Performance characteristics and criteria for non-targeted methods

Eurachem workshop in Tartu, Estonia

21 May 2019

Steffen Uhlig, Kapil Nichani, Bertrand Colson, Karina Hettwer, Kirsten Simon (QuoData.de, Germany)

Carsten Uhlig (Akees.com, Berlin)

Manfred Stoyke, Ulrike Steinacker, René Becker, Petra Gowik (BVL, Germany)

Introduction

In method validation, the **performance** of the method is **characterized** and then **assessed** against **criteria** derived from fitness-for-purpose considerations.

Performance characteristics (e.g. trueness, reproducibility precision, sensitivity, LOD)

Performance criteria (e.g. $s_R \leq 30\%$, $LOD \leq 1$ CFU per test portion)

Introduction

- In many modern applications, non-targeted methods are applied (food fraud, detection of all potentially toxic substances in a water sample, contamination via NIAS migration).

- The aim of this presentation is to discuss the characterization and assessment of *method performance* in connection with non-targeted methods.

Introduction

- Data from non-targeted workflows are typically used in connection with ***classification problems.***

- Typical examples include:
  - Food origin
  - Species identification

→ In this presentation, the discussion will be based on such a classification problem from the field of microbiology.

Identification of a particular type of *Staphylococcus aureus*

The discussion in this presentation will revolve around a concrete example:

a method for the distinction between *Staph. aureus* subtypes (Type R versus Type S).

MALDI-TOF/AI method and dataset

- The *method* being validated consists of two broad steps:
  - Obtaining a full-scan spectrum (e.g. MALDI-TOF)
  - Artificial Intelligence (AI) algorithm for spectrum analysis
    - ➢ The method will be referred to as MALDI-TOF/AI

- Data corresponding to 190 *Staph. aureus* isolates collected from diseased cattle were available for the validation of the method:
  - 162 Type S isolates
  - 28 Type R isolates

    → 380 MALDI_TOF duplicate (2018 and 2019) *Staph. aureus* spectra.

Validation of qualitative methods

- At the moment, no procedure has been set forth in an international standard or guideline for the validation of a qualitative method such as MALDI-TOF/AI.

- In the ISO 16140 series (validation of methods in food microbiology), the validation of qualitative methods is addressed – however, the question is not whether a sample can be assigned to a particular class but whether detection has taken place.

- Nonetheless, traditional performance characteristics for qualitative methods are – at least on the face of it – perfectly applicable to MALDI-TOF/AI.

→ first and foremost, sensitivity and specificity.

Validation of qualitative methods

The question is: how reliable is the characterization of method performance?

In other words: **how many samples** are required in order to ensure a reliable characterization?

		Isolate	
		+ (Type R)	– (Type S)
MALDI-TOF/AI Classification result	+ (Type R)	True positive	False positive
	– (Type S)	False negative	True negative

Validation of qualitative methods

- Take a random sample of 10 isolates for each class.
- What can be concluded if all isolates are correctly identified?
- False positive rate (FPR) is calculated as 0 %.
- However, the *upper limit* of the 95 % confidence interval for FPR is around 26 % (binomial distribution).

		Isolate	
		+ (Type R)	– (Type S)
MALDI-TOF/AI Classification result	+ (Type R)	10	0
	– (Type S)	0	10

Validation of qualitative methods

- Take a random sample of 20 isolates for each class.
- What can be concluded if all isolates are correctly identified?
- False positive rate (FPR) is calculated as 0 %.
- However, the *upper limit* of the 95 % confidence interval for FPR is around 14 % (binomial distribution).

		Isolate	
		+ (Type R)	– (Type S)
MALDI-TOF/AI Classification result	+ (Type R)	20	0
	– (Type S)	0	20

Validation of qualitative methods

- The last two slides show that, unless the sample size is large enough, the estimates of performance characteristics (such as False Positive Rate) have unacceptably large confidence intervals.

- One consequence could be: After a successful validation study, the performance of the method is characterized in terms of FPR = 0 %. However, the true FPR lies e.g. around 20 %.

- If it is not possible to increase the sample size (say, to 50 samples), turn to another approach:

→ method characterization in terms of the underlying quantitative values[1]

[1]For a discussion of qualitative results and underlying (or "latent") quantitative variables, see the following publications:

Uhlig et al. (2011) Can the usual validation standard series for quantitative methods, ISO 5725, be also applied for qualitative methods? Accreditation and Quality Assurance

Uhlig et al. (2013) A new profile likelihood confidence interval for the mean probability of detection in collaborative studies of binary test methods. Accreditation and Quality Assurance

Performance characterization on the basis of classification scores

The output of many common AI methods such as

- principal component analysis (PCA)
- nonlinear iterative partial least squares (NIPALS)
- logistic regression
- random forests
- artificial neural networks (ANN)
- support vector machines (SVM)

can be transformed in such a way as to obtain

quantitative results

which often follow a ***normal distribution***

These (hopefully normally-distributed) **quantitative** results will be called **classification scores**

Performance characterization on the basis of classification scores

Decision rules

The classification result often involves the application of a **decision rule.**

This decision rule typically involves comparison to a **cut-off.**

Decision rule – comparison to a cut-off:

If

Classification score ≥ Cut-off

then the corresponding sample is assigned to **Class A**

Otherwise

it is assigned to **Class NOT A**

Performance characterization on the basis of classification scores

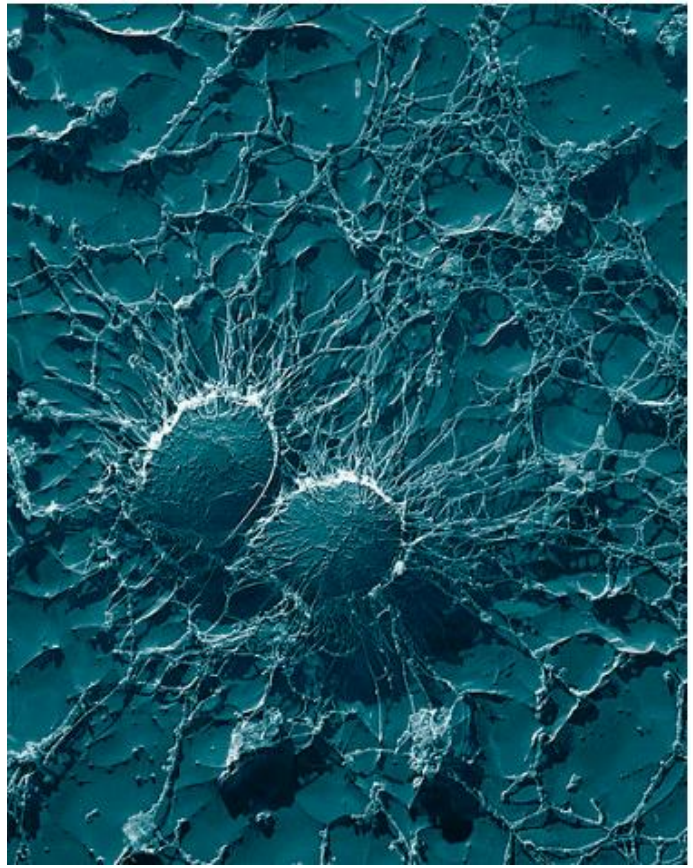
In the following,

the *characterization* and *assessment*

of the performance of MALDI-TOF/AI

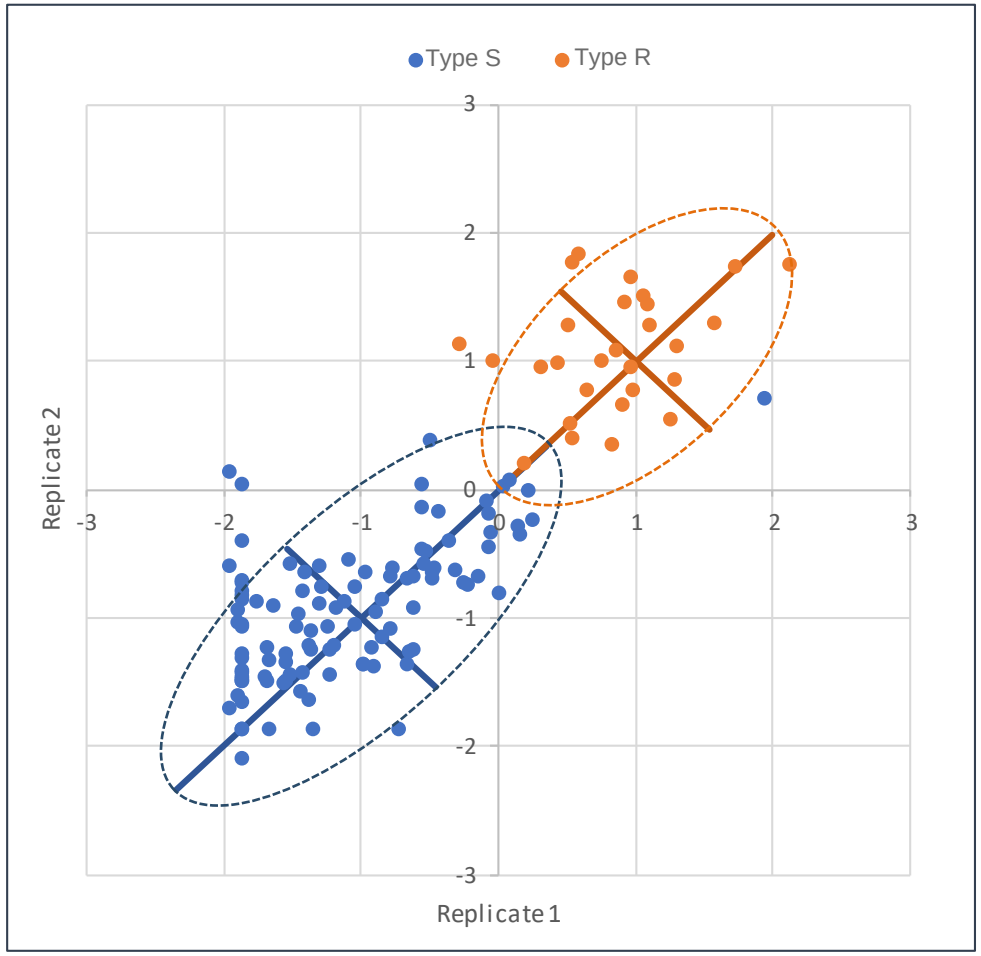
on the basis of classification scores

will be illustrated with the *Staph. Aureus* data.

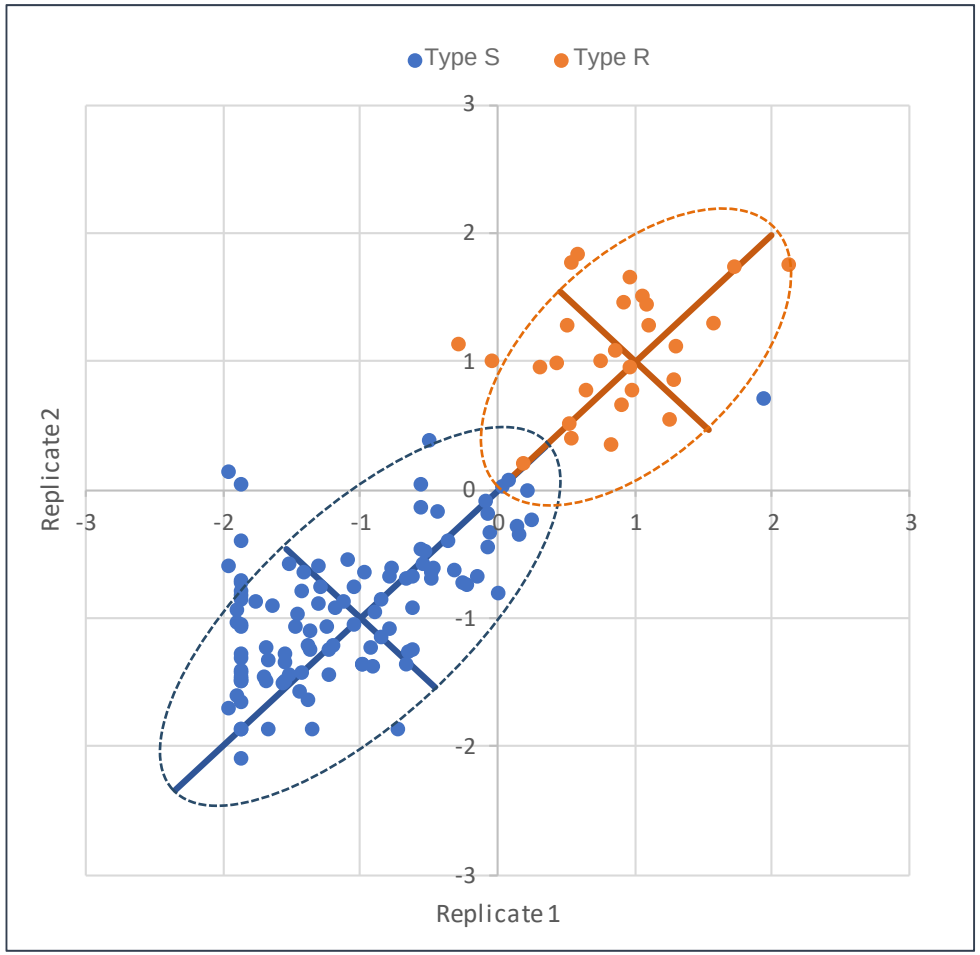


Performance characterization on the basis of classification scores

Youden plot of standardized classification scores

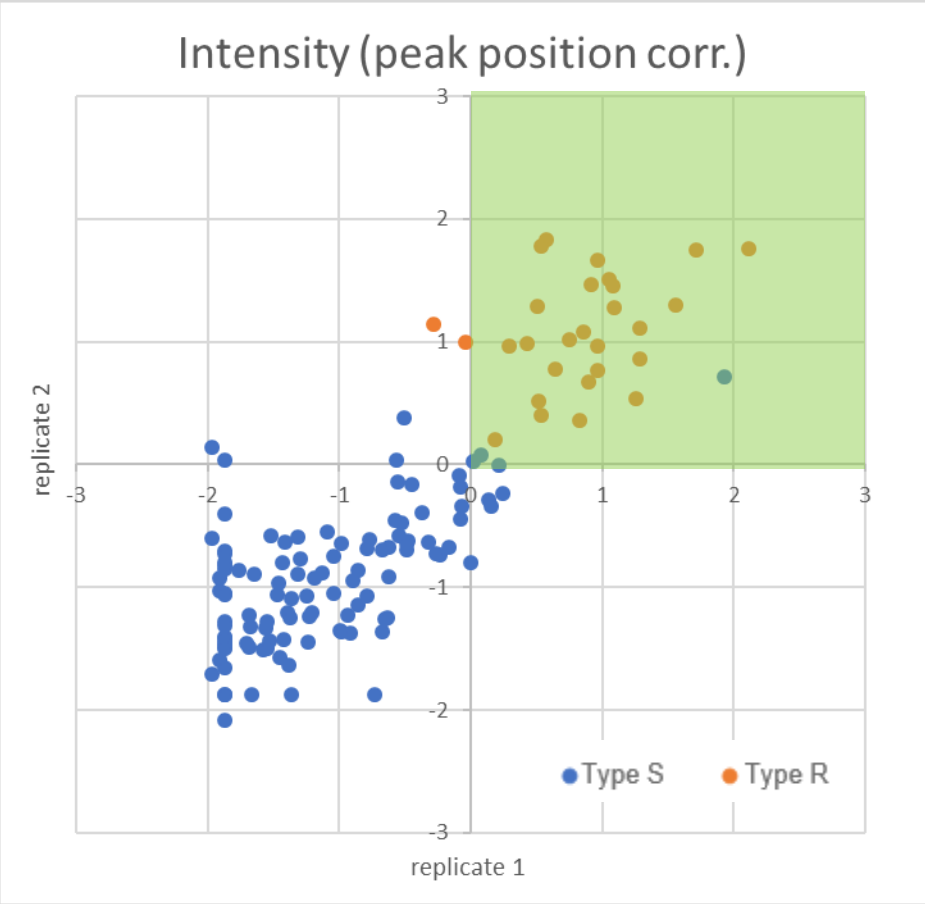


Performance characterization on the basis of classification scores



	Type S	Type R
Mean value	-1	1
Repeatability SD	0.26	-
Time SD	0.17	-
Intermediate SD	0.31	0.31
Population SD	0.50	0.34
Classification SD	0.59	0.46

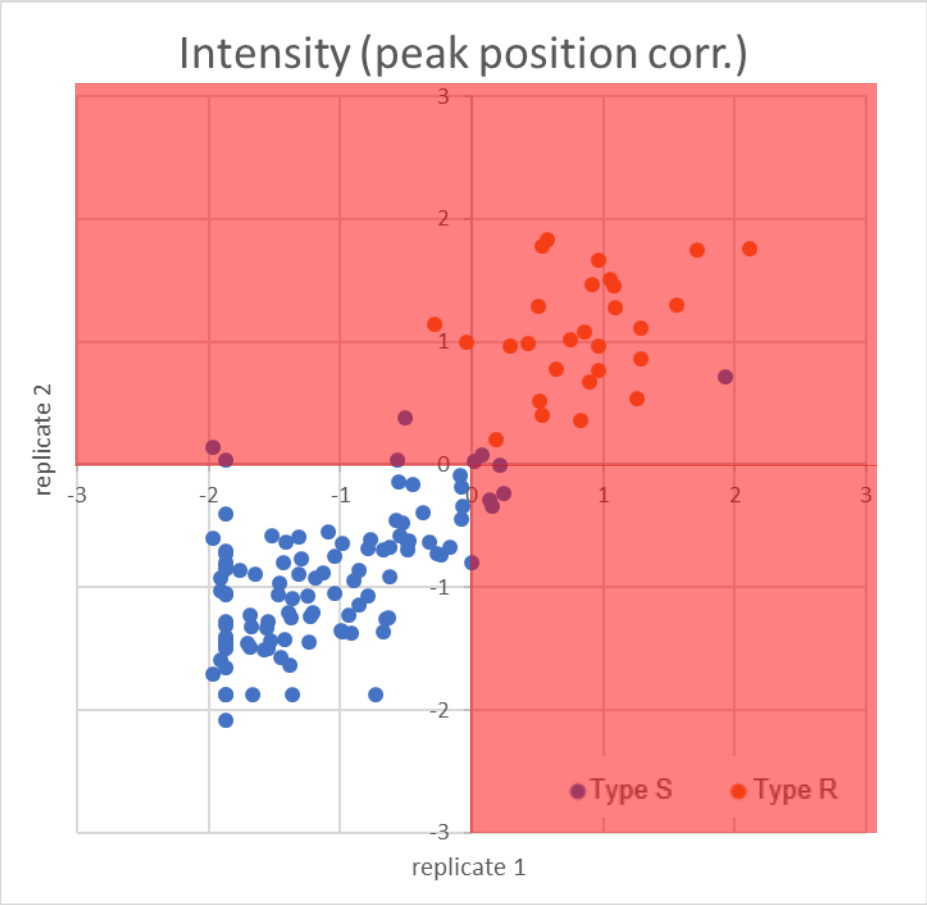
Performance characterization on the basis of classification scores



A cut-off level of 0 ensures that nearly all Type R spectra are identified.

Performance characterization on the basis of classification scores

The false positive rate is quite low:



Performance assessment on the basis of classification scores

A performance criterion for the classification SD

- A criterion for the classification SD ensuring acceptable false positive *and* false negative rates can be formulated as follows:

$$\sigma_{classification,\,Type\,S} + \sigma_{classification,\,Type\,R} \leq 1$$

- If this criterion is met, then it will always be possible to specify a cut-off such that both false positive and false negative rates are less than 5 %.

- Consider the case that both classification SD values are 0.5.

  Assuming a normal distribution, we then have:
  - 95 % of Type S classification scores will lie below $-1 + 1.64 \cdot 0.5 = -0.18$
  - 95 % of Type R classification scores will lie above $1 - 1.64 \cdot 0.5 = 0.18$

Population heterogeneity

- If none of the isolates from subpopulation 2 are represented in the validation study

  → the FPR will be much larger than the value calculated in the validation study.

- This may constitute an unacceptable risk.

- It must be emphasized that this risk depends on:
  - The numbers of isolates for each class
  - The representativeness of the isolates included in the validation study

Conclusions

- Even though methods such as MALDI-TOF/AI are qualitative, it is usually possible to base the characterization and assessment of method performance on normally distributed classification scores.

- Doing so allows a more reliable characterization of method performance. For instance, the uncertainty in the estimate of FPR can be quite large if the evaluation is based on the qualitative outcomes.

- In particular, the characterization of method performance can be conducted in terms of precision parameters which – upon prior standardization of the classification scores – are easily compared and interpreted. It is thus possible to identify the main sources of random error (intermediate, repeatability, etc.).

- A criterion for the assessment of method performance was formulated in terms of the total precision (classification SD) estimates.

Thanks for your attention!

*QUALITY & STATISTICS!

uhlig@quodata.de

Data Intelligence GmbH

carsten.uhlig@akees.com

Federal Office of Consumer Protection and Food Safety

manfred.stoyke@bvl.bund.de