Towards a mathematical framework for method validation and performance characteristics of non-targeted methods for food authenticity

Kapil Nichani [1], Steffen Uhlig [2], Manfred Stoyke [3], Sabine Kemmlein [3],

Franz Ulberth [4], Ilka Haase [5], Maik Döring [5], Stephan G Walch [6] and Petra Gowik [3]

The affiliations
[1]   QuoData GmbH, Prellerstr. 14, 01309 Dresden, Germany
[2]   QuoData GmbH, Fabeckstr. 43, 14195 Berlin, Germany
[3]   Bundesamt für Verbraucherschutz und Lebensmittelsicherheit (BVL),
      Diedersdorfer Weg 1, 12277, Berlin, Germany
[4]   European Commission, Joint Research Centre,
      Retieseweg 111, 2440 Geel, Belgium
[5]   Max Rubner-Institut (MRI) - Bundesforschungsinstitut für Ernährung und Lebensmittel,
      Nationales Referenzzentrum für authentische Lebensmittel,
      E-C-Baumannstr. 20, 95236, Kulmbach, Germany
[6]   Chemisches und Veterinäruntersuchungsamt (CVUA) Karlsruhe,
      Weißenburger Str. 3, 76187 Karlsruhe, Germany

kapil.nichani@quodata.de; steffen.uhlig@quodata.de

Abstract

Through its suggestive name, non-targeted methods (NTMs) do not aim at a predefined "needle in the haystack". Instead, they exploit all the constituents of the haystack. This new form of analytical methods is increasingly finding applications in food and feed testing. However, the concepts, terms, and considerations related to this burgeoning field of analytical testing needs to be propagated for the benefit of ones associated in academic research, commercial development, and official control. This paper addresses the frequently asked questions around notations and terminologies surrounding NTMs. The widespread development and adoption of these methods also necessitates the need to develop approaches to NTM validation, i.e., evaluating the performance characteristics of a method to determine if it is fit-for-purpose. This work aims to provide a roadmap to approaching NTM validation. In doing so, the paper deliberates on the different considerations that influence the approach to validation and provides suggestions thereof.

Keywords

Non-targeted methods; method validation; food fraud; food authenticity; mass spectrometry; NMR; spectroscopy; NGS

1　Introduction

A newer form of analytical methods known as non-targeted methods (NTMs) have emerged as powerful techniques in tackling problems of food authenticity (including food fraud) [1], food quality [2], food safety [3], water monitoring [4], microbial species subtyping [5], among others. There is a growing body of literature describing these new methods and some are already being used in routine testing and monitoring [6–8]. When referring to NTM, reported literature uses a mix of terms such as "untargeted method", "non-target testing", "non-target approaches", "fingerprinting methods", among others [9]. In order to describe and make the meaning of NTMs as clear as possible, in this paper we limit the vocabulary and choose to only use "non-targeted methods".

NTMs bring together the superiority of high-resolution analytical measurement instruments and advancements in chemometrics and machine learning algorithms. The acquired measurements from the instrument are typically large array(s) of values, which are sometimes referred to as the "fingerprints" of the measured samples. Owing to the immense potential of these methods, the past decade has seen their rapid development and adoption by researchers and laboratories, especially for testing food authenticity. As a result, there is a pressing need to devise ways to ensure the reliability of the obtained results. The key to doing so is method validation (MV) studies, which help provide objective evidence that the NTM is fit for its intended use. Harmonized protocols for MV help to compare results between different laboratories, which benefit producers, consumers, official control, and regulators alike [10,11]. European Union (EU) Official Controls Regulation requires official food control laboratories to apply, when available, standardized methods i.e., multi-lab validated methods (MLV) [12]. Only in case, a standardized method is not available, single lab validated (SLV) methods should be applied. In such cases, the comparability of results between official food control laboratories and commercial laboratories doing the counter analysis on behalf of the business operators becomes the challenge. Several researchers and experts have already drawn attention to the paucity of internationally accepted validation protocols for NTMs [13–15]. Thus, much has to be addressed regarding how to perform NTM validation [13,16,17].

To this end, the purpose of this paper is two-fold. First, to review and describe notations and terminologies surrounding NTMs. The burgeoning field of NTMs has been accompanied by an expanding set of terminologies. These terms are related to the new analytical technologies, machinations of artificial intelligence methods, information systems infrastructures, and statistical decision theory. Hence there is considerable room for misinterpretation of what different experts might be conveying. Second, it aims to investigate the factors that make devising an NTM validation protocol challenging. In doing so, it highlights the points of contention surrounding validation choices. This work proposes considerations for how NTM validation can be implemented in practice. Altogether, this work aims to lend to the limited body of work that is currently available. This will eventually help researchers of the scientific community, officials at control agencies and experts of standardization bodies to draft relevant guidelines, protocols, or standards.

The paper is organized in the following way. The first section describes the notation and terminology around NTMs. The second section discusses the meaning of validation and aggregates definitions for it. The following section then discusses the concepts that are currently available for NTM validation. Given the current understanding of NTMs and their validation, the first three sections segue into specific proposals for the NTM validation procedure, and the considerations to be made, delineated, in sections 5 and 6, respectively.

Next, we discuss the validation sample requirements in section 7. And finally, layout the stages for NTM development and validation for NTMs and how we see if to be differing from the traditional way.

2   Describing NTMs and the terminology around it

Devising MV concepts for NTMs such that it can bring under its ambit all the different methods that are available currently plus the new ones which will be introduced in the future – is a challenging task, to say the least. However, to tackle this, it is important to address what constitutes an NTM. In the following are listed eight questions to deduce what an NTM is. Instead of providing an unambiguous definition for NTM, the discussions around the questions describe the multitude of concepts embodied around NTMs.

2.1   What are the components of an NTM?

Figure 1 shows the general components of an NTM. All the steps involved in the NTM until the analytical measurements, performed on a lab bench are collectively referred hereafter as "wet lab" procedures. Some manner of chemometric / statistical / machine learning / artificial intelligence model is then responsible to parse this multi-dimensional dataset. This step is referred to hereafter as "dry lab" procedures. The copious amounts of measurement data are saved, processed, and retrieved with the help of electronic databases or reference databases. An NTM uses several features obtained by measurements (in the wet lab) in combination with data analytics (in the dry lab using reference database) to authenticate a food product (product characteristic). Food authenticity testing can be in connection to the species, origin, production/processing system, purity, etc. Ultimately, to verify the authenticity, an imminent decision is then usually based on a set of criteria.

It is necessary to state the above notations as it will help determine what can (and cannot) be considered as an NTM. It should be emphasized that the analytical method employed to collect measurements for building the reference database, must be validated regarding its analytical performance in a way (e.g., single lab validation) that it is poised for eventual standardization (further elaborated in Section 8). This is important to ensure that the data basis in the reference database is not rendered unusable, leaving any ensuing NTM development to no avail.
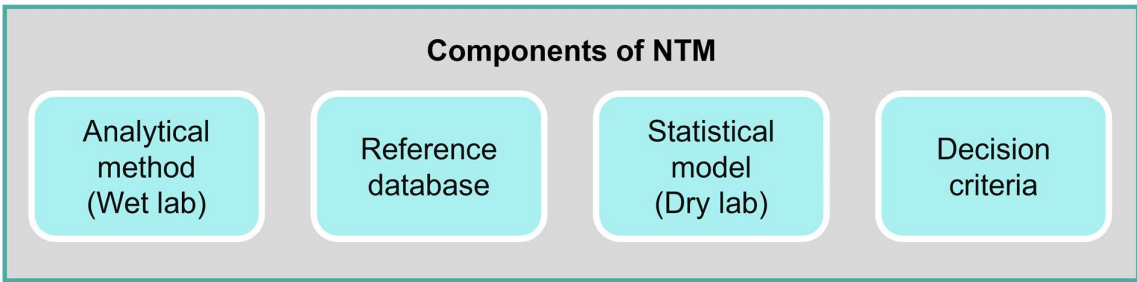
Figure 1: Components of a non-targeted method.

2.2   Are NTMs associated with a particular measurement platform or instrument or analytical technology?

In the simplest sense, a measurement platform refers to the analytical measurement procedure. The term platform encompasses the ensemble of instruments or equipment involved in the wet lab; that is not only highly complex and expensive but also of considerable size. Different analytical technologies that are typically part of NTMs are as follows. Methods involving chromatographic separation (like gas or liquid chromatography), mostly, but not exclusively, followed by (high resolution) mass spectrometry [18]. Alternatively, these can be other spectra generating methods such as nuclear magnetic resonance (NMR) [19], Fourier transform infrared (FTIR) [20], near-infrared (NIR) [21], or Raman spectroscopy [22]. In practice, NTMs are not specific to any instrument, platform, or analytical measurement principle. In fact, they can also be a combination of one or more analytical measurement platforms.

We believe that NTMs may not only involve spectra, chromatograms or identified and quantified chemical entities (elements, fatty acids, etc.) as measured signals from the instruments but can also be nucleotide sequences. This can be the case in metabarcoding methods involving next-generation sequencing (NGS) technologies, to identify several species or taxa, parallelly [23]. Whether these methods are also an NTM can be argued with the reasoning that a genomic region (barcode) is targeted by a defined primer pair. However, this primer pair needs to fit universally to the selected genetic region in all organisms of interest (e.g., all land vertebrates), allowing a non-targeted species identification by multiple parallel sequencing and subsequent assignment of the sequences by database comparison.

2.3   Are databases required for NTMs?

Yes. In the generic sense, large amounts of empirical data are acquired to define the sample populations (classes), e.g., wine from France and wine from Germany. However, databases may have a setup for different purposes (e.g., milk: databases for geographic origin, for production process [organic versus conventional]). The generation of data of well-defined samples is mandatory, together with meta-data capturing the related traceability information are the most common types of databases used by NTMs. Usually, the greater the amount of training data available, the more accurate is the model. The reference database can be used to 'train' supervised machine learning algorithms. Alternatively, the measured data can be compared against the reference sample data, using non-supervised approaches, e.g., similarity metrics or correlation coefficients.

2.4   Do NTMs always help answer a yes or no question?

Broadly, food authenticity testing aims to investigate whether claims made for a particular food item are correct (e.g., regarding the geographic origin, plant/animal species resp. varieties, ingredients, process). In other words, can we classify the sample to the classes as defined in a reference database. The dry lab procedure furnishes quantitative values such as probabilities, referred to herein as decision scores. The sample is then assigned to belong to a particular class (say class α) if their score is below a given univariate decision threshold (or decision limit). Conversely, a sample is assigned to the other class (say class β) if their score is larger than the decision threshold. Experience shows that a quantitative decision score exists for a large number of NTMs owing to the underlying chemometric or machine

learning model [24]. The outcome of the decision will undoubtedly either be yes or no.

2.5 What is the difference between 'fingerprinting' and 'profiling', and are both NTMs?

Considering the definitions reviewed and provided by Balin et al. profiling methods fall into the type of targeted methods [9]. An alternative perspective is that both, fingerprinting and profiling, are NTMs as both require a statistical model (dry lab) and a reference database for decision making (Figure 1). The main difference between fingerprinting and profiling relates to the output generated by the analytical method (wet lab), in other words, whether it targets specific entities. Fingerprints of a material are electronic records (e.g. whole or part of chromatograms or spectra) produced by an instrument without further information regarding the identities or quantities of entities represented by the record, whereas quantity values of defined entities constitute the profile of a material (e.g. elements, fatty acids, sugars, etc.) [25–27]. However, the profile itself does not allow to decide on the authenticity of the material but is used as the input of a multivariate decision model. Quite often fingerprinting methods are converted into profiling methods by attempting to identify the most relevant variables for discrimination. Usually, this variable reduction reduces noise and guards against model overfitting, but the biggest advantage is that the resulting targeted, profiling methods are independent of the measurement platform through calibration with reference materials.

2.6 Is 'suspect screening' also a type of NTM?

Suspect screening workflows are widely used in food, environmental and forensic chemistry [8–10]. In these types of pipelines, a large list of suspect compounds (n >> 1) is checked for presence or absence [12]. In food analysis, such methods are mainly used for food safety questions (e.g., pesticide screening) and are less known for authenticity questions. Therefore, in this paper such workflows may not be considered as NTMs, *stricto sensu*, but do make use of the information content from the multi-dimensional measurement data.

2.7 What is the role of calibration in an NTM?

The IUPAC definition of calibration 'the set of operations which establish, under specified conditions, the relationship between values indicated by the analytical instrument and the corresponding known values of an analyte is in principle applicable to NTMs if the state of material (authentic or non-authentic) is regarded as the quantity value to be determined [28]. However, in contrast to methods where the quantity of a targeted analyte (measurand) is estimated via a univariate calibration function (e.g., linear, or quadratic regression), NTMs use multivariate models for deciding whether the sample is authentic. Such models need to be calibrated as well but to avoid confusion, the term 'training' is frequently used instead, and the samples used for setting up the model are called 'training samples' or 'training set'.

2.8 What is the role of quality control materials in an NTM?

As with any other analytical method, quality control is a prerequisite for producing valid results, requiring the availability of quality control materials (QCMs). The main function of QCMs is to monitor the accuracy (precision, trueness) and stability of the analytical method during its application. (Certified) reference materials ((C)RMs) are key for establishing

metrological traceability and can also be used for creating quality control charts to keep a targeted method of analysis under statistical control over time. (C)RMs for NTMs are rare but a few exist, e.g., NIST SRM 1950 Metabolites in Human Plasma, which has values assigned for approximately 100 analytes. An anhydrous butter fat (BCR-519) and cocoa butter (IRMM-801) has been made available for checking the authenticity (purity) of milk fat and cocoa butter by triglyceride analysis in combination with multivariate statistical analysis. Other types of commercially available quality control materials include e.g. meat from different species (LGC Ltd.), or plant specimens obtained from botanical gardens.

For novel NTMs such (C)RMs do not exist in most cases, particularly not for fingerprinting methods [29]. Pooling aliquots of reference samples used in training the decision model and including them in the analytical sequence is a frequently used quality control technique. The obtained data are subjected to multivariate analysis (e.g., principal component analysis and inspection of the score plot) but also to generate Shewhart charts of principal components, extracted features, etc. Such tools are appropriate for setting up quality control in a single laboratory but may be insufficient if the NTM is to be applied in multiple laboratories. This situation frequently requires QCMs for the normalization of data produced in different laboratories and/or by different instrument brands. A QCM obtained by pooling of samples and making it available as a 'normalization sample' to interested laboratories is a potential solution but its long-term stability as well as the effect of renewing the normalization sample on data quality once the original batch is used up, have to be considered.

3   Validation definitions

A fundamental matter of contention is the perception of the term "validation". Depending on the problem, context, and application, it may have a different meaning to different stakeholders. MacNeil et al. aptly pointed out that validation, just like beauty, is in the eyes of the beholder [30]. In the context of NTMs, validation often refers to "model validation" (dry lab) [13,15], which can be done using techniques such as "nested cross-validation", or "k-fold validation" (details of these techniques are described elsewhere [31,32]). Thus, it is necessary here to revisit MV definitions and review what can be applicable to NTM validation.

The ISO/IEC 17025:2017 defines validation as the "provision of objective evidence that a given item fulfills specified requirements, where the specified requirements are adequate for an intended use" [33]. The Eurachem Guide on the fitness-for-purpose of analytical methods defines it as "the process of establishing the performance characteristics and limitations of a method, and the identification of the influence which may change these characteristics and to what extent" [10]. In ISO 16140-1:2016 MV is defined as, "Determine the performance characteristics of a process and provide objective evidence that the performance requirements for a specified intended application are met" [34]. Classical MV concepts involve the evaluation of method performance characteristics [11,35,36]. The process of MV usually consists of experiments that can be performed in a single lab or multiple labs to determine these performance characteristics. Although there can be different implications regarding the terminology surrounding single lab vs in-house validation or multi-lab vs collaborative validation, in this text for simplicity we use the former in both cases.

Methods can be validated for more than one analyte, for different matrices, or different instruments or platforms. Following the above-mentioned protocols, such validation experiments include e.g. (C)RMs or matrix spiked samples to determine recovery rates, matrix blank samples to determine background levels, e.g. blanks to determine the limit of detection,

and replicate analysis of a sample to determine precision. If available, (C)RMs can be used to determine the precision and trueness of a method. The performance characteristics for validation of a method strongly depend on the intended use, the type of the method (quantitative or qualitative), or in case of method extension (new analyte, new matrix, new platform, etc.), the degree to which it has been previously validated. We believe that there is a unanimous agreement, that performance characteristics of a newly developed quantitative method include – trueness, precision, selectivity, limit of detection, limit of quantitation, linearity (or other calibration models), working range, measurement uncertainty, ruggedness, confirmation of identity and recovery rates. That sensitivity, selectivity, false-positive rate (FPR), and false-negative rate (FNR) are necessary performance characteristics of the validation of a new qualitative method also seems undisputed. Some of these classical experiments can be transferred to NTMs, while other ones are simply not available or applicable (such as the use of (C)RMs, etc.).

Many NTMs are motivated and related to a binary decision problem (discussed in detail in the following sections). A validation procedure for such NTMs, in general, involves determining the risk of a false positive or false negative decision. In this paper, "validation procedure" and "validation approach" might be used interchangeably. The performance characteristics (or figures of merit) of non-targeted methods and the way they are estimated undoubtedly differ from the ones related to targeted methods; however, the ultimate objective remains the same, i.e., demonstrate the 'fitness-for-purpose' of the method, independent of the physico-chemical principles of the analytical method, the data evaluation, etc. For the method developer, this is important to objectively demonstrate the fitness for its intended use; for the method user it is important for quality assurance and accreditation.

4   Existing concepts for NTM validation

The guideline for the development and validation of non-targeted methods from US pharmacopeia (USP) has been a go-to resource in the absence of other harmonized guidelines or standards [37]. This is part of the USP Food Chemicals Codex. The guideline describes a procedure for methods entailing binary classification into adulterated (atypical) and unadulterated (typical). The USP guideline defines NTM as "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" [37]. It is noteworthy that the prescribed procedure in the USP guideline is independent of the analytical method principle or food type. This is beneficial to ensure horizontal applicability to a wide range of methods. However, the scope of the guideline remains limited to a subset of methods for food authenticity testing, i.e., to one-class classification methods for testing adulteration or mixing.

The performance characteristics recommended include evaluation of sensitivity as the correct identification of unacceptable samples as "atypical," and specificity as the correct identification of acceptable samples as "typical". In other words, the sensitivity of the method is the rate of detection of adulterated or fraudulent samples, and specificity is the rate of detecting safe or compliant food items.

The attractiveness of the USP guideline also alludes to the considerations for the

method development steps along with single lab validation. A generic thought process is described so that sufficient method suitability is established before going to the validation stage. The performance characteristics are checked against the criteria set upfront in the applicability statement. Furthermore, the guide defines the different sample sets necessary at different stages. In the method development stage, the mathematical model is developed using a reference dataset containing adequately represented samples which are unadulterated (authentic) samples. This model is optimized using a test set. In the validation phase, an independent sample set comprising typical and atypical samples is to be tested as unknown samples. However, the absence of a guideline or criteria for how many samples to consider in the validation set remains to be another important gap to be filled.

By virtue of relying on a specific version of the instrumentation hardware, reference databases, and complex mathematical models – NTMs might be required to be updated and re-validated. Intermittently, the necessary samples (of number and type) in the reference database might be influenced due to environmental or anthropogenic factors, leading to drifts in the mathematical model parameters. The USP guide also mentions such scenarios and recommends monitoring of the method. This needs to be addressed not only by the method developer (e.g., database maintenance), but also by the user (laboratory) in the framework of quality assurance. However, whether institutions would like to carry the resource burden of re-validating the method at intervals remains a contentious issue (see also section 8).

Apart from the USP guide, standard method performance requirements (SMPRs®) for non-targeted testing (NTT) are also available from AOAC [38]. At the time of writing, SMPR are available for methods testing for economically motivated adulteration in three food items, viz. extra virgin olive oil (EVOO) [39], honey [40], and bovine milk [41], and draft SMPRs are available for vanilla [42] and saffron [43]. These SMPRs define the NTT method as "any method generating a baseline fingerprint of the authentic material and comparing test sample fingerprints to assess differences will be considered. The final binary result identifies test samples as either authentic or potentially adulterated" [44]. Unlike the USP guide, the SMPRs do not describe generic steps to adhere to in the method development stage. But they do provide a number for the samples required to be tested in the validation stage. Furthermore, it is to be noted that the AOAC SMPRs describe procedures for single lab validation.

Together, these NTM validation resources possess several attractive features which can be used for a future harmonized NTM validation concept, such as (i) description of applicability statement such as "a non-targeted method for detecting the adulteration of honey with sugar syrups at a level > 10 % with a sensitivity rate of 90 % and a specificity rate of 95 %, both with a significance level of p = 0.05" (USP), (ii) proposed performance characteristics for binary non-targeted methods, (which are more straightforward) (USP) (iii) requirements for the number of samples that are needed to reach a certain level of confidence (AOAC), and (iv) importance of method monitoring and need for re-validation in case of drift (USP).

A harmonized protocol for the validation of NTM should ideally build on existing proposals and proven design principles of internationally accepted protocols. These can serve as a springboard to establish an NTM validation framework. Existing harmonization efforts are ongoing in Europe (CEN) and North America (AOAC). At the same time, they should be scrutinized for scientific validity and extended for practicability and applicability. Additional work is also required to merge philosophies and specify terminologies of the different working groups, communities, and existing documents. Furthermore, effort is required to describe the principles in more detail (nature of samples, number of replicates, number of laboratories, etc.) and data requirements (number of samples) for single as well as multi-lab validation. Building

on the existing concepts for NTM validation, the following section proposes four NTM validation procedures.

5   Extending the existing concepts to NTM validation

5.1   Considerations for one-class, two-class, and multi-class NTM

NTMs can be defined and structured in different ways. Figure 2 serves to illustrate this perspective. Consider methods related to testing the authenticity of olive oil as running examples throughout this section. Here, the starting point is how is the scope of the binary NTM specified. Suppose it is a two-class problem, e.g., if the method distinguishes between genuine olive oil and olive oil adulterated with seed oil (e.g., sunflower oil) at a level that is widely regarding as being economically justified. In this case, the reference database must contain entries for both classes, i.e., namely, reference database entries for olive oil samples with adulteration greater than 15% seed oil and samples with less than or equal to 15% seed oil. In contrast, only entries for one class in the reference database are required for a one-class problem, suppose, only for the authentic olive oils.

The advantage of the one-class problem is that only samples corresponding to one class (say, the authentic olive oils class) are required, which are usually accessible and available for the method developer. However, one-class problems do not consider sensitivity or FNR with respect to a specific class, and only consider specificity or FPR. One of the consequences of this is that sensitivity of an NTM for the one-class problem is typically lower than that of a two-class problem.

One option to overcome the limitations of one-class problems from the perspective of validation would be to nevertheless define a counter class. The counter class comprises samples representing a reasonable approximation of samples being "non-authentic" with regard to the initial classification question. Consider the example in Figure 2 – is the olive oil originating from Italy or not. The one class is all possible olive oils coming from Italy. Here the counter class (not originating from Italy) can be defined by measuring a bunch of other olive oil samples typically found in the market, which are from nearby countries. Defining a counter class in this way with samples likely to be candidates for fraud, can allow us to use the more efficient two-class problem validation.
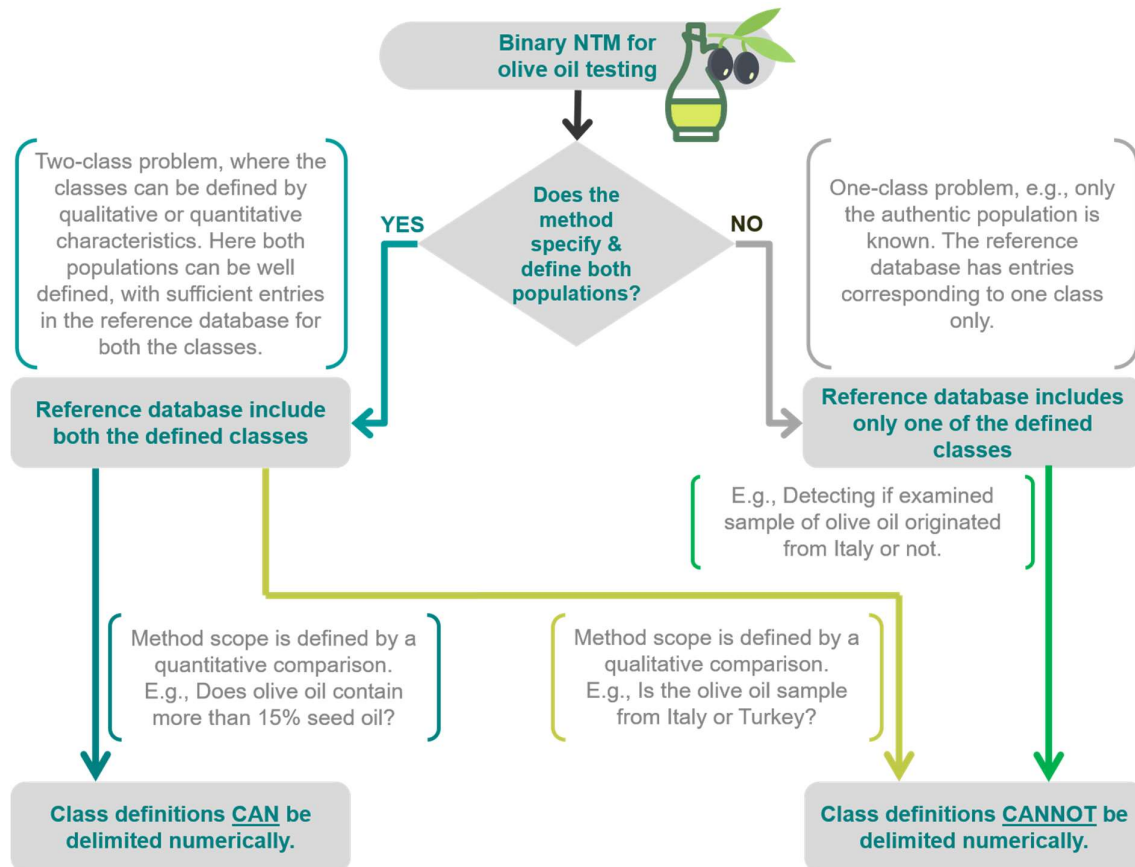
Figure 2: Different ways to formulate an NTM, focusing on examples around olive oil testing. The text in the white box adjoining each connecting arrow provides context and examples.

5.2   Ways to define classes

The underlying classes can be numerically delimited, i.e., an olive oil sample belongs to the class exactly when its seed oil content is below 15%. More generally, the numerical demarcation of class definition can be based on adulteration level, concentration, contamination, etc. Such numerically delimited class definitions also have been reported elsewhere in the literature [45–48]. Alternatively, suppose the method tests whether the examined olive oil sample originates from Italy or Turkey – the classes cannot be delimited numerically. We assume that most NTMs which do not involve measures of purity (mixture levels or adulteration levels) can be defined by qualitatively delimited classes. Methods with one-class problem descriptions characteristically will also only be defined by qualitatively delimited classes. Coming back to the example of an NTM to test the claim on a bottle of olive oil if it is originating from Italy or from elsewhere (i.e., not from Italy). The manifestation of the class is confined by the qualitative conditions imposed, i.e., olive oil and whether it originated in Italy or not.

In addition to binary classification NTMs, multi-class classification problems (more than 2 classes) can also be expected. Interestingly they can be broken down into several binary classification problems. Hence, the discussion around validation approaches thus far can be extended to the multi-class case. Consider an illustrative example – a method to check the claim if the olive oil is originating from Italy, Spain, or Greece. Here the reference database must include data for classes that can be described as Italy, Spain, and Greece. As a

motivating example only, such a multi-class classification can be broken down into sequential binary classification problems as follows. First, classify if the oil sample is from the EU region or not. Then classify if oil is from Italy or otherwise. Following this, determine if the oil is from Spain or not. After which classify if oil is from Greece or otherwise. Due to the practicalities of breaking down a multi-class problem into sequential binary classifications, it is foreseen that it can be used in the validation of NTMs in authenticity testing.

5.3   NTM validation approach

The above examples for the diverse ways in which a method for olive oil testing is structured are by no means comprehensive. But it is evident that they will have to be validated using differing approaches. We now turn to an illustration of what the different validation approaches can be.

In almost all cases, the decision step in a classification procedure of an NTM is based on a quantitative decision score value calculated in the dry lab step of the NTM. This score value is compared to a specified decision limit. Examples of quantitative decision scores include correlation coefficients, similarity metrics, probabilities of a class assignment, principal components scores, etc. These could also be proprietary scores provided by commercial software. By using this quantitative score value in the validation, a better validation result can be achieved with a significantly lower validation effort (number of samples and number of laboratories) in comparison to using only the yes or no decision. Depending on whether the class definitions are numerically delimited in Figure 3, and whether the quantitative decision scores are used – 4 different validation approaches can be used (named generically as A, B, C, D).

Figure 3 illustrates the different statistical bases for each of these approaches. We stick to the NTM examples for olive oil testing illustrated in Figure 2. The reader must note that the illustrations in Figure 3 do not cover the entire validation procedure. However, they graphically compare how a performance characteristic will be evaluated. Validation *approach A* is used when – (i) the classes are delimited by the level of seed oil adulteration in olive oil, and (ii) decisions scores from the dry lab statistical model are used. Figure 3 (top left) illustrates how the decision scores can be plotted as a function of the adulteration level of seed oil. A decision limit of 1.2 is considered here (exemplarily). The class definitions are delimited numerically, i.e., above and below 15% adulteration of seed oil. It can be evaluated that the FPR and FNR are below 5%, when the seed oil adulteration is below 13% and above 17% respectively. On similar lines, a recent study described a preliminary method performance characterization study, using quantitative decision scores (called D scores in the study [45]). Here the authors describe a potential method for distinguishing grain cultivars. Here the class definitions can be demarcated numerically for spelt bread containing 10% wheat (that is typically expected), and spelt bread greater than 10% wheat (case for food fraud).

Validation *approach B* (Figure 3, top right) is used when (i) classes are qualitatively delimited, i.e., olive oil from Italy (class α) and Turkey (class β), and (ii) quantitative decisions scores are used for validation. The statistical distributions for the decision scores associated with the classes are shown exemplarily. Assuming the decision limit is 2, about 10% of Italian olive oil samples are misidentified as originating from Turkey (FPR of 10%). Examples of validation procedures fitting to a general statistical model for this approach and corresponding validation parameters have been recently described by Uhlig and colleagues [49]. Such a construct for validation *approach B* was also one of the central themes of the work by Alewijn

et al. [24]. The authors describe in their work a case study to validate a method to detect organic and conventional eggs (two-class problem), where the decision score was utilized.
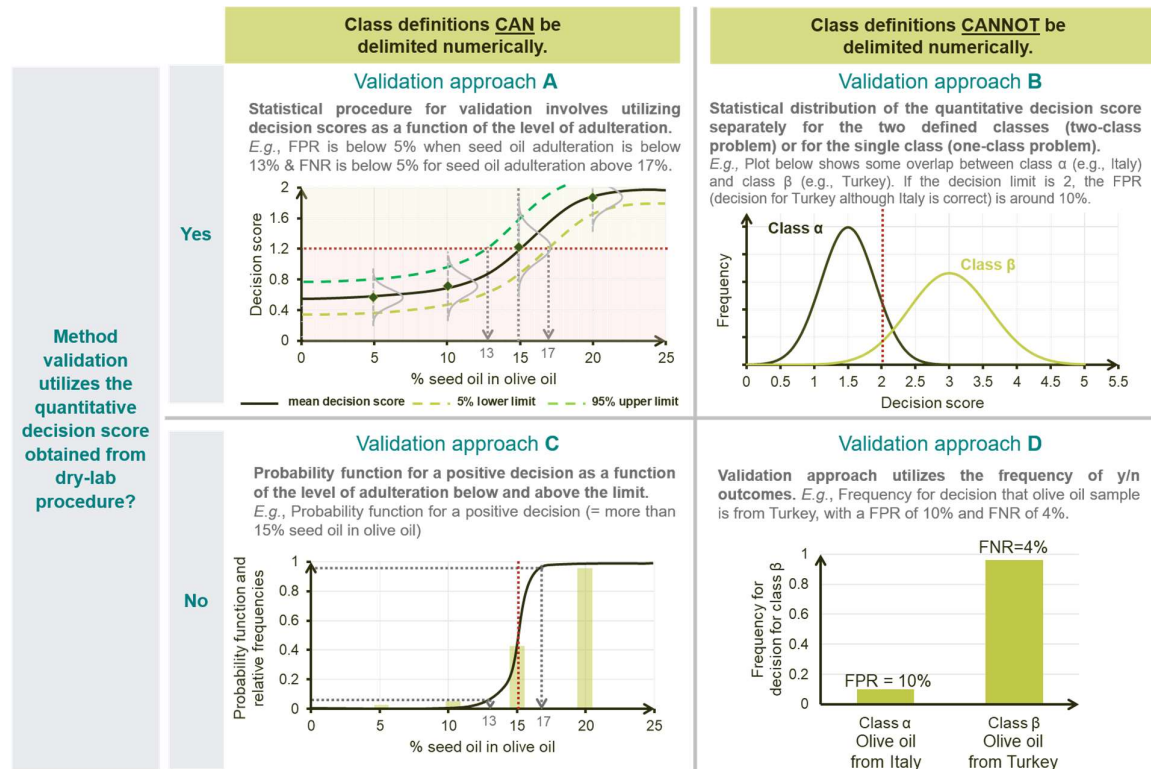
Figure 3: Options for validation approaches depending on whether the class definitions can be delimited numerically and whether the method validation utilizes quantitative decision scores or not.

Moving to validation *approach C* (Figure 3, bottom left). It is used when (i) the classes are delimited by the level of seed oil adulteration in olive oil, and (ii) y/n decision outcomes are used for validation. The figure exemplarily shows the relative frequencies of the samples tested at different levels of adulteration of vegetable oil, along with a probability of detection function for a decision that adulteration is greater than 15%. A statistical model for evaluating the probability of detection in collaborative studies of binary test methods has been presented by Uhlig and coworkers [50].

The probability of identification (POI) approach proposed by LaBudde and Harnly is an example for validation *approach C* [51]. The described approach attempts to discriminate between botanicals with acceptable limits of expected ingredients, and the ones with unacceptable limits. The POI is obtained from the measurement of specific superior test materials (SSTM) and specific inferior test materials (SITM) in different mixing ratios, so that the functional dependence on the respective mixing ratio can be determined.

Lastly, the validation *approach D* is used when - (i) classes are qualitatively delimited, i.e., olive oil from Italy (class α) and Turkey (class β), and (ii) y/n decision outcomes are used for validation. The example provided involves testing samples and finding the relative frequencies for deciding that the olive oil sample is from Turkey. The FPR thus obtained is 10% and FNR is 4%. The procedure described in the USP protocol is a befitting example of validation *approach D* [37].

5.4   Benefit for using quantitative decision scores for validation: simulation

To demonstrate the implications of using the quantitative decision scores instead of the positive or negative detect (y/n) decision, the following describes a simulated validation study. Figure 4 illustrates a comparison of calculated FPR (=1-specificity) according to validation *approaches B and D*. Three different simulations runs are shown under the assumption that the population of samples can be well described by a normal distribution. We consider 30 validation samples for class α (=negative) (shown as circles) and a decision limit of 2 (shown by vertical dotted line). Validation *approach B* makes use of the quantitative decision scores, whereas *approach D* is based on the counts of positive and negative results. In case that the classes can be considered homogeneous and normal distribution can be used to describe the distribution of quantitative decision scores, the arithmetic mean and standard deviation of decision scores can be used to calculate the expected FPR (which is represented by the red shaded area in Figure 4). For illustrative purposes, we chose a simple example where scores corresponding to only one class are shown (class α, suppose olive oil from Italy), and only compare one performance characteristic, namely, FPR.

Validation approach	False-positive rate	Validation approach	False-positive rate	Validation approach	False-positive rate
Approach B	< 1%	Approach B	6%	Approach B	10%
Approach D	0% (0/30 positive)	Approach D	0% (0/30 positive)	Approach D	10% (3/30 positive)
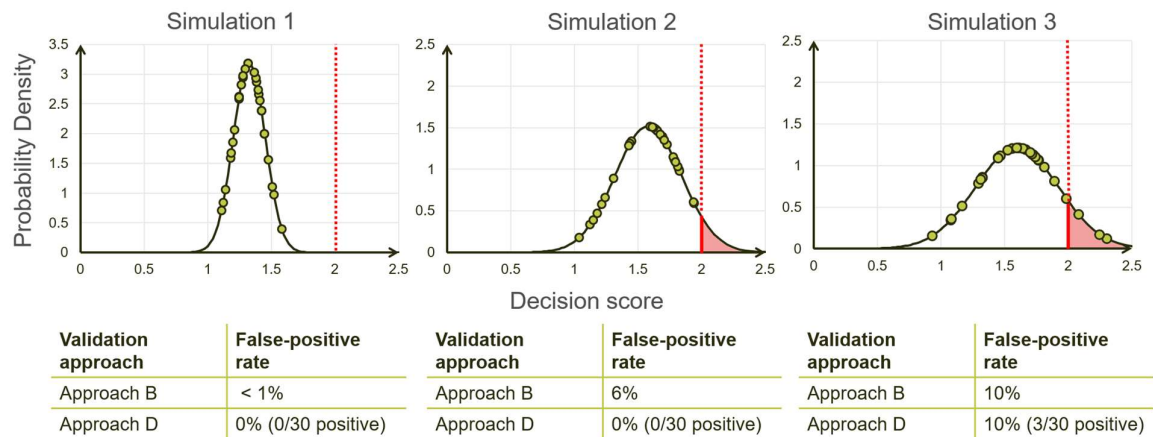
Figure 4: Simulations for illustrating the difference between approaches B and D. Consider the decision limit of 2 (red vertical dotted line) for all three simulations. For each simulation, the green circles represent the decision scores for 30 samples. The region beyond the decision limit of 2 and under the probability density curve (shaded red) is the probability of obtaining a false positive result.

Consider the scenario as shown in Figure 4 simulation 1. The decision scores of all 30 samples are below 2 and therefore the decision is for all samples: class α (=negative). In this situation, *approaches B and D* will come to the same conclusion. The FPR calculated with *approaches B and D* are <1% and 0% (i.e., 0/30 are positive), respectively. Since the FPR calculations are dependent on the underlying distribution of the data, it is a good statistical practice to not claim very small probabilities, and hence we only state that FPR < 1%. In another scenario as shown in Figure 4 simulation 2, the decision scores of all 30 samples are below 2 and therefore the decision is for samples class α (=negative). In this situation, *approaches B and D* will not come to the same conclusion. The FPR according to approach B is 6% (red shaded area) and according to approach D is still 0%. Alternatively, consider the scenario depicted in Figure 4 simulation 3. Here, 27 samples were detected as class α (=negative), and 3 samples were detected as not class α (=positive). In this scenario, again *approach B and D* will come to the same conclusion.

For *approach D*, even if the validation study result is perfect, i.e., FPR=0%, it is not

apparent whether this result is actually as clear-cut as in simulation 1, or borderline as in simulation 2. One cannot be sure with *approach D*, even without a single misclassification with 30 samples, that the actual FPR rate maybe 10% (see simulation 3). This shows that validation with 30 samples using *approach D* is very uncertain, and more samples are recommended. Roughly speaking, *approach B* might be leading to better results than *approach D* by considering the additional distance to the decision limit. It has to be noted however that *approach B* also has statistical uncertainties like the appropriate choice of the underlying distribution, but these are considerably lower so that a validation approach with only 30 samples per class seems justified.

The relationship highlighted in comparing validation *approaches B and D*, is also applicable for *approaches A and C*. Thus, there is considerable benefit in utilizing the decision scores in the validation procedure.

In the case of single lab validation, the decision scores may be used directly. However, for multi-lab validation, standardization of the decision scores across laboratories might be required, when different labs have minor differences in the dry lab procedure (e.g., software or algorithm is different for some labs). In the simplest sense, standardization of scores involves bringing all the decision scores to the same scale, such that they can be compared. Certainly, this additional step makes the validation procedure cumbersome. Furthermore, in many cases, the decision scores do not possess a physical interpretation and cannot be traced to a true physical value. Perhaps that are some reasons that existing validation procedures (USP protocol and AOAC SMPRs) only make use of the qualitative y/n outcomes.

6   Further considerations for applying NTM validation approaches

Experts need to make certain considerations when applying any validation approach. The considerations will influence (i) what performance characteristics to focus on, (ii) how to determine these performance characteristics, (iii) how to derive performance criteria, (iv) what data considerations need to be made, among others. In the following discussed are the different considerations.

6.1   Choice of considering the method as screening method or confirmatory method

It has to be distinguished whether the NTM method is to be used for screening purposes or as a confirmatory method. In the case of a screening method, the aim is to identify all samples that could be considered as suspect samples. In this case, one will try to minimize the FNR. In the case of a confirmatory method, on the other hand, one will try to prove that the sample is indeed positive. In this case, one will primarily try to minimize the FPR.

6.2   Dependence on the measurement platform

Due to their high resolution and high analytical sensitivity, NTMs with specific measurement platforms are increasingly used especially in the research field. As a result of the increasing popularity of a particular measurement platform, it might be useful to have a validation procedure tailored to it. In this case, the specific nuances of the method can be taken into consideration.

For example, NMR measurements have been used as part of an NTM for testing various types of beverages such as juice, coffee, wine, beer, and honey [52]. Platform-specific

(use of the instrument specific standard operating procedure and, where appropriate, use of the reference database of the instrument provider) validation of such proprietary methods allows specific aspects of the platform to be addressed in the validation process. On the other hand, for purposes of official food control, it seems more appropriate to adopt a platform-independent approach. It is to be expected that, similar to targeted methods, systematic differences between the platforms cannot be avoided. The use of different dry lab approaches can also lead to systematic differences in the results. Therefore, the influence of platform-specific effects as well as the influence of dry lab effects must be checked during validation.

6.3   Modular or comprehensive validation

Selecting a modular strategy for validating the wet lab and dry lab procedures separately [22,23] is another consideration to be made. Given the different parts of NTM, a modular validation can help to avoid *surprises* at the end. It can be performed only if there is evidence that the wet lab and dry lab performance can be considered independent. This can be possible when dealing with certain microbiological or molecular methods [34]. However, in most other cases, the final outcome will be affected by variations in the wet lab and the dry lab procedures.

In a comprehensive validation, all the steps until the final decision outcome are included. An argument in favor of comprehensive validation is that the method outcomes from the dry lab can seldom be considered independent of the measurement results from the wet lab. They are affected by variations from different steps. How these variations in the wet-lab data translate to the final outcome, is an important aspect to be examined. Thus, we believe comprehensive validation will be necessary for NTMs.

7   Samples for the validation study

Turning to the question of how many and which samples to be used in the validation study, the following lists the different criteria that must be fulfilled. Here the difference between samples used to develop the reference database (or train a machine learning model) and samples used in the validation study should be emphasized. The former is part of the method development phase, while the latter is part of the method validation phase (see section 8). First and foremost, the samples must be representative of the population of the food covered by the method. Consider an example of an NTM to detect if rice is basmati or not. In order to validate the method, it is crucial that rice samples must be sourced from the Indian subcontinent and not another region (e.g., Italy). This is important as basmati is largely grown in that region. Secondly, these samples must be independent and distinct from the ones that are there in the reference database. Additionally, it must be ensured (to the best possible extent) that the validation samples are not sourced from the same distributor/supplier or farm/location or processing plant as the samples used to build the reference database.

Next, is the question of how many samples are to be tested in the validation study. As discussed in Sections 5 and 6, NTMs can be formulated in a variety of ways. The validation approach and the choice of the number of samples depend on several factors such as (i) the desired confidence in the results, (ii) whether an NTM is to be employed as a screening method or confirmatory method, (iii) the scope of the NTM, (iv) the testing burden on the laboratory from an economical and practical perspective, (v) type of statistical study design adopted – conventional or factorial designs, (vi) considerations for matrix effects (vii) variations (e.g. seasonal effects) within the respective sample groups/cohort. Therefore, claims for an

exact number of samples required in a validation study for NTM not only need to be embedded in sound statistical theory but also must consider the details of the method to be validated.

However, a few proposals for the number of samples required for method validation have been previously reported. These numbers should be considered only in connection to the described validation scheme and the underlying statistical assumptions. The AOAC SMPRs suggest 30 validation samples for each adulterant [39–41]. Another recent report states for a binary NTM, at least 60 samples per class would be required to ensure with a statistical certainty of 95%, that there is the inclusion of at least one sample from a subpopulation with a small proportion of 5% [37]. This can be ensured by using a factorial approach, in which all subclasses or subpopulations resulting from differences in the cultivation, processing, packaging, and delivery of the food are equally taken into account.

8   Stages in the development and validation of an NTM

Owing to the complexity involved in the development of an NTM, with so many different components and steps, it can be foreseen that some manner of cooperative method development can alleviate the resource burden. Herein, the effort in the method development stage is distributed among multiple laboratories (or institutions). For instance, a simple split in the effort is made with respect to the wet lab and dry lab development, performed by separate labs (or institutions) respectively. The cooperative method development with multiple labs represents a new paradigm in method development because such an approach has ramifications on the existing procedures of laboratory accreditation and the establishment of an official method. Efforts on this front will be required to introduce procedures and protocols thereof. Until then, the "conventional way" of method development in one lab will likely continue (see Table 1). However, the concept of cooperative method development is gaining traction, especially for NTM development.

Table 1: Important phases of method development and method validation, to be performed in a conventional or cooperative way.

Method development	Method Validation
Wet lab + dry lab + reference database	
Single lab method development	Validation of the method in a single laboratory using a specific set of samples, followed by a multi-lab validation study with a smaller number of samples.
Multi-lab cooperative method development	Multi-lab validation studies

Figure 5 illustrates the different stages that are passed through until the standardization of an NTM for food authenticity testing. Once the wet lab and dry lab components of the method are developed, optimized, and perfected for the given method scope, the next natural phase step is the method validation. The method validation phase has several stages until a method can be adopted for official control. Here, while referring to method validation, once again it must be emphasized that it is used in the context of determining method performance characteristics. First, the unvalidated method enters the implementation or prevalidation stage. This stage is referred to as implementation and prevalidation because (i) it precedes the most important stage – the multi-lab validation study, and (ii) it allows to further finetune the method as a whole, as results from method implementation can be used to revisit the scope, or improve the analytical procedure, or edit spurious entries in the reference database.

Typically, after method development, SLV is performed, followed by a method validation interlaboratory test. But if the NTM is developed collaboratively, then single lab validation is inappropriate. SLV would be a suitable option to perform during this stage when a method validation interlaboratory test is followed (see Table 1). Thus, with NTMs, SLV can be performed as part of the prevalidation stage. The exact validation procedure to be adopted and the sample requirements will be based on the discussions made so far in the previous sections.

Another important function of the prevalidation phase is to identify the characteristics of challenging samples, i.e., to determine for which sample types the NTM has particular difficulty in finding the correct classification. Interestingly, since challenging is not an inherent characteristic of the sample, a set of challenging samples for one lab, may not be challenging for another lab. Evidently, the main outcome of the method implementation and prevalidation stage is to identify such challenging samples. Iterative cycles of development, implementation, and prevalidation can be conducted to improve the method.

The final step in method validation is the multi-lab method validation study, the design for which can be very different: conventional or more efficient factorial designs can be used here [53]. With factorial designs, there is a possibility of reducing the number of labs. Once the multi-laboratory validation study has been performed and the method performance characteristics have been established, the standardization process can be completed and the NTM can be introduced into routine practice.

Nevertheless, even after completion of the standardization procedure, changes in the underlying reference database, corresponding changes in the dry lab procedure or extended or modified questions can be expected at any time. For instance, the reference sample database can be updated to include a greater number of samples, broadening the scope of food types that can be assessed. Further, the algorithm or the software is updated with a newer version, which might lead to superior discrimination. In these scenarios, the results from the method can deviate drastically. Even though the modifications to the method are aimed at improving it, the validation data then may not adequately describe the method's performance. It is therefore necessary to implement a monitoring program for the NTM to control important performance parameters on a regular basis.
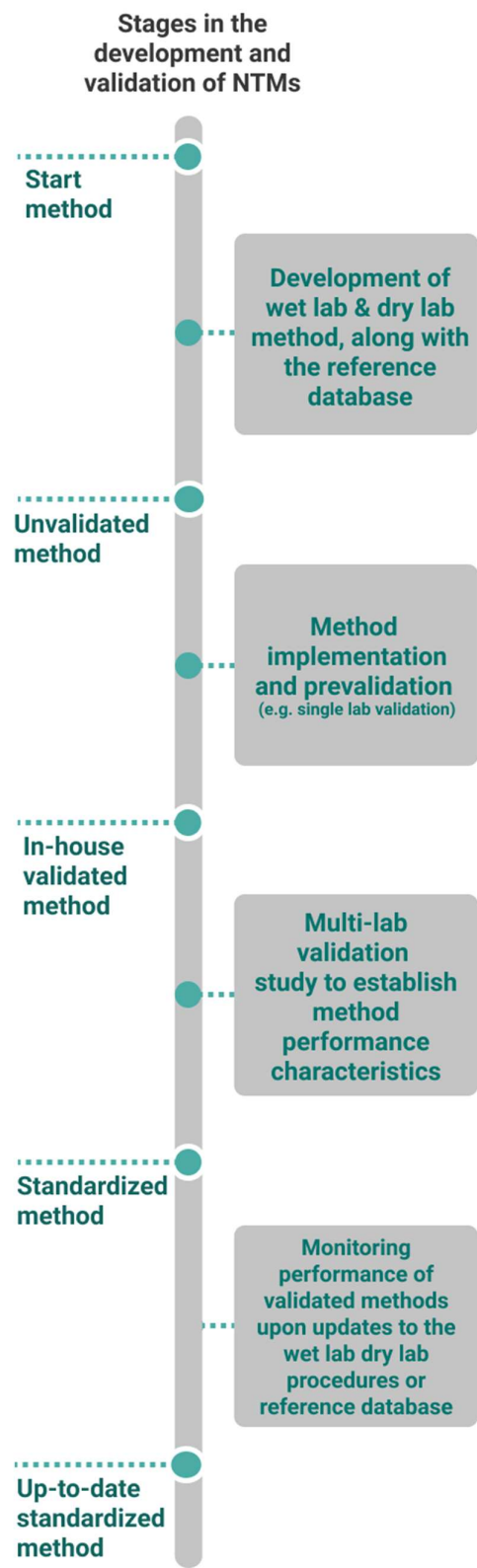
Figure 5: Stages in the development and validation of NTMs.

9   Final discussion and conclusions

Even with a simplified view, the notations around NTMs are dense. And hence, an attempt to deconvolve and disseminate the intricacies has been made in this paper. Our particular focus on the different considerations for NTM validation makes obvious the fact that devising a validation strategy will require their collective assessment. We discussed how to schematize NTMs for authenticity testing, with the aim of proposing a road map to approaching NTM validation. Ultimately, the intention of an NTM is to classify samples and thereby help in deciding whether a tested sample belong to a class. The method can be formulated as a single-class, two-class (binary NTM) or multi-class NTM problems. The definition of the underlying classes can be derived qualitatively, e.g., classifying by geographic origin. Alternatively, they can be derived quantitatively, e.g., deciding whether the weight percentage of a lower-priced food (adulterant) is below or above a certain threshold. In both cases, the classification is mostly based on a quantitative score value that is compared to a fixed decision limit. The aim of this comparison is either to decide whether a tested sample belongs to a defined group (single-class) or to decide to which of the previously defined groups the sample belongs (multi-class). This validation of a quantitative score value with a fixed decision limit finds its analog in the validation of a measured pollutant concentration with a legal maximum value.

The validation procedures discussed so far in this work make a differentiation according to whether the respective qualitative decision results of different laboratories for the validation samples are used in the assessment of the performance of the NTM, or whether the respective underlying quantitative scores values are also considered for the respective validation samples measured by different laboratories. The number of samples required for validation will have to determined accordingly, however some orientation values are discussed. We show the merits in the assessment of the NTM performance by using this quantitative score value in the validation. A superior validation result can be achieved with a significantly lower validation effort (number of samples and number of laboratories).

The particular challenge now is that the quantitative score value, which forms the basis for the respective classification, is not traceable to a specific reference standard. Unlike a content determination, it cannot be assumed that the quantitative score value fluctuates more or less randomly around a known, true value. When now different manufacturers develop NTMs for their respective instrument platforms, the corresponding quantitative score values are not directly comparable. It is a mathematically challenging task to develop a mathematical-statistical procedure that allows platform-specific quantitative score values to be compared with each other.

It also follows from the above that the planning of the NTM development and validation, should ideally be designed from the beginning to be platform-independent and multi-laboratory. Individual sub-steps (development of wet and dry lab methods, method implementation, and prevalidation - see Figure 5) can be carried out by individual laboratories. Furthermore, the planning of the validation of NTM should already include all components of the NTM (see Figure 1). The contents of the reference database (if necessary, already existing database can be used) and the intended decision criteria must be clearly planned and predefined from the beginning. Finally, it must be ensured that the analytical method (wet lab) must be suitably standardized (as few random error components as possible and as few systematic error components as possible) as well as applied comparably by all laboratories in order to generate suitable data for a reference database.

If necessary, different statistical models (workflows) are also possible, the comparability of which has been checked during the development and validation of the method. An important difference to targeted methods is that the effort for further quality assurance in the routine is significantly higher. All components of the NTM (see Figure 1) have to be evaluated (reference database and decision criteria) or quality assured (analytical method and statistical model) at regular intervals. In this context, it should be noted that it is in the nature of an NTM that questions may change (e.g., slight changes in the underlying classes), or that additions or extensions in the objectives are likely to be the rule rather than the exception. Thus, it is recommended that the (raw) data collected during method development and validation should therefore be stored centrally and in a structured form. We believe that embracing and elaborating the tenets around NTM validation outlined in this paper can guide to devise and adopt suitable validation procedures.

Acknowledgements

KN acknowledges the support from Harshadrai M. Rawel from Institute of Nutritional Science, University of Potsdam, and is thankful for the stimulating discussions. KN also thanks Kirsten Simon from QuoData for supporting and facilitating this work.

10 References

1. Ulberth, F. Tools to combat food fraud - a gap analysis. *Food Chem.* **2020**, 127044. 10.1016/j.foodchem.2020.127044.
2. Böhme, Karola, Calo-Mata, P.; Barros-Velázquez, J.; Ortea, I. Recent applications of omics-based technologies to main topics in food authentication. **2018**. https://doi.org/10.1016/j.trac.2018.11.005.
3. Medina, S.; Pereira, J.A.; Silva, P.; Perestrelo, R.; Câmara, J.S. Food fingerprints--A valuable tool to monitor food authenticity and safety. *Food Chem.* **2019**, *278*, 144–162.
4. Hollender, J.; Schymanski, E.L.; Singer, H.P.; Ferguson, P.L. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? *Environ. Sci. Technol.* **2017**, *51*, 11505–11512. 10.1021/acs.est.7b02184.
5. Singhal, N.; Kumar, M.; Kanaujia, P.K.; Virdi, J.S. MALDI-TOF mass spectrometry : an emerging technology for microbial identification and diagnosis. **2015**, *6*, 1–16. 10.3389/fmicb.2015.00791.
6. Monakova, Y.B.; Schütz, B.; Schäfer, H.; Spraul, M.; Kuballa, T.; Hahn, H.; Lachenmeier, D.W. Validation studies for multicomponent quantitative NMR analysis : the example of apple fruit juice. *Accredit. Qual. Assur.* **2014**, *19*, 17–29. 10.1007/s00769-013-1026-3.
7. PerkinElmer DairyGuard Milk Powder Analyzer Available online: https://www.perkinelmer.com/product/dairyguard-milk-powder-analyzer-l1280080 (accessed on Nov 25, 2020).
8. Solovyev, P.A.; Fauhl-Hassek, C.; Riedl, J.; Esslinger, S.; Bontempo, L.; Camin, F. NMR spectroscopy in wine authentication: An official control perspective. *Compr. Rev. Food Sci. Food Saf.* **2021**, *20*, 2040–2062. 10.1111/1541-4337.12700.
9. Ballin, N.Z.; Laursen, K.H. To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication. *Trends Food Sci. Technol.* **2019**, *86*, 537–543. 10.1016/j.tifs.2018.09.025.
10. Magnusson, B.; Örnemark, U. *Eurachem Guide: The Fitness for Purpose of Analytical Methods – A Laboratory Guide to Method Validation and Related Topics*; 2014; ISBN 978-91-87461-59-0.
11. Thompson, M.; Ellison, S.L.R.; Wood, R. Harmonized guidelines for single-laboratory validation of methods of analysis (IUPAC Technical Report). *Pure Appl. Chem.* **2002**, *74*, 835–855.
12. *EU Controls Regulation: REGULATION (EU) 2017/625 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 15 March 2017 on official controls and other official activities performed to ensure the application of food and feed law, rules on animal health and welfare*; European Union, 2017;
13. Riedl, J.; Esslinger, S.; Fauhl-Hassek, C. Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. *Anal. Chim. Acta* **2015**, *885*, 17–32.

10.1016/j.aca.2015.06.003.

14. Cavanna, D.; Righetti, L.; Elliott, C.; Suman, M. The scientific challenges in moving from targeted to non-targeted mass spectrometric methods for food fraud analysis : A proposed validation workflow to bring about a harmonized app ... Trends in Food Science & Technology The scienti fi c challenges in mo. *Trends Food Sci. Technol.* **2018**, *80*, 223–241. 10.1016/j.tifs.2018.08.007.

15. McGrath, T.F.; Haughey, S.A.; Patterson, J.; Fauhl-Hassek, C.; Donarski, J.; Alewijn, M.; van Ruth, S.; Elliott, C.T. What are the scientific challenges in moving from targeted to non-targeted methods for food fraud testing and how can they be addressed? – Spectroscopy case study. *Trends Food Sci. Technol.* **2018**, *76*, 38–55. 10.1016/j.tifs.2018.04.001.

16. Esslinger, S.; Riedl, J.; Fauhl-Hassek, C. Potential and limitations of non-targeted fingerprinting for authentication of food in of fi cial control. **2014**, *60*, 189–204.

17. Creydt, M.; Fischer, M. Food authentication in real life: How to link nontargeted approaches with routine analytics? *Electrophoresis* **2020**, 1665–1679. 10.1002/elps.202000030.

18. Creydt, M.; Fischer, M. Mass-Spectrometry-Based Food Metabolomics in Routine Applications: A Basic Standardization Approach Using Housekeeping Metabolites for the Authentication of Asparagus. **2020**. 10.1021/acs.jafc.0c01204.

19. Sobolev, A.P.; Thomas, F.; Donarski, J.; Ingallina, C.; Circi, S.; Cesare Marincola, F.; Capitani, D.; Mannina, L. Use of NMR applications to tackle future food fraud issues. *Trends Food Sci. Technol.* **2019**, *91*, 347–353. https://doi.org/10.1016/j.tifs.2019.07.035.

20. Segelke, T.; Schelm, S.; Ahlers, C.; Fischer, M. Food Authentication : Truffle (Tuber spp.) Species Differentiation by FT-NIR and Chemometrics. *Foods* **2020**, 922. 10.3390/foods9070922.

21. Grassi, S.; Casiraghi, E.; Alamprese, C. Handheld NIR device: A non-targeted approach to assess authenticity of fish fillets and patties. *Food Chem.* **2018**, *243*, 382–388. https://doi.org/10.1016/j.foodchem.2017.09.145.

22. Xu, Y.; Zhong, P.; Jiang, A.; Shen, X.; Li, X.; Xu, Z.; Shen, Y.; Sun, Y.; Lei, H. Raman spectroscopy coupled with chemometrics for food authentication: A review. *TrAC Trends Anal. Chem.* **2020**, *131*, 116017. https://doi.org/10.1016/j.trac.2020.116017.

23. Haynes, E.; Jimenez, E.; Pardo, M.A.; Helyar, S.J. The future of NGS (Next Generation Sequencing) analysis in testing food authenticity. *Food Control* **2019**, *101*, 134–143. 10.1016/j.foodcont.2019.02.010.

24. Alewijn, M.; van der Voet, H.; van Ruth, S. Validation of multivariate classification methods using analytical fingerprints – concept and case study on organic feed for laying hens. *J. Food Compos. Anal.* **2016**, *51*, 15–23. 10.1016/j.jfca.2016.06.003.

25. Ballabio, D.; Robotti, E.; Grisoni, F.; Quasso, F.; Bobba, M.; Vercelli, S.; Gosetti, F.; Calabrese, G.; Sangiorgi, E.; Orlandi, M.; et al. Chemical profiling and multivariate data fusion methods for the identification of the botanical origin of honey. *Food Chem.* **2018**, *266*, 79–89.

26. Danezis, G.P.; Pappas, A.C.; Zoidis, E.; Papadomichelakis, G.; Hadjigeorgiou, I.; Zhang, P.; Brusic, V.; Georgiou, C.A. Game meat authentication through rare earth elements fingerprinting. *Anal. Chim. Acta* **2017**, *991*, 46–57.

27. Sørensen, S.; Lund, K.H.; Cederberg, T.L.; Ballin, N.Z. Identification of Baltic Sea salmon based on PCB and dioxin profiles. *Food Control* **2016**, *61*, 165–171.

28. *IUPAC Compendium of Chemical Terminology*; IUPAC, 2014; ISBN 0-9678550-9-8. 10.1351/goldbook.

29. Gao, B.; Holroyd, S.E.; Moore, J.C.; Laurvick, K.; Gendel, S.M.; Xie, Z. Opportunities and Challenges Using Non-targeted Methods for Food Fraud Detection. *J. Agric. Food Chem.* **2019**, *67*, 8425–8430. 10.1021/acs.jafc.9b03085.

30. MacNeil, J.D.; Patterson, J.; Martz, V. Validation of Analytical Methods-Proving Your Method is' Fit for Purpose'. *Spec. Publ. R. Soc. Chem.* **2000**, *256*, 100–107.

31. Filzmoser, P.; Liebmann, B.; Varmuza, K. Repeated double cross validation. **2009**, 160–171. 10.1002/cem.1225.

32. Uhlig, S.; Colson, B.; Hettwer, K.; Simon, K.; Uhlig, C.; Wittke, S.; Stoyke, M.; Gowik, P. Valid machine learning algorithms for multiparameter methods. **2019**.

33. ISO *ISO 17025/IEC:2017 General requirements for the competence of testing and calibration laboratories*; 2017;

34. *ISO 16140-1:2016 Microbiology of the food chain — Method validation — Part 1: Vocabulary*; 2016;

35. ISO Accuracy (Trueness and Precision) of Measurement Methods and Results", Parts 1-6, International Standard ISO 5725-1:1994, 5725-2:1994, 5725-3:1994, 5725-4:1994, 5725-5:1994, and 5725-6:1994. Available online: https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:ed-1:v1:en (accessed on Aug 17, 2021).

36. FDA, U. *Guidelines for the Validation of Chemical Methods for the FDA Foods Program*; 2019;

37. US Pharmacopeia *Appendix XVIII: USP Guidance on Developing and Validating Non-Targeted*

Methods for; 2019;

38. AOAC Appendix F: Guidelines for Standard Method Performance Requirements. *AOAC Off. Methods Anal.* **2016**, 9.

39. AOAC AOAC SMPR 2020.007® Standard Method Performance Requirements (SMPRs®) for Nontargeted Testing (NTT) of Ingredients for Food Authenticity/Fraud Evaluation of Extra Virgin Olive Oil Intended Available online: https://www.aoac.org/resources/aoac-standard-method-performance-requirements-smpr-for-non-1-targeted-testing-ntt-of-ingredients-for-food-authenticity-fraud-evaluation-of-extra-virgin-olive-oil/ (accessed on Dec 15, 2021).

40. AOAC AOAC SMPR 2020.006® Standard Method Performance Requirements (SMPRs®) for Nontargeted Testing (NTT) of Ingredients for Food Authenticity/Fraud Evaluation of Honey Available online: https://www.aoac.org/resources/standard-method-performance-requirements-smpr-for-non-targeted-testing-ntt-of-ingredients-for-food-authenticity-fraud-evaluation-of-honey/ (accessed on Dec 15, 2021).

41. AOAC AOAC SMPR 2020.008® Standard Method Performance Requirements (SMPRs®) for Nontargeted Testing (NTT) of Ingredients for Food Authenticity/Fraud Evaluation of Pasteurized Whole Liquid Bovine Milk Available online: https://www.aoac.org/resources/standard-method-performance-requirements-smpr-for-non-targeted-testing-ntt-of-ingredients-for-food-authenticity-fraud-evaluation-of-liquid-raw-bovine-milk/ (accessed on Dec 15, 2021).

42. AOAC AOAC SMPR 2021.XXX; Draft AOAC Standard Method Performance Requirements (SMPRs) for Non-Targeted Testing (NTT) of Ingredients for Food Authenticity Methods Evaluation of Vanilla Powder and Extracts; 10/19/21 Available online: https://www.aoac.org/wp-content/uploads/2021/05/SMPR-NTT-Adulteration-of-Saffron-2021-04-29.pdf (accessed on Dec 15, 2021).

43. AOAC AOAC SMPR 2021.XXX; Draft AOAC Standard Method Performance Requirements (SMPRs) for Non-Targeted Testing (NTT) of Ingredients for Food Authenticity Methods Evaluation of Saffron; April 29, 2021 Available online: https://www.aoac.org/wp-content/uploads/2021/05/SMPR-NTT-Adulteration-of-Saffron-2021-04-29.pdf (accessed on Dec 15, 2021).

44. AOAC *Standard Method Performance Requirements (SMPRs) for Non-Targeted Testing (NTT) of Ingredients for Food Authenticity/Fraud Evaluation of Extra Virgin Olive Oil; version 5; April 21, 2020*; 2020;

45. Nichani, K.; Uhlig, S.; Colson, B.; Hettwer, K.; Simon, K.; Bönick, J.; Uhlig, C.; Rawel, H.M.; Stoyke, M.; Gowik, P.; et al. AI-based identification of grain cultivars via non-target mass spectrometry. *bioRxiv* **2020**, 2020.05.07.082065. 10.1101/2020.05.07.082065.

46. Weesepoel, Y.; Alewijn, M.; Wijtten, M.; Müller-Maatsch, J. Detecting food fraud in extra virgin olive oil using a prototype portable hyphenated photonics sensor. *J. AOAC Int.* **2020**. 10.1093/jaoacint/qsaa099.

47. Alamprese, C.; Manuel, J.; Casiraghi, E.; Balling, S. Identification and quantification of turkey meat adulteration in fresh, frozen-thawed and cooked minced beef by FT-NIR spectroscopy and chemometrics. **2016**, *121*, 175–181. 10.1016/j.meatsci.2016.06.018.

48. Botelho, B.G.; Reis, N.; Oliveira, L.S.; Sena, M.M. Development and analytical validation of a screening method for simultaneous detection of five adulterants in raw milk using mid-infrared spectroscopy and PLS-DA. *Food Chem.* **2015**, *181*, 31–37. 10.1016/j.foodchem.2015.02.077.

49. Uhlig, S.; Nichani, K.; Stoyke, M.; Gowik, P. Validation of binary non-targeted methods: mathematical framework and experimental designs. **2021**, 15–17.

50. Uhlig, S.; Krügener, S.; Gowik, P. A new profile likelihood confidence interval for the mean probability of detection in collaborative studies of binary test methods. *Accredit. Qual. Assur.* **2013**, *18*, 367–372.

51. Labudde, R.A.; Harnly, J.M. Probability of identification: A statistical model for the validation of qualitative botanical identification methods. *J. AOAC Int.* **2012**, *95*, 273–285. 10.5740/jaoacint.11-266.

52. Kuballa, T.; Brunner, T.S.; Thongpanchang, T.; Walch, S.G.; Lachenmeier, D.W. Application of NMR for authentication of honey, beer and spices. *Curr. Opin. Food Sci.* **2018**, *19*, 57–62. https://doi.org/10.1016/j.cofs.2018.01.007.

53. Uhlig, S.; Gowik, P. Efficient estimation of interlaboratory and in-house reproducibility standard deviation in factorial validation studies. *J. Consum. Prot. Food Saf.* **2018**, *13*, 315–322.