Bacterial and viral respiratory tract microbiota and host characteristics in adults with lower respiratory tract infections: a case-control study

Bastiaan W. Haak\textsuperscript{1,2}, Xanthe Brands\textsuperscript{1}, Mark Davids\textsuperscript{2}, Hessel Peters-Sengers\textsuperscript{1}, Robert F.J. Kullberg\textsuperscript{1,2},
Robin van Houdt\textsuperscript{3}, Floor Hugenholtz\textsuperscript{1,2}, Daniël R. Faber\textsuperscript{4}, Hans L. Zaat\textsuperscript{3}, Brendon P. Scicluna\textsuperscript{1,5}, Tom van der Poll\textsuperscript{1,6}, W. Joost Wiersinga\textsuperscript{1,2,6}

\textsuperscript{1}Center for Experimental and Molecular Medicine (CEMM), Amsterdam University Medical Centers - Location AMC, University of Amsterdam, Amsterdam, The Netherlands
\textsuperscript{2}Microbiota Center Amsterdam, Amsterdam University Medical Centers - Location AMC, University of Amsterdam, Amsterdam, The Netherlands
\textsuperscript{3}Department of Virology, Amsterdam University Medical Centers - Location AMC, University of Amsterdam, Amsterdam, The Netherlands
\textsuperscript{4}Department of Internal Medicine, BovenIJ hospital, Amsterdam, The Netherlands
\textsuperscript{5}Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam University Medical Centers - Location AMC, University of Amsterdam, Amsterdam, The Netherlands
\textsuperscript{6}Department of Internal Medicine, Division of Infectious Diseases, Amsterdam University Medical Centers - Location AMC, University of Amsterdam, Amsterdam, The Netherlands

Corresponding author: Bastiaan W. Haak, MD PhD, Amsterdam UMC, University of Amsterdam, Center for Experimental and Molecular Medicine, Meibergdreef 9, Room G2-130, 1105 AZ, Amsterdam, the Netherlands, Phone: +31-20-5669111, E-mail: b.w.haak@amsterdamumc.nl

© The Author(s) 2021. Published by Oxford University Press for the Infectious Diseases Society of America.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
**Summary**: Classification models using a small set of nasopharyngeal bacteria, viruses and clinical variables can be leveraged to accurately distinguish respiratory tract infections from health, with even better results when bacterial and viral infections were stratified.
ABSTRACT

**Background** Viruses and bacteria from the nasopharynx are capable of causing community-acquired pneumonia (CAP), which can be difficult to diagnose. We aimed to investigate whether shifts in the composition of these nasopharyngeal microbial communities can be used as diagnostic biomarkers for CAP in adults.

**Methods** We collected nasopharyngeal swabs from adult CAP patients and controls without infection in a prospective multicenter case-control study design. We generated bacterial and viral profiles using 16S ribosomal RNA gene sequencing and multiplex PCR, respectively. Bacterial, viral and clinical data were subsequently used as inputs for extremely randomized trees classification models aiming to distinguish subjects with CAP from healthy controls.

**Results** We enrolled 117 cases and 48 control subjects. Cases displayed significant beta diversity differences in nasopharyngeal microbiota ($P=0.016$, $R^2=0.01$) compared to healthy controls. Our extremely randomized trees classification models accurately discriminated CAP caused by bacteria (area under the curve (AUC) 0.83), viruses (AUC 0.95) or mixed origin (AUC 0.81) from healthy control subjects. We validated this approach using a dataset of nasopharyngeal samples from 140 influenza patients and 38 controls, which yielded highly accurate (AUC 0.93) separation between cases and controls.

**Conclusions** Relative proportions of different bacteria and viruses in the nasopharynx can be leveraged to diagnose CAP and identify etiologic agent(s) in adult patients. Such data can inform the development of a microbiota-based diagnostic panel used to identify CAP patients and causative agents from nasopharyngeal samples, potentially improving diagnostic specificity, efficiency, and antimicrobial stewardship practices.

**Keywords:** community-acquired pneumonia, influenza, 16S rRNA sequencing, microbiome, nasopharynx
INTRODUCTION

Community-acquired pneumonia (CAP) is the leading cause of hospitalization and death worldwide[1–3]. The diagnosis of CAP can often be challenging for clinicians, as symptoms of acute heart failure, chronic obstructive pulmonary disease (COPD), or pulmonary embolism may mimic the presentation of the disease[1]. In addition, patients with viral and bacterial lower respiratory tract infections clinically present similarly, and a confirmed microbiological diagnosis is only obtained in approximately half of all cases[1,4].

An increasing amount of research has focused on studying the composition and function of commensal micro-organisms in the nasopharynx, which has shown to be an ecological proxy for the lower airway microbiota[5–7]. While lower in total biomass compared to the gastrointestinal tract, the upper respiratory tract harbors a surprisingly diverse and stable ecosystem of bacteria, viruses and fungi. Common respiratory pathogens, such as Streptococcus (S.) pneumoniae and Haemophilus (H.) influenzae are frequently identified in the healthy nasopharynx, yet a balanced community of other commensal bacteria prevents the overgrowth of these pathobionts[8,9]. However, studies have shown that a loss of these mechanisms of protection by commensal bacteria, otherwise known as colonization resistance, facilitates the enrichment of single bacterial taxa, ultimately leading to respiratory infections[10–15].

The observed relationship between the loss of colonization resistance and respiratory infections has fueled the hypothesis that shifts in nasopharyngeal bacterial communities could potentially be employed to facilitate the diagnosis of CAP. A recent matched case-control study in neonates strengthened this hypothesis by demonstrating that a classifier model based on nasopharyngeal bacterial and viral communities can accurately identify lower respiratory tract infections[5]. These findings could have implications for future treatment protocols, potentially leading to a reduction in the inappropriate use of antibiotics, and while currently underexplored, are therefore of equal interest to adult CAP patients[16]. However, most studies have limited their scope to identifying the
absence or presence of a single nasopharyngeal bacterial species, which does not account for the role that shifts in bacterial communities as a whole could play in the acquisition of CAP[8]. Therefore, this study aimed to investigate if shifts in bacterial and viral communities of the nasopharynx can aid in diagnosis of CAP in adults who present to the hospital.

METHODS

Study design and patient recruitment

Details of recruitment have been published previously[17]. In brief, consecutive patients older than 18 years admitted to the Amsterdam UMC, location Academic Medical Center (AMC) or BovenIJ hospital in the Netherlands during the influenza seasons (October 2016 – June 2017 and October 2017 – June 2018) were screened by trained research physicians. Patients were included if they were admitted with a clinical suspicion of a community-acquired pneumonia. Patients exposed to antibiotics within 48 hours prior to hospital admission, with the clinical suspicion of an aspiration pneumonia or a hospital-associated pneumonia were excluded. Subjects of comparable age and sex, who presented for periodical control of cardiovascular management, diabetes care or cancer follow up at the outpatient clinic of the Amsterdam UMC, location AMC, served as controls without acute infection.

Data collection

Nasopharyngeal swabs in Universal Transport Medium (UTM™, Copan) were taken within 24 hours of hospital admission and one month thereafter and stored immediately at -80°C. Clinical data and host characteristics were retrieved from electronic medical records and standardized case report forms. We also collected microbiological data regarding causative pathogens (based on a combination of viral nasal/throat swab PCR, urine antigen tests, blood cultures and sputum
Microbiological results were deemed to be clinically relevant if the physician caring for the patient deemed an infection was present, and elicited a treatment plan based on these findings. Written informed consent was obtained from all eligible participants, or their legal representatives. The study protocol was approved by the local institutional review boards (ref number NL57847.018.16).

**Bacterial and viral analysis**

A detailed methodology of the bacterial sequencing procedure is described in the Supplementary Material and in prior publications of our group[18]. In short, 16S rRNA gene amplicons were generated using a single step PCR protocol targeting the V3-V4 region. The libraries were sequenced using a MiSeq platform using V3 chemistry with 2x251 cycles. Amplified Sequence Variants (ASVs) were inferred for each sample individually with a minimum abundance of 4 reads[19]. Previously detected contaminating sequences, identified using negative controls and the decontam package, were removed[20]. Nasopharyngeal viral communities were analyzed by multiplex real-time PCR (RespiFinder SMARTfast 22 [Maastricht, Netherlands]).

**Statistical analysis**

Statistical analysis was performed in R (Version 3.6.1, Vienna, Austria). To assess alpha diversity and richness, we calculated the Inverse Simpson Index and Observed Taxa Richness index with the phylloseq package[21]. Data were not normally distributed and were therefore analyzed using either a Wilcoxon rank sum or Kruskal-Wallis test. Beta diversity was assessed using the weighted and unweighted UniFrac distance metrics, and differences among groups were tested using PERMANOVA as implemented in the vegan package. DESeq2 analysis was used to identify differentially abundant bacterial genera[22]. Finally, extremely randomized trees classification models were used to assess...
the value of a combination of clinical variables and nasopharyngeal bacterial and viral communities to distinguish CAP of bacterial, viral and mixed etiology from health[23]. Extremely randomized trees classification models are considered one of the best models to identify bacterial taxa associated with disease[24,25]. The relative abundance of the top 40 nasopharyngeal bacterial genera, viral presence and host characteristics (for details see Supplementary Table 1) were used as input for the models. We performed 100 iterations of 5-fold cross validation on 75% of each of the datasets, with subsequent testing on the remaining 25% of the samples, and assessed the performance of these classifiers by calculating the mean area under the receiver operating characteristic curve (AUC-ROC) of all 100 shuffles. The models were implemented in Python (v. 3.7.4) using numpy (v. 1.16.4), pandas (v. 0.25.1), and scikit-learn (v. 0.21.2) packages. We validated this approach using a dataset of publicly available 16S rRNA V4 sequences (accession number SRP132207), of nasopharyngeal samples from 140 patients with influenza A virus admitted at New York Presbyterian Hospital and 38 healthy controls [12].

RESULTS

117 cases and 48 control subjects were enrolled in this study (CONSORT flow diagram is provided in Supplementary Figure 1). Median age of cases (69.0 years; IQR 60.0 – 78.0) and controls (70.5 years; IQR 63.8 – 75.0) were similar. Demographic characteristics, dietary habits, prior antibiotic exposure and comorbidities (diabetes, cardiovascular disease, malignancy, gastrointestinal disease and/or chronic renal disease) were comparable between CAP patients and controls. However, cases had a lower body mass index ($P=.014$) and a higher prevalence of COPD ($P=.004$) and immunocompromised status ($P=.022$) compared to controls. (Table 1) The median Pneumonia Severity Index (PSI) class[26] at admission was 4 (IQR 3–4). Intensive Care Unit admission was required for 9 CAP patients (7.8%) and 28-day mortality was 4.4%. Blood cultures, sputum cultures and urine antigen tests were obtained in 109 (93.2%), 68 (58.1%) and 73 (62.4%) patients,
respectively. Viral data were available for all CAP cases and 47 controls (97.9%). A causative pathogen was identified in 68 patients (59.0%), with 12 patients (10.3%) having co-pathogen infection. 19 patients (16.2%) were infected by S. pneumoniae, 11 patients (9.4%) by H. influenzae and 5 by Staphylococcus (S.) aureus (4.3%). 37 patients (31.6%) were diagnosed with a respiratory virus, of which 13 (17.9%) were attributed to Influenza A or Influenza B virus; other prevalent viruses were rhinovirus (10 patients), parainfluenza virus 1-4 (7 patients), and coronavirus NL63 (5 patients) (Figure 1). Two asymptomatic healthy control subjects displayed colonization with a respiratory virus; one with human metapneumovirus (hMPV) and one with rhinovirus.

16S rRNA gene sequencing of nasopharyngeal samples yielded 11,658,346 high-quality reads (average 45,187 per sample, range 2,181-135,753), classified into 397 Amplified Sequence Variants (ASVs). We first investigated the microbiota diversity profiles of the nasopharyngeal samples of control subjects and cases at admission and follow-up. Inverse Simpson diversity and Observed Taxa richness indices were comparable between all groups (Figure 2A-B). Unweighted UniFrac beta diversity of the nasopharyngeal microbiota of cases at admission differed significantly from controls (P=.016, R²=.01; Figure 2C), whereas the weighted UniFrac remained comparable between all groups. (Figure 2D). These findings are supported by the large heterogeneity of nasopharyngeal microbiota composition among all study participants, with varying degrees of nasopharyngeal domination of the genera Corynebacterium, Staphylococcus and Dolosigranulum (Supplementary Figure 2). Although it has been shown that COPD is associated with airway microbiota alterations[27], we found no interindividual dissimilarities in nasopharyngeal microbiota composition and diversity between cases at admission with and without COPD (Supplementary Figure 3).

Next, we verified if elevated nasopharyngeal relative abundances (proportion of total 16s rRNA reads) of common causative CAP agents corresponded with detection in blood cultures, sputum
cultures and urinary antigen tests. We observed that at admission, cases with a confirmed *H. influenzae* and *S. pneumoniae* pneumonia harbored higher nasopharyngeal relative abundances of *Haemophilus* and *Streptococcus* species, respectively, in comparison to cases with other causative pathogens and controls (*P*=.002 and *P*=.0099, **Figure 3**). These findings were independent of the microbial detection method by which the causative pathogen was identified (**Supplementary Figure 4**). Nasopharyngeal samples collected following antimicrobial treatment (one month following admission for CAP) displayed normalized abundances of the corresponding pathogens. This infers that the nasopharynx acts as a proxy of clinically meaningful lower respiratory tract infections.

Given these observations, we aimed to explore if nasopharyngeal bacterial community structures, rather than single bacterial taxa, were capable of distinguishing CAP patients at hospital admission from controls. We employed extremely randomized trees classification models using the relative abundance of the top 40 nasopharyngeal bacterial genera, viral presence and clinical variables (depicted in **Supplementary Table 1**) as input. The accuracy of the combined use of these three parameters on the classification of CAP versus health was high, with a mean cross validation AUC of 0.81 (SD ± 0.05; **Figure 4A**). Important bacterial discriminatory features of pneumonia were, among others, a high abundance of *Haemophilus, Streptococcus, Actinomyces,* and *Curvibacter* in the nasopharynx, and low abundance of several nasopharyngeal commensals, such as *Corynebacterium, Cutibacterium* and *Lawsonella* (**Figure 4B**). Most of these genera were among the most differentially abundant features between cases and controls (**Supplementary Figure 5**). Clinical discriminators of CAP were prior antibiotic treatment in the past three months, COPD, low BMI, immunosuppressive disease and diabetes (**Figure 4B**). This combined classification system outperformed the models based on bacterial microbiota (AUC 0.71 ± 0.08), viral microbiota (AUC 0.63 ± 0.04), and clinical characteristics alone (AUC 0.71 SD ± 0.06) (**Supplementary Figure 6**). Separate models for isolated bacterial CAP (33 cases) and viral CAP (27 cases) showed even higher accuracy, with an AUC of 0.83 (SD ± 0.07) and 0.95 (SD ± 0.04), respectively (**Figure 4A**). Of interest, most discriminating factors for bacterial CAP (**Figure 4C**), specifically the absence of commensal nasopharyngeal communities, such
as Corynebacterium, Staphylococcus, Cutibacterium and Lawsonella, were similar to the mixed analysis, with viral pathogens dominating the discriminatory power of the viral CAP model (Figure 4D).

Given the discriminatory power of these communities in distinguishing CAP from health, we validated our approach in a recently published dataset of hospitalized patients with influenza A infection and controls from New York, USA[12]. In correspondence with the results obtained in our data set, we observed that the use of nasopharyngeal bacterial community structures alone allowed for excellent distinguishing capacity between influenza patients and control subjects with a mean holdout cross validation AUC of 0.93 (SD ± 0.04; Supplementary Figure 7). Combined, these findings show that lower respiratory tract infections are associated with consistent shifts in nasopharyngeal community compositions. Such data can inform the development of a microbiota-based diagnostic panel used to identify CAP patients and causative agents from nasopharyngeal samples.

DISCUSSION

We demonstrate that CAP in adults can be robustly differentiated from health using a small set of nasopharyngeal bacteria, viruses and clinical variables. This study sheds light on the complexity of the composition of nasopharyngeal microbes during infections, as direct comparison of the microbiota between cases and controls revealed only subtle changes. Specifically, we observed no differences in alpha diversity and richness of nasopharyngeal communities of patients with CAP and controls in this study. In addition, CAP patients displayed small but significant beta diversity differences that were characterized by a higher prevalence of pathogenic bacterial taxa, such as S. pneumoniae and H. influenzae. The observed differences were not driven by altered exposure to antibiotics or altered dietary habits, as these exposures were similar between patients and controls. Patients with microbiologically confirmed S. pneumoniae or H. influenzae infection displayed increased abundance of the corresponding pathogens in the nasopharynx, which is in line with the
hypothesis that the nasopharynx can indeed be considered an important source of these microorganisms in respiratory infections\cite{5,14,15}. However, a large proportion of CAP patients displayed low levels of these pathogens, indicating that no one-size-fits-all community composition exists during CAP. In support of this theory, a recent human challenge model with influenza suggests that specific baseline microbiota communities, rather than single bacterial species alone, might be most relevant in controlling colonization and spread of \textit{S. pneumoniae}\cite{9}.

Given these observations, we aimed to investigate if shifts beyond microbiota diversity metrics and single bacterial taxa comparisons could be employed to discriminate CAP from health. Despite the observed heterogeneity in both clinical variables and causative agents underlying the disease, our model was capable of robustly discriminating health and disease. The community structures were driven by alterations in the core microbial taxa \textit{Staphylococcus}, \textit{Dolosigranulum}, \textit{Cutibacterium} and \textit{Corynebacterium} as well as rare and low abundant taxa, such as \textit{Paracoccus} and \textit{Gemella}, while the abundance of causative pathogens attributed to a lesser extent to the distinguishing capacity of the model. Upon validation of our approach in a publicly available dataset of nasopharyngeal 16S rRNA sequences\cite{12}, the distinguishing capacity of Influenza A virus infection based on nasopharyngeal bacterial markers was nearly perfect. The discrepancy in performance of both classification models could potentially be explained by the homogeneity of the validation cohort, which only consisted of patients infected with a single pathogen, compared to the heterogeneity of both viral and bacterial pathogens of the original cohort. However, the distinguishing capacity of the classification model remained robust even in a setting of CAP with a wide variety of viral, bacterial and unknown-causative pathogens.

Given the rapid decrease in sequencing costs and time in recent years\cite{28}, our findings in adults and those of others in children\cite{5}, indicate that microbiota-targeted tools could improve the diagnostic specificity and efficiency of identifying respiratory infections. Of interest, a proof-of-principle study
using rapid microbiota sequencing with MinION technology in adult intensive care unit patients with pneumonia has shown that such diagnostic tools hold promise and could be clinically applied[16,29].

There are several potential explanations for the observed shifts in nasopharyngeal microbiota composition during CAP. First, the changes observed in the nasopharyngeal microbiome may precede and contribute to increased susceptibility to infection. For example, recent preclinical evidence shows that mice display larger perturbations of the upper respiratory microbiome with age, leading to an enrichment and increased risk of infection by *S. pneumoniae*[30]. A recent cross-sectional study investigating the upper respiratory tract microbiota revealed that increasing age was associated with a loss of common nasopharyngeal commensals *Corynebacterium*, *Dolosigranulum*, *Staphylococcus* and *Cutibacterium*, and a relative enrichment of oral flora, such as *Actinomyces*, which could predispose to reduced colonization resistance and an increased risk of infections[10]. Longitudinal studies in neonates have shown that early airway colonization with pathogenic bacteria occurred prior to the microbiological detection of (viral) pathogens and acute symptoms, and this overgrowth is therefore considered an increased risk for respiratory infections[13–15]. In our data, this mode of action is supported by the finding that prior antibiotic use is a strong marker for bacterial CAP.

Secondly, the host response induced by the respiratory infection itself allows for selective enrichment of certain bacteria, therefore leading to an altered composition of the nasopharyngeal microbiota. For example, it has been shown that separate mechanisms involving Th17-cell responses and interferon-lambda production disrupt the nasopharyngeal microbiome and predispose the occurrence of *S. pneumoniae* and *S. aureus* colonization[31,32]. Support for these mechanisms has also been demonstrated in the context of viral infection, as two recent human rhinovirus challenge models and one study investigating the nasopharyngeal microbiome during COVID-19 pneumonia showed the potential of disturbing bacterial nasopharyngeal communities, further elucidating the close relationship between bacterial and viral kingdoms[33-35]. Longitudinal observational studies
should further elucidate the role of the observed communities in the development of respiratory infections. Regardless of the directionality of the observed shifts, this study adds to the growing body of work that supports a common pathway for the development of viral and bacterial CAP, in which a loss of microbial colonization resistance and individual host factors are more important drivers of disease than the characteristics of individual pathogens[5,36].

Strengths of this study include the use of validated classification models and the prospective case-control design. A limitation of this investigation is the use of amplicon-based 16S rRNA gene sequencing, which provides limited taxonomic resolution at species level. In addition, PCR bias did not allow us to directly validate the yielded models from our dataset (which sequenced the V3-V4 region of the 16S rRNA gene) to the validation dataset (which only addressed the V4 region)[37]. It is warranted to increase the taxonomic resolution and use reproducible metagenomic sequencing based tools in order to validate our approach across multiple longitudinal cohorts and clinical settings.

In conclusion, this study shows that the relative proportions of bacteria and viruses - based on 16S rRNA gene sequencing and multiplex PCR - can be leveraged to diagnose CAP and identify etiologic agents in adult patients. Future studies are warranted to validate these classification models in the clinical setting, specifically aiming to distinguish patients with respiratory infections from those presenting with heart failure, pulmonary embolism or inflammatory pneumonitis. In addition, further study is warranted to validate if nasopharyngeal communities could discriminate bacterial CAP from viral etiologies with sufficient sensitivity and specificity. If addressed, such data can inform the development of a microbiota-based diagnostic panel used to identify CAP patients and causative agents from nasopharyngeal samples, potentially improving diagnostic specificity, efficiency, and antimicrobial stewardship practices.
NOTES

Contributors

BWH, XB, BPS, TP and WJW conceived the original study. DRF, WJW and TvdP oversaw sample collection. BWH and XB acquired all samples. RvH and HLZ designed and performed the viral multiplex PCR. Microbiome sequencing and initial analysis was performed and facilitated by FH and MD. MD performed the extremely randomized tree analysis. Statistical analysis was overseen by HPS and BPS. BWH and RFK analyzed the data, wrote the original manuscript, and prepared the final figures. TvdP, BPS and WJW secured funding for this project. All authors have seen and approved the final version of the manuscript.

Acknowledgments

We thank all patients and their families who participated in the study. We also thank Jorn Hartman, Stijn Klarenbeek and Rosan van der Lee for their help with the workup of the nasopharyngeal microbiota samples. Raw sequencing data (bacterial ASVs) have been submitted to the European Nucleotide Archive (accession number PRJEB42265).

Funding

This work was supported by the Netherlands Organization for Scientific Research [VIDI grant number 91716475 to WJW], and the Netherlands Organization for Health Research and Development [ZonMW grant number 50-53000-98-139 to XB]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of interests

WJW served as consultant for GSK (DSMB committee), fees paid to institution. The other authors report no competing interests.
REFERENCES

1. Prina E, Ranzani OT, Torres A, Paulo S, Paulo S. Community-acquired pneumonia. Lancet 2015; 386: 1097–108.

2. Jain SS, Self WH, Wunderink RG, et al. Community-Acquired Pneumonia Requiring Hospitalization among U.S. Adults. N Engl J Med 2015; 373: 415–27.

3. World Health Organization. The top 10 causes of death. 2018. https://www.who.int/health-topics/news-room/fact-sheets/detail/the-top-10-causes-of-death. (accessed July 12, 2019).

4. Tonkin-Crine SK, Tan PS, van Hecke O, et al. Clinician-targeted interventions to influence antibiotic prescribing behaviour for acute respiratory infections in primary care: an overview of systematic reviews. Cochrane Database Syst Rev 2017; published online Sept 7. DOI:10.1002/14651858.CD012252.pub2.

5. Man WH, van Houten MA, Mérelle ME, et al. Bacterial and viral respiratory tract microbiota and host characteristics in children with lower respiratory tract infections: a matched case-control study. Lancet Respir Med 2019; 7: 417–26.

6. Demuri GP, Gern JE, Eickhoff JC, Lynch S V, Wald ER. Dynamics of Bacterial Colonization with Streptococcus pneumoniae, Haemophilus influenzae, and Moraxella catarrhalis during Symptomatic and Asymptomatic Viral Upper Respiratory Tract Infection. Clin Infect Dis 2018; 66: 1045–53.

7. Hanada S, Pirzadeh M, Carver KY, Deng JC. Respiratory Viral Infection-Induced Microbiome Alterations and Secondary Bacterial Pneumonia. Front Immunol 2018; 9. DOI:10.3389/fimmu.2018.02640.

8. Man WH, De Steenhuijisen Piters WAA, Bogaert D. The microbiota of the respiratory tract: Gatekeeper to respiratory health. Nat Rev Microbiol 2017; 15: 259–70.
9. de Steenhuijsen Piters WAA, Jochems SP, Mitsi E, et al. Interaction between the nasal microbiota and S. pneumoniae in the context of live-attenuated influenza vaccine. Nat Commun 2019; 10: 1–9.

10. De Steenhuijsen Piters WAA a, Huijskens EGWW, Wyllie AL, et al. Dysbiosis of upper respiratory tract microbiota in elderly pneumonia patients. ISME J 2015; 10: 1–12.

11. Edouard S, Million M, Bachar D, et al. The nasopharyngeal microbiota in patients with viral respiratory tract infections is enriched in bacterial pathogens. Eur J Clin Microbiol Infect Dis 2018; 37: 1725–33.

12. Ding T, Song T, Zhou B, et al. Microbial composition of the human nasopharynx varies according to influenza virus type and vaccination status. MBio 2019; 10. DOI:10.1128/mBio.01296-19.

13. Teo SM, Mok D, Pham K, et al. The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. Cell Host Microbe 2015; 17: 704–15.

14. Teo SM, Tang HHF, Mok D, et al. Airway Microbiota Dynamics Uncover a Critical Window for Interplay of Pathogenic Bacteria and Allergy in Childhood Respiratory Disease. Cell Host Microbe 2018; 24: 341-352.e5.

15. Biesbroek G, Tsivtsivadze E, Sanders EAM, et al. Early Respiratory Microbiota Composition Determines Bacterial Succession Patterns and Respiratory Health in Children. Am J Respir Crit Care Med 2014; 190: 1283–92.

16. Pendleton KM, Erb-Downward JR, Bao Y, et al. Rapid pathogen identification in bacterial pneumonia using real-time metagenomics. Am J Respir Crit Care Med 2017; 196: 1610–2.

17. Brands X, Haak BW, Klarenbeek AM, Otto NA, Faber DR, Lutter R, Scicluna B, Wiersinga WJ, van der Poll T. Concurrent Immune Suppression and Hyperinflammation in Patients With
Community-Acquired Pneumonia. Front Immunol 2020; 11: 796.

18. Haak BW, Lankelma JM, Hugenholtz F, Belzer C, De Vos WM, Wiersinga WJ. Long-term impact of oral vancomycin, ciprofloxacin and metronidazole on the gut microbiota in healthy humans. J Antimicrob Chemother 2019; 74: 782–6.

19. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 2016; 13: 581–3.

20. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 2018; 6: 226.

21. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS One 2013; 8: e61217.

22. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014; 15(12):550.

23. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning 2006; 63.1: 3-42.

24. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018; 16(7): 410-22.

25. Zhang Y, Pechal JL, Schmidt CJ, Jordan HR, Wang WW, Benbow ME. Machine learning performance in a microbial molecular autopsy context: A cross-sectional postmortem human population study. PLoS One. 2019; 14(4): e0213829.

26. Fine MJ, Auble TE, Yealy DM, et al. A Prediction Rule to Identify Low-Risk Patients with Community-Acquired Pneumonia. N Engl J Med 1997; 336: 243–50.

27. Einarsson GG, Comer DM, Mcllreavey L, et al. Community dynamics and the lower airway
microbiota in stable chronic obstructive pulmonary disease, smokers and healthy non-smokers. Thorax 2016; 71(9): 795-803.

28. Malla MA, Dubey A, Kumar A, Yadav S, Hashem A, Abd_Allah EF. Exploring the Human Microbiome: The Potential Future Role of Next-Generation Sequencing in Disease Diagnosis and Treatment. Front Immunol 2019; 9. DOI:10.3389/fimmu.2018.02868.

29. Dickson RP, Erb-Downward JR, Huffnagle GB. Towards an ecology of the lung: New conceptual models of pulmonary microbiology and pneumonia pathogenesis. Lancet Respir Med 2014; 2: 238–46.

30. Krone CL, Biesbroek G, Trzcinski K, Sanders EAM, Bogaert D. Respiratory Microbiota Dynamics following Streptococcus pneumoniae Acquisition in Young and Elderly Mice. Infect Immun 2014; 82: 1725–31.

31. Ritchie ND, Ijaz UZ, Evans TJ. IL-17 signalling restructures the nasal microbiome and drives dynamic changes following Streptococcus pneumoniae colonization. BMC Genomics 2017; 18: 807.

32. Planet PJ, Parker D, Cohen TS, et al. Lambda Interferon Restructures the Nasal Microbiome and Increases Susceptibility to Staphylococcus aureus Superinfection. MBio 2016; 7. DOI:10.1128/mBio.01939-15.

33. Rosas-Salazar C, Shilts MH, Tovchigrechko A, et al. Differences in the nasopharyngeal microbiome during acute respiratory tract infection with human rhinovirus and respiratory syncytial virus in infancy. J Infect Dis 2016; 214: 1924–8.

34. Hofstra JJ, Matamoros S, van de Pol MA, et al. Changes in microbiota during experimental human Rhinovirus infection. BMC Infect Dis 2015; 15. DOI:10.1186/s12879-015-1081-y.

35. Mostafa HH, Fissel JA, Bergman Y, et al. Metagenomic Next-Generation Sequencing of Nasopharyngeal Specimens Collected from Confirmed and Suspect COVID-19 Patients. mBio
36. Langelier C, Kalantar KL, Moazed F, et al. Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. Proc Natl Acad Sci USA 2018; 115: E12353–62.

37. Yang B, Wang Y, Qian P-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics 2016; 17: 135.
### Table 1. Baseline Characteristics of Cases and Controls

| Characteristic                        | Cases (n=117)   | Controls (n=48) | P value |
|---------------------------------------|----------------|----------------|---------|
| Age, y                                | 69.0 (60.0 – 78.0) | 70.5 (63.8 – 75.0) | .911    |
| Sex, male                             | 64 (54.7)       | 28 (58.3)      | .799    |
| Ethnicity, Caucasian                  | 85 (73.3)       | 40 (85.1)      | .520    |
| Body Mass Index                       | 25.29 (6.23)    | 27.83 (4.92)   | .014    |
| Influenza vaccination                 | 69 (60.0)       | 22 (45.8)      | .067    |
| Pneumococcal vaccination              | 1 (0.9)         | 1 (2.1)        | .181    |
| Past smoker                           | 64 (55.2)       | 23 (47.9)      | .253    |
| Flexitarian diet                      | 108 (93.9)      | 47 (97.9)      | .779    |
| Recent exposure to antibiotics<sup>5</sup> | 11 (9.4)       | 2 (4.2)        | .182    |
| Chronic comorbidity                   |                |                |         |
| COPD                                  | 36 (30.8)       | 4 (8.3)        | .004    |
| Cardiovascular disease                | 89 (76.1)       | 31 (64.6)      | .189    |
| Diabetes                              | 32 (27.4)       | 6 (12.5)       | .064    |
| Malignancy                            | 40 (34.2)       | 9 (18.8)       | .074    |
| Immunosuppressive disease*            | 30 (25.6)       | 4 (8.3)        | .022    |
| Gastrointestinal disease              | 18 (15.4)       | 2 (4.2)        | .081    |
| Chronic renal disease                 | 14 (12.0)       | 3 (6.2)        | .415    |
| Imaging and Microbiology              |                |                |         |
| Radiologically confirmed CAP          | 117 (100.0)     |               |         |
| Blood culture obtained                | 109 (93.2)      |               |         |
| Sputum culture obtained               | 68 (58.1)       |               |         |
| Viral nasal/throat swab PCR performed | 117 (100.0)     | 47 (97.9)      | >.999   |
| PUAT/LUAT performed                   | 73 (62.4)       |               |         |
| Severity of disease and outcome       |                |                |         |
| PSI class                             | 4.0 (3.0 - 4.0) |               |         |
| ICU admission                         | 9 (7.8)         |               |         |
| Length of hospital stay, days         | 4.0 (3.0, 7.8)  |               |         |
| 28-day mortality                      | 5 (4.4)         |               |         |

Data are no. (%) or median (IQR). Abbreviations: CAP, Community-acquired pneumonia; COPD, chronic obstructive pulmonary disease; ICU, intensive care unit; IQR, interquartile range; LUAT, Legionella urinary antigen test; PCR, polymerase chain reaction; PUAT, pneumococcal urinary antigen test; SD, standard deviation.<sup>5</sup> Recent antibiotics usage was defined as antibiotic administration from 90 days up to 48 hours prior to inclusion. *Immunosuppressive disease was defined as clinically suspected or proven immunodeficiency, the use of immunosuppressive therapy or immunomodulating medication in the past 3 months, including chemotherapy, or the use of more than 10 mg prednisone or equivalent each day for the past 3 months.
FIGURE LEGENDS

Figure 1. Overview of causative pathogens of CAP patients

Cumulative overview of causative pathogens (A) and the proportion of bacterial, viral and mixed cases within the cohort (B). *Other pathogens constitute of Rothia dentocariosa, Stenotrophomonas maltophilia, Moraxella osloensis and Streptococcus salivarius. Abbreviations: hMPV: human metapneumovirus; RSV: respiratory syncytial virus

Figure 2. Alpha and beta diversity of cases (n=117) and control subjects (n=48)

The Inverse Simpson Index (A) and the Observed Taxa (B) index were used to calculate the alpha diversity community and richness within each individual microbiota sample. Data are presented as box plot overlaid by a dot plot with a line at the median. P values were calculated using the Wilcoxon rank sum test. Beta diversity is depicted by unweighted (C) and weighted (D) UniFrac index in a PCoA representation. P values were calculated using Permutational Multivariate Analysis of Variance (PERMANOVA).

Figure 3. High nasopharyngeal abundance of Streptococcus species (top) and Haemophilus species (bottom) in patients with microbiological diagnosis of these pathogens as obtained via culture or urine antigen test.

Relative abundances (proportion of total 16s rRNA reads) within each individual microbiota sample are presented as box plot overlaid by a dot plot with a line at the median. P values were calculated using the Kruskal Wallis test.
Figure 4. ROC curves for extremely randomized trees classifying models aimed to discriminate cases from controls using of nasopharyngeal bacterial abundance, viral presence, and host characteristics

Depiction of Mean Area under Response Curve (AUC) in the entire CAP cohort (blue line), patients with bacterial CAP (gold line) and viral CAP (green line) cohort (A). Depiction of the 15 discriminatory variables with the highest feature importance in all CAP patients (B), patients with bacterial CAP only (C) and patients with viral CAP only (D) The relative feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability is depicted as a percentage, which can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value, the more important the feature.
Figure 2

A. Microbiota diversity

B. Microbiota richness

C. Unifrac PCoA (unweighted)

D. Unifrac PCoA (weighted)
Figure 3

Nasopharyngeal abundance of *Streptococcus* species

|          | Admission | One month | p-value |
|----------|-----------|-----------|---------|
|          | 0.001     | 0.002     | 0.002   |

Nasopharyngeal abundance of *Haemophilus* species

|          | Admission | One month | p-value |
|----------|-----------|-----------|---------|
|          | 0.009     | 0.009     | 0.009   |

Causative pathogen

- *Streptococcus pneumoniae*
- *Haemophilus influenzae*
- Other bacterial pathogen
- Respiratory virus
- No causative pathogen found
- Healthy controls
