This non-targeted method was originally designed to be used in conjunction with various targeted methods of detecting adulteration, thus success was defined by the project team as the detection of three or more adulterants for which targeted methods exist. In this example, this was interpreted as detection of adulterant concentrations at or above the notional boundary (i.e., samples S3, S4, and S5) for the three adulterants hydroxyproline, melamine, and dicyandiamide. In order to ensure adequate confidence in the results, the Sensitivity was chosen as 99% (confidence interval of 95%). To minimize false alerts, the Specificity was set at 95% (confidence interval of 95%).

Based on the successful detection of the three adulterants (dicyandiamide, hydroxyproline, and melamine) in raw milk, the validation was deemed pragmatically successful. The model was accepted into use, with the limitations for urea noted and accounted for.

The assessment of non-targeted methods is critical to the acceptance of a method before implementation, using careful experimental design and established good laboratory practices. Methods should be assessed against a broad variety of possible adulterants, at different concentration ranges and ensuring the full variety of measured product and measuring devices is covered.

Limitations of the method that may be uncovered in the assessment need not preclude its use, should it not meet all desired outcomes, as long as those limitations are known and accounted for by alternative means (e.g., targeted methods for known or possible adulterants).▲3S (FCC 11)

[1] For a description of one-class classification in terms of modellng, refer to *USP* general chapter *Chemometrics* ‹1039›, *USP 40–NF 35*, accessible at www.uspnf.com.

[2] 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.

[3] This guidance 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 guidance 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] .

[4] Adapted from the definition of "Food contaminant" from the *Codex Alimentarius* glossary at http://www.fao.org/docrep/006/y8705e/y8705e07.htm.

[5] http://www.fda.gov/downloads/Food/FoodDefense/FoodDefensePrograms/UCM478509.pdf.

[6] An incorrect Atypical result could lead to increased costs through unnecessary inventory management activities and reference testing.

[7] An incorrect Typical result would lead to a food safety concern should the true adulterant levels in the sample exceed food safety limits.

[8] See *Chemometrics* ‹1039›, *USP 40–NF 35*, accessible at www.uspnf.com.

[9] Richard G. Brerton, Chemometrics for Pattern Recognition, Chapter 5 Two Class Classifiers, pp 177-231, and Chapter 7 Multiclass Classifiers, pp. 289-309, 2009, John Wiley & Sons, Ltd., ISBN 978-0-470-98725-4 .

[10] Richard G. Brerton, Chemometrics for Pattern Recognition, Chapter 6 One Class Classifiers, pp 233-287, 2009, John Wiley & Sons, Ltd., ISBN 978-0-470-98725-4.

[11] For non-targeted methods this is the number of correctly identified Atypical samples divided by the total number of (verified) Atypical samples.

[12] For non-targeted methods this is the number of correctly identified Typical samples divided by the total number of (verified) Typical.

[13] Harnly, J., Chen, P., et al. (2015). "Comparison of Flow Injection MS, NMR, and DNA Sequencing: Methods for Identification and Authentication of Black Cohosh (Actaea racemosa)". *Planta Med*. 2016; 82: 250-262 .

▲[14] Rodionova, OY, Titova, AV, and Pomerantsev, AL. (2016). Discriminant analysis is an inappropriate method of authentication. *TrAC Trends in Analytical Chemistry*, 78: 17–22.▲3S (FCC 11)

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

[16] 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.

[17] A significance of p = 0.01 equates to a Confidence Interval of 99%.

[18] 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.

[19]

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

[20] Guidance on multivariate analysis and chemometrics can be found in many published articles including *Chemometrics* ‹1039›, *USP 40–NF 35*, accessible at www.uspnf.com.

[21] 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.

[22] Harnly J, Chen P, de B. Harrington P. 2013. Probability of Identification: Adulteration of American Ginseng with Asian Ginseng. *Journal of AOAC International*. 96(6):1258–1265.

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

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

[25] 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.

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

[27] 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 1000 µg/mL.

[28] Some Theory of Sampling, W.E. Deming, Dover Publications, ISBN-13: 978-0486646848, ISBN-10: 048664684X. See also Sherman, P.J. "Smart charting: Guidelines to manage processes effectively." *Quality Progress*, July 2012.

[29] Engineering Statistics (5th edition), D.C. Montgomery, G.C. Runger, N.F. Hubele, ISBN–13: 978–0470631478, ISBN–10: 0470631473, Wiley (2010).

[30] See *USP* general chapter *Analytical Data—Interpretation and Treatment* ‹1010›, *USP 40–NF 35*, accessible at www.uspnf.com for information on acceptable practices of data analysis of chemical and other analyses.

Please check for your question in the FAQs before contacting USP.

Topic/Question	Contact	Expert Committee
APPENDIX XVIII -- GUIDANCE ON DEVELOPING AND VALIDATING NON-TARGETED METHODS FOR ADULTERATION DETECTION	Kenny Xie Scientific Liaison +1 (240) 221-2052	FI2015 Food Ingredients 2015

Page Information

Not Applicable