**Identification of a Pediatric Acute Hypoxemic Respiratory Failure Signature in Peripheral Blood Leukocytes at 24 Hours Post-ICU Admission with Machine Learning**

Jocelyn R. Grunwell, MD, PhD\*1,2, Milad G. Rad, MS\*3, Nadir Yehya4,5, and Rishikesan Kamaleswaran, PhD2,3

1Children’s Healthcare of Atlanta, Egleston Hospital, Atlanta, GA, USA.

2Emory university School of Medicine, Department of Pediatrics, Atlanta, GA, USA.

3Georgia Institute of Technology, Department of Electrical and Computer Engineering, Atlanta, GA, USA.

4Department of Anesthesiology and Critical Care Medicine, 6040A Wood Building, Children’s Hospital of Philadelphia, 3401 Civic Center Boulevard, Philadelphia, PA 19104, USA.

5University of Pennsylvania, Philadelphia, PA, USA.

\*These authors contributed equally to this work.

**Correspondence:** [jgrunwe@emory.edu](mailto:jgrunwe@emory.edu); Emory University School of Medicine, Department of Pediatrics, Children’s Healthcare of Atlanta at Egleston, Division of Critical Care Medicine, 1405 Clifton Road NE, Atlanta, GA, USA, 30322; phone 404-785-1600; fax 404-785-6233

**Co-author Email Addresses:** [ghiasirad.milad@gatech.edu](mailto:ghiasirad.milad@gatech.edu), yehyan@email.chop.edu, [rkamaleswaran@gatech.edu](mailto:rkamaleswaran@gatech.edu)

**Institutions where work was performed:** Children’s Hospital of Philadelphia, University of Pennsylvania, Emory University School of Medicine, Children’s Healthcare of Atlanta, and Georgia Institute of Technology

**Reprints:** Reprints will not be ordered.

**Financial Support:** Funding was provided by NIH grants K12HD072245 (Atlanta Pediatric Scholars Program), K23 HL151897-01, and an Emory University Pediatrics Research Alliance Junior Faculty Focused Pilot award to JG. Funding was provided by the NIH grant K24 NR018866 to AF. RK is supported by the NIH under Award Numbers R01GM139967 and UL1TR002378.

**Keywords:** pediatric, acute respiratory distress syndrome, mechanical ventilation, gene expression profiling, machine learning, transcriptomics

**Competing Interests Statement:** The author(s) declare no competing interests.

**Authors’ contributions:** NY obtained the patient samples, performed the transcriptomics experiments. JG, NY, and RK conceived and developed the study, supervised the analysis and interpreted the data. JG drafted and edited the manuscript. RK, MR, and NY assisted with drafting and editing the manuscript. All authors edited and approved the final version of this manuscript.

**Words:** X,XXX

**Abstract (XXX words)**

**Importance:**

**Objectives:**

**Design, Setting and Participants:**

**Main Outcomes and Measures:**

**Results:**

**Conclusions and Relevance:**

**Introduction**

Acute hypoxemic respiratory failure is a defining feature of moderate and severe acute respiratory distress syndrome (ARDS) with a PaO2/FiO2 ratio of less than 200 accompanied by bilateral chest infiltrates on chest radiograph within seven days of a clinical trigger and not related to fluid overload or heart failure.(1) Heterogeneity is a hallmark feature of ARDS, and correlating clinical phenotypes with biological endotypes using plasma biomarkers and clinical variables using machine learning methods has led to the discovery of a hyperinflammatory (reactive) and hypoinflammatory (unreactive) ARDS endotype with differential responses to high PEEP, low tidal-volume and restrictive fluid therapies.(2-5) Systemic infections, such as sepsis, often caused by pneumonia, is a common trigger of ARDS; however, trauma, aspiration, near-drowning, burns, and blood transfusions are also instigators of ARDS.

Statistical and machine learning approaches have been used to perform a meta-analysis of publicly available whole-blood gene expression data to predict mortality from sepsis in adults and children.(6) While a multi-cohort analysis of adult and pediatric whole-blood gene expression data did not find a generalizable transcriptomics signature of ARDS,(7) machine learning random forest models applied to clinically available laboratory and physiologic data within the first two days of intensive care outperformed multivariable logistic regression at predicting prolonged acute hypoxemic respiratory failure of a week of more following influenza infection in a large multicenter cohort of children enrolled in the PICFLU observational study.(8) Furthermore, since publication of the ARDS transcriptomic meta-analysis incorporating adult ARDS and pediatric sepsis-triggered acute hypoxemic respiratory failure,(7) release of a pediatric ARDS-specific whole-blood microarray gene expression cohort is publicly available.(9)

The primary objective of this study was to determine a gene expression signature of mild and moderate/severe acute hypoxemic respiratory failure from publicly available whole blood gene expression microarray data. A secondary objective was to explore the gene networks common to both datasets.

**Methods**

**Data Collection**

The pediatric sepsis dataset, GSE66099, and the pediatric ARDS (PARDS) dataset, GSE147902 were downloaded from the NCBI Gene Expression Omnibus (GEO) repository.(9, 10) Both datasets contain gene expression profiles obtained from peripheral whole blood samples of patients who were admitted to the pediatric intensive care unit (PICU) within 24h of a diagnosis of sepsis or PARDS using the Berlin definition, respectively. Acute hypoxemic respiratory failure was defined in both cohorts as a PaO2/FiO2 (P/F) < 200.

**Normalization and Background Correction**

The Affy package from R was used to remove technical variations in the gene expression data.(11) Background noise correction and normalizations was performed using the R package gcrma.(12) Surrogate Variable Analysis using the R package sva was used to correct for batch effects variation between datasets.(13)

**Probe to Gene Mapping**

The GSE66099 and GSE147902 microarray dataset Affymetrix probes were matched to gene symbols using Affymetrix Human Genome U133 Plus 2.0 (hgu133plus2.db) and Human Gene 2.1 ST Array (hugene21sttranscriptcluster.db), respectively. Genes with multiple probes were matched to the same gene and the expressions were averaged.(6)

**Variable Selection Methods and Classification Models**

Candidate markers were selected using stability selection [Meinshausen, Nicolai, and Peter Bühlmann. "Stability selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4 (2010): 417-473.], a bootstrapping method of 100 simulations using logistic regression as a classifier, using 0.6 as the threshold. The top-ranked genes that contributed to the acute hypoxemic respiratory failure signature were selected in each dataset. Genes common to both of the top 1,500 ranked gene lists were selected for pathway analysis.

**Pathway analysis**

The most common pathways were explored using the 185 top-ranked genes common to both datasets using the Kyoto Encyclopedia of Genes and Genomes (KEGG).(14, 15) The identified KEGG pathways were fed into the R packagePathway Network Visualizer (PANEV) to identify and map the first- and second-level pathway networks associated with acute hypoxemic respiratory failure.(16) The Python Reactome package was used to perform an over-representation gene network analysis using the 185 top-ranked genes common to both datasets.(17, 18) The data analysis workflow using machine learning methods for common dataset gene selection and pathway analysis is shown in **Supplementary Figure 1**.

**Results**

**Cohort Description**

There were forty-seven healthy control children, sixty-five children without and ninety-six children with sepsis-related acute hypoxemic respiratory failure, defined as a P/F ≥ 200 from GSE66099. There were twenty-six children with mild PARDS and seventy children with moderate/severe ARDS, based on bilateral chest x-ray findings and a P/F = 200 – 300 (mild) and less than 200 (moderate/severe) in GSE147902. The available demographics and clinical characteristics of the participants with acute hypoxemic respiratory failure (GSE66099) and ARDS (GSE147902) stratified by P/F ratio are summarized in **Table 1**.

**Differential Gene Expression of PARDS Severity Compared to Healthy Controls**

We determined whether the differential gene expression profiles for children with PARDS by comparing children with mild ARDS and moderate/severe ARDS from GSE147902 with healthy controls from the GSE66099 data set. We chose GSE147902 as the ARDS data set because children in this cohort met both the Berlin and PALICC criteria to be enrolled. There were 16,221 genes included in the analyses, and there were 4313 genes down-regulated and 1063 genes up-regulated in the mild PARDS vs. healthy controls. There were 4008 genes down-regulated and 1190 genes up-regulated in the moderate/severe PARDS vs. healthy controls. Volcano plots of the mild PARDS vs. healthy controls and moderate/severe PARDS vs. healthy controls are shown in **Supplementary Figure 2A and 2B**, respectively. Pathway analysis was performed after controlling for genes with 95% or more collinearity followed by a stability selection algorithm to select genes by their importance in distinguishing mild or moderate/severe PARDS from controls. Over-representation analysis of differentially expressed genes corresponding to pathways within the Kyoto Encyclopedia of Genes and Genomes (KEGG) database associated with mild and moderate/severe PARDS are shown in **Figures 1A** and **1B**, respectively.(14, 15) The top five pathways common between the mild and moderate/severe PARDS groups include: metabolic pathways, pathways in cancer, Herpes simplex virus 1 infection, Cytokine-cytokine receptors interactions, and the phosphatidylinositol 3-kinaseand protein kinase B (PI3K-Akt) signaling pathway. While the first-level KEGG pathways are similar for mild PARDS vs. control and moderate/severe PARDS vs. control, there are more genes are represented in the moderate/severe PARDS group compared with the mild PARDS group and the rank-ordering of the top five pathways differ between the mild vs. moderate/severe PARDS groups.

The genes and corresponding first-level KEGG pathways for mild PARDS or moderate/severe PARDS vs. healthy controls were fed into the R package Pathway Network Visualizer (PANEV).(16) The PANEV network of first-level and second-level genes and pathways corresponding to mild PARDS vs. healthy controls and to moderate/severe PARDS vs. healthy controls were mapped. The genes corresponding the first- and second-level pathways and a summary of the number of genes within each second-level pathways are shown for the mild PARDS vs. healthy controls (**Supplementary Files 1-3**)and moderate/severe PARDS vs. healthy controls (**Supplementary Files 4-6**). Of the top 10 second-level pathways, there were nine in common between mild and moderate/severe PARDS vs. healthy controls with the top five pathways in the same rank-order including: antigen processing and presentation, natural killer cell cytotoxicity, endocytosis, cytokine-cytokine interactions, and regulation of actin cytoskeleton (compare **Supplementary Files 3 and 6**). The top ten second-level pathways that differed included the spliceosome network for mild PARDS vs. controls and oxidative phosphorylation in moderate/severe PARDS vs. controls.

**Selection of Differentially Expressed Genes in Moderate/Severe versus Mild PARDS**

While the comparison with healthy children was a starting point, we were also interested in the genes that differentiated mild from moderate/severe PARDS. We used a machine learning-based stability selection approach to compare the differences in gene expression between mild vs. moderate/severe PARDS within the GSE147902 group.(19, 20) Selected genes were mapped to KEGG pathways using an over-representation analysis (**Figure 2**). Top pathways differentiating mild from moderate/severe PARDS included metabolic pathways, pathways in cancer, pathways of neurodegeneration (multiple diseases). Pathways in cancer encompass many of the individual pathways listed including the cytokine-cytokine receptor interaction, cAMP, PI3K-Akt, MAPK, Janus kinase/signal transducer and activator of transcription proteins (JAK/STAT), peroxisome proliferator-activated receptor gamma (PPARγ) signaling, calcium signaling, mammalian target of rapamycin (mTOR), and transforming growth factor receptor (TGF-β) signaling pathways. The neurodegenerative pathway encompasses mitochondrial dysfunction, mitophagy and autophagy, oxidative stress and the formation of reactive oxygen species.

Second- and third-level pathways were discovered using the first-level KEGG pathways and corresponding top 200 genes of importance identified by the stability selection algorithm using PANEV (**Supplementary Files 7-10**). The top three second-level pathways included MAPK signaling, regulation of actin cytoskeleton, and the insulin signaling pathway. The top third-level pathways include natural killer cell mediated cytotoxicity, Parkinson disease, and non-alcoholic fatty liver disease, and AMP-activated protein kinase (AMPK) signaling. The non-alcoholic fatty liver disease involves pathways that signal through TNFα, interleukin-6 (IL-6), insulin, and PPAR to activate downstream PI3K-Akt signaling and fatty acid oxidation.

We next used the curated Reactome database of pathways and reactions in human biology to perform a pathway over-representation analysis using the 185 common genes defining the moderate/severe vs. mild PARDS transcriptomic signature.(17, 18) There were twenty-three enriched pathways shown in **Table 2** with a false discovery rate (FDR) < 0.05. The top pathways found in Reactome included nucleosome assembly, the deacetylation of histones by histone deacetylases (HDACs), RNA Polymerase I promoter opening, the formation of ATP by chemiosmotic coupling, and the formation of a pool of free 40S ribosomal subunits. Fundamental processes such as recognition and removal of damaged DNA bases, methylation of DNA, control of transcription and translation, and dysfunctional pyroptosis leading to cytokine storm were discovered using Reactome.

**Selection of Common Genes of an Acute Hypoxemic Respiratory Failure Signature**

Sepsis is a common trigger for acute hypoxemic respiratory failure that in its most severe form can lead to ARDS. We wanted to determine whether the differential gene expression pattern discovered by comparing mild (P/F 200 – 300) with moderate/severe PARDS (P/F < 200) from GSE147902 could be applied to a cohort of children with acute hypoxemic (20)overlap in differentially expressed genes for children with a P/F < 200 vs. a P/F ≥ 200 using traditional statistical methods. We therefore independently ranked the differentially expressed genes by P/F < 200 vs. P/F ≥ 200 using the machine learning stability select algorithm.(19, 20) From the top-ranked 1500 genes in each data set, genes common to both ranked gene sets were selected in ten gene increments as depicted in **Figure 3A**. The optimal number of common genes was determined by plotting the area under the receiver operative curve (ROC) characteristic for each incremental addition of ten genes to the acute hypoxemic respiratory failure model as shown in **Figure 3B**. The red line in **Figure 3B** shows the change in the ROC with the addition of common genes for the data set specific for PARDS (GSE147902). The black line in **Figure 3B** shows the change in the ROC with the addition of common genes for the pediatric sepsis acute hypoxemic respiratory failure data set (GSE66099). The ROC plateaus around 25 – 30 genes for each data set. The area under the receiver operating curves (AUROC) and area under the precision recall curves (AUPRC) for each GSE66099 (**Figure 3C, 3E**) and GSE147902 (**Figure 3D, F**) are shown for the optimal common genes for each incremental addition of ten genes to the model. For both datasets the AUROC was maximized at 92 genes with a value of 0.74 (95% CI: 0.66 – 0.80) for GSE66099 and 0.81 (95% CI: 0.72 – 0.89) for GSE147902.

Due to the imbalance in the proportion of mild to moderate/severe acute hypoxemic respiratory failure and PARDS cases in the cohorts, we report the area under the precision recall curve (AUPRC) for the stability select model of moderate/severe vs. mild ARDS. The AUPRC measures the ability of the models to correctly distinguish mild from moderate/severe acute hypoxemic respiratory failure (GSE66099) or PARDS (GSE147902). The sepsis GSE66099 data set consists of 40% of children with a P/F < 200 while the PARDS GSE147902 data set consists of 27% of children with a P/F < 200. The AUPRC was maximized at 92-genes with a value of 0.66 (95% CI: 0.5 – 0.76) for GSE66099 and 0.90 (95% CI: 0.75 – 0.91) for GSE147902 both of which are an improvement over the baseline percentage of children with a P/F < 200 in either cohort. As the inclusion criteria for the GSE147902 is specific for PARDS, while the GSE66099 cohort is specific for children with sepsis, it is not surprising that the AUROC and AUPRC perform better for the PARDS-specific compared to the pediatric sepsis cohort.

**Network Analysis**

The 185 genes selected using stability selection were ranked by normalized importance coefficient for each dataset. The selected genes, ranked by the number of genes involved in the KEGG pathway, are shown in **Figure 4**. The top 3 KEGG pathways included metabolic, ribosomal, and coronavirus/COVID-19 disease pathways. Metabolic pathways involve the intricate networks of glycolysis, the tricarboxylic acid (TCA) cycle, the pentose phosphate pathway, gluconeogenesis, fatty acid oxidation and lipid metabolism, one-carbon metabolism, amino acid and cofactor metabolism, and nucleic acid metabolism for the production of energy and substrates for fundamental biological reactions. The ribosomal pathway is comprised of protein components necessary to translate proteins. The coronavirus/COVID-19 disease pathway consists of anti-viral detection via the toll-like receptor (TLR) and retinoic acid-inducible gene I (RIG-I) pattern recognition receptor pathways, activation of the NLRP3 inflammasome, and signaling through the pro-inflammatory cytokines IL-6, JAK/STAT, PI3K, and MAPK signaling cascades, and activation of the transcription factor NF-KB.

We next visualized networks of pathways involved in the moderate/severe acute hypoxemic respiratory failure gene expression signature by imputing the list of 185 common genes into the PANEV package in R. PANEV uses the KEGG library to find first-level pathways corresponding to the imputed genes and then maps the links to second-level and third-level pathways creating a network of pathways important to moderate/severe acute hypoxemic respiratory failure. The first-, second-, and third-level pathways with the corresponding genes contributing to each pathway are shown in **Supplementary Files 11-14.** The top three second-level pathways include natural killer cell mediated cytotoxicity, endocytosis, and neutrophil extracellular trap formation. The top third-level pathways include RNA transport, synaptic vesicle formation, and regulation of the actin cytoskeleton.

Finally, we used the curated Reactome database to perform a pathway over-representation analysis using the 185 common genes defining the moderate/severe acute hypoxemic respiratory failure gene expression signature. There were nineteen significantly enriched pathways shown in **Table 3** with cytosolic and mitochondrial protein translation, selenoamino acid metabolism, selenocysteine synthesis, and nonsense codon mediated decay of messenger RNA.

**Discussion**

We used publicly available microarray gene expression data sets and machine learning methods to determine metabolic and immune pathways involved in the systemic response of children within 24 hours of a PARDS diagnosis. We first identified gene networks involved in distinguishing children with mild or moderate/severe PARD from healthy controls. Next, we determined pathways differentiating children with mild from moderate/severe PARDS. Finally, we evaluated the overlap amongst genes differentiating children with a P/F < 200 vs. a P/F ≥ 200 using a PARDS-specific cohort and a pediatric sepsis cohort with acute hypoxemic respiratory failure. Metabolic pathways, immune activation via cytokine-cytokine receptor interactions, networks to contain oxidative stress, and phosphorylation signaling cascades were the predominant pathways involved in distinguishing mild and moderate/severe PARDS from healthy controls and in differentiating mild from moderate/severe PARDS. We then determined that a core systemic gene expression signature of moderate/severe acute hypoxemic respiratory failure triggered by sepsis involved metabolic, mitochondrial, immunologic derangement involving fundamental cellular processes such as protein translation and quelling of oxidative stress. Our results support further investigation of gene networks and signaling pathways involved in differentiating PARDS severity and understanding PARDS heterogeneity by underlying biological processes.

An attempt at finding a generalizable diagnostic gene expression signature for ARDS using publicly available human whole blood gene expression arrays from adults with sepsis, trauma and burns and children with acute hypoxemic respiratory failure triggered by sepsis failed to find an ARDS-specific signature after adjusting for clinical severity.(7) Since this publication, a pediatric-specific ARDS transcriptomics dataset from whole blood collected within 24 hours of PARDS onset was used to identify three ARDS sub-phenotypes with divergent clinical characteristics and outcomes using k-means clustering.(9) A limitation of the transcriptomic PARDS phenotyping study was the lack of a non-ARDS mechanically ventilated cohort as a control group.(9) Using the definition of a P/F < 200 for moderate/severe acute hypoxemic respiratory failure from GSE66099 and ARDS from GSE147902, we reexamined whether a systemic transcriptomic signature of moderate/severe PARDS emerged using the GSE147902 data set that was specific for children with ARDS.(7, 9) We were able to find an overlapping pattern of gene pathways of importance to pediatric ARDS using stability selection modeling and determining the discriminatory ability of the model to predict moderate/severe acute hypoxemia using the intersection of the top 1500 ranked genes in the pediatric ARDS-specific and pediatric sepsis acute hypoxemic respiratory failure cohorts.

Machine learning-based analysis of data with class imbalances can be plagued by a number of issues, stemming from the inherent mathematical assumption of equal case to control distributions among many learning algorithms. As a result, these models when applied in a single train-test instance may produce random effects that are poorly generalized in external datasets. The traditional AUROC metrics poorly characterizes class-based performances, thus necessitating alternative metrics to evaluate the robustness of models. One of the methods to investigate the effects of class imbalance is by the use of metrics such as AUPRC, F1 (harmonic mean of the precision and recall), and bootstrap with replacement. In this work, we utilize AUPRC and stability selection utilizing bootstrap with replacement to identify a coherent set of ‘stable’ genes that indicate robust predictive performance consistently across 100 bootstrap iterations. We illustrate the AUPRC plot to show strong performance even in the minority class, thus emphasizing the robustness of the learned model.

Metabolic pathways were the top KEGG pathway in each analysis. Several recent studies have investigated metabolic changes in ARDS.(21) How metabolic derangements, such as mitochondrial dysfunction, decreased oxidative phosphorylation, and oxidative stress, lead to bioenergetic failure in ARDS and metabolic reprogramming of the immune system are active areas of investigation. Metabolomics studies of the plasma and airway fluid of patients with ARDS have revealed changes in tricarboxylic acid (TCA) cycle intermediates of glucose, alanine and glutamine due to energetic stress on lung epithelial cells.(21-27) Microarray analysis of whole-blood gene expression in adults with sepsis-triggered ARDS revealed that the “reactive” or hyperinflammatory subgroup is enriched for genes associated with oxidative phosphorylation, and that the “reactive” subgroup is also associated with high plasma lactate levels indicative of mitochondrial dysfunction.(28) Differential expression of the nuclear-encoded mitochondrial transcriptome in children with septic shock who had more organ failures and higher mortality (endotype A) compared to children with sepsis with a less complicated (endotype B or C) from the GSE66099 have been reported.(29) Serum lactate levels were the highest in pediatric sepsis endotype with the highest organ dysfunction and largest decrease in electron transport chain genes that may correspond to the “reactive” subgroup with decreased mitochondrial gene expression reported in adults with sepsis-triggered ARDS.(29, 30)

By contrast, the mitogen-activated protein kinase (MAPK) pathways that control cell proliferation, differentiation, motility, and survival were enriched in the comparatively “uninflamed” subgroup.(30) The MAPK pathway is a first-level or second-level pathway found in all of the analyses distinguishing PARDS from healthy controls and moderate/severe from mild PARDS. It is not known whether children with ARDS exhibit similar “reactive” or “hyperinflammed” and “uninflamed” phenotypes as shown in adults with ARDS and without phenotypic classification we cannot speculate on the pathobiology underlying class differences.(4, 5, 30, 31) The three Children’s Hospital of Philadelphia ARDS transcriptomic subtypes (CATS) described by clustering GSE147902 do not conform to the previously described adult phenotypes and are not yet externally validated.(9)

Several neurodegenerative pathways, including Alzheimer’s disease, Parkinson’s disease, and Huntington disease, emerged as first-level pathways in our analyses of moderate/severe PARDS. Examination of the processes comprising these neurodegenerative pathways revealed networks involving impairment of oxidative phosphorylation, calcium signaling, oxidative stress, autophagy and mitophagy, mitochondrial dysfunction, the RAGE signaling pathway, and endoplasmic reticulum stress.(32) In addition to these key mitochondrial bioenergetic functions mitochondrial protein synthesis emerged as a Reactome pathway in our analyses, and mitochondrial ribosomal proteins are known to be downregulated in the blood of early Alzheimer’s disease patients.(33) Mitochondria are an integral part of cell signaling pathways of the immune response, including cytokine release, inflammasome formation, and the formation of reactive oxygen species.(34) Bioenergetic failure from mitochondrial dysfunction is a prominent feature of the pathways and complex cellular appearing in our transcriptomic analysis of children with moderate/severe PARDS and acute hypoxemic respiratory failure.

Respiratory viral infections are a common trigger of ARDS in children. Pathogenic influenza and coronaviruses use host epigenetic reprogramming to evade the host immune response.(35, 36) In the current analysis, HDACs were found to be a top pathway associated moderate/severe ARDS. Histone deacetylases (HDACs) are a class of nuclear enzymes that regulate chromatin structure by deacetylation of histones that repress transcription by preventing transcription factors and RNA polymerase II from binding to DNA. For example, HDACs repress the production of proinflammatory cytokines in alveolar macrophages in chronic inflammatory lung conditions such as chronic obstructive pulmonary disease.(37, 38) Using a multi-omics approach with multiple publicly available datasets identified signaling events mediated by HDAC class I and chromatin modifying enzymes as top pathways associated with ARDS mortality.(39) DNA methylation is another epigenetic modification that emerged in the Reactome analysis distinguishing moderate/severe from mild PARDS. Alterations in DNA methylation sites in ARDS are related to an imbalance in inflammation, immunity, endothelial and epithelial function, and coagulation.(40)

We recently identified gene networks important to the pediatric acute respiratory distress syndrome airway immune response using semi-targeted transcriptomics from primary airway cells and a neutrophil reporter assay.(41) Several of the KEGG pathways were similar between the tracheal aspirate airway cell transcriptomics identified using a Nanostring array, and the whole-blood gene expression analysis herein reported. For example, cytokine-cytokine receptor interactions, human T-cell leukemia virus 1 pathway, and multiple kinase signaling cascades such as, PI3K-Akt signaling pathway MAPK, and JAK-STAT signaling pathways, are present in the airway and whole-blood transcriptome from children with ARDS.

Our study is limited by the *post hoc* analysis focused on finding a systemic transcriptomic signature of children with moderate/severe ARDS. Only one single-center study of pediatric ARDS exists without a non-hypoxemic respiratory failure control group. We attempted to validate our findings using a multi-center pediatric sepsis-triggered acute hypoxemic respiratory failure cohort using the same definitions as in the PARDS gene array cohort without the bilateral infiltrate radiographic findings and the more stringent requirement of two arterial blood gases at least four hours apart with a P/F < 200. These cohorts capture gene expression differences early in the course of acute hypoxemic respiratory failure, and temporal changes in gene expression are not available for analysis. As remarked on in prior attempts to discover an ARDS transcriptomic signature, whole-blood derived gene expression was used that may be different from the airway-specific transcriptomic response. Concomitant changes in metabolites are not readily available for analysis in these cohorts.

In summary, our analysis demonstrated that changes in metabolic pathway involved in energy balance, fundamental cellular processes such as protein translation, mitochondrial function, oxidative stress, immune signaling, and inflammation are differentially regulated early in pediatric ARDS and sepsis-induced acute hypoxemic respiratory failure compared to both healthy controls and to milder acute hypoxemia. Our findings support the hypothesis that differential regulation of metabolic pathways involved in cellular energetics and metabolic reprogramming of the immune response are important mechanisms to consider to further our understanding of the heterogeneity and underlying pathobiology of moderate and severe pediatric acute respiratory distress syndrome.

**References**

1. Matthay MA, Zemans RL, Zimmerman GA, et al: Acute respiratory distress syndrome. *Nat Rev Dis Primers* 2019; 5(1):18

2. Bos LD, Schouten LR, van Vught LA, et al: Identification and validation of distinct biological phenotypes in patients with acute respiratory distress syndrome by cluster analysis. *Thorax* 2017; 72(10):876-883

3. Calfee CS, Delucchi K, Parsons PE, et al: Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. *Lancet Respir Med* 2014; 2(8):611-620

4. Sinha P, Churpek MM, Calfee CS: Machine Learning Classifier Models Can Identify Acute Respiratory Distress Syndrome Phenotypes Using Readily Available Clinical Data. *Am J Respir Crit Care Med* 2020; 202(7):996-1004

5. Sinha P, Delucchi KL, Thompson BT, et al: Latent class analysis of ARDS subphenotypes: a secondary analysis of the statins for acutely injured lungs from sepsis (SAILS) study. *Intensive Care Med* 2018; 44(11):1859-1869

6. Banerjee S, Mohammed A, Wong HR, et al: Machine Learning Identifies Complicated Sepsis Course and Subsequent Mortality Based on 20 Genes in Peripheral Blood Immune Cells at 24 H Post-ICU Admission. *Front Immunol* 2021; 12:592303

7. Sweeney TE, Thomas NJ, Howrylak JA, et al: Multicohort Analysis of Whole-Blood Gene Expression Data Does Not Form a Robust Diagnostic for Acute Respiratory Distress Syndrome. *Crit Care Med* 2018; 46(2):244-251

8. Sauthier MS, Jouvet PA, Newhams MM, et al: Machine Learning Predicts Prolonged Acute Hypoxemic Respiratory Failure in Pediatric Severe Influenza. *Crit Care Explor* 2020; 2(8):e0175

9. Yehya N, Varisco BM, Thomas NJ, et al: Peripheral blood transcriptomic sub-phenotypes of pediatric acute respiratory distress syndrome. *Crit Care* 2020; 24(1):681

10. Sweeney TE, Shidham A, Wong HR, et al: A comprehensive time-course-based multicohort analysis of sepsis and sterile inflammation reveals a robust diagnostic gene set. *Sci Transl Med* 2015; 7(287):287ra271

11. Gautier L, Cope L, Bolstad BM, et al: affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 2004; 20(3):307-315

12. Wu J, Irizarry R, MacDonald J, et al: Gcrma: background adjustment using sequence information. . Version. 2200:3–10. ed. 2012, p^pp R Pack.

13. Leek JT, Johnson WE, Parker HS, et al: sva: Surrogate Variable Analysis. 3.40.0 ed. 2021, p^pp R package

14. Kanehisa M, Furumichi M, Sato Y, et al: KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 2021; 49(D1):D545-D551

15. Kanehisa M, Goto S: KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000; 28(1):27-30

16. Palombo V, Milanesi M, Sferra G, et al: PANEV: an R package for a pathway-based network visualization. *BMC Bioinformatics* 2020; 21(1):46

17. Fabregat A, Jupe S, Matthews L, et al: The Reactome Pathway Knowledgebase. *Nucleic Acids Res* 2018; 46(D1):D649-D655

18. Fabregat A, Sidiropoulos K, Viteri G, et al: Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* 2017; 18(1):142

19. Meinshausen N, Buhlmann P: Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2010; 72(4):417-473

20. Shah RD, Samworth RJ: Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2013; 75(1):55-80

21. Robinson MJ, Krasnodembskaya AD: Therapeutic targeting of metabolic alterations in acute respiratory distress syndrome. *Eur Respir Rev* 2020; 29(156)

22. Izquierdo-Garcia JL, Nin N, Cardinal-Fernandez P, et al: Metabolomic profile of acute respiratory distress syndrome of different etiologies. *Intensive Care Med* 2019; 45(9):1318-1320

23. Izquierdo-Garcia JL, Nin N, Jimenez-Clemente J, et al: Metabolomic Profile of ARDS by Nuclear Magnetic Resonance Spectroscopy in Patients With H1N1 Influenza Virus Pneumonia. *Shock* 2018; 50(5):504-510

24. Rogers AJ, Matthay MA: Applying metabolomics to uncover novel biology in ARDS. *Am J Physiol Lung Cell Mol Physiol* 2014; 306(11):L957-961

25. Viswan A, Ghosh P, Gupta D, et al: Distinct Metabolic Endotype Mirroring Acute Respiratory Distress Syndrome (ARDS) Subphenotype and its Heterogeneous Biology. *Sci Rep* 2019; 9(1):2108

26. Viswan A, Singh C, Kayastha AM, et al: An NMR based panorama of the heterogeneous biology of acute respiratory distress syndrome (ARDS) from the standpoint of metabolic biomarkers. *NMR Biomed* 2020; 33(2):e4192

27. Bos LDJ: Diagnosis of acute respiratory distress syndrome by exhaled breath analysis. *Ann Transl Med* 2018; 6(2):33

28. Bos LDJ, Scicluna BP, Ong DSY, et al: Understanding Heterogeneity in Biologic Phenotypes of Acute Respiratory Distress Syndrome by Leukocyte Expression Profiles. *Am J Respir Crit Care Med* 2019; 200(1):42-50

29. Weiss SL, Cvijanovich NZ, Allen GL, et al: Differential expression of the nuclear-encoded mitochondrial transcriptome in pediatric septic shock. *Crit Care* 2014; 18(6):623

30. Bos LD, Scicluna BP, Ong DY, et al: Understanding Heterogeneity in Biological Phenotypes of ARDS by Leukocyte Expression Profiles. *Am J Respir Crit Care Med* 2019

31. Sinha P, Delucchi KL, McAuley DF, et al: Development and validation of parsimonious algorithms to classify acute respiratory distress syndrome phenotypes: a secondary analysis of randomised controlled trials. *Lancet Respir Med* 2020; 8(3):247-257

32. Genomes KEoGa: Pathways of Neurodegeneration - Multiple Diseases. Available at: <https://www.genome.jp/pathway/hsa05022+5595>. Accessed October 4, 2021

33. Lunnon K, Keohane A, Pidsley R, et al: Mitochondrial genes are altered in blood early in Alzheimer's disease. *Neurobiol Aging* 2017; 53:36-47

34. Tait SW, Green DR: Mitochondria and cell signalling. *J Cell Sci* 2012; 125(Pt 4):807-815

35. Menachery VD, Eisfeld AJ, Schafer A, et al: Pathogenic influenza viruses and coronaviruses utilize similar and contrasting approaches to control interferon-stimulated gene responses. *mBio* 2014; 5(3):e01174-01114

36. Menachery VD, Schafer A, Burnum-Johnson KE, et al: MERS-CoV and H5N1 influenza virus antagonize antigen presentation by altering the epigenetic landscape. *Proc Natl Acad Sci U S A* 2018; 115(5):E1012-E1021

37. Barnes PJ, Adcock IM, Ito K: Histone acetylation and deacetylation: importance in inflammatory lung diseases. *Eur Respir J* 2005; 25(3):552-563

38. Ito K, Ito M, Elliott WM, et al: Decreased histone deacetylase activity in chronic obstructive pulmonary disease. *N Engl J Med* 2005; 352(19):1967-1976

39. Liao SY, Casanova NG, Bime C, et al: Identification of early and intermediate biomarkers for ARDS mortality by multi-omic approaches. *Sci Rep* 2021; 11(1):18874

40. Zhang S, Wu Z, Xie J, et al: DNA methylation exploration for ARDS: a multi-omics and multi-microarray interrelated analysis. *J Transl Med* 2019; 17(1):345

41. Grunwell JR, Rad MG, Stephenson ST, et al: Machine Learning-Based Discovery of a Gene Expression Signature in Pediatric Acute Respiratory Distress Syndrome. *Crit Care Explor* 2021; 3(6):e0431

**Figure Legends**

**Figure 1.** Bar chart with number of genes represented in each KEGG pathway following differential gene expression analysis of pediatric acute respiratory distress syndrome (PARDS) versus healthy controls followed by stability selection using a collinearity threshold of 0.95. **A)** Mild PARDS (P/F = 200-300) from GSE147902 compared with healthy controls from GSE66099. **B)** Moderate/Severe PARDS (P/F < 200) from GSE147902 compared with healthy controls from GSE66099.

**Figure 2.** Bar chart with number of genes represented in each KEGG pathway following stability selection of moderate/severe pediatric acute respiratory distress syndrome (PARDS; P/F < 200) versus mild PARDS (P/F = 200-300) from GSE147902.

**Figure 3.** AUROC and AUPRC curves for different number of genes (12, 21, 48, 69, 92, 185) selected to model moderate/severe pediatric acute respiratory distress syndrome (PARDS; P/F < 200) versus mild PARDS (P/F = 200-300). **A)** AUROC and **B)** AUPRC from GSE147902. **C)** AUROC and **D)** AUPRC from GSE66099.

**Figure 4.** KEGG pathway analysis bar chart with number of genes for the overlapping 185 stability selected ranked genes from GSE147902 and GSE66099 comparing children with a P/F < 200 with a P/F = ≥ 200.

**Table 1. Demographic and Clinical Characteristics of Cohort Participants**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | GSE66099  PaO2/FiO2 | | GSE147902  PaO2/FiO2 | |
| Characteristics | **< 200**  ***n* = 65** | **≥ 200**  ***n* = 99** | **< 200**  ***n* = 26** | **200-299**  ***n* = 70** |
| Age (yr), median (IQR) | 2.4  (0.8, 5.9) | 2.4  (1.15, 5.75) | 6.7  (0.8, 14.1) | 9.4  (2.9, 14.5) |
| Sex, *n* (%)  Female  Male | 29 (45)  36 (55) | 39 (39)  60 (61) | 7 (27)  19 (73) | 31 (44)  39 (56) |
| Race/Ethnicity, *n* (%)  Black  White  Hispanic  Asian/Pacific Islander  Other | 11 (17)  43 (66)  2 (3.1)  2 (3.1)  9 (9) | 21 (21)  73 (74)  2 (2)  1 (1)  4 (4) | 4 (15)  15 (58)  6 (23)  0 (0)  1 (3.8) | 26 (37)\*  24 (34)\*  8 (11)\*  3 (4.3)\*  9 (13)\* |
| Severity of illness,  median (Q1, Q3)  Pediatric Risk of Mortality III  Non-pulmonary organ failures | 16 (11, 22)  1 (1, 6) | 13 (9, 19)  2 (1, 6) | 10 (7, 17)  1 (1, 2.75) | 11 (6, 20)  1 (1, 2) |
| Comorbidities, *n* (%)  Any comorbidity  Malignancy  Immunocompromised  Bone Marrow or Stem Cell Transplant | 21 (32.3)  3 (4.6)  7 (10.8)  2 (3.1) | 43 (44.3)  6 (6.2)  9 (9.3)  4 (4.1) | 21 (81)  6 (29)  8 (31)  5 (19) | 44 (63)  14 (32)  22 (31)  10 (14) |
| Cause of Acute Hypoxemic Respiratory Failureb  Infectious  Fungal  Bacterial  Gram negative  Gram positive  Viral  None detected  Non-infectious | 65 (100)  2 (3.1)  13 (20)  26 (40)  4 (6.2)  20 (32)  NA | 99 (100)  1 (1.0)  21 (21)  25 (25)  6 (6.1)  46 (46)  NA | 1 (3.8)  4 (15)  –a  –a  8 (31)  –a  4 (15.3) | 1 (1.4)  12 (17)  –  –  30 (43)  –  17 (24.3) |
| Cause of Acute Hypoxemic Respiratory Failure  Infectious pneumonia  Non-pulmonary sepsis  Aspiration pneumonia  Trauma  Other | 17 (26.2)  73 (73.8)  NA  NA  NA | 14 (14.1)  85 (85.9)  NA  NA  NA | 13 (50)  9 (35)  1 (3.8)  0 (0)  3 (11.5) | 43 (61)  10 (14)  9 (12.9)  3 (4.3)  5 (7.1) |
| Outcomes, *n* (%)  PICU mortality  Complicated course  28-day VFD, median (Q1, Q3) | 13 (20.0%)  29 (45)  –a | 9 (9.1)  16 (17)  – | 8 (31)  9 (35)  20 (0, 23) | 12 (17)  19 (27)  18 (1, 22) |

\*Wilcoxon rank sum test; Fisher’s exact test; Pearson’s Chi-squared test, *p* < 0.05.

aData is not available for reporting

bCause of pulmonary infection triggering PARDS for GSE147902

**Table 2. Reactome pathways sorted by p-value for the top 200 overlapping stability selected ranked genes from GSE147902 comparing children with a P/F < 200 with a P/F ≥ 200 - 300.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Entities | | | Reactions | |
| Pathway name | **Found** | **Ratio** | **FDR\*** | **Found** | **Ratio** |
| Nucleosome assembly | 8 / 54 | 0.002 | 3.55e-07 | 4 / 4 | 2.96e-04 |
| Deposition of new CEPNA-containing nucleosomes at the centromere | 8 / 54 | 0.002 | 3.55e-04 | 4 / 4 | 2.96e-04 |
| HDACs deacetylate histones | 8 / 73 | 0.003 | 0.002 | 5 / 5 | 3.70e-04 |
| RNA Polymerase I Promoter Opening | 6 / 38 | 0.002 | 0.003 | 1 / 2 | 1.48e-04 |
| Formation of ATP by chemiosmotic coupling | 5 / 23 | 0.001 | 0.003 | 3 / 3 | 2.22e-04 |
| Formation of a pool of free 40S subunits | 9 / 111 | 0.005 | 0.003 | 2 / 2 | 1.48e-04 |
| Recognition and association of DNA glycosylase with site containing an affected purine | 6 / 41 | 0.002 | 0.003 | 2 / 10 | 7.40e-04 |
| Cleavage of the damaged purine | 6 / 45 | 0.002 | 0.004 | 2 / 9 | 6.66e-04 |
| L13a-mediated translational silencing of Ceruloplasmin expression | 9 / 124 | 0.006 | 0.004 | 2 / 3 | 2.22e-04 |
| PRC2 methylates histones and DNA | 6 / 48 | 0.002 | 0.004 | 4 / 4 | 2.96e-04 |
| Depurination | 6 / 48 | 0.002 | 0.004 | 4 / 19 | 0.001 |
| Cleavage of the damaged pyrimidine | 6 / 51 | 0.002 | 0.006 | 1 / 20 | 0.001 |
| Defective pyroptosis | 6 / 56 | 0.003 | 0.008 | 1 / 3 | 2.22e-04 |
| DNA methylation | 6 / 60 | 0.003 | 0.012 | 7 / 7 | 5.18e-04 |
| RNA Polymerase I Promoter Escape | 6 / 64 | 0.003 | 0.015 | 1 / 2 | 1.48e-04 |
| Ribosomal scanning and start codon recognition | 6 / 70 | 0.003 | 0.022 | 2 / 2 | 1.48e-04 |
| GTP hydrolysis and joining of the 60S ribosomal subunit | 11/ 166 | 0.008 | 0.022 | 3 / 3 | 2.22e-04 |
| Packaging of telomere ends | 6 / 71 | 0.003 | 0.022 | 2 / 2 | 1.48e-04 |
| Recognition and associated of DNA glycosylase with site containing as affected pyrimidine | 6/ 76 | 0.004 | 0.03 | 1 / 21 | 0.002 |
| Selenocysteine synthesis | 7/115 | 0.005 | 0.044 | 3 / 7 | 5.18e-04 |
| Depyrimidination | 6 / 85 | 0.004 | 0.047 | 2 / 41 | 0.003 |
| Formation of the ternary complex, and subsequently, the 43S complex | 6 / 86 | 0.004 | 0.047 | 1 / 3 | 2.22e-04 |
| SRP-dependent cotranslational protein targeting to membrane | 7/ 119 | 0.005 | 0.047 | 5 /5 | 3.70e-04 |

\*False Discovery Rate

**Table 3. Reactome pathways sorted by p-value for the top 185 overlapping stability selected ranked genes from GSE147902 and GSE66099 comparing children with a P/F < 200 with a P/F ≥ 200 - 300.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Entities | | | Reactions | |
| Pathway name | **Found** | **Ratio** | **FDR\*** | **Found** | **Ratio** |
| Peptide chain elongation | 12 / 134 | 0.006 | 8.99e-04 | 5 / 5 | 3.70e-04 |
| Selenoamino acid metabolism | 12 / 198 | 0.009 | 0.003 | 5 / 33 | 0.002 |
| Formation of a pool of free 40S subunits | 9 / 111 | 0.005 | 0.003 | 2 / 2 | 1.48e-04 |
| Eukaryotic Translation Termination | 9 / 113 | 0.005 | 0.003 | 3 / 5 | 3.70e-04 |
| Selenocysteine synthesis | 9 / 115 | 0.005 | 0.003 | 2 / 7 | 5.18e-04 |
| SRP-dependent cotranslational protein targeting to membrane | 9 / 119 | 0.005 | 0.003 | 5 / 5 | 3.70e-04 |
| Nonsense mediated Decay independent of the Exon Junction Complex | 9 / 121 | 0.006 | 0.003 | 1 / 1 | 7.40E-05 |
| Eukaryotic Translational Elongation | 12 / 187 | 0.009 | 0.003 | 8 / 9 | 6.66e-04 |
| L13a-mediated translational silencing of Ceruloplasmin expression | 9 / 124 | 0.006 | 0.003 | 2 / 3 | 2.22e-04 |
| Viral mRNA Translation | 10 / 135 | 0.006 | 0.005 | 2 / 2 | 1.48e-04 |
| Translation | 24 / 755 | 0.035 | 0.011 | 52 / 99 | 0.007 |
| GTP hydrolysis and joining of the 60S ribosomal subunit | 9 / 166 | 0.008 | 0.018 | 3 / 3 | 2.22e-04 |
| Ribosomal scanning and start codon recognition | 6 / 70 | 0.003 | 0.019 | 2 / 2 | 1.48e-04 |
| Mitochondrial translation elongation | 8 / 135 | 0.006 | 0.02 | 5 / 8 | 5.92e-04 |
| Mitochondrial translation termination | 8 / 137 | 0.006 | 0.02 | 5 / 5 | 3.70e-04 |
| Mitochondrial translation initiation | 9 / 181 | 0.008 | 0.025 | 3 / 4 | 2.96e-04 |
| Formation of the ternary complex, and subsequently, the 43S complex | 6 / 86 | 0.004 | 0.039 | 1 / 3 | 2.22e-04 |
| Nonsense-Mediated Decay | 11 / 197 | 0.009 | 0.039 | 5 / 6 | 4.44e-04 |
| Nonsense Mediated Decay enhanced by the Exon Junction Complex | 11 / 197 | 0.009 | 0.039 | 4 / 5 | 3.70e-04 |

\*False Discovery Rate