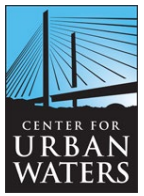
Approaches for Assessing Performance of HRMS-based NTA Methods

Christine M. Fisher (O'Donnell)[1], **Katherine T. Peter[2]**, Seth R. Newton[3], Andrew J. Schaub[4], and Jon R. Sobus[3]

[1]FDA, [2]UW Tacoma (current); NIST (former), [3]EPA, [4]SwRI

May 23, 2022

SETAC NTA FTM



Acknowledgements

BP4NTA:

Co-chairs 2020-21
Ben Place *(NIST)*
Elin Ulrich *(EPA)*

Co-chairs 2021-22
Christine Fisher (O'Donnell) *(FDA)*
Ruth Marfil-Vega *(Shimadzu)*

Clearance Reviewers:

Benjamin Place *(NIST)*
James McCord *(EPA)*
Ann Knolhoff *(FDA)*

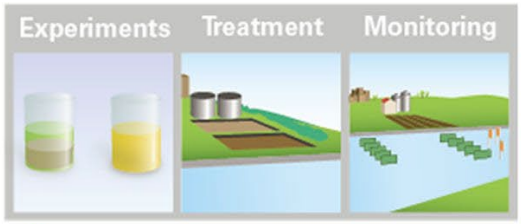
Additional members available at:
www.nontargetedanalysis.org

(as well as many other great NTA resources!)

Why NTA performance assessment?

Targeted analytical methods	Performance assessment well established (accreditation!)	Familiar terms and metrics to quantify uncertainty – precision, sensitivity, LOD, selectivity…	Readily interpreted, quantitative "final" output
Non-targeted analytical methods	Lab-by-lab approaches to QA/QC & performance metrics	Lots of uncertainty without clear descriptors [e.g., why was a compound not detected?]	Results often considered "non-final" – funneling toward targeted methods

Performance assessment vs. QA/QC



Commercial / Open Software

Databases

In Silico Prediction

Sampling
Space & Time
Before & After
Hot spots

Analysis
(Sample enrichment)
Chromatography (GC, LC)
Appropriate Ionization
HR-tandem MS

Data Pre-processing
Detect Peaks
Subtract Blind
Replicates & Grouping
Build Profiles, Series

Prioritization
Isotopes, Mass Defect
Time/ Spatial Trends
PCA/Clustering
Processes, Effects

Identification
Database, Suspect Search
Assign Formula, Fragments
Analytical (RT, Ionization,..) &
Meta Information

Field blanks, method blanks

Spiked compound RTs, mass error, relative abundance?

Spiked compounds are "peak picked"?

Known/expected compounds retained with filters used?

Correct formulas for spiked compounds?

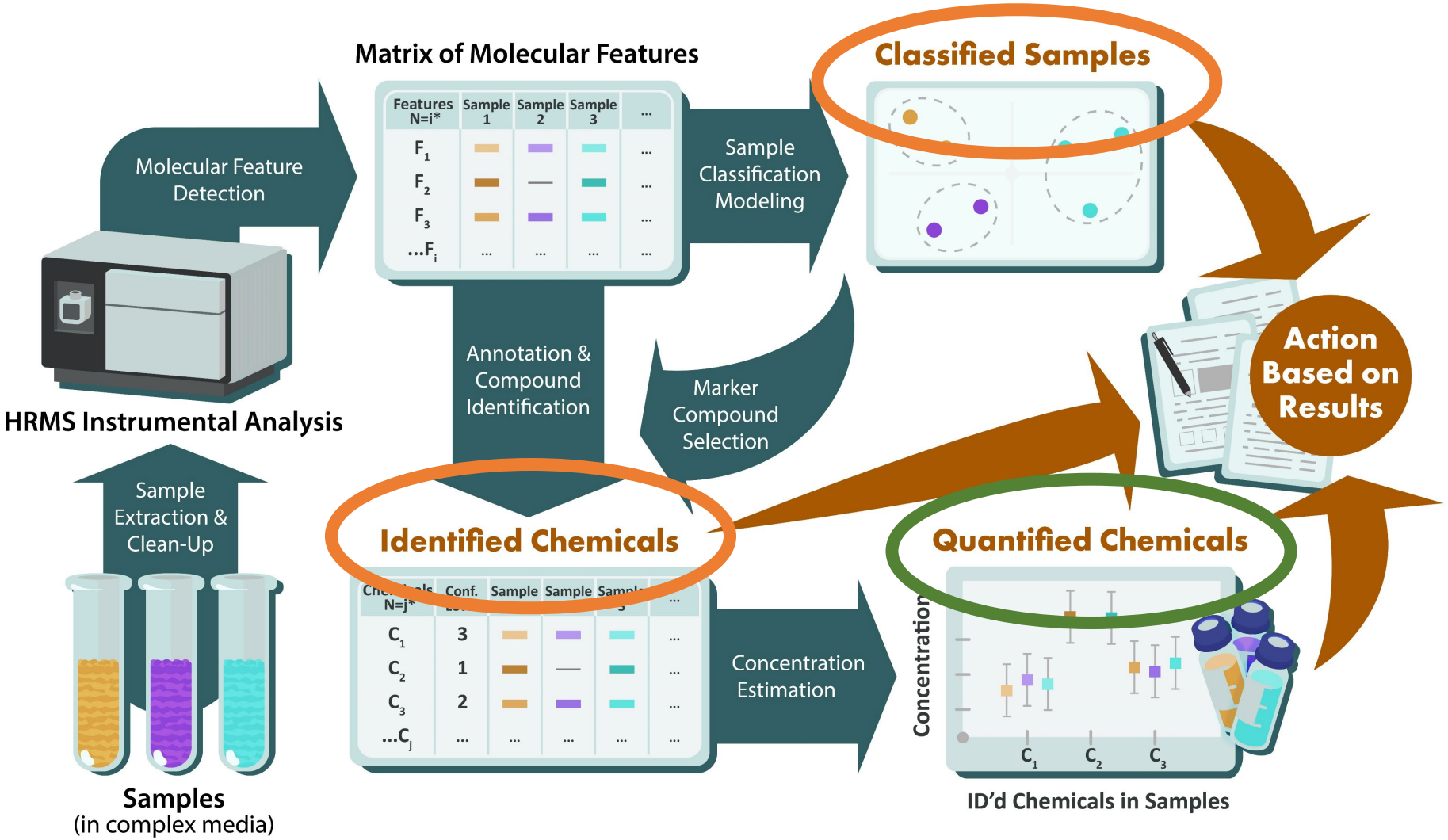
QA/QC happens throughout the NTA workflow…

Performance assessment happens at the END of the workflow!

(Can always do a "look back" to figure out root cause of poor performance)

Hollender et al., *ES&T,* 2017 (with modifications)

Let's focus on 3 NTA study objectives



Qualitative outputs:
the confusion matrix (with caveats…)

Quantitative outputs:
echo targeted analysis approaches?

Fisher et al. Submitted, 2022

Setting the stage: Confusion matrix 101

Reported Condition: Honey Classification:	Real Condition: Honey Classification: Adulterated	Real Condition: Honey Classification: Authentic
Reported Adulterated	True Positive (TP)	False Positive (FP) Type I Error
	TP = 10	FP = 8
Reported Authentic	False Negative (FN) Type II Error	True Negative (TN)
	FN = 2	TN = 30

N = 50 honeys (12 adulterated, 38 authentic)

Confusion matrices are routinely applied to assess performance of tests with discrete, often binary, qualitative outputs

Requirements:

- Define positive and negative conditions (positive = rarer)
- A discrete “boundary”
- Sufficient statistical power (# of samples)

Fisher et al. Submitted, 2022

Setting the stage: Confusion matrix 101

Reported Condition: Honey Classification:	Real Condition: Honey Classification: Adulterated	Real Condition: Honey Classification: Authentic		
Reported Adulterated	True Positive (TP) TP = 10	False Positive (FP) Type I Error FP = 8	Precision $\frac{TP}{TP + FP}$ Precision = 0.56	False Discovery Rate (FDR) $\frac{FP}{TP + FP}$ FDR = 0.44
Reported Authentic	False Negative (FN) Type II Error FN = 2	True Negative (TN) TN = 30		
	True Positive Rate (TPR; Recall, Sensitivity) $\frac{TP}{TP + FN}$ TPR = 0.83	False Positive Rate (FPR; Fall-out Rate) $\frac{FP}{FP + TN}$ FPR = 0.21	F_1 Score $2 \times \frac{Precision \times TPR}{Precision + TPR}$ F_1 = 0.67	Accuracy $\frac{TP + TN}{TP + FP + FN + TN}$ Accuracy = 0.80
	False Negative Rate (FNR; Miss Rate) $\frac{FN}{TP + FN}$ FNR = 0.17	True Negative Rate (TNR; Specificity, Selectivity) $\frac{TN}{FP + TN}$ TNR = 0.79	Matthew's Correlation Coefficient (MCC) $\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$ MCC = 0.55	

Fisher et al. Submitted, 2022

LOTS of associated metrics – can be viewed as pairs/groups used interchangeably or in tandem

<u>Note of caution:</u> There is overlapping terminology for targeted analytical methods vs. the confusion matrix!

Assessing sample classification performance with the confusion matrix

Good news: Given a test set of samples, there's a clearly defined CM boundary!

Developing robust sample classification models is challenging – introduces caveats to consider during performance assessment

- Is the test/training set representative? Large enough N? Well-matched to potential variability? Balanced positive/negative condition?
- Risk of overfitting (1000's of HRMS features, 10's of samples)
- Are feature lists reproducible? Changes over time/across instruments impact model outputs…

https://www.fillmorecontainer.com/blog/2015/02/19/can-i-re-use-my-baby-food-jars/

Using the confusion matrix for chemical identification performance is trickier…

Need a discrete number of considered chemicals to define the boundary of the confusion matrix

Boundary n = X		Chemical is… spiked into sample	Chemical is… not spiked into sample
Chemical is…	reported in sample	TP	FP
Chemical is…	not reported in sample	FN	TN

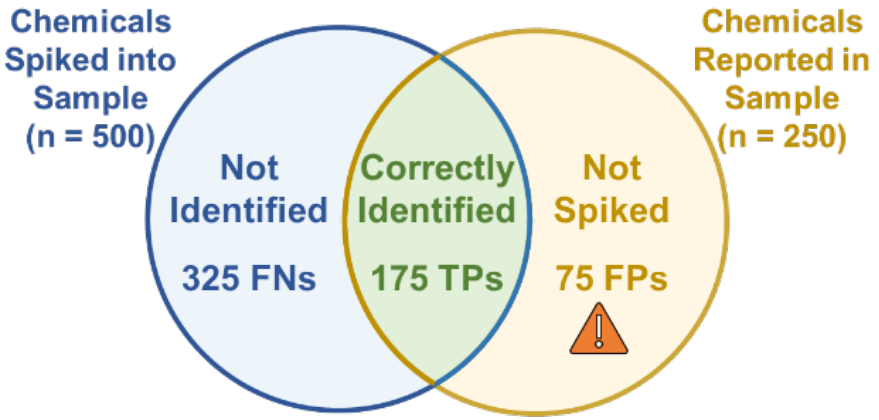
The "chemical universe" is too large…TNs become infinite!

We propose two options for discrete boundaries:

Chemicals known to be present and/or reported in a sample

A suspect screening database

Boundary 1: Chemicals known to be present/reported in a test sample



- Relies on a spiked test sample
- No TNs! Can't define # chemicals NOT in the sample
- Smaller suite of performance metrics…

Boundary n = 575		Chemical is… spiked into sample	Chemical is… not spiked into sample		
Chemical is…	reported in sample	TP 175	FP 75	Precision 0.70	FDR 0.30
	not reported in sample	FN 325			
		TPR 0.35		F_1 0.47	
		FNR 0.65			

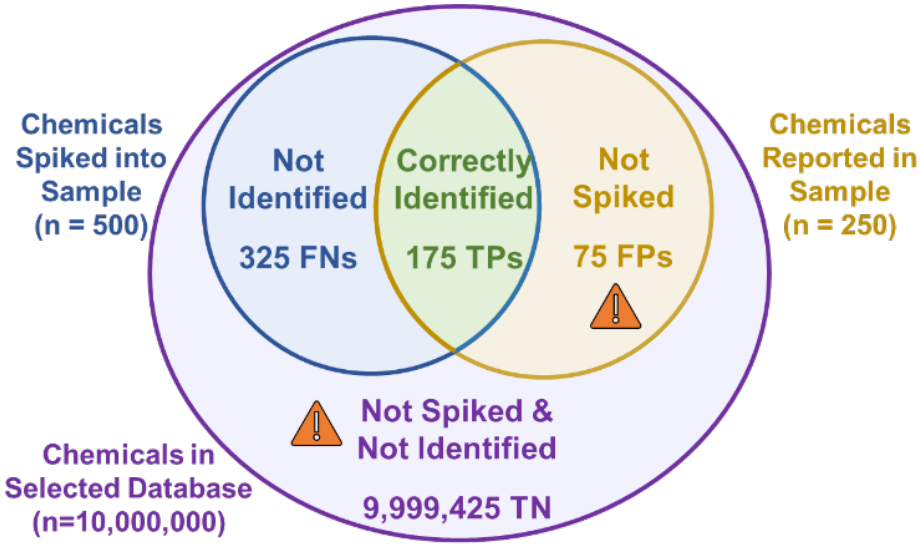
Provides balance to assessment, because FPs are considered

Bounded by known chemicals, but no penalty for over-reporting…

Useful when FPs and FNs of equal importance

Fisher et al. Submitted, 2022

Boundary 2: A suspect screening database



Boundary n = 10,000,000		Chemical is… spiked into sample	Chemical is… not spiked into sample		
Chemical is…	reported in sample	TP 175	FP 75	Precision 0.70	FDR 0.30
	not reported in sample	FN 325	TN 9,999,425		
		TPR 0.35	FPR 0.00001	F_1 0.47	Accuracy 0.99996
		FNR 0.65	TNR 0.99999	MCC 0.49	

- Relies on a spiked test sample
- Can count TNs! But watch out for huge databases – lots of TNs might not be method-amenable!
- Watch out for metrics that use TN! And note that using F1 Score and MCC alongside Accuracy provides better indication of method performance

Fisher et al. Submitted, 2022

More to consider with ID performance…

- **What's actually in the sample??**

  Rely on initial characterization of the test sample to define TPs.
- **What about impurities?? FPs or "unintentional TPs" (uTPs)?**

  Treat them all as FPs to be conservative.
- **How to address poor performance?**

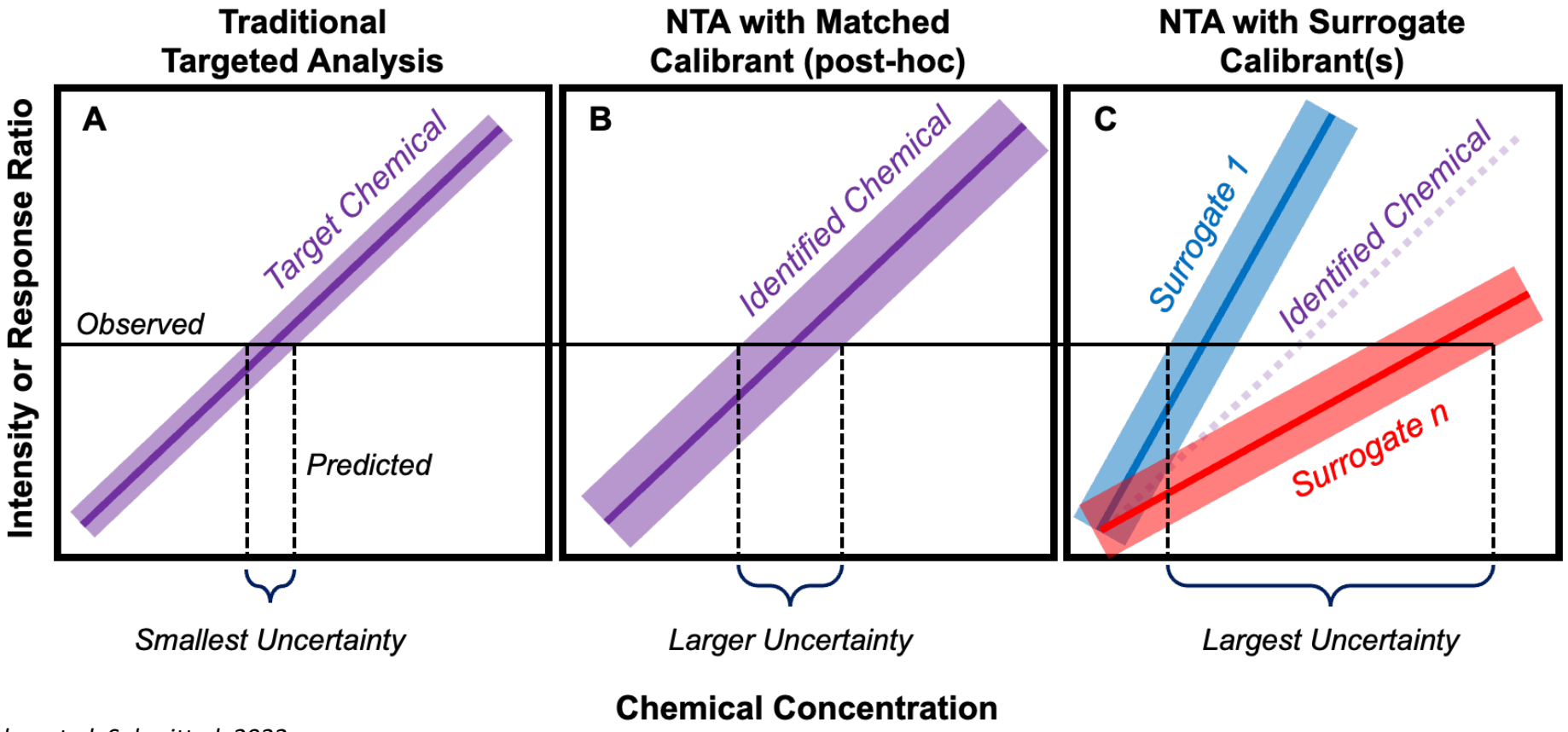
  Start backtracking to evaluate individual workflow steps, use QA/QC to inform where things went wrong…
- **What about identification confidence?**

  Enumerate # of TPs/FPs at each confidence level. Doesn't change metrics, but gives nuance to their interpretation.
- **What if there are two potential structures for a given feature?**

  Pick one to report – necessary for confusion matrix. Much room for development….

Defining Quantitative NTA (qNTA)

Generating concentration estimates in the absence of reference standards for the compounds of interest



Fisher et al. Submitted, 2022

The key to assessing qNTA performance: evaluating uncertainty

Easy, but limited predictive value:

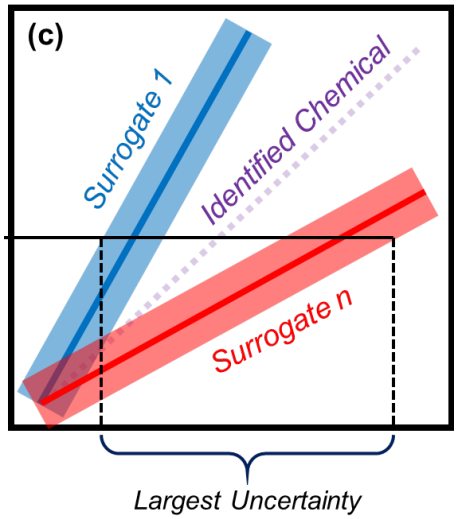
Calculate error in known test samples, apply to all subsequent results:

$$\frac{[\text{Estimated}]}{[\text{Known}]}$$

Harder, but has more predictive value:

- Develop a distribution of response factors for surrogate "calibrants"
- Incorporate additional info (e.g., predicted ionization efficiency) to narrow which surrogates are used

[Groff et al. In review, 2022]



Many remaining challenges, e.g.:

- Accounting for sample recovery & matrix effects?
- Transferability across instruments?

Where do we go from here?

- Current BP4NTA goal is to improve awareness of needs and challenges in NTA performance assessment – not an end-point, these are not hard & fast recommendations

- What is “good” performance? No one-size-fits-all answer... Different projects and outputs will merit different approaches and stringencies

- Community discussion and further development needed
  - Standardized sample/chemical test sets? Or guidelines?
  - Evaluate transferability of sample classification models
  - Guidance for performance assessment across ID confidence levels
  - Methods to bound qNTA predictions

Thank you!

Kathy Peter
ktpeter@uw.edu

Other BP4NTA talks/posters to check out:
Ann Knolhoff – *QC chemical mixture*
Ruth Marfil-Vega – *Harmonization efforts*
Gabby Black – *Chemical space coverage*
Jon Sobus – *SRT poster*

Interested in joining BP4NTA or receiving our mailing list?

Head to our website: www.nontargetedanalysis.org
Or reach out directly to leadership:
Christine Fisher (Christine.ODonnell@fda.hhs.gov)
Ruth Marfil-Vega (rmmarfilvega@shimadzu.com)

BP4NTA
BENCHMARKING AND PUBLICATIONS FOR NON-TARGETED ANALYSIS