Approaches for Assessing Performance of High-Resolution Mass Spectrometry-Based Non-Targeted Analysis Methods

Christine M. Fisher (O'Donnell), Katherine T. Peter, Seth R. Newton, Andrew J. Schaub, and Jon R. Sobus

Manuscript will be submitted to Analytical and Bioanalytical Chemistry

(currently in authors' internal reviews)

Introduction

Goals:

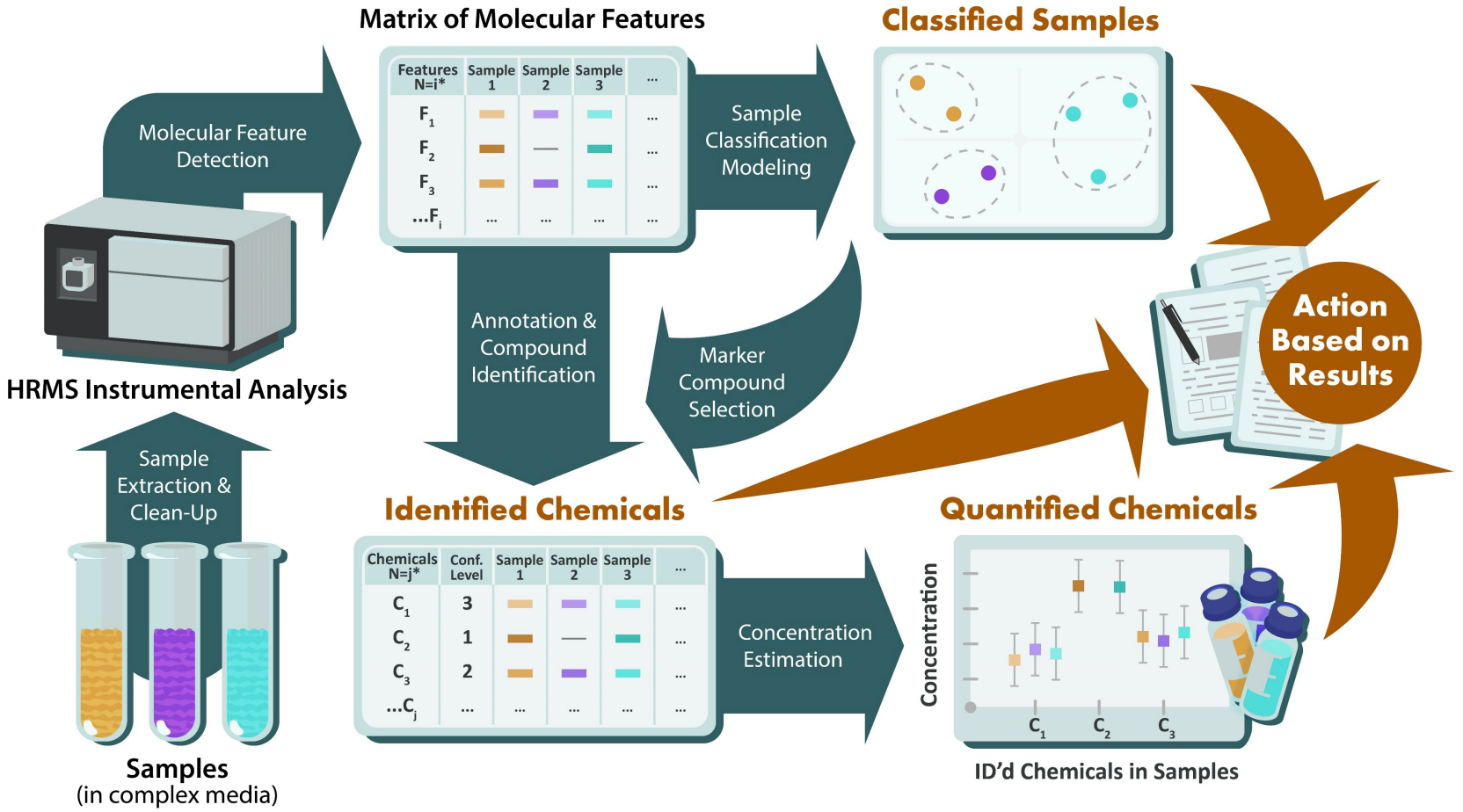
- Assess currently available approaches for performance assessment
- Stimulate discussion in the community
- Supports efforts to improve communication of NTA results to stakeholders

3 Main Study Objectives/Results:

1. Classified Samples
2. Identified Chemicals
3. Quantified Chemicals

Scope:

- Uses:
  - Compare methods
  - Toward Lab accreditation
- Although important and will impact performance metrics, data quality metrics outside of scope



Sample Classification Performance Evaluation

- Confusion Matrix
- Example: Classifying Adulterated vs. Authentic Honeys
  - 50 total samples (Boundary of confusion matrix)
  - 12 adulterated
  - 38 authentic

		Real Condition	
		Honey Classification:	
		Adulterated	Authentic
Reported Condition — Honey Classification:	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

https://honey.com/

- True Positives: 10 samples correctly reported as adulterated
- False Positives: 8 samples incorrectly reported as adulterated
- False Negatives: 2 samples incorrectly reported as authentic
- True Negatives: 30 samples correctly reported as adulterated

Sample Classification Performance Evaluation using the Confusion Matrix
Example = Adulterated vs. Authentic Honey samples

		Real Condition: Honey Classification: Adulterated	Real Condition: Honey Classification: Authentic		
Reported Condition: Honey Classification:	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	

Note: Similar Performance Metric Terminology for Targeted vs. NTA

Sample Classification Performance Evaluation: Challenges

- Domain of Applicability
- Biased Datasets:
  - 12 adulterated vs. 38 authentic

F_1 Score	Accuracy
$2 \times \frac{\text{Precision} \times \text{TPR}}{\text{Precision} + \text{TPR}}$	$\frac{TP + TN}{TP + FP + FN + TN}$
$F_1 = 0.67$	Accuracy = 0.80

Matthew's Correlation Coefficient (MCC)

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC = 0.55

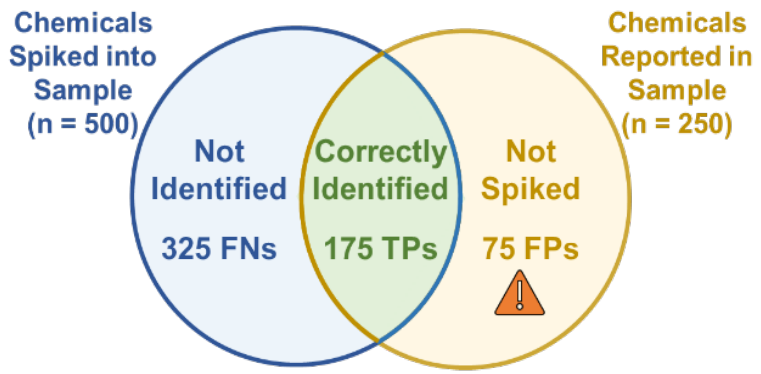
- Comparing models between instruments/labs
- Robust models overtime (same instrument)
- Risk of overfitting:
  - #variables >> #samples

Chemical Identification Performance Evaluation: Example 1

- Confusion Matrix:
  - Boundary: Unique chemicals known and/or reported to be present in a sample (n = 575)
  - TP: 175 spiked chemicals reported as present
  - FP: 75 reported chemicals that were not spiked
  - FN: 325 spiked chemicals that were not reported
  - TN: not defined

- Challenges/Considerations:
  - Some FPs may be unintentional TPs (uTPs)
  - No TNs
  - Precision/FDR = provide “penalty” for over-reporting
  - F1-Score: useful “overall” metric if minimizing both FNs and FPs is of equal importance
  - Domain of applicability

A.

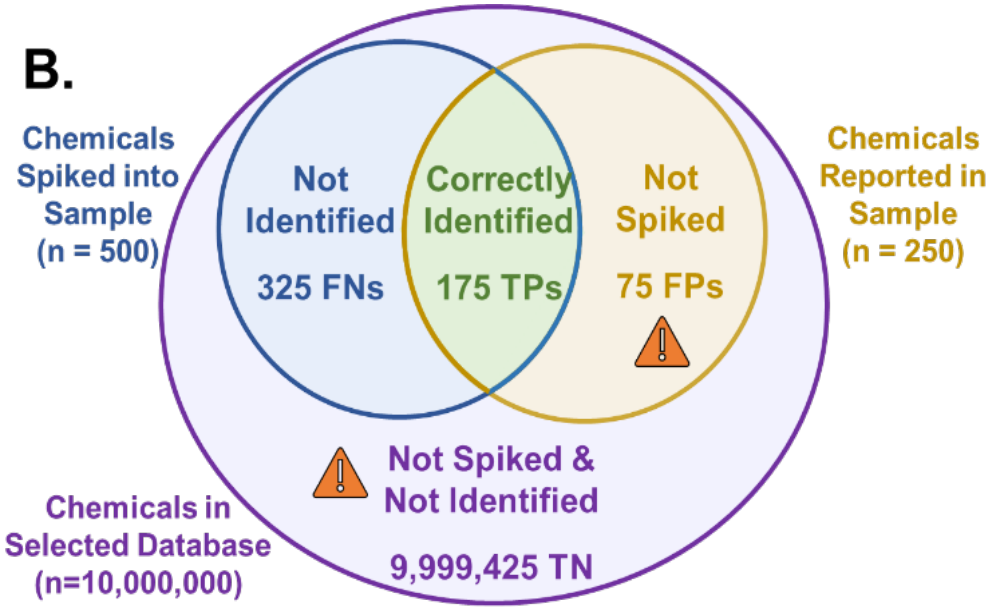


Boundary n = 575		Chemical is… spiked into sample	Chemical is… not spiked into sample		
Chemical is…	reported in sample	TP 175	FP 75	Precision 0.70	FDR 0.30
	not reported in sample	FN 325			
		TPR 0.35		F_1 0.47	
		FNR 0.65			

Chemical Identification Performance Evaluation: Example 2

- Confusion Matrix:
  - Boundary: Suspect screening database (n = 10,000,000)
  - TP: 175 spiked chemicals reported as present
  - FP: 75 reported chemicals that were not spiked
  - FN: 325 spiked chemicals that were not reported
  - TN: 9,999,425 database chemicals that were not spiked nor reported

- Challenges/Considerations:
  - Some FPs may be unintentional TPs (uTPs)
  - All spiked/reported chemicals must be present in the suspect screening database
  - TNs and TN-derived metrics:
    - Impacted by database size (can bias metrics)
    - Should database chemicals not detectable/ identifiable by method count as TNs?
  - Domain of applicability

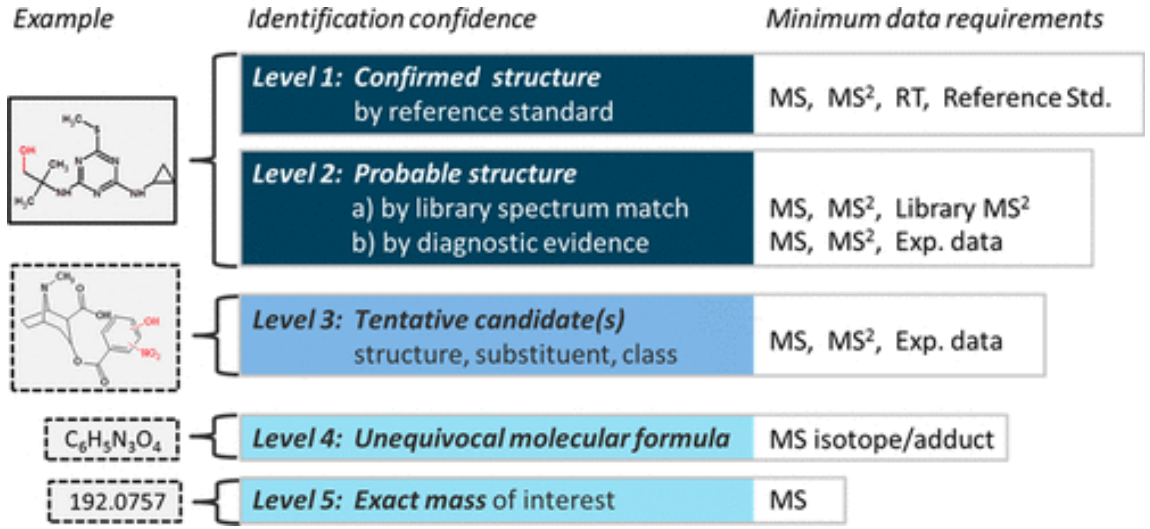


Boundary n = 10,000,000		Chemical is... spiked into sample	Chemical is... not spiked into sample		
Chemical is...	reported in sample	TP 175	FP 75	Precision 0.70	FDR 0.30
	not reported in sample	FN 325	TN 9,999,425		
		TPR 0.35	FPR 0.00001	F_1 0.47	Accuracy 0.99996
		FNR 0.65	TNR 0.99999	MCC 0.49	

Literature example: Nunez et al. J Chem Inf Model. 2019; 59(9); 4052-60

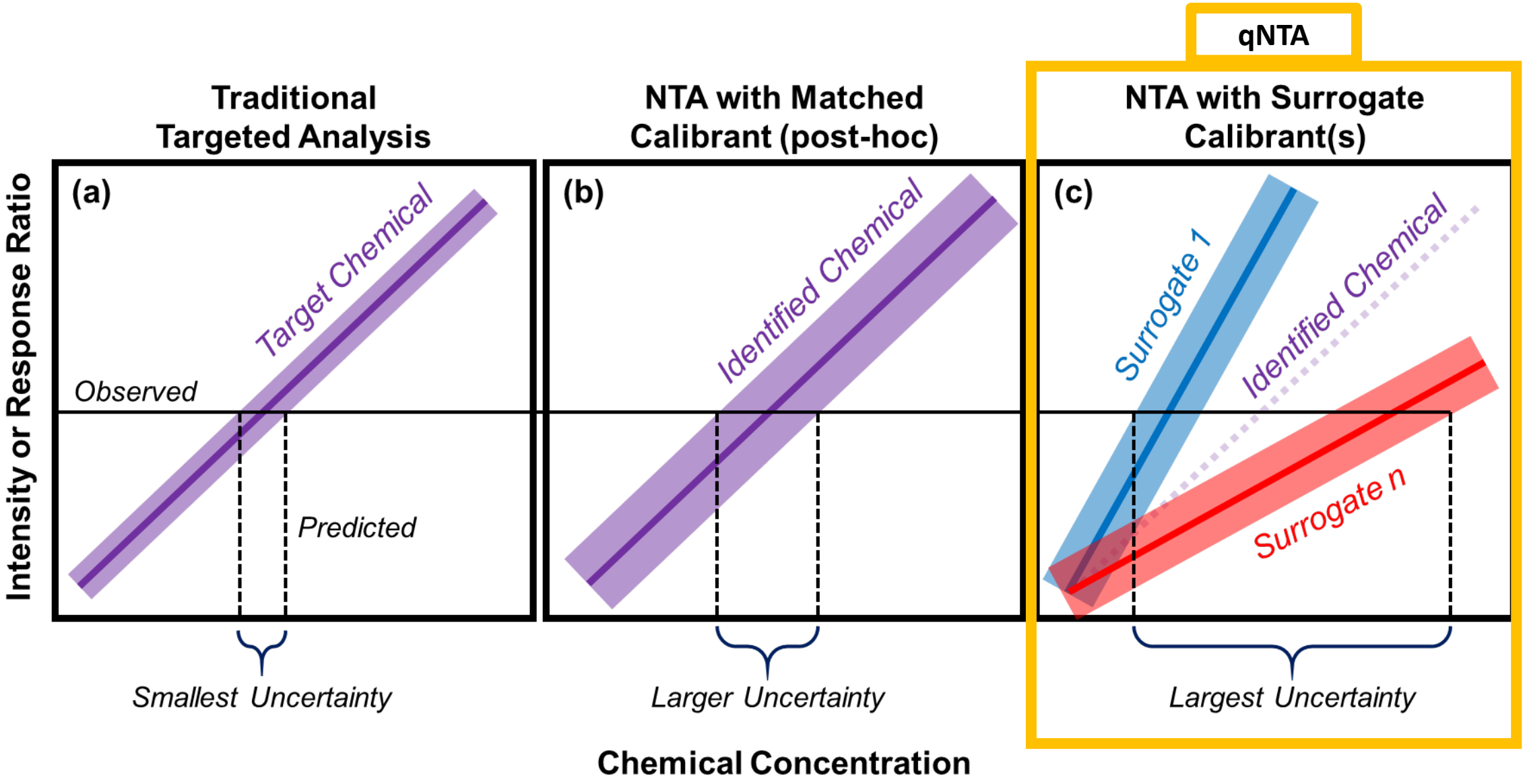
Chemical Identification Performance Evaluation: Incorporating Confidence Level

- Separate confusion matrices cannot be built for each confidence level: # of chemicals not reported are not associated with any confidence level
- Should report the proportion of TPs and FPs at each confidence level
- Can report separate precision/FDR for individual confidence levels
  - Literature example: Nunez et al. J Chem Inf Model. 2019; 59(9); 4052-60
- For performance assessment as we describe: can only consider one reported chemical per feature



Schymanski et al. Environ Sci Technol. 2014; 48(4); 2097-98

Chemical Quantitation: Targeted vs. NTA

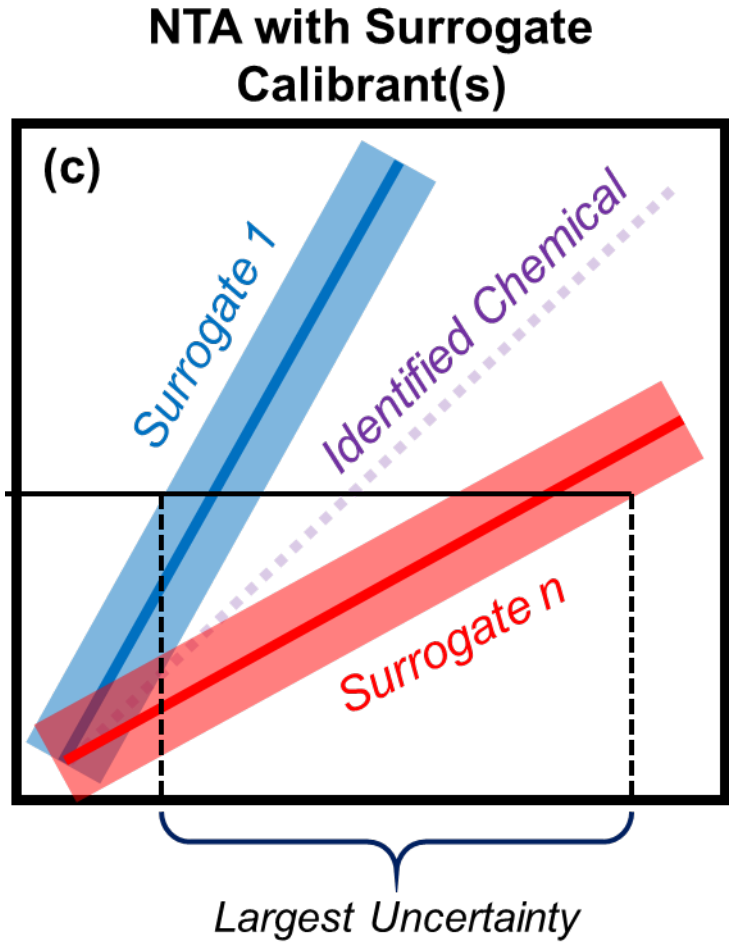


Chemical Quantitation: qNTA Approaches and Performance Evaluation Considerations

- Approaches to determine response factor (RF) for identified chemical:
  - Single surrogate analyte
  - Optimized surrogate analyte based on identification
  - Model-based (structural descriptors used to predict ionization efficiency; see Kruve et al.)

- Performance Evaluation:
  - Common:
    - Error for each QC spike: $\frac{Estimated\ Concentration}{Known\ Concentration}$
    - Mean absolute error, maximum observed error, R^2, or Q^2
  - Confidence intervals should be included!
    - Groff et al. publication in process

- These estimates only apply to concentration in prepared sample extract!
  - Currently no models for predicting matrix effects/extraction efficiencies
  - Real-world concentration estimates remain largely unconstrained



Conclusions

- Sample Classification Performance:
  - Can use confusion matrix
  - Challenging to develop robust/reproducible models over time/across instruments; biased data sets

- Chemical Identification Performance:
  - Can use confusion matrix
  - Challenging to bound confusion matrix
  - Metrics should be interpreted with caution

- Chemical Quantitation Performance:
  - Important to bound estimates with confidence intervals
  - Additional efforts needed to estimate concentration in real-world samples