Interindividual Variability Assessment Through Application of Machine Learning with In Vitro Molecular Profiles to Understand Key Mechanisms of Emerging Inhaled Toxicants

Elise Hickman
Postdoctoral Fellow, Rager Lab
The University of North Carolina at Chapel Hill
Human health risks are known to vary across and within populations.

Current questions/challenges in risk assessment include:

1. How can we improve assessment of human interindividual variability?

2. How can we improving linkages between exposures that include multiple stressors and disease outcomes across the full range of human responses?

3. How can we determine uncertainty factors that are applicable to specific endpoints and exposures and that capture interindividual variability?
Technological advances have made measuring molecular signatures in experimental samples more feasible and affordable.

**Pros:**
- Increased accessibility of measuring a wide range of molecular signatures
- Opportunity for broader investigation of the effects of toxicants
- Higher sensitivity in capturing molecular signatures
- Ability to obtain more data from a single sample

**Challenges:**
- Sufficiently powering studies
- Distilling meaningful biological conclusions AND communicating them clearly
- Data science training
Outline of Presentation

1. Share examples of recent efforts leveraging supervised and unsupervised machine learning to understand key biological mechanisms of inhaled toxicants in human clinical studies.

2. Highlight a study leveraging an organotypic in vitro co-culture model of the respiratory system to understand variables underlying interindividual variability in response to acrolein.

3. Discuss major takeaways, upcoming data science training efforts, and future studies.
1. Share examples of recent efforts leveraging supervised and unsupervised machine learning to understand key biological mechanisms of inhaled toxicants in human clinical studies.

2. Highlight a study leveraging an organotypic *in vitro* co-culture model of the respiratory system to understand variables underlying interindividual variability in response to acrolein.

3. Discuss major takeaways, upcoming data science training efforts, and future studies.
Example Studies

1. Are there overall differences in human respiratory protein profiles in users of different types of e-cigarette devices?

2. Are human respiratory protein profiles in e-cigarette users similar to those found in people with chronic obstructive pulmonary disease (COPD)?
What are e-cigarettes?

E-cigarettes heat and aerosolize an e-liquid, allowing users to inhale nicotine and other chemicals.

E-cigarettes were originally touted as a “safer” alternative to cigarettes but are used by both former cigarette smokers and nonsmokers.

E-liquids typically contain:

• Nicotine or Nicotine Salts, 0-7% (0-70 mg/mL)
• Flavoring Chemicals
• Propylene Glycol (throat hit)
• Vegetable Glycerin (sweetness, cloud)
E-Cigarette Device Evolution

Constant evolution of e-cigarette devices is a major challenge in the field of e-cigarette toxicology, particularly with popular devices such as JUUL and disposables.

- **Previous (3rd) Generation E-Cigs**
  - Freebase Nicotine (pH ~8)
  - User control over settings, can be high power

- **Pod and Disposable E-Cigs (4th Gen)**
  - Nicotine Salts (pH ~4-6)
  - Fewer settings, lower power

What biomarkers are altered in 4th generation e-cigarette users?

- Free-base Nicotine
- Nicotine Salt
- E.g. lactic, benzoic, and levulinic acids

User control over settings, can be high power

Puffing patterns? User population?
Demographic Summary:
• n = 21-28 participants per group
• 4th generation e-cigarette users were significantly younger
• Each group had a mixture of male and female participants, but ratio was not always even
Soluble Mediator Expression is Significantly Decreased in 4th Generation E-Cig Users

n = 12 mediators significant for exposure group variable

Hickman et al. 2022 American Journal of Respiratory and Critical Care Medicine

Soluble mediators that were significantly different between exposure groups after adjusting for age, sex, and race differences between exposure groups. Results are presented as mean ± standard error of log2 transformed mediator concentrations. * p < 0.05, ** p < 0.01, *** p < 0.001 using ANCOVA followed by Dunnett’s (comparisons with NS/NV) and Tukey’s (3rd v. 4th Gen) post-hoc tests. NS/NV = non-smoker/non-vaper, SM = smoker.
Machine learning demonstrates best separation for 4th generation e-cigarette users

Variable Selection
(Best Subsets Regression)

IL10  MIP1a  MMP9  Eotaxin3  IL6  TARC  VEGFD  MPO  Tie2

4th Gen  3rd Gen  NS/NV  SM

Machine Learning Model Performance

Variable Selection

- IL10
- MIP1a
- MMP9
- Eotaxin3
- IL6
- TARC
- VEGFD
- MPO
- Tie2

4th Gen  3rd Gen  NS/NV  SM

Hickman et al. 2022 American Journal of Respiratory and Critical Care Medicine
Conclusions

Suggestive of dysregulated immune homeostasis in the form of overall immune suppression in 4th generation e-cigarette users, which could result in impaired response to infection or vaccination.

*Observed notable interindividual variability between participants.*
Example Studies

1. Are there overall differences in human respiratory protein profiles in users of different types of e-cigarette devices?

2. Are human respiratory protein profiles in e-cigarette users similar to those found in people with chronic obstructive pulmonary disease (COPD)?
Example Studies

1. Are there overall differences in human respiratory protein profiles in users of different types of e-cigarette devices?

2. Are human respiratory protein profiles in e-cigarette users similar to those found in people with chronic obstructive pulmonary disease (COPD)?
Background on COPD

- Chronic obstructive pulmonary disease (COPD) is a highly prevalent, progressive condition marked by an altered airway inflammatory and immune milieu that encompasses emphysema and chronic bronchitis.

- In industrialized nations, cigarette smoking is the primary risk factor for COPD, and smoking is estimated to account for 8 in 10 COPD deaths.

- E-cig use has been associated with chronic bronchitis, increased airway proteases, inflammation, and altered immune markers in sputum, which are also found in COPD.

Do e-cig users have sputum soluble mediator profiles that resemble specific stages of COPD?
GOLD Stage | Description | Lung Function
---|---|---
0 | Pre-COPD; individuals with respiratory and/or structural or physiological abnormalities without airflow obstruction | ![Lung Function](image1)
1/2 | Mild COPD (FEV1 ≥ 80% predicted), Moderate COPD (50% ≤ FEV1 < 80% predicted) | ![Lung Function](image2)
3 | Severe COPD (30% ≤ FEV1 < 50%) | ![Lung Function](image3)
Study Design

SPIROMICS: SubPopulations and InteRmediate Outcome Measures In COPD Study

Demographic Summary:
- n = 25-29 participants per group
- Balanced male/female in each group
- Balanced current smokers vs. non-smokers in each group
- Older on average than e-cig study cohort

Dr. Neil Alexis
Dr. Julia Rager
Similarity in Soluble Mediator Profiles Between Groups: Hierarchical Clustering

All Mediators Together (Group Averages)

All Mediators Together (Individual Participants)
Similarity in Soluble Mediator Profiles Between Groups: Hierarchical Clustering

Hierarchical Clustering

A. Inflammatory Markers
B. Chemotactic Mediators
C. Proteases & Enzymes

“Semi-supervised machine learning”

Hickman et al. manuscript submitted to Am J Resp Crit Care Med
Similarity in Soluble Mediator Profiles Between Groups: Mahalanobis Distance

Mahalanobis distance is calculated between the multivariate mean and the datapoints after rescaling (using eigenvectors and eigenvalues) to remove covariance.

Distance metrics such as Mahalanobis and Jaccard can serve as complementary approaches to machine learning.

A

<table>
<thead>
<tr>
<th></th>
<th>All Mediators</th>
<th>Inflammatory</th>
<th>Chemotactic</th>
<th>Proteases/Enzymes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>0.039</td>
<td>0.103</td>
<td>0.062</td>
<td>0.02</td>
</tr>
<tr>
<td>4th Gen EC</td>
<td>0.069</td>
<td>0.155</td>
<td>0.013</td>
<td>0.034</td>
</tr>
<tr>
<td>Pre-COPD</td>
<td>0.117</td>
<td>0.169</td>
<td>0.031</td>
<td>0.076</td>
</tr>
<tr>
<td>GOLD 1/2</td>
<td>0.032</td>
<td>0.183</td>
<td>0.031</td>
<td>0.007</td>
</tr>
<tr>
<td>GOLD 3</td>
<td>0.114</td>
<td>0.007</td>
<td>0.026</td>
<td>0.182</td>
</tr>
</tbody>
</table>

B

<table>
<thead>
<tr>
<th></th>
<th>All Mediators</th>
<th>Inflammatory</th>
<th>Chemotactic</th>
<th>Proteases/Enzymes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>0.13</td>
<td>0.044</td>
<td>0.004</td>
<td>0.195</td>
</tr>
<tr>
<td>4th Gen EC</td>
<td>0.069</td>
<td>0.155</td>
<td>0.013</td>
<td>0.034</td>
</tr>
<tr>
<td>Pre-COPD</td>
<td>0.359</td>
<td>0.073</td>
<td>0.001</td>
<td>0.541</td>
</tr>
<tr>
<td>GOLD 1/2</td>
<td>0.154</td>
<td>0.082</td>
<td>0.001</td>
<td>0.21</td>
</tr>
<tr>
<td>GOLD 3</td>
<td>0.054</td>
<td>0.01</td>
<td>0</td>
<td>0.053</td>
</tr>
</tbody>
</table>
Conclusions

Taken together, our results demonstrate partial overlap between e-cig user and COPD soluble mediator profiles, warranting further investigation into the relationship between e-cigarette use and airway disease.

Continued to observe notable interindividual variability between participants.
Outline of Presentation

1. Share examples of recent efforts leveraging supervised and unsupervised machine learning to understand key biological mechanisms of inhaled toxicants in human clinical studies.

2. Highlight a study leveraging an organotypic *in vitro* co-culture model of the respiratory system to understand variables underlying interindividual variability in response to acrolein.

3. Discuss major takeaways, upcoming data science training efforts, and future studies.
• Acrolein is a ubiquitous volatile aldehyde that is emitted from the combustion of fossil fuels, tobacco, wood, and plastic.

• Exposure to acrolein is associated with irritation throughout the respiratory tract, pulmonary edema, and dysregulation of immune responses.

• Primary human bronchial epithelial cell + fibroblast co-cultures represent sophisticated organotypic in vitro models that can inform interindividual variability.

Moghe et al 2015 DOI: 10.1093/toxsci/kfu233
Cell Culture Model & Exposure

**Primary HBEC Differentiation**

Primary HBEC Culture

24 days

**Co-Culture**

n = 14 donor-matched co-cultures

48 hours

**Acrolein Exposure**

Groups: Incubator Control; 0, 0.6, 1, 2, and 4 ppm

2 hours

**Endpoint Collection**

Apical Wash Cytokines/Growth Factors (IL-10, IL-1β, IL-6, IL-8, TNF-α)

Phenotypic Endpoints (Barrier integrity, ciliary function, mucin production)

Basolateral Media Cytokines/Growth Factors (IL-10, IL-1β, IL-6, IL-8, TNF-α, VEGF)
Initial Observations

1. Significant interindividual variability between physical characteristics of pHBEC cultures.

![H&E and Alcain Blue PAS (Mucins) images]

2. Significant interindividual variability in responsivity of co-culture system to acrolein exposures.

3. Significant increase in cytokine/growth factor production alongside decreased barrier integrity with higher doses of acrolein.

Can we leverage benchmark dose-response modeling and machine learning to assess interindividual variability in response to acrolein?
Benchmark dose-response modeling is an established tool to inform human health risk calculations that can leverage both phenotypic and molecular-level response signatures.
BMDs Were Lower and More Variable When Analyzing Trends on a Per-Donor Basis
Benchmark Doses Vary by Donor and Cluster by Sex for Cytokines and Secreted Growth Factors

Cytokines & Secreted Growth Factors

<table>
<thead>
<tr>
<th>Sex</th>
<th>Donor</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>23R</td>
<td>80</td>
</tr>
<tr>
<td>F</td>
<td>48M</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>70M</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>60P</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>03N</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>26M</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>50Q</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>44P</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>23P</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>58P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40I</td>
<td></td>
</tr>
<tr>
<td></td>
<td>46L</td>
<td></td>
</tr>
</tbody>
</table>

BMD (ppm)

20  30  40  50  60  70  80

Sex

M
F

Age

0  1  2  3  4  5  6  7

IL1B_AW
TNF_AW
VEGF_BLM
IL1B_BLM
IL8_AW
IL6_AW
TNF_BLM
IL10_AW
IL10_BLM
IL6_BLM
IL8_BLM

Phenotypic Endpoints

<table>
<thead>
<tr>
<th>Sex</th>
<th>Donor</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>23R</td>
<td>80</td>
</tr>
<tr>
<td>F</td>
<td>48M</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>70M</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>60P</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>03N</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>26M</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>50Q</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>44P</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>23P</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>58P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40I</td>
<td></td>
</tr>
<tr>
<td></td>
<td>46L</td>
<td></td>
</tr>
</tbody>
</table>

CBF_24hr
MemPerm
TEER_24hr
TEER_4hr
CBF_4hr
MUC5AC
Potential Sex-Based Differences in BMD Model Parameters Were Identified Using K-Means Clustering

**K-means Clusters**

![K-means Clusters Diagram]

*Input: Power curve model fit parameters for cytokine and growth factor data*

<table>
<thead>
<tr>
<th></th>
<th>Cluster 1 (N=3)</th>
<th>Cluster 2 (N=5)</th>
<th>Cluster 3 (N=6)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>0 (0%)</td>
<td>1 (20.0%)</td>
<td>4 (66.7%)</td>
<td>0.115</td>
</tr>
<tr>
<td>Male</td>
<td>3 (100%)</td>
<td>4 (80.0%)</td>
<td>2 (33.3%)</td>
<td></td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>26.0 (15.7)</td>
<td>41.0 (17.6)</td>
<td>28.4 (33.3)</td>
<td>0.652</td>
</tr>
<tr>
<td>Median [Min, Max]</td>
<td>19.0 [15.0, 44.0]</td>
<td>46.0 [13.0, 58.0]</td>
<td>13.5 [0.330, 91.0]</td>
<td></td>
</tr>
<tr>
<td><strong>BMD (Model Avg)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>2.90 (0.434)</td>
<td>4.87 (2.73)</td>
<td>3.09 (0.640)</td>
<td>0.203</td>
</tr>
</tbody>
</table>
This study is impactful because it is among the first to combine in vitro primary co-culture models with advanced computational modeling to expand human response variability assessments in new approach methods (NAMs)-based risk assessment.

**We detected factors underlying interindividual variability using machine learning.**
Outline of Presentation

1. Share examples of recent efforts leveraging supervised and unsupervised machine learning to understand key biological mechanisms of inhaled toxicants in human clinical studies.

2. Highlight a study leveraging an organotypic *in vitro* co-culture model of the respiratory system to understand variables underlying interindividual variability in response to acrolein.

3. Discuss major takeaways, upcoming data science training efforts, and future studies.
Overarching Conclusions

Themes across all projects:
• Human respiratory toxicology data
• High interindividual variability
• Relatively small N and number of endpoints
• Goal of quantifying endpoints as a whole

Supervised and unsupervised machine learning represent methods that can aid in understanding key biological mechanisms of inhaled toxicants and interindividual variability in response to inhaled toxicant exposure.

Ongoing challenges:
• Sample size
• Human variability
• Batch effects
• Covariates
• Data pre-processing
• Selection and interpretation of ML
• Biases in analysis
• Data analysis training
Training the Next Generation of Toxicologists

- **inTelligence And Machine LEarning (TAME)** Toolkit, promoting didactic data generation, management, and analysis methods to “TAME” data in environmental health studies
- Development led by Dr. Julia Rager

**Development of the InTelligence And Machine LEarning (TAME) Toolkit for Introductory Data Science, Chemical-Biological Analyses, Predictive Modeling, and Database Mining for Environmental Health Research**

- Kyle Roelle
- Lauren E. Koval
- Rebecca Boyle
- Grace Patlewicz
- Caroline Ringer
- Cynthia V. Rider
- Kevin Ward-Caviness
- Ilona Jaspers
- Rebecca C. Fry
- Julia L. Rager

**The inTelligence And Machine LEarning (TAME) Toolkit for Introductory Data Science, Chemical-Biological Analyses, Predictive Modeling, and Database Mining for Environmental Health**

**Research**

- Kyle Roelle
- Lauren E. Koval
- Rebecca Boyle
- Grace Patlewicz
- Caroline Ringer
- Cynthia V. Rider
- Kevin Ward-Caviness
- Ilona Jaspers
- Rebecca C. Fry
- Julia L. Rager

**Background**

Research in exposure science, toxicology, and environmental health is becoming increasingly reliant upon data science and computational methods that can more efficiently extract information from complex datasets. These methods can be leveraged to better identify relationships between exposures to chemicals in the environment and human disease outcomes. Still, there remains a critical gap surrounding the training of researchers on these in silico methods.

**Objectives**

The aim is to address this critical gap by developing the InTelligence And Machine LEarning (TAME) Toolkit, promoting hands-on-driven data generation, management, and analysis methods to “TAME” data in environmental health studies. This toolkit encompasses training modules, organized as chapters within this Github repository site. All underlying code is in Matlab, Python, R, and other languages and may be found at the parent UNC-SPR Github page.

**Module Development Overview**

Training modules were developed to provide applications-driven examples of data organization and analysis methods that can be used to address environmental health questions. Target audiences for...
TAME is a Publicly Available, Online Bookdown Site

2.4 -Omics Analyses and Systems Biology

This training module was developed by Lauren Kovil, Dr. Kyle Reed, and Dr. Julia E. Regar
Fall 2021

The Field of “-omics”

The field of “-omics” has rapidly evolved since its inception in the mid-1990s, initiated from information obtained through sequencing of the human genome (see the Human Genome Project) as well as the advent of high-content technologies. High-content technologies have allowed the rapid and economical assessment of genome-wide, or -omics-based, endpoints.

Traditional molecular biology techniques typically evaluate the function(s) of individual genes and gene products. -omics-based methods, on the other hand, utilize non-targeted technologies to identify many to all genes or gene products in a given environmental/biological sample. These non-targeted approaches allow for the unbiased investigation of potentially unknown or understudied molecular mediators involved in regulating cell health and disease. These molecular profiles have the potential of being altered in response to toxicant exposures and/or during disease initiation/progression.

To further understand the molecular consequences of -omics-based alterations, molecules can be overlaid onto molecular networks to uncover biological pathways and molecular functions that are perturbed at the systems biology level. An overview of these generally methods, starting with high-content technologies and ending of systems biology, is provided in the below figure (created with BioRender.com).

TAME 2.0 Coming Soon!
TAME 2.0 Chapter 4: Converting Wet Lab Data Into Dry Lab Analyses

Experimental Design

Typically Considered Technical Replicates
- N = 3 cells
- N = 3 cell culture wells

Typically Considered Biological Replicates
- N = 3 independent experiments
- N = 3 cell lines

Data Processing & Transformation

Basic Statistical Testing & Improved Visualizations

The overall p-value comes from the main statistical test (e.g., t-test, Wilcoxon test, ANOVA, Kruskal-Wallis, Friedman Test).

Pairwise p-values are derived from post-hoc tests such as pairwise t-tests, pairwise Wilcoxon tests, Tukey’s HSD, and Dunn’s test.
• General historical context and taxonomy of modern AI/ML, including ChatGPT!

• Application of machine learning in environmental health science
  • Why do we need machine learning?
  • Machine learning vs. traditional statistical methods
  • Predictive modeling in the context of environmental health science
  • Additional applications of machine learning in environmental health science

• Scripted examples of supervised and unsupervised machine learning in the following modules
"Mechanisms of wildfire smoke toxicity and susceptibility involving extracellular vesicles in humans"

**Goal:** Determine differential responses to wildfire smoke exposure in asthmatics and non-asthmatics through the novel integration of EV signatures obtained from epithelial *in vitro* studies with clinical human *in vivo* studies on biomass smoke exposures.

We hypothesize that the hypoxia inducible factor 1 subunit alpha (HIF-1α) pathway mediates differential inflammatory responsiveness to biomass smoke exposure between asthmatics vs non-asthmatics through extracellular vesicle (EV)-mediated communication.
Acknowledgements

Rager Lab
Julia Rager, PhD
Alexis Payton

Jaspers Lab
Ilona Jaspers, PhD
Parker Duffney, PhD (now EPA)
Stephanie Brocke
Aleah Bailey

CEMALB Study Coordinators
Carole Robinette
Martha Almond
Noelle Knight
Brian Ring

*Study Participants*

Other Collaborators
Agatha Ceppe, PhD (UNC)
Meghan Rebuli, PhD (UNC)
David Reif, PhD (NIEHS)
Shaun McCullough, PhD (RTI International)
Alysha Simmons, PhD (UNC)

Groups
SPIROMICS
TAME Contributors

Funding
CiTEM Training Grant (T32 ES007126)
National Heart, Lung, and Blood Institute
(R01 HL139369, P50 HL120100, F31 HL154758,
P50 HL120100)
Thank you! Questions?

Contact: ehickman@email.unc.edu