Metabolomics in community-acquired pneumonia: exploring metabolomics-based biomarkers for diagnosis and treatment response monitoring of community-acquired pneumonia

Hartog, I. den

CHAPTER 2

Metabolomic profiling of microbial disease etiology in community-acquired pneumonia

Ilona den Hartog, Laura B. Zwep, Stefan M.T. Vestjens, Amy C. Harms, G. Paul Voorn, Dylan W. de Lange, Willem J.W. Bos, Thomas Hankemeier, Ewoudt M.W. van de Garde, J.G. Coen van Hasselt. Metabolomic profiling of microbial disease etiology in community-acquired pneumonia, *PLoS One* **16:6** (2021).

Abstract

Diagnosis of microbial disease etiology in community-acquired pneumonia (CAP) remains challenging. We undertook a large-scale metabolomics study of serum samples in hospitalized CAP patients to determine if host-response associated metabolites can enable diagnosis of microbial etiology, with a specific focus on discrimination between the major CAP pathogen groups *S. pneumoniae*, atypical bacteria, and respiratory viruses. Targeted metabolomic profiling of serum samples was performed for three groups of hospitalized CAP patients with confirmed microbial etiologies: *S. pneumoniae* (n=48), atypical bacteria (n=47), or viral infections (n=30). A wide range of 347 metabolites was targeted, including amines, acylcarnitines, organic acids, and lipids. Single discriminating metabolites were selected using Student's T-test and their predictive performance was analyzed using logistic regression. Elastic net regression models were employed to discover metabolite signatures with predictive value for discrimination between pathogen groups. Metabolites to discriminate *S. pneumoniae* or viral pathogens from the other groups showed poor predictive capability, whereas discrimination of atypical pathogens from the other groups was found to be possible. Classification of atypical pathogens using elastic net regression models was associated with a predictive performance of 61% sensitivity, 86% specificity, and an AUC of 0.81. Targeted profiling of the host metabolic response revealed metabolites that can support diagnosis of microbial etiology in CAP patients with atypical bacterial pathogens compared to patients with *S. pneumoniae* or viral infections.

2.1 Introduction

Community-acquired pneumonia (CAP) is a commonly occurring respiratory tract infection caused by bacterial or viral pathogens that can lead to severe disease, especially in elderly patients [4]. The predominant pathogens found in hospitalized CAP patients are *Streptococcus pneumoniae* and to a lesser extent, *Haemophilus influenzae*, *Legionella pneumophila,* and respiratory viruses [29, 30]. Patients hospitalized with severe CAP typically receive empirical antibiotic treatment with broad-spectrum antibiotics until the microbial etiology is determined [31, 8]. Current standard diagnostic methods for microbial identification are pathogen-targeted and include culturing, antigen testing, and molecular diagnostics such as PCR [8]. In over 60% of CAP patients, no causative pathogen can be identified with these pathogen-targeted diagnostic techniques [29, 32]. As a consequence, broad-spectrum antibiotics are over-used, which facilitates the emergence of antimicrobial resistance [1, 33]. To this end, a need exists to explore innovative methods to enhance the diagnostic performance for the detection of microbial pathogens in CAP.

Evaluation of differences in the host-response to CAP-associated pathogens may be an alternative approach to improve diagnosis [34]. There is growing evidence that the host, i.e. the patient, metabolic response to infections can be a relevant source of novel host immune response biomarkers to infections [35, 36]. Several small studies have reported differences in metabolite profiles in blood and urine samples in patients with different types of infections (Table 2.4) [37, 38, 23, 20, 39, 22, 40]. For instance, studies comparing metabolomic changes in CAP and tuberculosis (TB) patients show increased levels of plasma lipids and decreased levels of metabolites involved in cholesterol synthesis [37, 20]. A study comparing viral and bacterial respiratory tract infections showed that plasma metabolite profiles of patients with influenza A and bacterial pneumonia differed significantly [22]. In another study, urine samples of patients with a respiratory syncytial virus (RSV) or a bacterial respiratory tract infection showed differences in metabolite levels as well [40]. An important limitation of these studies is that the comparisons made cannot yet support the etiological diagnosis of CAP but merely focus on differences between diseases such as TB versus CAP. The studies that compared viral and bacterial causative pathogen groups of CAP used an untargeted metabolomics approach. While an untargeted approach is especially useful for the discovery of new features and hypothesis-free analysis, a targeted approach that can be fully quantified to clinical laboratory standards may be preferable for clinical implementation. Furthermore, these studies have the limitation that they focus on the comparison of pediatric patients while most hospitalized CAP patients are adults. No studies have evaluated differences in metabolite profiles of CAP patients comparing different microbial etiologies relevant for treatment of CAP, i.e. *S. pneumoniae*, atypical pathogens, and viral infections.

In the current study, we performed extensive targeted metabolomic profiling for three groups of hospitalized CAP patients with confirmed microbial etiologies of *S.*

pneumoniae, atypical bacteria, or viral infections. We aimed to determine whether host-response associated metabolites can enable diagnosis of microbial etiology, focusing on discrimination between the pathogen groups *S. pneumoniae*, atypical bacteria, and respiratory viruses in patients hospitalized with CAP.

2.2 Materials and methods

2.2.1 Study population

Serum samples were taken from 505 patients that were diagnosed with CAP in two previously conducted clinical studies that were executed between October 2004 and September 2010. [29, 30]. The samples were taken from CAP patients within 24 hours after hospital admission. In 57% of these patient samples, the causative pathogen could be identified using conventional diagnostic methods such as culturing, PCR, and urinary antigen tests. The most commonly found causative pathogen in these patients was *S. pneumoniae*, followed by atypical bacterial and viral pathogens. A minority of patients was diagnosed with other bacteria.

From the selection of patients in which a causative pathogen was identified, we excluded patients with mixed infections. Furthermore, we constructed three distinctive groups of patients with *Streptococcus pneumoniae*, atypical (*Coxiella burnetii*, *Chlamydophila psittaci*, *Legionella pneumophila* or *Mycoplasma pneumoniae*), or viral (influenza virus, herpes simplex virus (HSV), respiratory syncytial virus (RSV), parainfluenza virus, or another respiratory virus) infections. The number of available samples for the patient group with confirmed viral CAP infection was limited (n=31). The patients included in the *S. pneumoniae* and atypical bacterial groups were randomly drawn from the remaining study population in an iterating fashion until the bacterial groups were composed in such a way that three groups showed comparable means for sex and pneumonia severity index scores. This resulted in a group of 49 patients with *S. pneumoniae* and a group of 50 patients with atypical infections (Figure 2.1). No matching of individual samples was performed. An overview of patient characteristics is provided in Table 2.1 and Table 2.5. Patient characteristics that might be considered as possible covariates were: age, sex, nursing home resident, renal disease, congestive heart failure, CNS disease, malignancy, COPD, diabetes, altered mental status, respiratory rate, systolic blood pressure, temperature, pulse, pH, BUN, sodium, glucose, hematocrit, partial pressure of oxygen, pleural effusion on x-ray, duration of symptoms before admission, antibiotic treatment before admission. The analyses performed in this study were executed conform the informed consent given by the patients. The clinical data was anonymized before use.

2.2.2 Bioanalytical procedures

Serum samples were analyzed with five liquid chromatography methods and one gas chromatography, mass spectrometry-based, targeted, metabolomics method. The

Table 2.1 Patient characteristics per pathogen group.

	S. pneumoniae (n=48)	Atypical (n=47)	Viral (n=30)	P-value
Age (years)				
Mean (SD)	62.2 (18.9)	54.7 (14.6)	70.1 (16.4)	<0.01
Median [Min, Max]	63.5 [18.0, 98.0]	52.0 [26.0, 81.0]	74.0 [29.0, 95.0]	
Sex				
Male	22 (45.8%)	34 (72.3%)	21 (70.0%)	0.12
PSI score				
<50	9 (18.8%)	9 (19.1%)	2 (6.7%)	0.33
51-70	7 (14.6%)	13 (27.7%)	6 (20.0%)	
71-90	5 (10.4%)	10 (21.3%)	7 (23.3%)	
91-130	23 (47.9%)	12 (25.5%)	11 (36.7%)	
>131	4 (8.3%)	3 (6.4%)	4 (13.3%)	
Liver disease				
No	48 (100%)	47 (100%)	30 (100%)	
Kidney disease				
Yes	3 (6.2%)	1 (2.1%)	4 (13.3%)	0.30
Cardiovascular disease				
Yes	6 (12.5%)	5 (10.6%)	3 (10.0%)	0.93
CNS disease				
No	46 (95.8%)	44 (93.6%)	28 (93.3%)	0.66
Yes	1 (2.1%)	3 (6.4%)	2 (6.7%)	
Missing	1 (2.1%)	0 (0%)	0 (0%)	
Malignancy				
No	44 (91.7%)	46 (97.9%)	28 (93.3%)	0.66
Yes	3 (6.2%)	1 (2.1%)	2 (6.7%)	
Missing	1 (2.1%)	0 (0%)	0 (0%)	
COPD				
No	24 (50.0%)	44 (93.6%)	25 (83.3%)	0.16
Yes	9 (18.8%)	3 (6.4%)	5 (16.7%)	
Missing	15 (31.2%)	0 (0%)	0 (0%)	
Diabetes				
No	26 (54.2%)	45 (95.7%)	26 (86.7%)	0.17
Yes	7 (14.6%)	2 (4.3%)	4 (13.3%)	
Missing	15 (31.2%)	0 (0%)	0 (0%)	
Duration of symptoms before admission (days)				
Mean (SD)	4.06 (3.03)	5.83 (5.65)	4.70 (3.21)	0.33
Median [Min, Max]	3.50 [1.00, 14.0]	5.00 [1.00, 42.0]	4.00 [0.00, 14.0]	
Missing	16 (33.3%)	0 (0%)	0 (0%)	
Antibiotic treatment before admission				
No	27 (56.2%)	29 (61.7%)	23 (76.7%)	0.17
Yes	5 (10.4%)	18 (38.3%)	7 (23.3%)	
Missing	16 (33.3%)	0 (0%)	0 (0%)	
Corticosteroid use before admission				
No	29 (60.4%)	46 (97.9%)	29 (96.7%)	0.67
Yes	2 (4.2%)	1 (2.1%)	1 (3.3%)	
Missing	17 (35.4%)	0 (0%)	0 (0%)	

Data are presented as number (%) or mean (SD). Abbreviations: PSI: pneumonia severity index; CNS: central nervous system; COPD: chronic obstructive pulmonary disease.

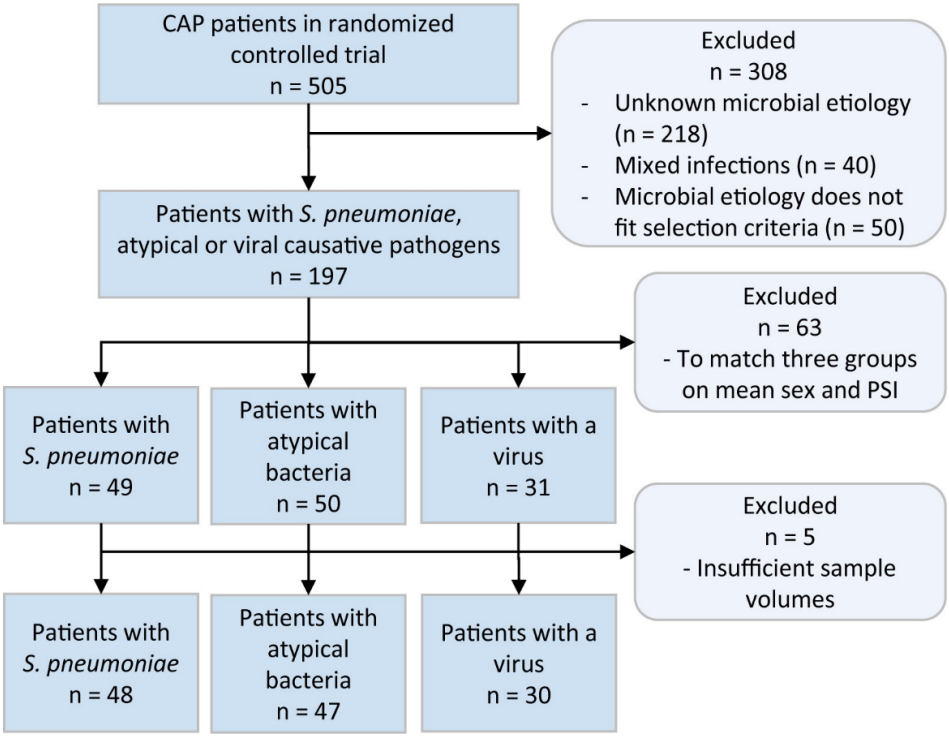

Figure 2.1 Flow chart of the formation of the three studied patient groups.

metabolomics profiling covered 596 metabolite targets from 25 metabolite classes, including amino acids, biogenic amines, acylcarnitines, organic acids, and multiple classes of lipids (Table 2.6). Levels of 374 unique metabolites were detected in the samples. The metabolomic profiling was performed within the Biomedical Metabolomics Facility of Leiden University in Leiden, The Netherlands. Details of the metabolomic analysis methods used are provided in section 2.5.

2.2.3 Data analysis

The data resulting from the metabolomic profiling was cleaned by removing patient samples with more than 10 missing metabolite values, for example, if results from one measurement platform were missing because of too low sample volumes, and by removing metabolites with missing patient samples, for example, because of a sample preparation error. The clean dataset consisted of 347 metabolite levels (Table 2.7) for 125 patients diagnosed with the microbial etiology *S. pneumoniae* (n=48), atypical (n=47), or viral (n=30). The pathogens identified in each group are shown in Table 2.2. The resulting metabolite levels were preprocessed by applying log transformation and

standardized to correct for heteroscedasticity. The preprocessed metabolomics dataset was visually inspected using a principal component analysis.

Data imputation was performed for patient characteristics that were to be evaluated as covariates in the statistical analysis and showed missingness in the data. Five times repeated imputation using predictive mean matching was performed with the 'mice' package for R to impute the patient data for the covariates with less than 25% missing data. Predictive mean matching is suitable for both numeric and binary covariates. Patient characteristics with >25% missing data were excluded from further analysis.

We performed logistic regression and elastic net regression modeling to determine if patients in one pathogen group could be discriminated from patients in the remaining two groups. Also, we aimed to determine which metabolites were important for prediction of the causative pathogen. In both methods, five-fold cross-validation was used to make the most efficient use of the available data for estimation of the predictive performance of the models and its associated metabolites [41]. Furthermore, the model generation was repeated 100 times to obtain robust estimates of the predictive performance of the models.

To identify single discriminative metabolites, Student's T-tests with false discovery rate (FDR) multiple testing corrections were performed ($p < 0.05$). Then, significant metabolites and a combination of significant metabolites were modeled using logistic regression. Also, models containing covariates age and sex and all covariates were generated. The predictive logistic regression models were analyzed by comparison of their area under the curve (AUC), sensitivity, specificity, balanced error rate (BER), and receiver operating characteristic (ROC) curve.

Elastic net regression was performed to test if the predictive power of the metabolite data could be increased by including correlations between metabolites in addition to evaluating single metabolites. In elastic net regression, metabolites that have no explanatory power can be set to zero, as in a lasso regression, and metabolites that explain the same amount of variance can all be included with balanced coefficient sizes, as in a ridge regression [42].

To obtain robust estimates of the predictive performance of the elastic net model, hyperparameters were optimized in a five-fold nested-cross validation, where the hyperparameters were selected truly independent of the calculation of the predictive performance, as is schematically shown in Figure 2.2 [43]. In the inner cross-validation loop, the model optimization loop, optimal values for model hyperparameters α and λ were determined. In the outer cross-validation loop, the model performance loop, the optimal model for the training fold was built on the set hyperparameters α and λ (Figure 2.5). Hyperparameter selection was performed using the balanced error rate (BER), which can be calculated from the true- and false positive (TP, FP), and true- and false-negative rates (TN, FN, Equation 2.1). The BER accounts for different group sizes

per model and therefore gives an accurate picture of the performance of models in the model optimization and model performance loop.

$$\mathrm{BER} = 0.5 * \left(\frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} + \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} \right) \tag{2.1}$$

The overall predictive diagnostic performance was evaluated using sensitivity and specificity performance measures, generated from the confusion matrix that represents the number of samples falling into each possible outcome (Equation 2.2-2.3). The average sensitivity and specificity of all 500 generated models and its standard deviation were used to compare the assay performance to currently used methods.

$$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.2}$$

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{2.3}$$

The relative contribution of metabolites to provide predictions of the expected pathogen group were quantified using the variable importance in prediction (VIP) score, expressed as a percentage. The VIP score was calculated per metabolite per fold or repeat as follows:

$$\mathrm{VIP}\ (\%) = \frac{\beta_j}{\sum_{i=0}^{p} |\beta_i|} \cdot 100\% \tag{2.4}$$

where β_j is the regression coefficient for fold j over the sum of all regression coefficient values in the model. Metabolites were arranged based on their mean VIP score over all folds and repeats. Metabolites with an absolute VIP > 1% were considered to be most important. Furthermore, to determine the need to include age and sex, or all covariates in the models we compared the BER for models with and without age and sex, or all covariates included. Finally, mean AUC values and ROC curves were calculated and generated to compare the performance of the elastic net models to the logistic regression models.

The scripts used for the statistical analyses were deposited in Github at http://github.com/vanhasseltlab/MetabolomicsEtiologyCAP.

Table 2.2 Distribution of causative microbial agents per pathogen group for statistical data analysis.

Causative pathogen	*S. pneumoniae* (n=48)	Atypical bacterial (n=47)	Viral (n=30)
S. pneumoniae	48 (100%)	0 (0%)	0 (0%)
Legionella pneumophila	0 (0%)	18 (38.3%)	0 (0%)
Coxiella burnetii	0 (0%)	17 (36.2%)	0 (0%)
Chlamydophila psittaci	0 (0%)	7 (14.9%)	0 (0%)
Mycoplasma pneumoniae	0 (0%)	5 (10.6%)	0 (0%)
Influenza virus	0 (0%)	0 (0%)	11 (36.7%)
HSV	0 (0%)	0 (0%)	6 (20.0%)
RSV	0 (0%)	0 (0%)	4 (13.3%)
Parainfluenza virus	0 (0%)	0 (0%)	3 (10.0%)
Other viruses	0 (0%)	0 (0%)	6 (20.0%)

Data are presented as number (%). Abbreviations: S. pneumoniae: Streptococcus pneumoniae; HSV: herpes simplex virus; RSV: respiratory syncytial virus.

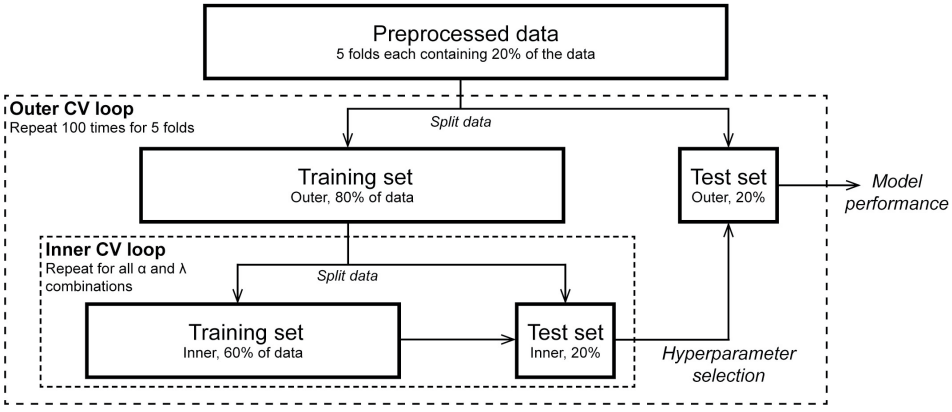

Figure 2.2 Schematic representation of stratified nested cross-validation for elastic net regression model optimization and performance [43]. Abbreviations: CV: cross-validation.

2.3 Results

2.3.1 Metabolomics profiling and exploratory analysis of metabolomics data

Metabolomics profiling was performed for 130 patients and 596 metabolite targets. Preprocessing of the metabolomics dataset resulted in a reduced dataset including 125 patients and 347 metabolites (Figure 2.1). The patient characteristics of these 125 patients are displayed in Table 2.1. The patients were diagnosed with the microbial etiology *S. pneumoniae* (n=48), atypical bacteria (n=47), or respiratory virus (n=30) (Table 2.2). A list of all targeted and detected metabolites and their identifiers can be found in Table 2.7. Unsupervised principal component analysis showed no clear separation between pathogen groups (Figure 2.6).

2.3.2 Single discriminating metabolites for pathogen groups

Three significant metabolites were found for the discrimination of atypical pathogens from *S. pneumoniae* and viral pathogens using a Student's T-test with FDR multiple testing correction ($p < 0.05$): glycylglycine, symmetric dimethylarginine (SDMA), and lysophosphatidylinositol (18:1) (LPI (18:1)). For the other comparisons, no significantly discriminating metabolites were found.

The significantly differentiating metabolites were included in logistic regression models to differentiate patients with atypical pathogens from patients suffering from CAP caused by *S. pneumoniae* or viral pathogens. The logistic regression models were evaluated based on their AUC, sensitivity, specificity, BER, and ROC curve after fivefold cross-validation with 100 repeats (Table 2.3, Figure 2.3). They show that logistic regression models of the individual metabolites glycylglycine, SDMA, and LPI(18:1) can differentiate atypical pathogens from *S. pneumoniae* and viral pathogens with AUCs between 0.70-0.72, sensitivities between 0.32-0.36, sensitivities between 0.83-0.85, and BERs of 0.39-0.41. A logistic regression model including all three significantly discriminating metabolites yields a more successful separation with an AUC of 0.78, sensitivity of 0.57, specificity of 0.83, and BER of 0.30. Addition of the covariates age and sex to the three metabolite model, slightly improved the predictive performance of the model resulting in a sensitivity of 0.63 and a specificity of 0.84. This model also showed the highest AUC (0.79) and lowest BER (0.26) of the tested logistic regression models. The addition of other covariates to the logistic regression model resulted in lower performance, probably due to overfitting of the model. The ROC curves emphasize the increased model performance upon the addition of more discriminating metabolites to the logistic regression model (Figure 2.3).

2.3.3 Predictive metabolites for diagnosis of CAP-associated pathogens

Elastic net models including multiple metabolites were fit to discriminate *S. pneumoniae*, atypical bacterial, and viral pathogens from the remaining two groups (e.g., *S. pneumoniae* versus atypical bacterial and viral pathogens). Elastic net models separating patients with atypical bacterial pathogens from patients with *S. pneumoniae* and viral infections resulted in a mean AUC of 0.81, a sensitivity of 0.61, a specificity of 0.86, and a BER of 0.26. Prediction of *S. pneumoniae* or viral infection etiologies showed lower predictive capabilities with AUC's of 0.74 and 0.63, high sensitivities of 0.83 and 0.89, but low specificities of 0.5 and 0.23, and BER's of 0.33 and 0.44, respectively (Table 2.3).

We included the covariates age and sex, and all covariates in the elastic net models to account for potential confounding effects. The addition of these covariates showed no improved performance of the elastic net models for differentiation of atypical pathogens or *S. pneumoniae* from the other groups. For the differentiation of viral pathogens from the other two pathogen groups, a slight performance improvement was seen upon the addition of the covariates age and sex resulting in an AUC of 0.63, a sensitivity of 0.89, a specificity of 0.23, and a BER of 0.44 (Table 2.3).

The ROC curves for the separation of atypical pathogens from *S. pneumoniae* and viral pathogens show that elastic net models perform better than the logistic regression models for single metabolites. However, the logistic regression model including the three significant metabolites and the covariates age and sex shows similar performance as the elastic net regression which included 100 metabolites on average (Figure 2.3).

2.3.4 Metabolite classes predictive for atypical bacterial pathogens

Focusing on the metabolites that have shown to be predictive for atypical bacterial pathogens, i.e., the only comparison with clinically relevant predictive performance, we identified 26 metabolites with an absolute VIP > 1% using elastic net regression (Figure 2.4). The metabolites originated from multiple metabolite classes. However, the classes of biogenic amines and lysophospholipids were well represented (4-5 metabolites per class), compared to the other classes. The number of metabolites included in the models varied across folds without a clear correlation to the BER. Commonly, models including all metabolites were favored, followed by models including 20-100 metabolites (Figure 2.7). We visualized the separation of the different pathogens in the atypical pathogen group using an unsupervised PCA analysis including all metabolites. The PCA plot indicated that no clear sub-group is present within the atypical group that would prominently drive the separation from the *S. pneumoniae* and viral infections (Figure 2.8).

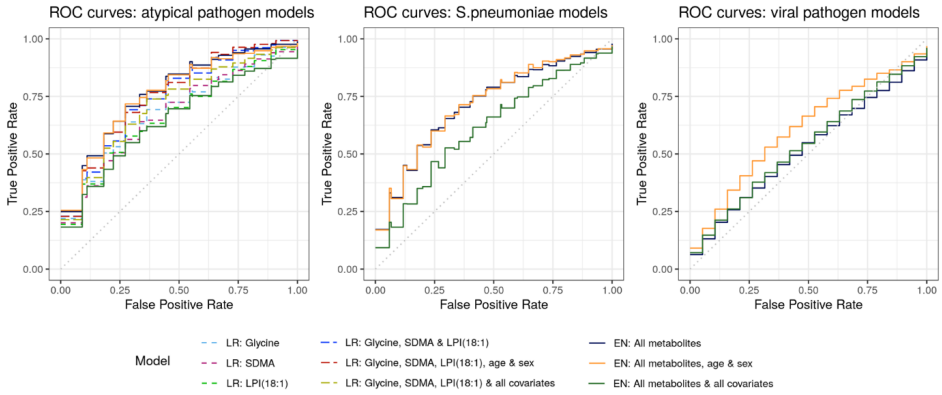

Figure 2.3 ROC curves of the results from logistic regression and elastic net regression models that were tested in five-fold cross-validation with 100 repeats for the comparisons: atypical versus *S. pneumoniae* and viral pathogens; *S. pneumoniae* pathogens versus atypical and viral pathogens; and viral versus *S. pneumoniae* and atypical pathogens. Abbreviations: LR: logistic regression, EN: elastic net regression, SDMA: symmetric dimethylarginine, LPI (18:1): lysophosphatidylinositol (18:1).

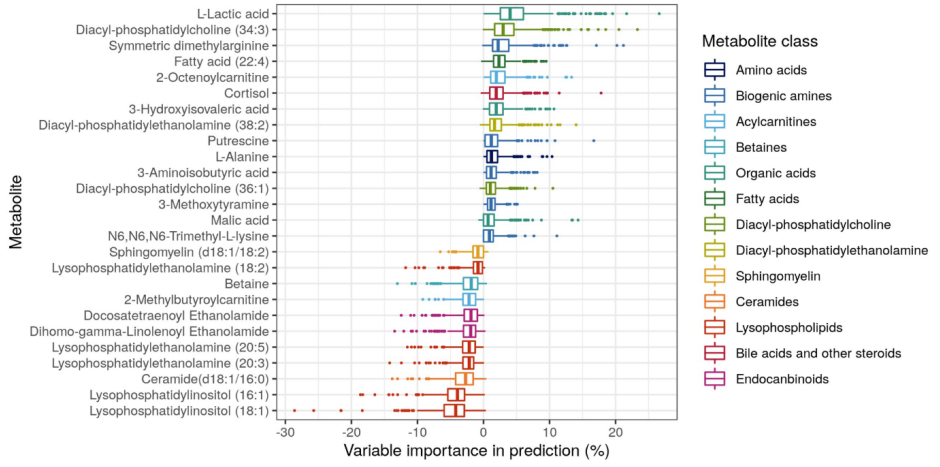

Figure 2.4 Variable importance of metabolites for the prediction of an atypical bacterial infection versus *S. pneumoniae* and viral infections. Only metabolites with an absolute mean percentage of influence > 1% are visualized.

Table 2.3 Results from the logistic regression and elastic net regression models that were tested in a fivefold cross-validation with 100 repeats. The table displays the performance of the models for the three comparisons: atypical versus *S. pneumoniae* and viral pathogens; *S. pneumoniae* pathogens versus atypical and viral pathogens; and viral versus *S. pneumoniae* and atypical pathogens. Logistic regression is only included for the comparison of atypical versus *S. pneumoniae* and viral pathogens because no significant single metabolites were found for the other comparisons. The performance is evaluated using the mean area under the curve (AUC), the mean sensitivity, the mean specificity, and the mean balanced error rate (BER) over all folds and repeats. All performances result from the test sets within the cross-validation. The best performing model per comparison and evaluation measure is displayed in bold and underlined.

Model	Variables	AUC	Sensitivity	Specificity	BER
	Atypical – (S. pneumoniae + viral)				
LR	Glycylglycine	0.72 (0.094)	0.36 (0.14)	0.83 (0.110)	0.40 (0.084)
LR	SDMA	0.72 (0.093)	0.36 (0.15)	0.86 (0.100)	0.39 (0.082)
LR	LPI.18.1.	0.70 (0.099)	0.32 (0.14)	0.85 (0.100)	0.41 (0.082)
LR	Age + sex	0.71 (0.097)	0.39 (0.15)	0.85 (0.090)	0.38 (0.071)
LR	All covariates	0.65 (0.098)	0.52 (0.15)	0.68 (0.120)	0.40 (0.087)
LR	Glycylglycine + SDMA + LPI.18.1.	0.78 (0.094)	0.57 (0.16)	0.83 (0.100)	0.30 (0.090)
LR	Glycylglycine + SDMA + LPI.18.1. + age + sex	0.79 (0.089)	0.63 (0.16)	0.84 (0.095)	0.26 (0.085)
LR	Glycylglycine + SDMA + LPI.18.1. + all covariates	0.75 (0.097)	0.60 (0.16)	0.78 (0.110)	0.31 (0.093)
ENR	100 (82)	0.81 (0.087)	0.61 (0.18)	0.86 (0.092)	0.27 (0.094)
ENR	110 (91) incl. age & sex	0.80 (0.094)	0.61 (0.17)	0.84 (0.096)	0.28 (0.090)
ENR	270 (140) incl. all covariates	0.69 (0.100)	0.58 (0.17)	0.70 (0.120)	0.36 (0.098)
	S. pneumoniae – (atypical + viral)				
ENR	210 (120)	0.74 (0.091)	0.83 (0.10)	0.50 (0.160)	0.33 (0.087)
ENR	240 (130) incl. age & sex	0.74 (0.095)	0.80 (0.10)	0.52 (0.160)	0.34 (0.084)
ENR	290 (120) incl. all covariates	0.63 (0.110)	0.69 (0.13)	0.51 (0.17)	0.40 (0.098)
	Viral – (S. pneumoniae + atypical)				
ENR	170 (140)	0.54 (0.120)	0.88 (0.11)	0.16 (0.170)	0.48 (0.075)
ENR	130 (130) incl. age & sex	0.63 (0.130)	0.89 (0.08)	0.23 (0.160)	0.44 (0.082)
ENR	180 (160) incl. all covariates	0.56 (0.130)	0.79 (0.11)	0.31 (0.190)	0.45 (0.099)

Data are presented as mean (SD). Variables are presented as variable names or as the number of variables that are included in the model. Abbreviations: LR: Linear regression, ENR: Elastic net regression, SDMA: symmetric dimethylarginine, LPI (18:1): lysophosphatidylinositol (18:1), AUC: area under the curve, BER: balanced error rate.

2.4 Discussion

Targeted profiling of the host metabolic response revealed metabolites that can support the diagnosis of microbial etiology in CAP patients with atypical bacterial pathogens compared to patients with *S. pneumoniae* or viral infections. CAP patients suffering from *S. pneumoniae* and viral infection could not be as successfully discriminated from the other groups based on the metabolic host-response.

The currently used clinical assays still outperform the metabolomics host-response assays developed in this study. For atypical pathogens, the sensitivity of 63% and specificity of 86% reported in this study are lower than the current urinary antigen tests for detection of *Legionella pneumophila* which shows a sensitivity of approximately 70% and a specificity up to 96% [44]. For detection of *S. pneumoniae*, the 83% sensitivity reached with the metabolomics-based assay outperforms the current antigen tests that show 70% sensitivity. However, the specificity of the metabolomics-based assay is only 50% while antigen tests reach specificity up to 96% [45, 46]. PCR assays of nasopharyngeal swabs for viral pathogens show sensitivities of up to 96% for influenza viruses A and B [47]. Our viral metabolomics-based assay shows a good sensitivity of 89% as well. However, the specificity of this assay is with 23% very low. The expected clinical utility of the studied metabolite classes as host-response biomarkers for etiological diagnosis of CAP may therefore be considered limited.

The combination of the metabolites glycylglycine, SDMA, and LPI (18:1) and the covariates age and sex showed predictive capacities similar to elastic net models including 100 metabolites in the comparison of atypical pathogens versus *S. pneumoniae* and viral pathogens. This result suggests that a simple model might perform as well as a more complex elastic net model, which is an important finding when considering the use of these biomarkers for clinical diagnostic applications, e.g., where a limited set of 3 metabolites is preferable.

Glycylglycine, a biogenic amine, showed to be significantly contributing to the differentiation of atypical pathogens from the other pathogens, but was not often included in elastic net models. In contrast, SDMA and LPI (18:1) were often included in the elastic net models as was shown in the overview of the 26 most influential metabolites. Metabolites of the classes biogenic amines and lysophospholipids, to which SDMA and LPI (18:1) have been assigned, were most represented in the 26 most influential metabolites compared to other metabolite classes in the comparison of atypical versus *S. pneumoniae* and viral pathogens. A comparison of the most influential metabolites in this study to metabolites of interest reported in previous studies of metabolomics in CAP patients shows limited overlap. Major reasons for this could be that (i) not all studies measured the same set of metabolic classes; (ii) some other studies poorly controlled patient comparator groups; and (iii) difference in bioanalytical methodologies, e.g. the use of NMR or MS as analytical method with their respective (dis)advantages might provide different results [48]. For example, most lipids found to be predictive in this study have not been reported previously, most

likely because the applied bioanalytical methodologies did not allow their detection. However, some overlap was found between the most influential metabolites for the comparison of atypical versus *S. pneumoniae* and viral pathogens in this study, and the metabolites of interest from other metabolomics studies involving CAP patients. The amino acid alanine was found in multiple studies [23, 39, 22]. Ceramide (d18:1/16:0), two diacyl-phosphatidylcholines, and diacyl-phosphatidylethanolamine (38:2) were found in other studies as well, the latter in the form of choline and ethanolamine [20, 39, 40]. Lactic acid was identified by several other metabolomics studies to respiratory bacterial and viral infections [37, 23, 22]. Lactic acid levels are also known to rise in case of severe disease. However, because the three pathogen groups were balanced in terms of disease severity and, for example, did not show significant differences in pH levels, we hypothesize that the differences in lactate levels are, in this case, an effect of the pathogen-specific host-response to infection. The result showed that models including disease severity covariates do not perform better than models without these confounders, thus supporting this hypothesis. Finally, 3-hydroxyisovaleric acid and betaine have been reported in a previous study comparing viral and bacterial pneumonia [40]. The overlap in these findings may provide insights into common metabolic responses to pathogens involved in CAP.

Multiple biological processes besides infection can influence metabolic processes in patients. Inclusion of age and sex in the models did not improve the predictive performance of the elastic net models for atypical bacteria and *S. pneumoniae* but did improve the model for viral pathogens. The average age in the viral pathogen group was higher than in the other groups, which could explain this result. For the other comparisons, we see that a model including age and sex or more covariates does not outperform models without these possible confounders. This doesn't imply there is no metabolomic effect of age in the bacterial pathogen groups but implies that the separation between bacterial pathogen groups is more dependent on the metabolomic host-response to the infection than on the age-related metabolomic changes. In this study, we included patients with mild to severe CAP, reflecting the target patient population for which improvements in a diagnostic assay are required. However, the combination of samples from patients with different disease severities may negatively influence the predictive capabilities of the model because the effect from the causative pathogen on the host-metabolism may be less pronounced for less severe disease [49]. However, separating the patients into groups with comparable disease severity scores would decrease the power for statistical analysis. Furthermore, no standardization of sampling times and conditions was applied, e.g., patients had not fasted before blood sampling, which may influence the metabolite patterns found. Since variations in sampling conditions were unknown, we were unable to consider these in our analyses. However, we expect that the impact of not standardizing and correcting for these factors is limited because the noise in metabolite levels introduced by these factors is expected to be random with regard to the pathogen groups compared in this study. A standardized sampling approach could improve the sensitivity of the models to detect predictive metabolites because some noise is reduced. However, the specificity of the models with

respect to the prediction of specific pathogens would be unchanged, since no correlation with pathogen groups is likely.

The sample size of this study (n = 125) was relatively large compared to studies researching metabolomic differences between causative pathogens of CAP that included approximately 70 patients [22, 40]. The compared groups *S. pneumoniae*, atypical bacteria, and viruses were chosen because antibiotic treatment strategies differ between these three groups. Ideally, we would have further investigated differences within studied groups, e.g. to identify metabolic responses to specific pathogens within the atypical pathogens and viral infection groups. For example, it would be of interest to study Legionella species more in-depth because their intracellular growth might result in a differentiated host-response. However, this was considered not feasible in this study due to sample size restrictions. The heterogeneous pathogen population in the atypical bacterial and viral pathogen groups might have lowered the predictive performance of the metabolomic analysis. Studying the individual pathogens in bigger sample sizes might reveal more characteristic metabolite signatures. In this study, no control group was included because the goal of the study was to provide a faster and optimal diagnostic method and a guide for antibiotic treatment in hospitalized CAP patients. In further studies, it would be preferable to include patients with all causes of CAP, including the remaining microorganisms, which were excluded in the current study because of their low frequency, to enable a more comprehensive comparison with current clinical assays. In this study, CAP patients with unknown pathogens were excluded. In a follow-up study, the metabolite pattern of the patients with unknown causative pathogens could be compared to the metabolite patterns of the distinguished pathogen groups to gain more information about the metabolomic resemblance of the samples in which pathogens could and could not be identified using the conventional diagnostic techniques.

Metabolomics analysis resulted in some missing data because of sample preparation errors or the limited volume of the samples. Because the measurement platforms covered multiple metabolites within one pathway, metabolites with missing data could be removed without influencing the final results. Some patient samples had to be removed because of multiple missing metabolite levels, for example, if the results from a whole metabolomics platform were missing. Data imputation was not performed for the metabolomics data, because the wide range of patients included in the dataset did, in our opinion, not provide enough information for accurate data imputation.

In summary, this comprehensive analysis of the host metabolic response across multiple metabolic classes and based on a well-balanced study cohort of CAP patients has shown the possibility to identify atypical pathogens in CAP and limited utility of predicting *S. pneumoniae* and viral infection disease etiologies.

2.5 Supporting information

Details on metabolomic sample analysis

Batch design: Aliquoted samples were run in a randomized fashion in several batches together with quality control (QC) samples (every 10 samples), sample replicates (every 7 samples), internal standards (ISTDs), blanks, and calibration lines.

Quality control: Blank samples were used to determine the blank effect. Replicate samples were used to check the instrument for repeatability. In-house developed algorithms were applied using the pooled QC samples to compensate for shifts in the sensitivity of the mass spectrometer over the batches.

Reported results: After quality control correction the metabolites that complied with the acceptance criteria of a relative standard deviation of the quality control samples (RSDqc) <15% were reported. The data was reported as relative response ratio (analyte signal area / ISTD area; unit free) of the metabolites after QC correction. Metabolites that did not comply with the acceptance criteria of the quality control, but have been included in the results present RSDs up to 30% and should be handled with caution.

Amine profiling: Amine profiling was performed according to the validated amine profiling analytical platform with minor optimization [50]. The amine platform covers amino acids and biogenic amines employing an Accq-Tag derivatization strategy adapted from the protocol supplied by Waters. 5,0 μL sample was spiked with an internal standard solution**.** Protein precipitation was performed by addition of MeOH and the sample was dried in a speedvac. The residue was reconstituted in borate buffer (pH 8.5) with AQC reagent. The prepared samples were transferred to autosampler vials and placed in an autosampler tray. The vials were cooled at 4^{o}C upon injection. 1,0 μL prepared sample was injected in a UPLC-MS/MS system. Chromatographic separation was achieved by an Agilent 1290 Infinity II LC System on an Accq-Tag Ultra column (Waters) with a flow of 0.7 mL/min over an 11 min gradient. The UPLC was coupled to electrospray ionization on a triple quadrupole mass spectrometer (AB SCIEX Qtrap 6500). Analytes were detected in the positive ion mode and monitored in Multiple Reaction Monitoring (MRM) using nominal mass resolution. Acquired data was evaluated using MultiQuant Software for Quantitative Analysis (AB SCIEX, Version 3.0.2), by the integration of assigned MRM peaks and normalization using proper internal standards. For analysis of amino acids, their 13C15N-labeled analogs were used. For other amines, the closest-eluting internal standard was employed. After quality control correction 48 amines complied with the acceptance criteria of RSDqc <15%. Additionally, 7 amines presented an RSDqc between 15 and 30%. They are included in the results but these compounds should be considered with caution.

Acylcarnitine profiling: The acylcarnitine platform covers acylcarnitines as well as trimethylamine-N-oxide, choline, betaine, deoxycarnitine, and carnitine. 10 µL sample was spiked with an internal standard solution. Protein precipitation was performed by

addition of MeOH. The supernatant was transferred to an autosampler vial and placed into an autosampler. The vials were cooled at 10°C upon injection. 1.0 µL of the prepared sample was injected into a triple quadrupole mass spectrometer. Chromatographic separation was achieved by UPLC (Agilent 1290, San Jose, CA, USA) on an Accq-Tag Ultra column (Waters) with a flow of 0.7 mL/min over an 11 min gradient. The UPLC was coupled to electrospray ionization on a triple quadrupole mass spectrometer (Agilent 6460, San Jose, CA, USA). Analytes were detected in the positive ion mode and monitored in Multiple Reaction Monitoring (MRM) using nominal mass resolution. Acquired data was evaluated using Agilent MassHunter Quantitative Analysis software (Agilent, Version B.05.01), by integration of assigned MRM peaks and normalization using proper internal standards. The closest-eluting internal standard was employed. After quality control correction 24 acylcarnitines complied with the acceptance criteria of RSDqc <15%. Additionally, 4 acylcarnitines presented an RSDqc between 15 and 30%. They are included in the results but these compounds should be considered with caution.

Organic acid profiling: The organic acid platform covers 28 organic acids. 50 µL sample was spiked with an internal standard solution. Protein precipitation was performed by addition of MeOH. After centrifugation, the supernatant was transferred and the sample was dried using a speedvac. Then, two-step derivatization procedures were performed on-line: oximation using methoxyamine hydrochloride (MeOX, 15 mg/mL in pyridine) as the first reaction and silylation using N-Methyl-N-(trimethylsilyl)- trifluoroacetamide (MSTFA) as the second reaction. 1 µL of each sample was directly after its derivatization injected on GC-MS. Gas chromatography was performed on an Agilent Technologies 7890A equipped with an Agilent Technologies mass selective detector (MSD 5975C) and MultiPurpose Sampler (MPS, MXY016-02A, GERSTEL). Chromatographic separations were performed on an HP-5MS UI (5% Phenyl Methyl Silox), 30 m × 0.25 m ID column with a film thickness of 25 µm, using helium as the carrier gas at a flow rate of 1,7 mL/min. A single-quadrupole mass spectrometer with electron impact ionization (EI, 70 eV) was used. The mass spectrometer was operated in SCAN mode mass range 50-500. Acquired data was evaluated using Agilent MassHunter Quantitative Analysis software (Agilent, Version B.05.01). After quality control correction and considering blank effects, 9 organic acid compounds complied with the acceptance criteria RSDqc <15% and blank effect <20%. 4 organic acids reported an RSDqc between 15 and 30% and should be considered with caution.

Negative lipid profiling: The negative lipid platform is a semi-target methodology for the identification of 30 fatty acids. 50 µL sample was spiked with 50 μL of an internal standard solution. Protein precipitation was performed by addition of 550 μL MeOH. After centrifugation, 600 μL supernatant was transferred and the sample was dried using a speedvac. The residue was reconstituted in 300 μL of isopropanol with 0,1% formic acid. The prepared samples were transferred to autosampler vials and placed in an autosampler tray. 8,0 µL of the prepared sample was injected into an LC-MS. The analysis was performed on an ACQUITY UPLC™ (Waters, the Netherlands) coupled to a high-resolution mass spectrometer with a Synapt G2 Q-TOF system (Waters, the

Netherlands) using reference lock mass correction. Lipids were detected in full scan in the negative ion mode. Chromatographic separation was achieved using an HSS T3 column (1.8 μm, 2.1 * 100 mm) with a flow of 0.4 mL/min over a 16-minute gradient. Acquired data was preprocessed using Targetlynx software (Masslynx, V4.1, SCN916). After quality control correction, 10 compounds complied with the acceptance criteria RSDqc <15%. 6 compounds reported an RSDqc between 15 and 30% and should be considered with caution.

Positive lipid profiling: The positive lipid platform covers 185 compounds including triglycerides (TGs, n=85) and non-triglycerides (non-TGs, n=100). 10 μL preprocessed sample was spiked with 1000 μL IPA containing internal standards and vortexed for 30 sec. Prepared samples were transferred to autosampler vials for LC-MS analysis. In total 2.5 μL prepared sample was injected for analysis. Chromatographic separation was achieved on an ACQUITY UPLC™ (Waters, Ettenleur, the Netherlands) with an HSS T3 column (1.8 μm, 2.1 * 100 mm) with a flow of 0.4 mL/min over a 16 min gradient. The lipid analysis is performed on a UPLC-ESI-Q-TOF (Agilent 6530, Jose, CA, USA) high-resolution mass spectrometer using reference mass correction. Lipids were detected in full scan in the positive ion mode. The raw data were preprocessed using Agilent MassHunter Quantitative Analysis software (Agilent, Version B.04.00). After quality control correction, 56 TGs and 39 non-TGs compounds complied with the acceptance criteria RSDqc<15% and blank effect <40 %. 1 TG and 53 non-TGs reported an RSDqc between 15 and 30% and should be considered with caution.

Signaling lipid profiling: The signaling lipids platform covers various isoprostane classes together with their respective prostaglandin isomers from different poly unsaturated fatty acids (PUFA), including n-6 and n-3 PUFAs such as dihomo-γ-linoleic acid (DGLA) and arachidonic acid (both n-6) and eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA) (both n-3). Also included in this platform are endocannabinoids, bile acids, and signaling lipids from the sphingosine and sphinganine classes and their phosphorylated forms, as well as three classes of lysophosphatidic acids. The three lysophosphatidic acid classes include lysophosphatidic acids (LPAs), lysophosphatidylglycerol (LPG), lysophosphatidylinositol (LPI), lysophosphatidyserine (LPS), lysophosphatidylethanolamines (LPE), cyclic-phosphatidic acids(cLPA), and fatty acid all ranging from C14 to C22 chain length species. The signaling and peroxidized lipids platform is divided into two chromatographic methods: low and high pH. In the low pH method, isoprostanes, prostaglandins, nitro-fatty acids, lyso-sphingolipids, endocannabinoids, and bile acids are analyzed. The high pH method covers lyso-sphingolipids, lysophosphatidic acids, lysophosphatidylglycerol, lysophosphatidylinositol, lysophosphatidyserine, lysophosphatidylethanolamines, cyclic-phosphatidic acids, and fatty acid. Each sample was spiked with antioxidant and internal standard solution. The extraction of the compounds is performed via liquid-liquid extraction (LLE) with butanol and methyl tert-butyl ether (MTBE). After collection, the organic phase is concentrated by first drying followed by reconstituted in a smaller volume. After reconstitution, the extract is transferred into amber autosampler vials and used for high and low pH injection. A Shimadzu system, formed

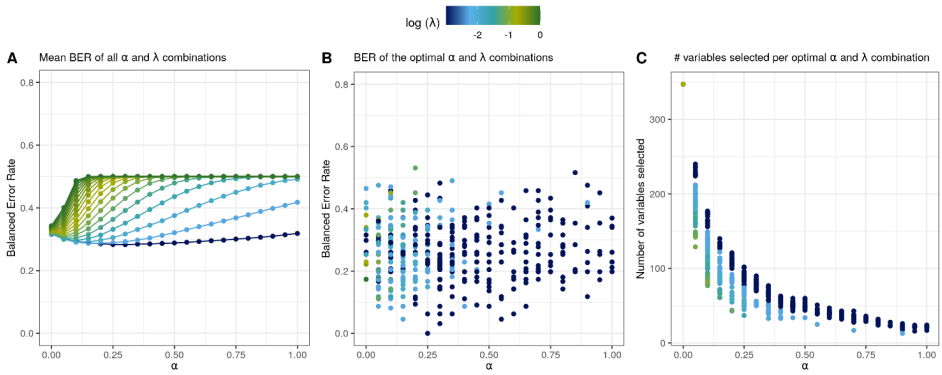

Figure 2.5 Optimization of α and λ in the inner cross-validation (CV) to reach a minimal balanced error rate (BER) in the outer CV. (A) Shows all α and λ values tested in inner CV against mean BER of the inner CV. (B) A plot of the optimal α and λ combinations chosen in the inner CV against their BER in the outer CV shows a variety of favorable α and λ concentrations. (C) A plot of the number of variables selected in the elastic net model in outer CV shows that with increasing alpha, the number of variables decreases as is expected in an elastic net model. The data shown in the figure is a result of the comparison Atypical – (*S. pneumoniae* + viral).

by three high-pressure pumps (LC-30AD), a controller (CBM-20Alite), an autosampler (SIL-30AC), and an oven (CTO-30A) from Shimadzu Benelux, was coupled online with an LCMS-8050 triple quadrupole mass spectrometer (Shimadzu) for high pH measurements. An LCMS-8060 triple quadrupole mass spectrometer (Shimadzu) was coupled to the Shimadzu system for low pH measurements. Both systems were operated using LabSolutions data acquisition software (Version 5.89, Shimadzu). The samples were analyzed by UPLC-MS/MS. An Acquity UPLC BEH C18 column (Waters) was used to measure the samples in the low pH method. For the high pH method, a Kinetex EVO column by Phenomenex was used. The triple quadrupole mass spectrometer was used in polarity switching mode and all analytes were monitored in dynamic Multiple Reaction Monitoring (dMRM). The acquired data was evaluated using LabSolutions Insight software (Version 3.1 SP1, Shimadzu), by integration of assigned MRM peaks and normalization using accordingly selected internal standards. When available, a deuterated version of the target compound was used as an internal standard. For the other compounds, the closest-eluting internal standard was employed. For low pH mode, after quality control correction, 46 metabolites complied with the acceptance criteria of RSDqc <15% and blank effect <40%. 6 compounds reported an RSDqc between 15 and 30% and should be considered with caution. For high pH mode, after quality control correction, 43 metabolites complied with the acceptance criteria of RSDqc <15% and blank effect <40%. Additionally, 18 compounds reported an RSDqc between 15 and 30% and should be considered with caution.

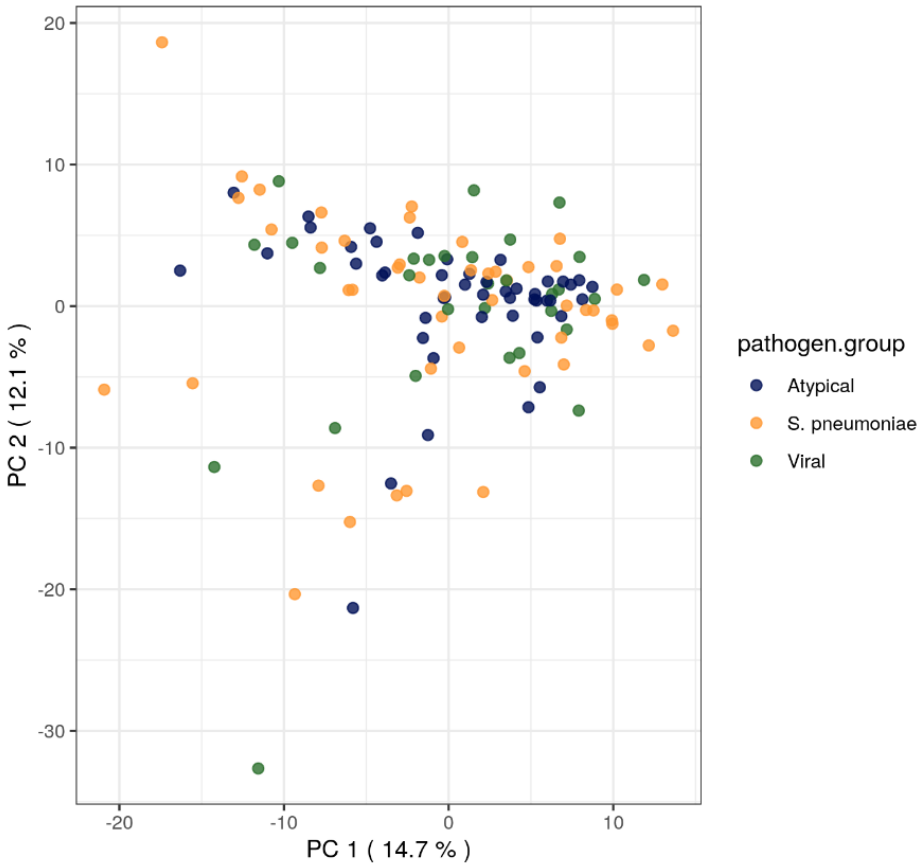

Figure 2.6 Unsupervised principal component analysis (PCA) plot of all pathogen groups.

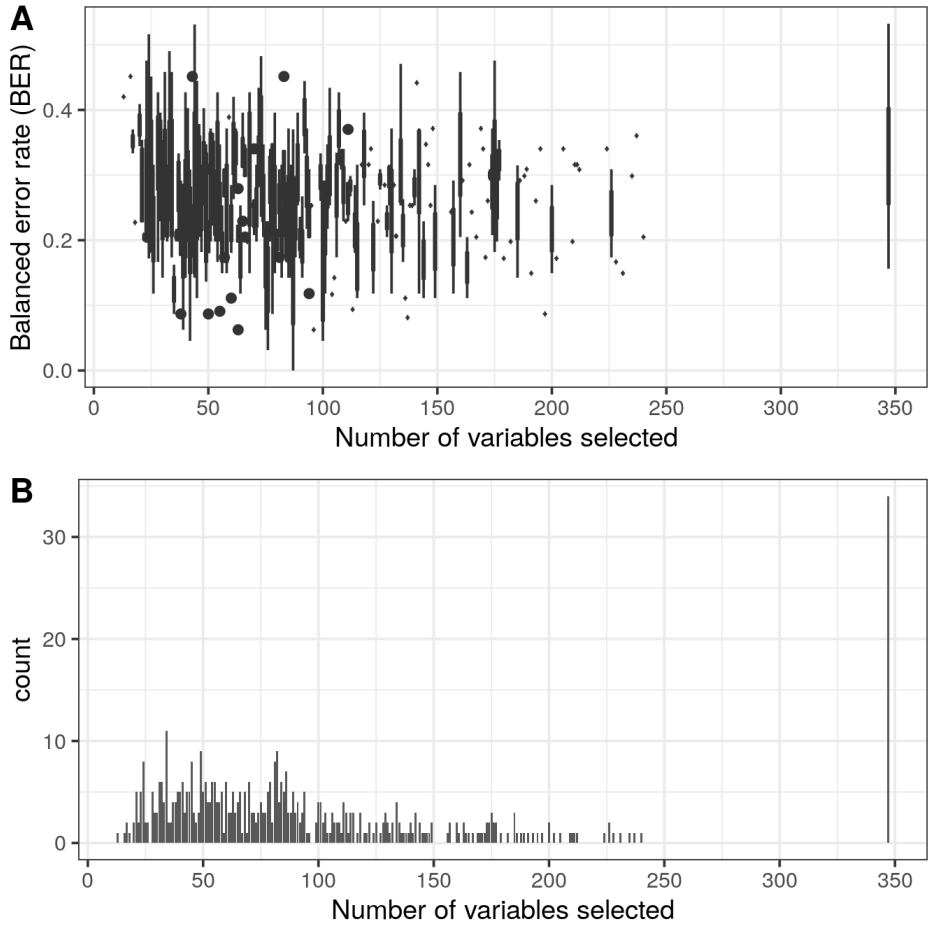

Figure 2.7 Boxplot of BER per number of variables selected shows no clear relation between the number of variables selected and model performance. (B) Histogram of the number of variables selected shows that a model with all metabolites included is favored, followed by models including 34, 49, 82, 24, or 45 metabolites. Both Figs contain the data of all folds and repeats (n=500) for the comparison between atypical versus *S. pneumoniae* and viral infections.

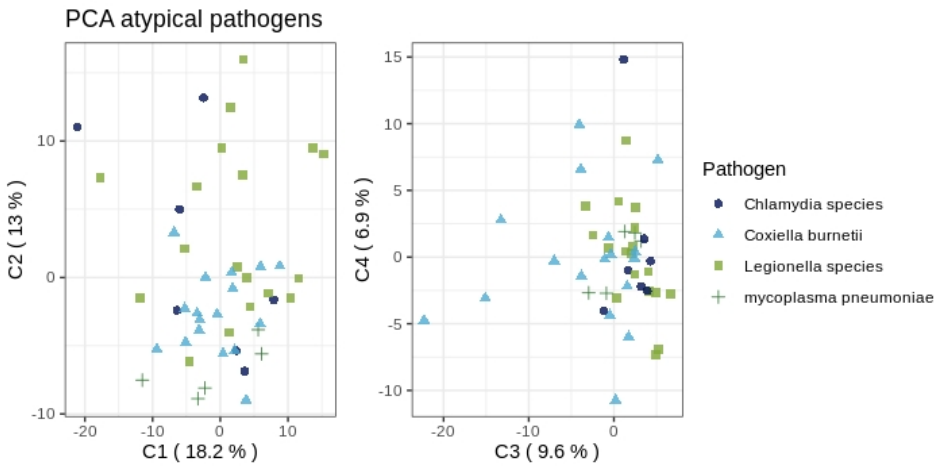

Figure 2.8 Principal component analysis (PCA) of the atypical pathogen group (log-transformed and standardized data) shows that there is no clear subgroup within the atypical group that would prominently drive the separation from the *S. pneumoniae* and viral infections.

Table 2.4 Summary of previous studies focusing on bacterial and viral respiratory tract infections and related metabolites.

Compared groups	Matrix	Analytical method	Significantly altered metabolites		Reference
			Upregulated	Downregulated	
30 CAP vs 30 HC	Plasma	NMR	1-methylhistidine	lactate, ketone bodies	Zhou et al. (2015)
30 CAP vs 38 TB	Plasma	NMR	lactate, pyruvate, lipids, ketone bodies	amino acids[leucine, isoleucine, valine], 1-methylhistidine, glucose, nicotinate, GPC	Zhou et al. (2015)
11 pneumonia vs 11 HC (children)	Plasma	UPLC-TOF-MS	uric acid, hypoxanthine, glutamic acid	L-tryptophan, adenosine-diphosphate	Laiakis et al. (2010)
11 pneumonia vs 11 HC (children)	Urine	UPLC-TOF-MS	uric acid, L-histidine		Laiakis et al. (2010)
47 pneumonia vs 47 HC	Urine	NMR	glucose, lactate, ketone bodies, amino acids [alanine, asparagine, isoleucine, leucine, lysine, serine, threonine, tryptophan, tyrosine, valine], carnitine, acetylcarnitine, hypoxanthine, fucose, myo-inositol, taurine, quinolinate, adipate, dimethylamine, creatine, 2-oxoglutarate, fumarate	citrate, trigonelline, 1-methylnicotinamide, succinate, levoglucosan, 1-methylhistidine	Slupsky et al. (2009)
30 CAP vs 46 TB	Plasma	UPLC-QTOF-MS	12(R)-hydroxyeicosatetraenoic acid, ceramide (d18:1/16:0), cholesterol sulfate, 4a-formyl-4b-methyl-5a-cholesta-8-en-3b-ol		Lau et al. (2015)
42 Influenza A vs 30 Bacterial CAP	Plasma	NMR, GC-MS	3-Methyl-2-Isovalerate, 3-Methyl-2-oxovalerate, 4-Hydroxybutyrate, Adipate, Alanine, Arabinonic acid, Asparagine, Aspartic Acid, Citrate, Citric acid, Fumerate, Histidine, Lysine, Methionine, Myoinositol, Phenylalanine, Serine, Threonic Acid, Threonine, Tyrosine, Uric acid, Urea	2-amino Butanoic acid, Acetoacetate, Alkane, Benzoic acid, Beta-alanine, Carnitine, Dimethylamine, Formate, Glycine, Gulonic acid, Hexanoic acid, Leucine, Lactic acid, Pentadecane, Pyruvic acid, Quinic acid	Banoei et al. (2017)
55 RVS vs 24 Bacterial pneumonia vs 37 HC (children)	Urine	NMR	3-Hydroxyisovalerate, 3-Indoxylsulfate, Acetoacetate, Betaine, Blue 1.06, Ethanolamine, Glutamate, N,N-Dimethylglycine, Pantothenate, Succinate, Tartrate, Uracil	Hippurate, Serine, Threonine	Adamko et al. (2016)

Abbreviations: CAP: community-acquired pneumonia; VAP: ventilator-associated pneumonia; HAP: hospital-acquired pneumonia; TB: tuberculosis; RSV: respiratory syncytial virus; HC: healthy control; NMR: nuclear magnetic resonance; UPLC: ultra-performance liquid chromatography; GC: gas chromatography; TOF: time-of-flight; QTOF: quadrupole time-of-flight; MS: mass spectrometry.

Table 2.5 Additional patient characteristics per pathogen group.

	S. pneumoniae (N=48)	Atypical (N=47)	Viral (N=30)	P-value
Race				
Other	1 (2.1%)	1 (2.1%)	0 (0%)	
White	31 (64.6%)	46 (97.9%)	30 (100%)	0.81
Missing	16 (33.3%)	0 (0%)	0 (0%)	
Nursing home resident				
No	46 (95.8%)	47 (100%)	25 (83.3%)	
Yes	1 (2.1%)	0 (0%)	4 (13.3%)	0.07
Missing	1 (2.1%)	0 (0%)	1 (3.3%)	
Altered mental status				
No	43 (89.6%)	42 (89.4%)	27 (90.0%)	
Yes	3 (6.2%)	5 (10.6%)	3 (10.0%)	0.85
Missing	2 (4.2%)	0 (0%)	0 (0%)	
Respiratory rate				
Mean (SD)	25.3 (6.64)	25.5 (6.44)	26.9 (7.32)	
Median [Min, Max]	25.5 [12.0, 40.0]	26.0 [14.0, 40.0]	29.0 [12.0, 44.0]	0.81
Missing	8 (16.7%)	8 (17.0%)	6 (20.0%)	
Systolic blood pressure				
Mean (SD)	131 (25.3)	133 (15.8)	137 (23.2)	
Median [Min, Max]	130 [88.0, 226]	130 [99.0, 161]	135 [90.0, 186]	0.81
Missing	1 (2.1%)	0 (0%)	1 (3.3%)	
temperature				
Mean (SD)	23.5 (8.41)	24.2 (11.5)	19.9 (9.40)	
Median [Min, Max]	24.0 [6.00, 42.0]	24.0 [1.00, 41.0]	20.0 [3.00, 39.0]	0.37
Missing	1 (2.1%)	0 (0%)	0 (0%)	
pulse				
Mean (SD)	104 (21.6)	96.6 (18.6)	94.3 (17.9)	
Median [Min, Max]	109 [60.0, 144]	93.0 [50.0, 140]	96.0 [60.0, 120]	0.28
Missing	1 (2.1%)	0 (0%)	0 (0%)	
pH				
Mean (SD)	12.7 (4.64)	14.5 (4.28)	12.0 (4.66)	
Median [Min, Max]	14.0 [3.00, 21.0]	14.0 [3.00, 22.0]	13.0 [1.00, 19.0]	0.31
Missing	8 (16.7%)	19 (40.4%)	5 (16.7%)	
BUN				
Mean (SD)	38.9 (24.9)	46.0 (21.6)	46.5 (24.1)	
Median [Min, Max]	35.0 [2.00, 81.0]	48.0 [1.00, 84.0]	52.0 [4.00, 82.0]	0.50
Missing	1 (2.1%)	0 (0%)	1 (3.3%)	
sodium				
Mean (SD)	132 (4.68)	131 (5.59)	136 (5.03)	
Median [Min, Max]	132 [117, 141]	132 [119, 141]	136 [125, 152]	0
Missing	1 (2.1%)	0 (0%)	0 (0%)	
glucose				
Mean (SD)	34.6 (13.9)	34.7 (14.0)	36.9 (17.6)	
Median [Min, Max]	35.0 [1.00, 58.0]	35.0 [5.00, 59.0]	42.0 [2.00, 59.0]	0.85
Missing	3 (6.2%)	0 (0%)	5 (16.7%)	
hematocrit				
Mean (SD)	11.0 (4.29)	11.5 (3.75)	10.5 (3.75)	
Median [Min, Max]	11.0 [1.00, 19.0]	12.0 [1.00, 17.0]	11.5 [1.00, 16.0]	0.81
Missing	2 (4.2%)	0 (0%)	2 (6.7%)	

	S. pneumoniae (N=48)	Atypical (N=47)	Viral (N=30)	P-value
Partial pressure of oxygen				
Mean (SD)	30.6 (16.5)	31.8 (17.7)	30.1 (16.3)	
Median [Min, Max]	34.5 [2.00, 56.0]	33.0 [1.00, 55.0]	37.0 [1.00, 50.0]	0.92
Missing	8 (16.7%)	19 (40.4%)	5 (16.7%)	
Pleural effusion on x ray				
No	39 (81.2%)	45 (95.7%)	25 (83.3%)	
Yes	8 (16.7%)	2 (4.3%)	5 (16.7%)	0.31
Missing	1 (2.1%)	0 (0%)	0 (0%)	

Data are presented as number (%) or mean (SD). Abbreviations: BUN: blood urea nitrogen.

Table 2.6 Overview of the number of metabolites included in the metabolomics platforms, measured in the samples and included in the data analysis.

Measurement platform	Number of metabolites included in platform	Number of metabolites measured in samples	Number of metabolites included in data analysis
Amines	74	55	55
Acylcarnitines	48	28	28
Organic acids	28	13	13
Negative lipids	30	16	16
Signaling lipids	231	113	91
Positive lipids	185	149	144
Total	**596**	**374**	**347**

Table 2.7 Information on measurement platforms used, metabolite classes targeted per platform, targeted metabolites, their abbreviations and names in R (if detected) and identifiers (if available).

This table is available as Excel file (S4 Table) on the website of the publisher at https://doi.org/10.1371/journal.pone.0252378.

Table 2.8 Metabolomics data after quality control.

This table is available as csv file (S5 Table) on the website of the publisher at https://doi.org/10.1371/journal.pone.0252378.