SCIENTIFIC REPORTS
natureresearch

OPEN

# Discovery and predictive modeling of urine microbiome, metabolite and cytokine biomarkers in hospitalized patients with community acquired pneumonia

Joseph F. Pierre[1,3,6]✉, Oguz Akbilgic[2,6], Heather Smallwood[1], Xueyuan Cao[5], Elizabeth A. Fitzpatrick[3], Senen Pena[4], Stephen P. Furmanek[4], Julio A. Ramirez[4] & Colleen B. Jonsson[3]✉

**Pneumonia is the leading cause of infectious related death costing 12 billion dollars annually in the United States alone. Despite improvements in clinical care, total mortality remains around 4%, with inpatient mortality reaching 5–10%. For unknown reasons, mortality risk remains high even after hospital discharge and there is a need to identify those patients most at risk. Also of importance, clinical symptoms alone do not distinguish viral from bacterial infection which may delay appropriate treatment and may contribute to short-term and long-term mortality. Biomarkers have the potential to provide point of care diagnosis, identify high-risk patients, and increase our understanding of the biology of disease. However, there have been mixed results on the diagnostic performance of many of the analytes tested to date. Urine represents a largely untapped source for biomarker discovery and is highly accessible. To test this hypothesis, we collected urine from hospitalized patients with community-acquired pneumonia (CAP) and performed a comprehensive screen for urinary tract microbiota signatures, metabolite, and cytokine profiles. CAP patients were diagnosed with influenza or bacterial (*Streptococcus pneumoniae* and *Staphylococcus aureus*) etiologies and compared with healthy volunteers. Microbiome signatures showed marked shifts in taxonomic levels in patients with bacterial etiology versus influenza and CAP versus normal. Predictive modeling of 291 microbial and metabolite values achieved a + 90% accuracy with LASSO in predicting specific pneumonia etiology. This study demonstrates that urine from patients hospitalized with pneumonia may serve as a reliable and accessible sample to evaluate biomarkers that may diagnose etiology and predict clinical outcomes.**

Community-acquired pneumonia (CAP) is the leading cause of infectious disease-related death and together with Influenza, the eight-leading cause of death in the USA[1]. The annual incidence of CAP worldwide is approximately 5–11 per 1,000 and the estimated annual CAP-associated costs in the US is over 12 billion dollars[2]. Even with

[1]Department of Pediatrics, College of Medicine, University of Tennessee Health Science Center (UTHSC), 425 Translational Science Research Building, 71 S Manassas St., Memphis, TN 28103, USA. [2]Department of Health Informatics and Data Sciences, Parkinson School of Health Informatics and Public Health, Loyola University Chicago, Maywood, IL 60153, USA. [3]Department of Microbiology, Immunology, and Biochemistry, College of Medicine, UTHSC, 801B Molecular Sciences Building, 858 Madison Ave, Memphis, TN 38163, USA. [4]Division of Infectious Diseases, School of Medicine, University of Louisville, Louisville, USA. [5]Department of Acute and Tertiary Care, College of Nursing, UTHSC, Memphis, USA. [6]These authors contributed equally: Joseph F. Pierre and Oguz Akbilgic. ✉email: Jpierre1@uthsc.edu; cjonsson@uthsc.edu

appropriate antibiotic and Supportive therapy, some hospitalized patients with CAP progress to clinical failure and death[3]. Intriguingly, even patients that survive the initial respiratory infection have significantly higher 1, 3, and 5-year mortality rates compared to other chronic diseases (reviewed in[4]). Therefore, there is a critical need to improve treatment and gain a deeper understanding into the factors contributing to short and long-term morbidity and mortality.

Both bacterial and viral pathogens cause CAP and both etiologies are associated with significant mortality[3,5,6]. Prompt identification of the pathogen causing pneumonia is critical for prescribing appropriate therapy. However, the tests necessary to identify the pathogen in blood or bronchoalveolar lavage (BAL) suffer from sensitivity, specificity, cost and availability issues[5,7–9]. Frequently, no pathogen is identified making treatment decisions exceedingly difficult and adversely affecting patients; delays in antibiotic treatment are associated with increased mortality. Additionally, the symptoms of viral and bacterial pneumonia overlap and it is difficult to distinguish between the two based on clinical and radiographic findings[10–13]. Early identification of the etiology is critical for prescribing appropriate treatment; unfortunately, there is no standard diagnostic criterion for distinguishing between viral vs bacterial pneumonia. There have been numerous attempts to identify biomarkers that will distinguish viral versus bacterial pneumonia however none have become part of standardized hospital diagnosis practice.

Many studies have quantified cytokines or eicosanoids in serum as potential biomarkers for pneumonia severity. According to the current paradigm, high cytokine levels produce an exaggerated systemic response—termed the "cytokine storm"—and this dysregulated systemic inflammation drives poor clinical outcomes[14]. However, using cytokines alone as biomarkers for severity has not been well established or incorporated into standard diagnostic criteria. In addition to cytokines, the advent of next generation sequencing has allowed rapid identification of specific microbiome compositions in multiple body sites. While urine was classically considered sterile, recent reports suggest a unique microbiome is detectable under healthy and diseased conditions[15,16]. These microbes, along with the mammalian host, produce a milieu of metabolites. Metabolites are functional outputs of various biological processes and they are an end point that incorporates biological state of the patient (e.g. age and genetic factors) with disease state and external environmental influences (e.g. nutrition and drug treatment) and internal microbiome influences. Metabolites are dynamic analytes present in biological fluids, including urine, producing unique signatures that are readily detected with mass spectrometry and are currently being studied in patients with pneumonia[17–20]. While these studies have made great strides in elucidating underlying mechanisms of pneumonia and determined some promising biomarkers for specific causative agents, none have combined metabolomics and microbial-omics nor expanded from the discovery phase to algorithmic predictive modeling.

The metabolome and microbiome[21] are also unique portraits of the individual patient as they are not only influenced by genetics and disease state but also by the environment, nutrition, age, and lifestyle[19,22–31]. Thus the microbiome and metabolome are very different from transcriptomic or proteomic biomarkers. Unlike biomarkers that individually vary with health status, meta-biomarkers by definition are so co-related and interwoven that they produce a precise disease signature that evolves with the pathophysiological state of the individual[31–34]. Thus meta-biomarkers are less likely to produce false identifications as they do not rely on changes to a single analyte[31,34]. Temporal and dynamic changes to the microbiome and metabolome along with their connection to phenotype and meta-biomarker characteristics leads us to select these quantifiable components in the urine as indicators of CAP. Therefore, here we set out to identify whether unique signatures of patients with CAP could be identified in urine by incorporating cytokines, the microbiome, and metabolites in our predictive models.

## Results

### Patient characteristics.

Patients were selected from the University of Louisville Pneumonia Study (ULPS) biorepository with IRB approval. The ULPS was a population-based cohort study of 7,449 unique patients hospitalized with CAP between June 1, 2014 and May 31, 2016. From this biorepository, we selected 30 urine samples from patients, ten each with a confirmed etiology of influenza A virus (IAV), *Streptococcus pneumoniae,* or *Staphylococcus aureus* infection. Supplementary Table 1 shows a comparison of clinical data for patients hospitalized with CAP for each etiologic agent. We also selected urine samples from ten healthy volunteers from the University of Louisville Infectious Diseases biorepository. These volunteers were majority female (90%) and aged 28–58, in contrast to the clinical groups which represented by (30–50%) females and aged 44–75.

### Urine cytokines.

We interrogated 34 cytokines in the urine of healthy controls, patients infected by influenza, *S. pneumo,* or *S. aureus*. Out of 34 cytokines tested, we detected 17 cytokines in the urine and 11 of those cytokines showed differential presentation among the four groups of participants (Table 1). Pneumonia caused by *S. aureus* differed from the healthy controls for all 11 cytokines and in general the level of cytokines detected in influenza patients was consistently lower than the levels in patients with bacterial pneumonia and only differed from healthy volunteers for IFNγ, IL-6, IL-18, eotaxin, IP-10 and MCP-1. Four of the cytokines demonstrated a significant difference between the 3 types of pathogens; IFNγ ($P = 0.005$), IL-18 ($P = 0.0052$), MCP-1 ($P = 0.0029$) and SDF-1 ($P = 0.0451$). The remaining cytokines that were present in the urine but did not differ from each other or healthy volunteers were IL-1β, IL-1α, IL-1RA, IL-22, IL-27 and IL-8. We did not detect IL-12p70, IL-13, IL-2, IL-5, GM-CSF, IL-10, IL-17A, IL-21, IL-23, IL-9, IFNα, IL-31, IL-7, TNFβ, MIP-1β, RANTES or TNFα in the urine in any of the groups.

### Urine microbiome.

Total DNA was extracted from urine samples for quantitative PCR of 16S copy numbers, which demonstrated that healthy volunteers exhibited higher bacterial DNA copy numbers compared with bacterial or viral pneumonia patient urine samples (Fig. 1A); potentially influenced by the initial administration of antibiotics that all cause pneumonia patients receive upon admission. Similarly, compared with healthy

| Cytokine | Pathogen | Group[a] (pg/mg creatinine) | Level P[b] | P value[c] |
|---|---|---|---|---|
| IFNγ | *S. aureus* | 2,935 (1718.5–8,199.75) | 2.00E−04 | 0 |
| | *S. pneumoniae* | 1,060 (772.5–1932.25) | 0.001 | |
| | Influenza | 995.5 (780.5–1,252) | 0.0013 | |
| | Healthy volunteers | 282 (199–513.75) | | |
| IL-4 | *S. aureus* | 560 (372.5–758.5) | 0.0041 | 0.0218 |
| | *S. pneumoniae* | 3.5 (0–473.25) | 0.1433 | |
| | Influenza | 77 (0–200) | 0.1692 | |
| | Healthy volunteers | 0 (0–0) | | |
| IL-6 | *S. aureus* | 1,404 (780.25–8,821.5) | 0.0046 | 0.0116 |
| | *S. pneumoniae* | 813 (273.5–1862.5) | 0.0263 | |
| | Influenza | 637 (393.5–1,005) | 0.0386 | |
| | Healthy volunteers | 32 (0–251.25) | | |
| IL-18 | *S. aureus* | 12,111 (7,826.25–22,594.5) | 2.00E−04 | 0 |
| | *S. pneumoniae* | 3,680.5 (2,578.5–6,105.75) | 0.0016 | |
| | Influenza | 3,158 (2,482.75–4,977.5) | 0.0016 | |
| | Healthy volunteers | 132.5 (0–1,377) | | |
| IL-15 | *S. aureus* | 441 (0–1,145.75) | 0.0403 | 0.0318 |
| | *S. pneumoniae* | 262 (0–1,150.25) | 0.0503 | |
| | Influenza | 0 (0–0) | 1 | |
| | Healthy volunteers | 0 (0–0) | | |
| Eotaxin | *S. aureus* | 873 (236.5–1,354) | 0.001 | 0.0031 |
| | *S. pneumoniae* | 364 (128.25–849.25) | 0.0448 | |
| | Influenza | 280.5 (213–336.75) | 0.0028 | |
| | Healthy volunteers | 91.5 (70.75–139) | | |
| Gro-α | *S. aureus* | 1,388.5 (342.5–3,895.5) | 0.0057 | 0.0313 |
| | *S. pneumoniae* | 273 (37–4,501.5) | 0.2694 | |
| | Influenza | 304 (182–342) | 0.0535 | |
| | Healthy volunteers | 123.5 (88.5–204.5) | | |
| IP-10 | *S. aureus* | 1,139.5 (849.25–1655.5) | 1.00E−04 | 0.0016 |
| | *S. pneumoniae* | 866.5 (323–4,605.25) | 0.0113 | |
| | Influenza | 420.5 (373.75–1696.75) | 0.0029 | |
| | Healthy volunteers | 130 (101.5–278.25) | | |
| MCP-1 | *S. aureus* | 62,335 (30,133–126,803) | 0 | 0 |
| | *S. pneumoniae* | 28,888.5 (20,015–50,736) | 2.00E−04 | |
| | Influenza | 15,638 (8,220.5–20,452.5) | 0.0147 | |
| | Healthy volunteers | 6,788 (4,850–8,002) | | |
| MIP-1α | *S. aureus* | 1665 (1,351.25–5,977.25) | 0.0147 | 0.047 |
| | *S. pneumoniae* | 1,406.5 (136.75–3,459) | 0.4495 | |
| | Influenza | 1,020.5 (780.5–1,068) | 0.9705 | |
| | Healthy volunteers | 767 (579.5–1,176) | | |
| SDF-1 | *S. aureus* | 18,199 (17,012–27,997.25) | 0.0021 | 0.0113 |
| | *S. pneumoniae* | 11,854.5 (7,832.5–25,720.25) | 0.0524 | |
| | Influenza | 10,195.5 (7,984.75–12,298.75) | 0.0892 | |
| | Healthy volunteers | 6,293 (2,706–8,618.75) | | |

**Table 1.** Cytokines detected in the urine of CAP patients. Cytokines were measured using the ProCarta Plex multiplex immunoassay and normalized to creatinine; normalized values were used for analysis. [a]The data are represented as median (interquartile range) of n = 10 samples/group. [b]Comparison of median cytokine value to healthy volunteers. [c]Comparison of median cytokine value among the four groups.

volunteers, pneumonia patients tended to have elevated alpha diversity, assessed by Shannon index (Anova: F = 2.5, P = 0.076) (Fig. 1B) and significantly elevated Evenness (Anova: F = 3.3, P = 0.03) (Fig. 1C). Relative abundanced of microbiome taxonomic composition across individuals displayed taxonomic signatures of pneumonia, with elevated phyla Proteobacteria and Chloroflexi and fewer Firmicutes compared with healthy controls (Figs. 1D, S1A). Genera level taxa are shown in Fig. 1E across individuals. Principal coordinate analysis (PCoA) of beta diversity assessed by Bray–Curtis broadly demonstrated clustering of pneumonia patients compared with healthy volunteers (Fig. 2A). Redundancy Analysis (RDA) further demonstrated similar distinct cluster-

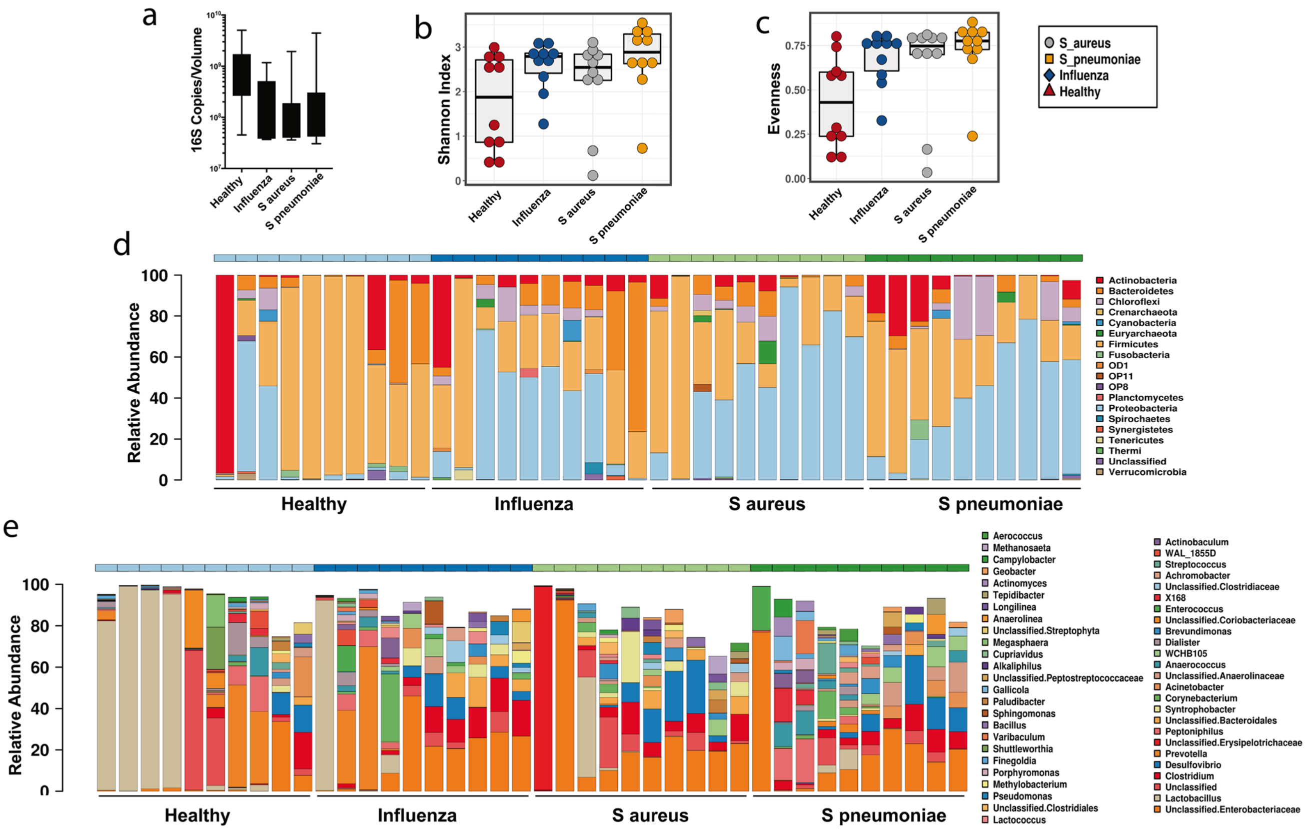

**Figure 1.** Urine microbiome alpha diversity of taxonomic analysis. (**a**) 16S copy numbers detected per ml of urine. (**b**) Shannon and (**c**) evenness indexes for assessment of microbiome alpha diversity. Taxonomic community structure of each patient at the phylum (**d**) and genus (**e**) levels. N = 10/group.

ing between healthy and influenza samples, while patients infected with bacterial pathogens, *S. aureus* and *S. pneumoniae*, clustered even more distinctly from healthy volunteer samples (RDA significance: Variance = 12.74, F = 1.13, P = 0.05) (Fig. 2B). A network analysis of detected taxa also demonstrated more tightly clustering of taxa found in healthy volunteer associated (Fig. 2C). Specific taxa associated with healthy volunteers included the genus *Lactobacillus* within Firmicutes, while pneumonia patients demonstrated elevated levels of the class *Gammaproteobacteria,* family *Enterobacteriaceae,* and the genera *Clostridium* and *Streptococcus* (Figs. 2D; S1B, S2A,B). Hierarchical taxonomic composition for each patient group are summarized in Fig. S1.

To identify specific taxanomic differences between groups, we further employed linear discriminate analysis of effect size (LEfSe) between experimental groups. Initially, we compared all 30 patients with CAP to healthy volunteers (Fig. S3A). At the phylum level, Proteobacteria was identified as significant in pneumonia samples while *Synergistetes* was identified in healthy controls based on LDA scores. At the genus level, *Clostridium* and *Sutterella* were identified in case with CAP while *Lactobacillus, Prevotella, Magasphaera, Dorea, Vibrio*, and *Coprococcus* were the most significantly enriched. Cladogram projection of these differences demonstrated the CAP case samples clustered primarily within the phyla Proteobacteria, while the healthy volunteers were more taxonomically distributed (Fig. S3B). The analysis was repeated after regrouping samples based on viral vs bacterial pathogen, which showed changes at the order level, where *Bifidobacteriales* was abundant in healthy controls, *Enterobacteriales* was abundant in bacterial pneumonia samples, and *Sphingomonadales* was abundant in influenza samples (Fig. S3C). A final regrouping determined comparisons of healthy vs *S. aureus* and *S. pneumoniae* (Fig. S3D) and healthy vs influenza (Fig. S3E), where healthy samples consistently displayed greater levels of the family Rikenneliaceae and the order Bifidobacterium. Interestingly, the level of detectable *Streptococcus* was most elevated in *S. pneumoniae* case samples, while *Syntrophobacter* and *Delftia* were most elevated in patients with *S. aureus* (Fig. S2B).

**Urine metabolites.** Metabolites were extracted from 50 μl of urine and subjected to ultra-high-performance liquid chromatography coupled high resolution mass spectrometry (UPLC-HRMS) 87 known metabolites were detected in the forty urine samples and identified using known masses (± 5 ppm) and retention times ($\Delta \leq 1.5$ min). Creatinine is considered the best internal standard to correct for urine volume variations as its rate of elimination is independent of urine flow and urine volume and creatinine concentration are inversely proportional[35, 36]. Thus, to ensure that observations were directly comparable peak intensity was normalized to creatinine. Then these data were compared to unnormalized data to make sure there was no masking of biologically relevant changes by normalization (DNS). As is the convention in metabolomics we first used unsupervised multivariate statistical analysis to determine the dataset structure and relationships between groups.

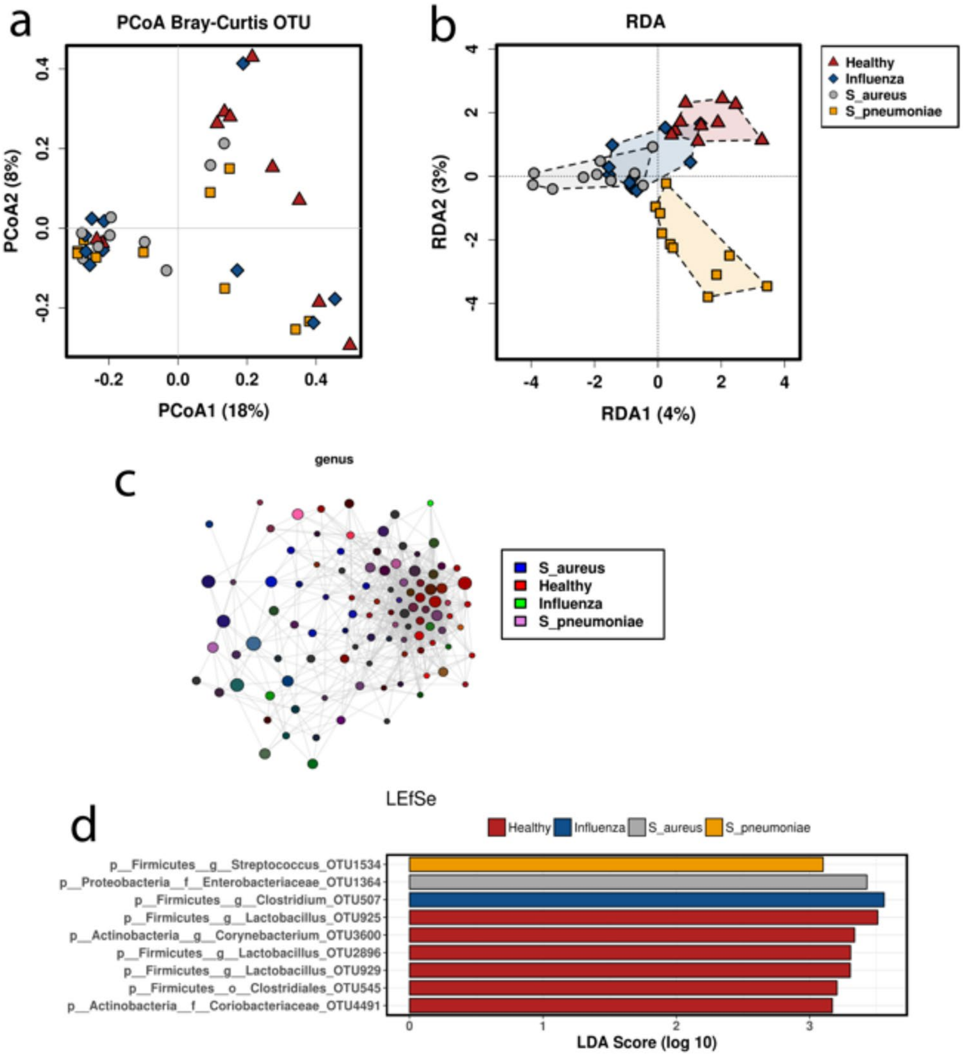

**Figure 2.** Urine microbiome beta diversity, network clustering, and LEfSe. (**a**) Principal component analysis of Bray–Curtis beta diversity of urine OTUs. (**b**) Redundancy analysis of urine OTUs. (**c**) Network analysis of genus detected in urine, color coded by group. (**d**) Linear Discriminant Analysis of Effect Size (LEfSe) of OTUs enriched in each experimental group. N = 10/group.

To evaluate the group trends, sample uniformity and identify potential outliers, multivariant principal component analysis (PCA). The variation were explained by F1 and F2 with a cumulative percent variability of 78.56% spread among the patient groups (i.e. Healthy 53.5 and 6.1, IAV 13.8 and 70.0, *S. aureus* 8.6 and 22.4, and *S. pneumoniae* 30.0 and 1.2 percent per F1 and F2 respectively). Adding a third component marginally increased the cumulative percent variability to 85.96%. The two component PCA analysis shows good separation between CAP patients and the healthy group (Fig. 3, circles). Likewise, the high-risk classes (IV–V) and low risk (I–III) centroids showed clear separation (Fig. 3, squares). We then used unsupervised clustering of both metabolites and individuals, they were clustered independently using k-means clustering followed by ascendant hierarchical clustering based on Euclidian distances. The data matrix's was rearranged according to the corresponding clustering with spatial relationship proportional to similarity among patient samples or metabolites (Fig. 4). These clusters were also represented via a dendrogram displayed vertically for metabolites and another horizontally for patients. We find the healthy volunteers centered and groups nicely together (red) as did the IAV (Blue) while the bacterial pneumonia samples were interspersed together (gray and orange) (Fig. 4). Consistent with the PCA analysis the high-risk groups tended to be close together on the far left or right (Fig. 4, brown bars).

Next, we employed a simple one-way ANOVA with Tukey's honestly significant difference test (Tukey's HSD) with Benjamini–Hochberg post hoc correction (XLSTAT) to identify 6 metabolites with significant differences among patient groups (Table 2). Adenosine 5′-phosphosulfate (APS) was the most significant metabolite with differential concentration based on pneumonia, it was significantly higher in the urine of healthy volunteers. In humans, all APS is converted to 3′-phosphoadenosine 5′-phosphosulfate (PAPS) for the sulfonation of glycosaminoglycans, proteins, peptides, lipids, bile acids, xenobiotics and steroids[37–39]. Guanidoacetic acid was also significantly higher in healthy volunteers and is a precursor to creatine, metabolite in the Urea cycle as well as metabolism

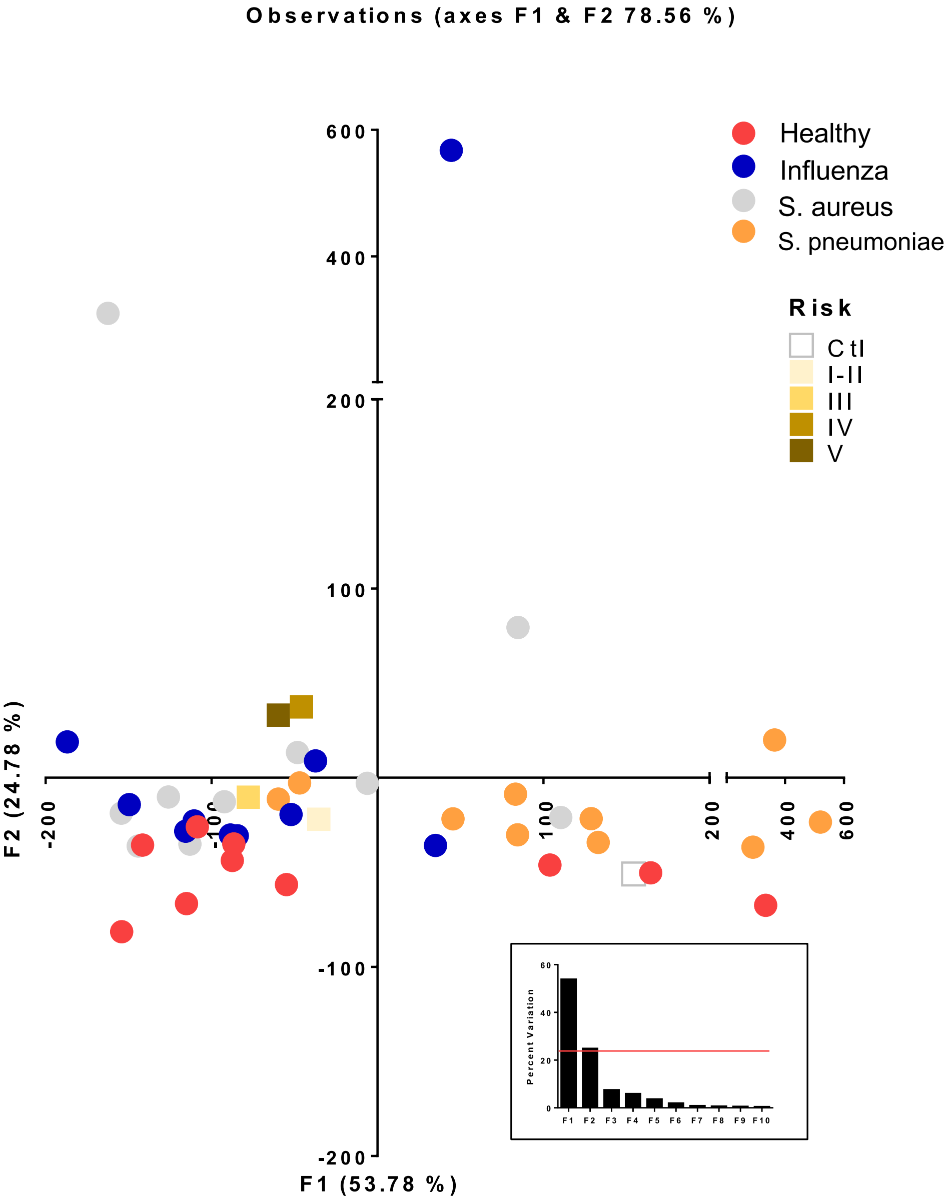

**Figure 3.** Principal component analysis of urine metabolites. Metabolites were extracted from 50 µl urine and subjected to UPLC–HRMS metabolomics analysis three times per sample. Metabolites were manually identified and integrated using known masses (± 5 ppm mass tolerance) and retention times ($\Delta \leq 1.5$ min). Peak intensity was normalized to creatinine followed by unsupervised multivariant principal component analysis (PCA) resulting in F1 and F2 with a cumulative percent variability of 78.56% Each circle represents the average of a patient and the centroids of the corresponding risk groups are represented by squares.

of amino groups of several amino acids including glycine, serine, threonine, arginine and proline. 2,3-dihydroxybenzoate is a conjugate base of 2,3-dihydroxybenzoic acid that is increased after consumption of nutrients (e.g. cranberry juice) or aspirin and is also a biomarker of OH radicals[40, 41]. Succinate was significantly decreased in patients with pneumonia in our studies as well as two other metabolite profiling studies of pneumonia from human pleural fluid and mouse urine infected with *S. pneumoniae*[18, 42]. We found citrate and succinate, metabolites related to the citric acid cycle, to be significantly reduced in all three groups with CAP (Table 2). Reduced citrate levels have previously been reported in plasma of patients with pneumonia and in mouse urine[18, 27]. Likewise, Adamko et al. observed decreases in both citrate and succinate in urine from children with bacterial and viral respiratory infections[18, 27, 43]. Conversely, we found uridine to be significantly increased in the urine of all patients with pneumonia. This is in keeping with previous reports of uridine transiently increasing in the lung and BAL fluid of mice with viral pneumonia from influenza[42]. It is worth noting these are highly abundant analytes whose values are relative to the peak intensity of creatinine in each sample (creatinine mean across samples was 2.753e + 009). Taken together these metabolites represent likely candidates for including among the signature biomarkers of pneumonia.

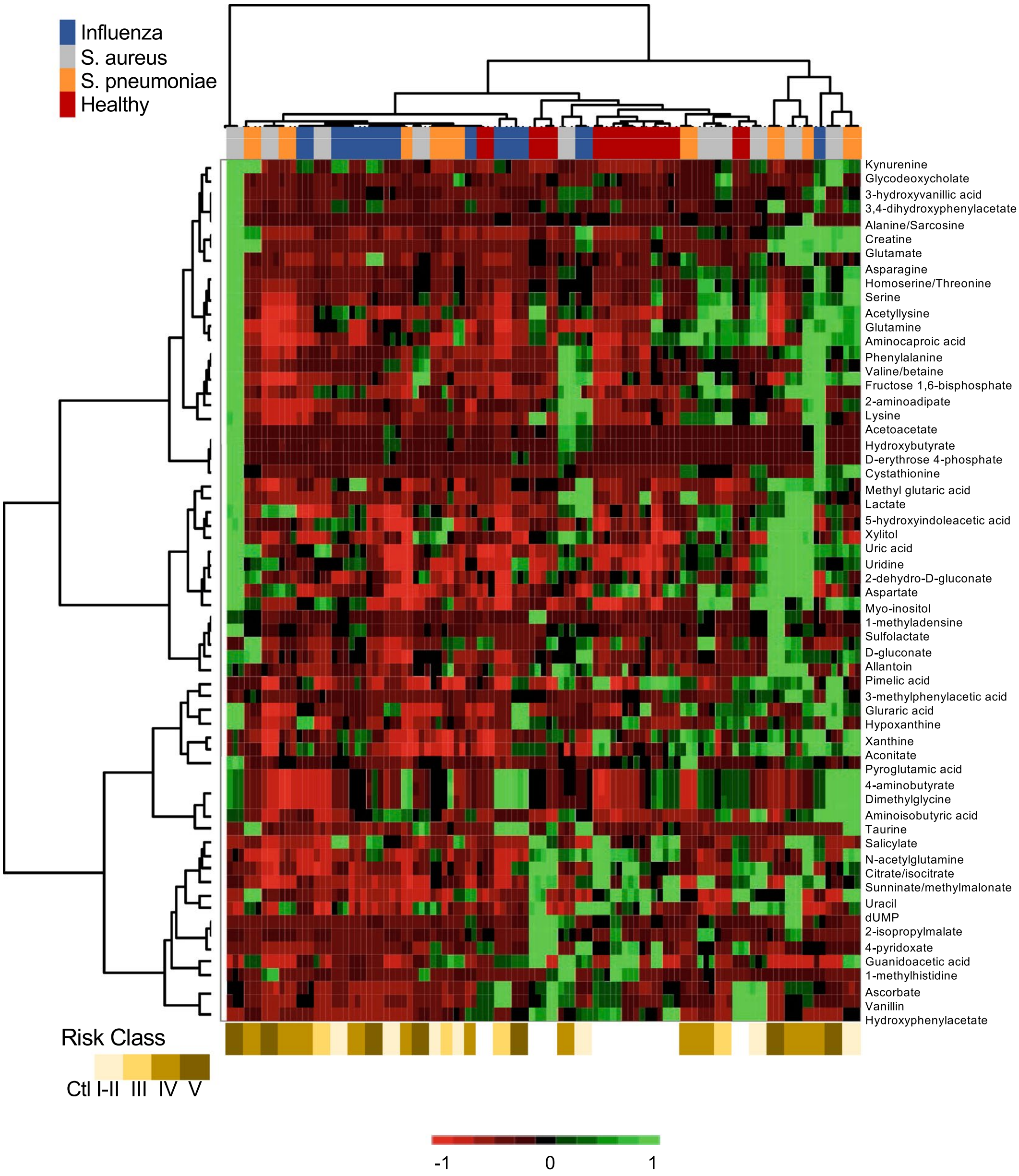

**Figure 4.** Comparison of urine metabolites by patient and risk group. Metabolites were K-means clustered followed by ascendant hierarchical clustering based on Euclidian distances with twenty-one metabolites excluded (0.25 < std dev). Metabolite clusters were also represented via a dendrogram displayed vertically for metabolites and another horizontally for patients. The data values of the permuted matrix were replaced by corresponding color intensities based on interquartile range with color scale of red to green through black resulting in a heat map. Patient identifiers and risk categories were replaced by color bars. Color bars on the top of the graph denote patient groups and bottom risk class.

| Features | P-value | *S. pneumoniae* | Influenza | *S. aureus* | Healthy volunteers |
|---|---|---|---|---|---|
| Adenosine 5′-phosphosulfate | 0.0003 | 0.201 (a) | 0.100 (a) | 0.116 (a) | 0.495 (b) |
| Guanidoacetic acid | 0.0216 | 0.154 (a) | 0.193 (a) | 0.104 (a) | 0.949 (b) |
| 2,3-Dihydroxybenzoate | 0.0216 | 0.571 (a) | 0.642 (a) | 0.908 (a) | 2.998 (b) |
| Succinate/methylmalonate | 0.0216 | 3.297 (a) | 1.939 (a) | 2.513 (a) | 6.613 (b) |
| Citrate/isocitrate | 0.0216 | 191.125 (a) | 145.455 (a) | 130.248 (a) | 363.235 (b) |
| Uridine | 0.0216 | 7.469 (bc) | 5.286 (ab) | 8.888 (c) | 3.988 (a) |

**Table 2.** Metabolites differentially observed between groups. Different letters denote significant differences between groups ($P < 0.05$).

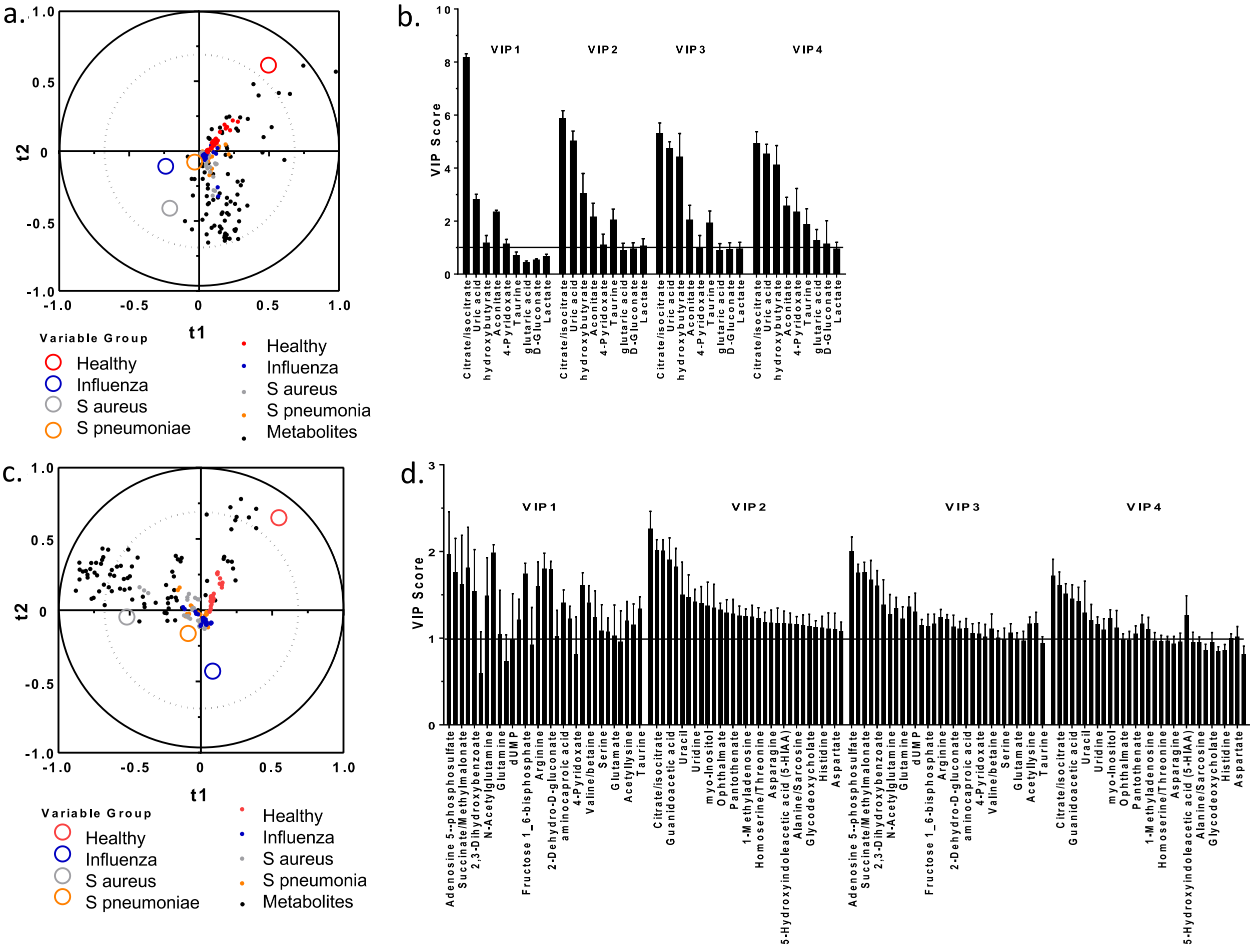

**Figure 5.** Identification of metabolites of import. (**a**) Four component partial least squares discriminant analysis (PLS-DA) was used to identify metabolites that reveal a clear separation of the healthy and pneumonia patient groups. (**b**) The index values of the Variable Importance in Projection (VIP) from the PLS-DA were then used to identify metabolites with VIP scores over one. (**c**) Revised PLS-DA (PLS-DA$^{VCR}$) by first centering and reducing the explanatory variables before starting the PLS-DA calculations. The quality of the PLS-DA$^{VCR}$ was improved (i.e. $Q^2$ cumulative 0.083–0.378). (**d**) VIP scores obtained from PLS-DA$^{VCR}$ that were over one.

We applied a supervised four component partial least squares discriminant analysis (PLS-DA) to distinguish between patient groups and identify differentially expressed variables. The correlation map of the first two components reveals a clear separation of the healthy individuals and group (solid and open grey circles respectively) from the pneumonia patients (Fig. 5A). The index values of the Variable Importance in Projection (VIP) from the PLS-DA were then used to identify 9 metabolites with VIP scores over one (Fig. 5B). However, the overall fit of this model was not robust (i.e. low $Q^2$ values), indicating the quality of the fit varies a lot depending on the metabolite. Likewise, the $R^2$ values were around 0.3 suggesting the components generated by the PLS regression did not summarize either the X or Y variables well. Thus, we revised this analysis by first centering and reducing the explanatory variables before starting the PLS-DA calculations (PLS-DA$^{VCR}$). The quality of the PLS-DA$^{VCR}$ was improved (i.e. $Q^2$ cumulative 0.083–0.378). While the $Q^2$ value is positive, thus has predictive relevance, it remains somewhat low suggesting the quality of the fit of this model varies a lot depending on the metabolite. The PLS-DA$^{VCR}$ also improved the regression's ability to summarize both the X and Y variables (i.e. $R^2Y$ 0.552 and $R^2X$ 0.483) resulting in better separation of pathogen groups (Fig. 5C). However, this produced a large number of metabolites, 35% of those identified, with VIP scores above 1 (Fig. 5D).

**Predictive modeling.** There were 291 variables including two demographics such as gender and age and 185 OTUs detected in urine samples, 17 cytokines, and 87 metabolites of 40 subjects. Note that we included OTUs that were observed for at least two subjects. First, we implemented multi-class classification with 5-folds cross validation to distinguish between four subject categories using a total of 291 predictors. However, none of the machine learning model provided desirable accuracy (all <47.5%). While none of the 6 metabolites identified by Tukey HSD failed the Dunnett's test, this method overlooked 12 metabolites that passed (Supp. Table 2; Fig. S4A). Given the post hoc correction method was for false discovery rate, it is not surprising that this expression analysis resulted in no Type I error but high levels of Type II Errors (Supp. Table 3). The initial PLS-DA

| | | Predicted | | |
|---|---|---|---|---|
| **Model 1** | **Categories** | **Healthy volunteers** | ***S. aureus* + *S. pneumoniae* + Influenza** | **Accuracy (%)** |
| Actual | Healthy volunteers | 9 | 1 | Specificity = 90.0 |
| | *S. aureus* + *S. pnemoniae* + Influenza | 2 | 28 | Sensitivity = 93.3 |
| | Predictive value (%) | 81.8 | 96.7 | Overall = 92.5 |
| | | **Predicted** | | |
| **Model 2** | **Categories** | ***S. aureus* + *S. pneumoniae*** | **Influenza** | **Accuracy (%)** |
| Actual | *S. aureus* + *S. pnemoniae* | 20 | 0 | Specificity = 100.0 |
| | Influenza | 0 | 10 | Sensitivity = 100.0 |
| | Predictive value (%) | 100.0 | 100.0 | Overall = 100.0 |
| | | **Predicted** | | |
| **Model 3** | **Categories** | ***S. aureus*** | ***S. pnemonia*** | **Accuracy (%)** |
| Actual | *S. aureus* | 10 | 0 | Specificity = 100.0 |
| | *S. pnemoniae* | 0 | 10 | Sensitivity = 100.0 |
| | Predictive Value (%) | 100.0 | 100.0 | Overall = 100.0 |

**Table 3.** 5-Folds cross-validation LASSO model performances.

| | Predicted | | | | |
|---|---|---|---|---|---|
| | **Healthy volunteers** | **Influenza** | ***S. pnemoniae*** | ***S. aureus*** | **Accuracy (%)** |
| **Actual** | | | | | |
| Healthy volunteers | 8 | 0 | 1 | 1 | 80.0 |
| Influenza | 0 | 9 | 1 | 0 | 90.0 |
| *S. pnemoniae* | 0 | 0 | 10 | 0 | 100.0 |
| *S. aureus* | 0 | 0 | 3 | 7 | 70.0 |
| Predictive value (%) | 100.0 | 100.0 | 66.7 | 87.5 | Overall = 85.0 |

**Table 4.** 5-Folds cross-validation ensemble method performance.

identified nine metabolites with VIP > 1, of these only citrate and taurine showed significant differences (Supp. Table 2 and Fig. S4A). Further, the PLS-DA analysis produced the most, 16, Type II errors (Supp. Table 3). The PLS-DA$^{VCR}$ analysis identified 31 metabolites with VIP > 1 (Supp. Table 2). Nineteen of the metabolites identified with PLS-DA$^{VCR}$ analysis passed the individual analysis thereby improving the Type II errors when compared to the PLS-DA. However, it misidentified 13 metabolites resulting in the largest number of Type I errors of any of the models (Supp. Table 3). Lasso Model 1 identified seven potential biomarkers that distinguished healthy from CAP patients, all of which passed the Dunnett's test (Supp. Table 2). Model 1 also produced the least errors with more positive identifications (Supp. Table 3). Thus Lasso model 1 did not require reducing explanatory variables as was done in the PLS-DA$^{VCR}$ analysis, that resulted in the greatest level of Type I errors, while producing the least Type I or II errors. It is important to note that in the first iteration, our model pulled out several predictions that are significantly altered and represent abundant metabolite markers in the urine.

We then implemented LASSO logistic regression with fivefold cross-validation and a total of 13 OTUs, 2 cytokines, and 13 metabolites were found to be discriminating between different subject categories as listed in Supp. Table 4.

Model 1 identified one OTUs and three metabolites to distinguish heathy subjects from *S. aureus*, *S. pneumoniae*, and Influenza (Model 1 providing AUC with 95% CI of 0.98; 0.94–1.00). Model 2 identified six OTUs, two cytokines, and four metabolites to distinguish Influenza from *S. aureus* and *S. pneumoniae* (Model 2 providing AUC of 1.00). Model 3 identified six OTUs, one cytokine, and six metabolites to distinguish *S. aureus* from *S. pneumoniae* (Model 3 providing AUC of 1.00). The confusion matrices with performance indicators for each model is presented in Table 3.

In recursive implementation of LASSO regression in three steps, we identified a total of 28 OTUs, cytokines, and metabolites to classify subjects into their actual categories. However, model 1 assumes the subject is not healthy and model 3 assumes the subjects not healthy nor influenza. Therefore, to develop a model that can be implemented on subjects without any assumption on their status, by using these selected 28 predictors, which is significantly smaller than the original dataset with 291 predictors, we readdressed the multi-class classification problem. We implemented various machine learning algorithms and found that Ensemble Method (Ensemble Method: Subspace, Learner Type: Discriminant, Number of Learners: 30, Subspace Dimension: 13) provided the highest overall classification accuracy of 85.0% (Table 4). Note that the parameter setting of the final model was fixed across folds and there was no parameter optimization implemented.

| | Predicted | | | |
|---|---|---|---|---|
| | Healthy volunteers | Influenza | Bacteria | Accuracy (%) |
| Actual | | | | |
| Healthy volunteers | 8 | 0 | 2 | 80.0 |
| Influenza | 0 | 9 | 1 | 90.0 |
| Bacteria | 0 | 0 | 20 | 100.0 |
| Predictive value (%) | 100.0 | 100.0 | 87.0 | Overall = 92.5 |

**Table 5.** Performance of the final model when *S. aureus* and *S. pneumoniae* are merged into one group.

Table 4 shows that most of the classification error is due to misclassification between two bacterial groups. When we merge two bacterial groups (Table 5), we found that one can distinguish between healthy, influenza and bacterial categories with very high accuracy of 92.5% and with perfect positive predictive values for healthy and influenza subjects.

## Discussion

The search for accurate predictors of infection and disease remains an important frontier in the era of omics and personalized medicine. In the setting of CAP, which is the leading cause of infectious related death in the United States, various methods have been employed using serum and other clinical samples to confirm and determine severity of respiratory infection, including the quantification of cytokines and eicosanoids. However, these approaches have achieved limited sucess and are not widely implemented for patient care. For the clinician, even the basic differential diagnosis between viral vs bacterial pathogens remains difficult in CAP cases due to the overlapping symptomatic presentations. Considering the dynamic nexus of host immunological, metabolic, and microbial networks, we moved beyond the search for single or limited numbers of biomarkers by instead comprehensively profiling urine cytokines, the microbiome, and the metabolome in samples collected from newly admitted pneumonia patients with either influenza or bacterial pathogens compared with healthy volunteers. Urine was chosen as a non-invasive sample since it is readily obtainable in the in-patient and out-patient setting.

There have been few studies to determine the utility of measuring cytokine markers in urine as potential biomarkers of infectious disease states. Out of the 34 cytokines we measured in urine, 11 were significantly different between groups. Patients with bacterial pneumonia exhibited the greatest elevation and number of cytokines in urine that differed from healthy controls; *S. aureus* pneumonia patients differed in all 11 cytokines with IL-4, IL-15, Gro-α, MIP-1α and SDF-1 uniquely elevated in those patients. Whereas patients with influenza exhibited the lowest levels of cytokines detectable in urine that differed compared to healthy controls. In several studies IL-6 levels in the serum have been found to correlate with either bacterial infections or disease severity in patients with CAP[44–47]. Our results also detected increased IL-6 in the urine which was elevated in all pneumonias compared to healthy controls. However, based on the predictive modeling, IL-15 and IL-18 in combination with the metabolome and microbiome data may be more useful for distinguishing bacterial vs viral pneumonias.

Analysis of the urine microbiome demonstrated a complex community under both healthy and infectious states, including the presence of slow growing *Lactobacillus* and *Corynebacterium*, consistent with recent reports on urine microbiome. While urinary tract bacterial populations were historically overlooked outside of the context of infection, next generation sequencing techniques enabled culture independent insights into these communities. Recent findings demonstrate major shifts in the urine microbiome community under diseases of the urinary tract, including[48] urolithiasis and certain cancers, however investigation of the urine microbiome as signatures of disease at distant body sites has not been employed. We observed elevated diversity and evenness in urine samples from pneumonia patients compared with healthy volunteers, including consistently elevated *Enterobacteriaceae* and *Clostridium*. Furthermore, community composition data suggested greater dissimilarity in the urine microbiome in patients with bacterial pneumonia than in those with influenza (Fig. 2A–C).

Similar to the microbiome, the metabalome of CAP patients clustered from healthy volunteers (Figs. 3, 4), suggesting a divergence in metabolite profiles under an infectious state. Drivers of this differential clustering included loss of numerous metabolites, including citrate/isocitrate, succinate, guanidoacetic acid, *N*-acetylglutamine, among others, compared with healthy volunteers. No metabolites were specific to influenza infection alone compared with the other groups. However, methyladenosine, uridine and 2-dehydro-D-gluconate were elevated under bacterial infection compared with influenza and healthy controls. These divergent profiles could be useful in determining pathogen kingdom.

Since the microbiome and metabolite profiles are influenced by the environment, nutrition, age, and lifestyle of the host, in addition to genetics, these concatenated profiles provide a unique snapshot of individual health[19, 22–31] Indeed, while these complex profiles can be examined independently, changes in the collective abundance patterns of metabolites and microbes may indicate deeper homeostatic disturbances, which may be reflected through changes in interleukin signaling. The membership of the bodies microbial communities have dynamic interconnected relationships with one another and the host that change under states of disease and stress. Therefore, the microbiome and metabolome complement to serve as personalized readout of individual health. The ability to detect rapid measurable changes in these profiles in response to challenges, such as infection, would be a novel systems biology approach to personalized medicine. For instance, the components of the metabolome and microbiome are physically or stoichiometrically co-related, leaving precise abundance patterns that may accurately reflect discreet pathophysiological states[23, 34]. Utilization of meta-biomarkers, such as the microbiome and metabolome, would represent a distinct shift away from the majority of clinical biomarkers currently in use, even in the era of genomics, transcriptomics, and proteomics[23, 34].

After combining all urine meta-biomarkers, totaling 385 data points, we performed machine learning models by implementing LASSO logistic regression in three unique models. Model 1 aimed to distinguish healthy subjects from pneumonia; Model 2 to distinguish between bacterial (*S. aureus* and *S. pneumoniae*) pathogens from Influenza; and Model 3 to distinguish between *S. aureus* and *S. pneumoniae*. For each predictive model, we implemented a fivefold cross validation process to avoid overfitting. Specifically, the data were split into five distinct folds where 4 folds were used for model testing and the remaining for validation. By repeating this process five times by changing the test fold, we identified a total of 28 predictors, including two cytokines, 13 microbial taxa, and 13 metabolites that provided a predictive power of 92.5% in distinguishing patient groups.

There are several limitations to our study, including the total sample size of 40 individuals. While we were able to detect consistent changes in our meta-biomarkers, larger studies with greater numbers and groups that included other pathogens may improve the resolution of our predictive models in determining unique signatures of pneumonia or other respiratory diseases. Moreover, despite cross-validation yielded high predictive accuracy, there is a need to both validation of the data on a larger cohort and on a more diverse external cohort to be able to claim broader generalizability. Another limitation was related to a characteristic of the clinical standard of care, where all pneumonia patients in this study were placed on antibiotics upon admission to the hospital. Future studies may attempt to compare urine samples collected from individual before and after the implementation of antibiotics. On the other hand, the inclusion of patients with influenza acted as a unique control group for the bacterial groups, since all patients were placed on antibiotics. We observed large perturbations in the meta-biomarkers in bacterial groups compared with the influenza group, suggesting that the changes were indeed driven by the pathogen and not a general response to infection. A final limitation that should be noted was the imbalance of gender between experimental groups, where the healthy volunteers were 90% female while the pneumonia groups were 30–50% female, and future work should place emphasis on larger and balanced gender composition between groups.

Here we describe a comprehensive profile of urine meta-biomarkers, including the microbiome, metabolome, and cytokines in pneumonia patients. Using these biomarkers, we achieved high success in predicting pneumonia pathogens. Depending on the infectious pathogen identified in each patient, distinctly different immune profiles were observed in cytokine profiles, and even larger shifts were observed in the metabolite and microbiome profile, especially in response to bacterial infections. This study provides a proof of concept that urine samples, which are easily accessible in outpatient and inpatient settings, could provide additional diagnostic insights to patient infectious status and future risk factor for complication.

## Materials and methods

### Sample processing.

Urine samples were collected using sterile technique and were aliquoted separately for cytokines, microbiome analysis, and metabolites[15, 16, 36]. For multiplex cytokine assays and microbiome analysis, the urine samples were centrifuged at 10,000 × g for 10 min and then analyzed as described below. For Metabolite analysis, samples (50 µl for urine) were extracted with 1.3 ml of extraction solvent (40:40:20 HPLC grade methanol, acetonitrile, water with 0.1% formic acid) pre-chilled to 4 °C in a cold room and incubated for 20 min at – 20 °C. The samples were centrifuged for 5 min (16.1 rcf) at 4 °C. The supernatants were transferred to new 1.5 ml centrifuge tubes and pellets were resuspended with 200 µl of extraction solvent. Extraction was allowed to proceed for 20 min at – 20 °C and all supernatants collected in glass vials. Vials containing the collected supernatant were dried under a stream of N2 until all the extraction solvent had been evaporated. Residue was resuspended in 300 µl of sterile water and transferred to 300 µl autosampler vials. Samples were immediately placed in autosampler trays for mass spectrometric analysis.

### Multiplex for cytokines and statistical analysis.

The levels of a panel of inflammatory mediators in urine samples were measured using a 34-plex ProcartaPlex Multiplex Immunoassay according to manufacturer's instructions (Invitrogen, Carlsbad CA, USA). Cytokine standards were prepared to determine the concentration of cytokines in the samples. The samples were run on a Millipore Magpix instrument and analyzed with xPONENT 4.2 software. For data analysis, a five-parameter logistic curve fitting method was applied to the standards and the sample concentrations extrapolated from the standard curve. The results were normalized to creatinine as measured by Creatinine Detection Kit (Thermofisher, Waltham, MA).

Kruskal–Wallis test was used to compare median level of cytokine among the four groups of samples. The median cytokine levels between urine samples from health volunteers and those from influenza, *S. pneumo* or *S. aureus* were tested via Wilcoxon sum rank test. The p values were not adjusted for multiple comparisons. All analyses were performed using R-3.4.0 (https://www.R-project.org/).

### Microbial DNA isolation.

Human urine samples (100 µl) were centrifuged at 10,000 × g for 10 min, supernatant was carefully removed, and 500 uL of extraction buffer (50 mM Tris (pH 7.4), 100 mM EDTA (pH 8.0), 400 mM NaCl, 0.5% SDS) containing 20 µl proteinase K (20 mg/ml, Cat# 03115887001, Roche) was added to each tube[15, 49]. 0.1-mm-diameter zirconia/silica beads (BioSpec Products, Bartlesville, OK, USA) were added to the extraction tubes and a Mini-Beadbeater-8 cell disrupter (BioSpec Products) for 2 × 1 min to lyse cells. After overnight incubation at 55 °C with agitation, extraction with phenol:chloroform:isoamyl alcohol, and precipitation with ethanol were performed. Isolated DNA was dissolved in nuclease-free water and stored at – 80 °C.

### 16S rRNA-based PCR, ilumina library preparation, and data analysis.

To assess total 16S copy numbers, 2 µl of isolated DNA was used in quantitative PCR analysis using 16S primers (Forward: 5′-TCCTAC

GGGAGGCAGCAGT-3′; Reverse: 5′-GGACTACCAGGGTATCTAATCCTGTT-3′) and an in-house standard to generate a standard curve. To assess bacterial community structure, primers specific for 16S rRNA V4-V5 region (Forward: 338F: 5′-GTGCCAGCMGCCGCGGTAA-3′ and Reverse: 806R: 5′-GGACTACHVGGGTWT CTAAT-3′) that contained Illumina 3′ adapter sequences, as well as a 12-bp barcode, were used. Sequences were generated by an Illumina MiSeq DNA platform at Argonne National Laboratory and analyzed by the program Quantitative Insights Into Microbial Ecology (QIIME)[50]. Operational Taxonomic Units (OTUs) were picked at 97% sequence identity using open reference OTU picking against the GreenGenes database. OTUs were quality filtered based on default parameters set in the open-reference OTU command in QIIME and sequences were rarified to an average sampling depth of 7,084 reads per sample. Representative sequences were aligned via PyNAST and taxonomy was assigned using the RDP Classifier. Processed data were then imported into Calypso 8.84 for further analysis and data visualization[51]. Alpha diversity was assessed using observed Shannon index and Eveness. Network analyses were generated with Speaman's correlations. Positive correlations were FDR-adjusted at P < 0.05 and presented as network edges. OTUs generated in QIIME were finally analyzed using linear discriminant analysis (LDA) effect size (LEfSe) where non-parametric factorial Kruskal–Wallis sum-rank testing ($P$ < 0.05) identified significantly abundant taxa followed by unpaired Wilcoxon rank-sum test to determine LDA scores > 2 was considered significant. Dendrograms of LEfSe display taxonomic distribution of significant taxa[52]. Microbiome raw sequence reads are deposited at NCBI Sequence Read Archive, SUB7620442: https://submit.ncbi.nlm.nih.gov/subs/sra/SUB7620442/overview.

**Metabolite analysis.** *UPLC–HRMS metabolomics analysis.* Samples placed in an autosampler tray were kept at 4 °C. A 10 µl aliquot was injected through a Synergi 2.5 micron reverse-phase Hydro-RP 100, 100 × 2.00 mm LC column (Phenomenex, Torrance, CA) kept at 25 °C. The eluent was introduced into the MS via an electrospray ionization source conjoined to an Exactive Plus Orbitrap Mass Spectrometer (Thermo Scientific, Waltham, MA) through a 0.1 mm internal diameter fused silica capillary tube. The mass spectrometer was run in full scan mode with negative ionization mode with a window from 85 to 1,000 m/z. with a method adapted from Lu et al.[53]. The samples were run with a spray voltage was 3 kV. The nitrogen sheath gas was set to a flow rate of 10 psi with a capillary temperature of 320 °C. AGC (acquisition gain control) target was set to 3e6. The samples were analyzed with a resolution of 140,000 and a scan window of 85–800 m/z for from 0 to 9 min and 110–1,000 m/z from 9 to 25 min. Solvent A consisted of 97:3 water:methanol, 10 mM tributylamine, and 15 mM acetic acid. Solvent B was methanol. The gradient from 0 to 5 min is 0% B, from 5 to 13 min is 20% B, from 13 to 15.5 min is 55% B, from 15.5 to 19 min is 95% B, and from 19 to 25 min is 0% B with a flow rate of 200 µl/min[53].

Files generated by Xcalibur (RAW) were converted to the open-source mzML format[54] via the open-source msconvert software as part of the ProteoWizard package[55]. Maven (mzroll) software, Princeton University[56, 57] was used to automatically correct the total ion chromatograms based on the retention times for each sample. Metabolites were manually identified and integrated using known masses (± 5 ppm mass tolerance) and retention times (Δ ≤ 1.5 min). Unknown peaks were automatically selected via Maven's automated peak detection algorithms.

Multivariate statistical analysis for the MS/MS data was performed using XLSTAT software (Addinsoft, New York, NY) interfaced with Excel (Microsoft Corporation, Redmond, WA). The average coefficient of variation (C.V.) was 0.395 (± 0.211). Thus an inclusion criterion for technical replicates were applied based on C.V. ≤ 0.5 resulting in 11 exclusion (i.e. 11 technical replicates in duplicate and 29 in triplicate) resulting in C.V. average of 0.288 (± 0.114). To ensure that observations were directly comparable and to account for the biofluid concentration peak intensity was normalized to creatinine (these data were compared to unnormalized data to make sure there was no masking of biologically relevant changes by normalization). To evaluate the group trends, sample uniformity and identify potential outliers unsupervised multivariant principal component analysis (PCA) was employed. The variation were explained by F1 and F2 with a cumulative percent variability of 78.558 and a marginal increase to 85.958% with the addition of F3. These data were then independently k-means clustered followed by ascendant hierarchical clustering based on Euclidian distances. The data matrix's was rearranged according to the corresponding clustering with similarity proportional to a closer spatial relationship for patient sample columns and metabolite rows. 21 metabolites with less than 0.25 standard deviation were eliminated to simplify the graph. These clusters were also represented via a dendrogram displayed vertically for metabolites and another horizontally for patients. The data values of the permuted matrix were replaced by corresponding color intensities based on interquartile range with color scale of red to green through black resulting in a heat map. Patient identifiers and risk categories were replaced by color bars. XLSTAT expression analysis was then used to determine metabolite significance between groups using one-way ANOVA with Benjamini–Hochberg post hoc correction and found to have with significant differences using Tukey's honest significance test (Tukey HSD) for multiple comparisons. Partial least squares discriminant analysis (PLS-DA) was then applied to separate patient groups and identify metabolites with corresponding variable importance in the projection (VIP) values above 1. The four component PLS-DA was then rerun with the variables centered and reduced prior to analysis to improve the model quality and identify the corresponding VIP. Each identified metabolite was then analyzed individually using ANOVA then the means of pneumonia samples were then compared to healthy controls (Dunnett's multiple comparisons test) prior to and post outlier removal performed with PRISM software (Graphpad, San Diego, CA). Raw metabolite data are available in the Metabolights Database at https://www.ebi.ac.uk/metabolights/MTBLS1722.

**Predictive modeling.** We implement predictive modeling approach to distinguish between four subject categories (healthy, *S. Aureus, S. Pneumoniae*, and Influenza) using identified operational taxonomic units (OTUs) and metabolites from urine samples. First, all 40 subjects were analyzed to identify OTUs. Sample decompositions were normalized in a way that sum of all detected OTUs are equal to 1 for each subject. We then combined identified OTUs, cytokines and metabolites as predictors of four subject categories. Our first approach is to implement multi-class classification using various machine learning algorithms such as random forest, ensemble trees, support vector machines, k-nearest neighborhood. However, small sample size (n = 40), multiple output categories (four subject groups) and expected larger number of predictors (OTUs and metabolites) made predictive modeling very challenging. Therefore, as an alternative approach, we implemented recursive binary classification in three steps and obtained three different models at each step. First, Model 1 is to distinguish healthy subjects from the other three disease categories, Model 2 to distinguish between bacterial *(S. Aureus* and *S. Pneumoniae*) disease from Influenza, and Model 3 to distinguish between S. Aureus and *S. Pneumoniae*. Considering the small sample size and large number of predictors, for each of the three models, we first implement Least Absolute Shrinkage and Selection Operator (LASSO)[58] logistic regression[59]. LASSO is statistical method retraining strong features of both subset selection and ridge regression. It implements ordinary least squares subject to sum of absolute values of the regression coefficients being less than a predetermined constant value. Logistic regression LASSO is an extension of LASSO for an output variable with binomial distribution.

By implementing LASSO, some of the regression coefficients are shrink to take a valued of zero therefore only variables with non-zero regression coefficients remain in the model. By taking advantage of LASSO, we will first identify OTUs and metabolites that are the most effective in distinguishing between our subject categories in Model 1, Model 2, and Model 3. Next, we combined all selected OTUs and metabolites in readdress multi-class classification problem using the machine learning approaches mentioned.

For each predictive model, we implemented a stratified 5-folds cross validation process to avoid overfitting. To ensure unbiasedness of our cross-validation strategy, we split the entire cohort into five distinct each including a same number of subjects from each category. Next, a model built on using four folds of data and tested on the remaining fold. By repeating this process five times by changing the test fold, we obtain a predict class labels for each subject using a model that is trained on other subjects. We did not implement cross-validation with the goal of hyper-parameter tuning and optimization, instead, we used cross-validation (1) to evaluate the variability (or stability) of the predictive models from one subset to another (2) to evaluate the model performance on an unseen dataset. Therefore, we did not transfer parameter setting from one fold to another, instead, we fixed model parameters across each folds. We compared different machine learning algorithms based on model performance statistics such as specificity, sensitivity, and positive and negative predictive values.

**Ethics statement.** The usage of human samples was approved and performed in accordance with the regulations and guidelines set by the Univeristy of Louisville Insitutional Review Boards and the Human Subjects Protection Program. Samples were obtained from the University of Louisville Infectious Disease Biorepository (IRB # 13.0001) and de-identified metadata were used for analysis under the Biomarkers study (IRB # 17.0601). All patients provided written informed consent for sample biorepository storage and subsequent use in research studies.

## Data availability

The datasets generated during microbiome analysis in this study are available as raw sequence reads at NCBI Sequence Read Archive: https://submit.ncbi.nlm.nih.gov/subs/sra/SUB7620442/overview. Datasets generated during metabolite analysis in this study are available in the Metabolights Database at https://www.ebi.ac.uk/metabolights/MTBLS1722.

Received: 30 October 2019; Accepted: 26 June 2020
Published online: 07 August 2020

## References

1. Heron, M. & Tejada-Vera, B. Deaths: Leading causes for 2005. *Natl. Vital Stat. Rep.* **58**, 1–97 (2009).
2. Colice, G. L., Morley, M. A., Asche, C. & Birnbaum, H. G. Treatment costs of community-acquired pneumonia in an employed population. *Chest* **125**, 2140–2145 (2004).
3. Mandell, L. A. *et al.* Infectious Diseases Society of America/American Thoracic Society Consensus Guidelines on the management of community-acquired pneumonia in adults. *Clin. Infect. Dis.* **44**, S27–S72 (2007).
4. Restrepo, M. I., Faverio, P. & Anzueto, A. Long-term prognosis in community-acquired pneumonia. *Curr. Opin. Infect. Dis.* **26**, 151–158 (2013).
5. Ramirez, J. A. & Anzueto, A. R. Changing needs of community-acquired pneumonia. *J. Antimicrob. Chemother.* **66**(Suppl 3), iii3-9 (2011).
6. Muller, M. P. *et al.* Evaluation of pneumonia severity and acute physiology scores to predict ICU admission and mortality in patients hospitalized for influenza. *PLoS ONE* **5**, e9563 (2010).
7. Avni, T., Mansur, N., Leibovici, L. & Paul, M. PCR using blood for diagnosis of invasive pneumococcal disease: Systematic review and meta-analysis. *J. Clin. Microbiol.* **48**, 489–496 (2010).
8. Gonsalves, W. I., Cornish, N., Moore, M., Chen, A. & Varman, M. Effects of volume and site of blood draw on blood culture results. *J. Clin. Microbiol.* **47**, 3482–3485 (2009).
9. Resti, M. *et al.* Community-acquired bacteremic pneumococcal pneumonia in children: Diagnosis and serotyping by real-time polymerase chain reaction using blood samples. *Clin. Infect. Dis.* **51**, 1042–1049 (2010).
10. Bradley, P. J. Treatment of hospital-acquired pneumonia. *Lancet. Infect. Dis.* **11**, 730–731 (2011) (**author reply 731–732**).
11. Edwards, M. O., Kotecha, S. J. & Kotecha, S. Respiratory distress of the term newborn infant. *Paediatr. Respir. Rev.* **14**, 29–37 (2013).
12. Hagaman, J. T., Rouan, G. W., Shipley, R. T. & Panos, R. J. Admission chest radiograph lacks sensitivity in the diagnosis of community-acquired pneumonia. *Am. J. Med. Sci.* **337**, 236–240 (2009).
13. Boersma, W. G., Daniels, J. M. A., Löwenberg, A., Boeve, W.-J. & van de Jagt, E. J. Reliability of radiographic findings and the relation to etiologic agents in community-acquired pneumonia. *Respir. Med.* **100**, 926–932 (2006).
14. Kellum, J. A. *et al.* Understanding the inflammatory cytokine response in pneumonia and sepsis. *Arch. Intern. Med.* **167**, 1655 (2007).

15. Colas, L. *et al.* Unique and specific proteobacteria diversity in urinary microbiota of tolerant kidney transplanted recipients. *Am. J. Transpl.* https://doi.org/10.1111/ajt.15549 (2019).
16. Bučević Popović, V. *et al.* The urinary microbiome associated with bladder cancer. *Sci. Rep.* **8**, 12157 (2018).
17. Slupsky, C. M. *et al.* *Streptococcus pneumoniae* and *Staphylococcus aureus* pneumonia induce distinct metabolic responses. *J. Proteome Res.* **8**, 3029–3036 (2009).
18. Slupsky, C. M. *et al.* Pneumococcal pneumonia: Potential for diagnosis through a urinary metabolic profile. *J. Proteome Res.* **8**, 5550–5558 (2009).
19. Seymour, C. W. *et al.* Metabolomics in pneumonia and sepsis: An analysis of the GenIMS cohort study. *Intensive Care Med.* **39**, 1423–1434 (2013).
20. Ning, P. *et al.* Metabolic profiles in community-acquired pneumonia: Developing assessment tools for disease severity. *Crit. Care* **22**, 130 (2018).
21. Zmora, N., Zeevi, D., Korem, T., Segal, E. & Elinav, E. Taking it personally: personalized utilization of the human microbiome in health and disease. *Cell Host Microbe* **19**, 12–20 (2016).
22. Viant, M. R. Recent developments in environmental metabolomics. *Mol. Biosyst.* **4**, 980 (2008).
23. Nicholson, J. K., Everett, J. R. & Lindon, J. C. Longitudinal pharmacometabonomics for predicting patient responses to therapy: Drug metabolism, toxicity and efficacy. *Expert Opin. Drug Metab. Toxicol.* **8**, 135–139 (2012).
24. Montoliu, I. *et al.* Current status on genome–metabolome-wide associations: An opportunity in nutrition research. *Genes Nutr.* **8**, 19–27 (2013).
25. Gieger, C. *et al.* Genetics meets metabolomics: A genome-wide association study of metabolite profiles in human serum. *PLoS Genet.* **4**, e1000282 (2008).
26. Altmaier, E. *et al.* Bioinformatics analysis of targeted metabolomics—Uncovering old and new tales of diabetic mice under medication. *Endocrinology* **149**, 3478–3489 (2008).
27. Banoei, M. M. *et al.* Plasma metabolomics for the diagnosis and prognosis of H1N1 influenza pneumonia. *Crit. Care* **21**, 97 (2017).
28. Kuehne, A. *et al.* An integrative metabolomics and transcriptomics study to identify metabolic alterations in aged skin of humans in vivo. *BMC Genom.* **18**, 169 (2017).
29. Gale, T. V., Horton, T. M., Grant, D. S. & Garry, R. F. Metabolomics analyses identify platelet activating factors and heme breakdown products as Lassa fever biomarkers. *PLoS Negl. Trop. Dis.* **11**, e0005943 (2017).
30. Wishart, D. S. *et al.* HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Res.* **37**, D603–D610 (2009).
31. Rattner, J. & Bathe, O. F. Monitoring for response to antineoplastic drugs: The potential of a metabolomic approach. *Metabolites* **7**, 60 (2017).
32. Farshidfar, F. *et al.* Serum metabolomic profile as a means to distinguish stage of colorectal cancer. *Genome Med.* **4**, 42 (2012).
33. Farshidfar, F. *et al.* A validated metabolomic signature for colorectal cancer: Exploration of the clinical value of metabolomics. *Br. J. Cancer* **115**, 848–857 (2016).
34. Marchand, C., Farshidfar, F., Rattner, J. & Bathe, O. A framework for development of useful metabolomic biomarkers and their effective knowledge translation. *Metabolites* **8**, 59 (2018).
35. Boeniger, M. F., Lowry, L. K. & Rosenberg, J. Interpretation of urine results used to assess chemical exposure with emphasis on creatinine adjustments: a review. *Am. Ind. Hyg. Assoc. J.* **54**, 615–627 (1993).
36. Payan, J. P., Viau, C. & Lafontaine, M. Creatinine normalization in biological monitoring revisited: The case of 1-hydroxypyrene. *Int. Arch. Occup. Environ. Health* **77**, 177–185 (2004).
37. Dawson, P. A. Sulfate in fetal development. *Semin. Cell Dev. Biol.* **22**, 653–659 (2011).
38. Klaassen, C. D. & Boles, J. W. Sulfation and sulfotransferases 5: The importance of 3′-phosphoadenosine 5′-phosphosulfate (PAPS) in the regulation of sulfation. *FASEB J.* **11**, 404–418 (1997).
39. Venkatachalam, K. V. Human 3′-phosphoadenosine 5′-phosphosulfate (PAPS) Synthase: Biochemistry, molecular biology and genetic deficiency. *IUBMB Life (Int. Union Biochem. Mol. Biol. Life)* **55**, 1–11 (2003).
40. Grootveld, M. & Halliwell, B. 2,3-Dihydroxybenzoic acid is a product of human aspirin metabolism. *Biochem. Pharmacol.* **37**, 271–280 (1988).
41. Haque, M. F., Aghabeighi, B., Wasil, M., Hodges, S. & Harris, M. Oxygen free radicals in idiopathic facial pain. *Bangl. Med. Res. Counc. Bull.* **20**, 104–116 (1994).
42. Chiu, C.-Y. *et al.* Metabolomic profiling of infectious parapneumonic effusions reveals biomarkers for guiding management of children with streptococcus pneumoniae pneumonia. *Sci. Rep.* **6**, 24930 (2016).
43. Adamko, D. J., Saude, E., Bear, M., Regush, S. & Robinson, J. L. Urine metabolomic profiling of children with respiratory tract infections in the emergency department: A pilot study. *BMC Infect. Dis.* **16**, 439 (2016).
44. de Brito, R. *et al.* The balance between the serum levels of IL-6 and IL-10 cytokines discriminates mild and severe acute pneumonia. *BMC Pulm. Med.* **16**, 170 (2016).
45. Bacci, M. R. *et al.* IL-6 and TNF-α serum levels are associated with early death in community-acquired pneumonia patients. *Braz. J. Med. Biol. Res.* **48**, 427–432 (2015).
46. Mira, J.-P., Max, A. & Burgel, P.-R. The role of biomarkers in community-acquired pneumonia: Predicting mortality and response to adjunctive therapy. *Crit. Care* **12**, S5 (2008).
47. Siljan, W. W. *et al.* Cytokine responses, microbial aetiology and short-term outcome in community-acquired pneumonia. *Eur. J. Clin. Invest.* **48**, e12865 (2018).
48. Whiteside, S. A., Razvi, H., Dave, S., Reid, G. & Burton, J. P. The microbiome of the urinary tract—A role beyond infection. *Nat. Rev. Urol.* **12**, 81–90 (2015).
49. Pierre, J. F. *et al.* Activation of bile acid signaling improves metabolic phenotypes in high-fat diet-induced obese mice. *Am. J. Physiol. Gastrointest. Liver Physiol.* **311**, G286–G304 (2016).
50. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
51. Zakrzewski, M. *et al.* Calypso: a user-friendly web-server for mining and visualizing microbiome-environment interactions. *Bioinformatics* **33**, 782–783 (2017).
52. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).
53. Lu, W. *et al.* Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone orbitrap mass spectrometer. *Anal. Chem.* **82**, 3212–3221 (2010).
54. Martens, L. *et al.* mzML—A community standard for mass spectrometry data. *Mol. Cell. Proteomics* **10**, R110.000133 (2011).
55. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
56. Melamud, E., Vastag, L. & Rabinowitz, J. D. Metabolomic analysis and visualization engine for LC−MS data. *Anal. Chem.* **82**, 9818–9826 (2010).
57. Clasquin, M. F., Melamud, E. & Rabinowitz, J. D. LC-MS data processing with MAVEN: A metabolomic analysis and visualization engine. in *Current Protocols in Bioinformatics*, Chapter 14, Unit 14.11 (Wiley, 2012). https://doi.org/10.1002/0471250953.bi1411s37.
58. Tibshirani, R. Regression selection and shrinkage via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
59. Lee, S., Lee, H., Abbeel, P. & Ng, A. Efficient L 1 regularized logistic regression. *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)* (2006).

## Acknowledgements

Metabolomic extraction and mass spectrometric analyses were performed at the Biological and Small Molecule Mass Spectrometry Core, University of Tennessee, Knoxville, TN with the assistance of Dr. Shawn R. Campagna, Dr. Hector F. Castro, Sara Howard and Eric Tague. We thank members of the ISHPS Pneumonia Working Group, Reba Umberger, Amber Smith and Robert Williams for their input into experimental design and data analysis. This work was supported by the Institute for the Study of Host Pathogen Systems (ISHPS), Offices of the Chancellor and Vice Chancellor for Research, UTHSC (CJ), Tennessee Govenor's Fund for Pediatric Recruitment (JP), and the Children's Foundation Research Institute, Memphis, TN (JP).

## Author contributions

C.J. and J.R. planned the study design, S.P. and S.F. organized samples and metadata, J.P., H.S., and E.F. generated data and analyzed results, O.A. and X.C. performed computational modeling, J.P. and C.J. drafted the manuscript. All authors reviewed and approved the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41598-020-70461-9.

**Correspondence** and requests for materials should be addressed to J.F.P. or C.B.J.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.