Using machine learning to identify metabolomic signatures based on pediatric chronic kidney disease etiology

Arthur Lee
Alison Abraham
Yunwen Xu
Jian Hu
Xia Or

See next page for additional authors

Follow this and additional works at: https://scholarlyexchange.childrensmercy.org/presentations
Part of the Nephrology Commons, and the Pediatrics Commons
Creators
Arthur Lee, Alison Abraham, Yunwen Xu, Jian Hu, Xia Or, Josef Coresh, Casey Rebholz, J Chen, Eugene Ree, Harold Feldman, Paul Kimmel, Bradley A. Warady, Susan Furth, and Michelle Denburg
House Officer Research Award

Arthur Lee
The Children's Hospital of Philadelphia
Philadelphia, PA
Using machine learning to identify metabolomic signatures based on pediatric chronic kidney disease etiology

Arthur M. Lee, MD

CHOP Pediatrics Residency Program, PGY-3
Penn-NHLBI StARR R38 Program, appointed scholar
I have no conflicts of interest to disclose
Pediatric chronic kidney disease (CKD)

- The prevalence of pediatric CKD is rising and is associated with significant morbidities
  - Progression to end-stage-renal disease (ESRD)
  - Cardiovascular disease
  - Poor developmental and neurocognitive outcomes
- Different etiologies have distinct clinical features, but less is known regarding biochemical pathophysiology
  - Limited ability to progression to ESRD
  - Limited targeted therapies for pediatric CKD
Metabolomics

- Metabolomics is the study of circulating small molecules that are metabolism intermediates.
- High dimensional – omics data poses challenges to traditional biostatistics.
Machine learning (ML)

Input data
Think of metabolites as individual pixels that form a composite matrix

Machine learning
Mathematical algorithms examine the composite and look for patterns among the individual pixels

Classification/regression task
Is this a cat?
Are there FSGS metabolite patterns that give insight to pathomechanism?
Hypotheses

• Different pediatric CKD etiologies will be associated with distinct metabolomic profiles: focal segmental glomerular sclerosis (FSGS), obstructive uropathy, reflux nephropathy, and the aplasia/dysplasia/hypoplasia spectrum

• Machine learning models can be successfully trained to recognize CKD etiology based on metabolomic differences

• Applying ML in conjunction with traditional biostatistics will improve clinical insight gained than either approach alone
The Chronic Kidney Disease in Children (CKiD) study

- Largest longitudinal cohort of pediatric CKD in North America
- Enrolled children aged 1-16 years with estimated GFR 30-90ml/min/1.73m²
- 702 participants with 842 named metabolites
Limitations of traditional biostatistics in metabolomics

- High number of input metabolites
- Metabolites are highly interrelated, not independent variables
- Flawed significance designations
- Limited inferences from a very rich data source
842 metabolites

Lasso analysis
With adjustment for:
age, sex, BMI z-score, race, CKD
duration, hypertension diagnosis,
ACE/ARB usage, proteinuria, & GFR

Selected metabolites
FSGS n=56
DI n=43
A/D/H n=69
RN n=78

Logistic regression
Significance by Bonferroni threshold,
p < 0.05 / (in Lasso-selected metabolites)

Support vector machine
Significance by top 10% most important metabolites in >5/10 training iterations

Random forest
Significance by top 10% most important metabolites in >5/10 training iterations

Extreme gradient boosting
Significance by top 10% most important metabolites in >5/10 training iterations

Significant metabolites
Metabolite significant in >2/4 of the modeling approaches
Application of ML tools

- Feature selection:
  - Lasso penalized logistic regression
- Multiple approaches to detect signals/patterns:
  - Logistic regression, support vector machine, random forest, extreme gradient boosting

![Volcano plot of Lasso-selected FSGS metabolites](image)
### Results: implicated metabolites

#### Focal segmental glomerular sclerosis (n=63) (Lasso metabolites = 56)

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Sub-pathway</th>
<th>Metabolite</th>
<th>Modeling approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipid</td>
<td>Sphingomyelin</td>
<td>Sphingomyelin (d18:1/18:1, d18:2/18:0)</td>
<td>LR, SVM, RF, XGB</td>
</tr>
<tr>
<td>Lipid</td>
<td>Sphingomyelin</td>
<td>Sphingomyelin (d18:2/24:2)</td>
<td>LR, SVM</td>
</tr>
<tr>
<td>Lipid</td>
<td>Plasmalogen</td>
<td>1-(1-etyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0)</td>
<td>LR, SVM, RF</td>
</tr>
<tr>
<td>Lipid</td>
<td>Plasmalogen</td>
<td>1-(1-etyl-palmitoyl)-2-arachidonoyl-GPC (P-16:0/20:4)</td>
<td>LR, SVM, RF, XGB</td>
</tr>
<tr>
<td>Lipid</td>
<td>Lysophospholipid</td>
<td>1-arachidonoyl-GPI* (20:4)*</td>
<td>LR, RF, XGB</td>
</tr>
<tr>
<td>Lipid</td>
<td>Diacylglycerol</td>
<td>Palmitoyl-arachidonoyl-glycerol (16:0/20:4) [1]</td>
<td>LR, XGB</td>
</tr>
<tr>
<td>Amino Acid</td>
<td>Glutamate</td>
<td>N-acetyl-aspartyl-glutamate (NAAG)</td>
<td>LR, SVM</td>
</tr>
<tr>
<td>Amino Acid</td>
<td>Urea Cycle</td>
<td>Homoarginine</td>
<td>LR, RF</td>
</tr>
<tr>
<td>Amino Acid</td>
<td>Tryptophan</td>
<td>6-bromotryptophan</td>
<td>LR, RF</td>
</tr>
<tr>
<td>Cofactors</td>
<td>Panthothene</td>
<td>Pantothenate</td>
<td>LR, RF</td>
</tr>
</tbody>
</table>

#### Obstructive uropathy (n=122) (Lasso metabolites = 43)

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Pathway</th>
<th>Metabolite</th>
<th>Modeling approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino Acid</td>
<td>Histidine</td>
<td>Trans-urocanate</td>
<td>LR, SVM, RF, XGB</td>
</tr>
<tr>
<td>Amino Acid</td>
<td>Histidine</td>
<td>Imidazole propionate</td>
<td>LR, SVM, RF, XGB</td>
</tr>
<tr>
<td>Amino Acid</td>
<td>Tyrosine</td>
<td>4-methoxyphenol sulfate</td>
<td>LR, SVM, XGB</td>
</tr>
</tbody>
</table>

#### Reflex nephropathy (n=56) (Lasso metabolites = 78)

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Pathway</th>
<th>Metabolite</th>
<th>Modeling approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino Acid</td>
<td>Tryptophan</td>
<td>Indolepropionate</td>
<td>LR, SVM, RF, XGB</td>
</tr>
<tr>
<td>Amino Acid</td>
<td>Phenylalanine</td>
<td>Phenylpyruvate</td>
<td>SVM, XGB</td>
</tr>
<tr>
<td>Amino Acid</td>
<td>Glycine</td>
<td>Dimethylglycine</td>
<td>SVM, XGB</td>
</tr>
<tr>
<td>Xenobiotics</td>
<td>Benzothia</td>
<td>5-vinylphenol sulfite</td>
<td>LR, SVM, RF, XGB</td>
</tr>
</tbody>
</table>

#### Aplasia, dysplasia, hypoplasia (n=109) (Lasso metabolites = 69)

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Pathway</th>
<th>Metabolite</th>
<th>Modeling approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino Acid</td>
<td>Glutamate</td>
<td>Citraulinate</td>
<td>LR, SVM, RF, XGB</td>
</tr>
<tr>
<td>Amino Acid</td>
<td>Urea cycle</td>
<td>N-delta-acetylmethane</td>
<td>LR, SVM, RF, XGB</td>
</tr>
<tr>
<td>Amino Acid</td>
<td>Methionine</td>
<td>Cysteine sulfonic acid</td>
<td>SVM, XGB</td>
</tr>
<tr>
<td>Lipid</td>
<td>Sphingomyelin</td>
<td>Sphingomyelin (d18:2/24:2)</td>
<td>LR, SVM, RF, XGB</td>
</tr>
<tr>
<td>Lipid</td>
<td>Ceramides</td>
<td>Ceramide (d18:1/20:0, d16:1/22:0, d20:1/18:0)</td>
<td>LR, SVM, RF, XGB</td>
</tr>
</tbody>
</table>
How do we know our ML models are actually detecting metabolite signals for CKD etiology classification?
Does sphingomyelin dysmetabolism induce FSGS pathology?

- Sphingomyelin metabolites have been shown to induce free radical damage in the kidney
- Abnormal sphingomyelin deposition implicated in HIV-related kidney & brain disease

Is sphingomyelin dysmetabolism secondary to primary FSGS?

- Dyslipidemia has been characterized in CKD patients
- Unclear therapeutic benefits of lipopheresis in pediatric nephrotic syndrome

What can we learn about potential therapeutics?

- Rituximab has been shown to alter sphingomyelin levels and associated with disease remission
Conclusions

• Different pediatric CKD etiologies are associated with distinct metabolomic profiles

• Machine learning models can be successfully trained on pediatric metabolomics data

• Machine learning can be used as pattern-recognition tools to augment traditional biostatistics to gain improved clinical insight
Thank you
Michelle R. Denburg (mentor)
Susan L. Furth
Jeanine Ronan
Erum Hartung
Peter Klein
Jian Hu
Rui Xiao
Yunwen Xu
Mark D. DeBoer
Jennifer Charlton
Helena Ritchie

CHOP Pediatric Residency program
CHOP Division of Nephrology
Penn-NHLBI StARR program
(Grant R38 HL143613-03)
CKD Biomarkers Consortium
(Grant U01DK106982)
Society for Pediatric Research

glm(super_bowl_champions ~ eagles,
family = “binomial”,
data = NFL_history)
p<0.05
Supplemental
Data analysis: challenges

- Useful to think about challenges related to data-analysis in relation to the 3 framing topics

**Pediatric CKD**
- Unbalanced data: relatively low number of test cases versus controls based on CKD etiology
- Differences in participant characteristics based on CKD etiology

**Metabolomics**
- Wide data: number of metabolites > number of study participants
- Metabolites are not independent variables, have significant interactions with other metabolites

**Machine learning**
- Interpretability: Unfamiliar output compared to established biostatistics
- How to determine if model was successfully trained
- How to designate a significant metabolite
### Table 1: Participant characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Total cohort</th>
<th>FSGS</th>
<th>Obstructive uropathy</th>
<th>Aplasia/dysplasia/hypoplasia</th>
<th>Reflux nephropathy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>702</td>
<td>63</td>
<td>122</td>
<td>109</td>
<td>86</td>
</tr>
<tr>
<td><strong>Categorical</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex (male)</td>
<td>432</td>
<td>35</td>
<td>103*</td>
<td>57*</td>
<td>49</td>
</tr>
<tr>
<td>Hypertension</td>
<td>95</td>
<td>12</td>
<td>12</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>ACEI/ARB usage</td>
<td>396</td>
<td>54*</td>
<td>46*</td>
<td>36*</td>
<td>47</td>
</tr>
<tr>
<td><strong>Numerical</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>11.3 (4.3)</td>
<td>13.6 (3.0)**</td>
<td>10.1 (4.2)**</td>
<td>9.8 (4.7)**</td>
<td>11.6 (4.1)</td>
</tr>
<tr>
<td>BMI z-score</td>
<td>0.5 (1.1)</td>
<td>1.0 (1.3)**</td>
<td>0.4 (1.2)</td>
<td>0.3 (1.0)**</td>
<td>0.4 (1.0)</td>
</tr>
<tr>
<td>GFR (ml/min/1.73m^2)</td>
<td>51.1 (1.4)</td>
<td>55.1 (1.5)</td>
<td>47.3 (1.4)**</td>
<td>47.4 (1.5)**</td>
<td>50.2 (1.4)</td>
</tr>
<tr>
<td>CKD duration (years)</td>
<td>8.5 (4.9)</td>
<td>5.4 (3.5)**</td>
<td>10.1 (4.2)**</td>
<td>9.8 (4.7)**</td>
<td>11.6 (4.1)**</td>
</tr>
<tr>
<td>Urine protein:creatinine ratio</td>
<td>0.4 (3.9)</td>
<td>0.8 (4.6)**</td>
<td>0.4 (2.9)</td>
<td>0.3 (3.6)</td>
<td>0.2 (3.2)**</td>
</tr>
</tbody>
</table>
Feature selection:
Lasso penalized logistic regression
Machine learning algorithms

**Support vector machine**
Plots participants to high dimensional space based on all input metabolites, then determines optimal hyperplace to separate classifications.

**Random forest**
Aggregated tree model in which a random number of metabolites are sampled at each branch point to perform classification.

**Extreme gradient boosting**
Aggregated regression models applied in sequence optimized for reduce previous misclassification error.
Training models

Total CKiD Cohort

80% training set
Models trained with repeated k-fold cross validation
Determined feature weighting

20% validation set
Evaluated performance with receiver-operator & precision-recall area-under-the-curve

Repeat x10
Volcano plots of Lasso-selected metabolites

FSGS
- 6-bromotryptophan
- N-acetyl-aspartyl-glutamate
- Pantothenate
- Diacylglycerol
- Homoarginine
- 1-arachidonoyl-GPI (20:4)

Obstructive uropathy
- Sphingomyelin (d18:1/18:1, d18:2/18:0)
- 1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0)
- Sphingomyelin (d18:1/18:1, d18:2/18:0)
- 1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/20:4)
- Pantothenate
- Bonferroni P<0.05/56
- Benjamini-Hochberg FDR=0.05

- Imidazole propionate
- Trans-urocanate
- 4-methoxyphenol sulfate
- Bonferroni P<0.05/43
- Benjamini-Hochberg FDR=0.05

Metabolite implicated by ML approach
ML performance evaluation metrics

F-1 score

- Harmonic mean of precision & recall
  \[ F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{tp}{tp + \frac{1}{2}(fp + fn)}. \]

Matthews correlation coefficient

- Includes true negative predictions
  \[ \text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \]
<table>
<thead>
<tr>
<th>Etiology</th>
<th>Model</th>
<th>ROC-AUC</th>
<th>PR-AUC</th>
<th>F-1 score</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No-skill</td>
<td>0.5</td>
<td>Prevalence</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FSGS</td>
<td>SVM</td>
<td>0.92 (0.92, 0.93)</td>
<td>0.60 (0.57, 0.63)</td>
<td>0.51 (0.49, 0.53)</td>
<td>0.50 (0.48, 0.52)</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.89 (0.88, 0.90)</td>
<td>0.50 (0.48, 0.51)</td>
<td>0.47 (0.46, 0.48)</td>
<td>0.45 (0.44, 0.46)</td>
</tr>
<tr>
<td></td>
<td>XGB</td>
<td>0.91 (0.90, 0.91)</td>
<td>0.54 (0.53, 0.56)</td>
<td>0.48 (0.47, 0.49)</td>
<td>0.47 (0.46, 0.48)</td>
</tr>
<tr>
<td>OU</td>
<td>SVM</td>
<td>0.84 (0.84, 0.85)</td>
<td>0.52 (0.51, 0.53)</td>
<td>0.54 (0.53, 0.54)</td>
<td>0.44 (0.44, 0.45)</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.73 (0.73, 0.74)</td>
<td>0.39 (0.38, 0.40)</td>
<td>0.42 (0.41, 0.42)</td>
<td>0.28 (0.27, 0.29)</td>
</tr>
<tr>
<td></td>
<td>XGB</td>
<td>0.79 (0.79, 0.80)</td>
<td>0.45 (0.43, 0.46)</td>
<td>0.48 (0.47, 0.48)</td>
<td>0.37 (0.37, 0.38)</td>
</tr>
<tr>
<td>A/D/H</td>
<td>SVM</td>
<td>0.84 (0.83, 0.85)</td>
<td>0.51 (0.50, 0.52)</td>
<td>0.53 (0.51, 0.54)</td>
<td>0.44 (0.42, 0.45)</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.68 (0.68, 0.69)</td>
<td>0.30 (0.29, 0.31)</td>
<td>0.38 (0.38, 0.39)</td>
<td>0.24 (0.23, 0.25)</td>
</tr>
<tr>
<td></td>
<td>XGB</td>
<td>0.75 (0.75, 0.76)</td>
<td>0.38 (0.37, 0.39)</td>
<td>0.43 (0.42, 0.44)</td>
<td>0.32 (0.31, 0.33)</td>
</tr>
<tr>
<td>RN</td>
<td>SVM</td>
<td>0.80 (0.79, 0.81)</td>
<td>0.37 (0.36, 0.38)</td>
<td>0.41 (0.40, 0.42)</td>
<td>0.34 (0.33, 0.35)</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.66 (0.65, 0.66)</td>
<td>0.19 (0.19, 0.20)</td>
<td>0.31 (0.30, 0.31)</td>
<td>0.20 (0.19, 0.21)</td>
</tr>
<tr>
<td></td>
<td>XGB</td>
<td>0.73 (0.72, 0.73)</td>
<td>0.25 (0.25, 0.26)</td>
<td>0.33 (0.33, 0.34)</td>
<td>0.25 (0.25, 0.26)</td>
</tr>
</tbody>
</table>
Results: implicated metabolites

- FSGS: sphingomyelin & plasmalogen metabolites
- Obstructive uropathy: histidine metabolites
Key questions:

Does feature selection improve ML performance and increase our confidence in the signals detected?

Does a Lasso feature selection approach significantly improve ML performance or alter the signals detected compared to a traditional biostatistics feature selection approach?