PLOS ONE

RESEARCH ARTICLE

The diagnostic value of nasal microbiota and clinical parameters in a multi-parametric prediction model to differentiate bacterial versus viral infections in lower respiratory tract infections

Yunlei Li[1], Chantal B. van Houten[2], Stefan A. Boers[3], Ruud Jansen[4], Asi Cohen[5], Dan Engelhard[6], Robert Kraaij[7], Saskia D. Hiltemann[1], Jie Ju[1], David Fernández[8], Cristian Mankoc[8], Eva González[8], Wouter J. de Waal[9], Karin M. de Winter-de Groot[10], Tom F. W. Wolfs[2], Pieter Meijers[11], Bart Luijk[12], Jan Jelrik Oosterheert[13], Sanjay U. C. Sankatsing[14], Aik W. J. Bossink[15], Michal Stein[16], Adi Klein[16], Jalal Ashkar[16], Ellen Bamberger[5,17], Isaac Srugo[17], Majed Odeh[18], Yaniv Dotan[19], Olga Boico[5], Liat Etshtein[5], Meital Paz[5], Roy Navon[5], Tom Friedman[5], Einav Simon[5], Tanya M. Gottlieb[5], Ester Pri-Or[5], Gali Kronenfeld[5], Kfir Oved[5], Eran Eden[5], Andrew P. Stubbs[1], Louis J. Bont[2], John P. Hays[3]*

1 Department of Pathology & Clinical Bioinformatics, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, The Netherlands, **2** Division of Paediatric Immunology and Infectious Diseases, University Medical Centre Utrecht, Utrecht University, Utrecht, The Netherlands, **3** Department of Medical Microbiology and Infectious Diseases, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, The Netherlands, **4** Streeklab Haarlem, Haarlem, The Netherlands, **5** MeMed, Tirat Carmel, Israel, **6** Division of Paediatric Infectious Disease Unit, Hadassah-Hebrew University Medical Centre, Jerusalem, Israel, **7** Department of Internal Medicine, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, The Netherlands, **8** Noray Bioinformatics, Derio, Spain, **9** Department of Paediatrics, Diakonessenhuis, Utrecht, The Netherlands, **10** Department of Paediatric Respiratory Medicine, University Medical Centre Utrecht, Utrecht University, Utrecht, The Netherlands, **11** Department of Paediatrics, Gelderse Vallei Hospital, Ede, The Netherlands, **12** Department of Respiratory Medicine, University Medical Centre Utrecht, Utrecht University, Utrecht, The Netherlands, **13** Department of Internal Medicine and Infectious Diseases, University Medical Centre Utrecht, Utrecht University, Utrecht, The Netherlands, **14** Department of Internal Medicine, Diakonessenhuis Utrecht, Utrecht, The Netherlands, **15** Department of Respiratory Medicine, Diakonessenhuis Utrecht, Utrecht, The Netherlands, **16** Department of Paediatrics, Hillel Yaffe Medical Centre, Hadera, Israel, **17** Department of Paediatrics, Bnai Zion Medical Centre, Haifa, Israel, **18** Department of Internal Medicine A, Bnai Zion Medical Centre, Haifa, Israel, **19** Pulmonary Division, Rambam Health Care Campus, Haifa, Israel

* j.hays@erasmusmc.nl

OPEN ACCESS

Citation: Li Y, van Houten CB, Boers SA, Jansen R, Cohen A, Engelhard D, et al. (2022) The diagnostic value of nasal microbiota and clinical parameters in a multi-parametric prediction model to differentiate bacterial versus viral infections in lower respiratory tract infections. PLoS ONE 17(4): e0267140. https://doi.org/10.1371/journal.pone.0267140

Editor: Mao-Shui Wang, Shandong Public Health Clinical Center: Shandong Provincial Chest Hospital, CHINA

Received: December 4, 2021

Accepted: April 4, 2022

Published: April 18, 2022

Data Availability Statement: The 16S rRNA gene sequencing dataset for the 293 patients generated during the current study is available in the European Nucleotide Archive (ENA) repository at https://www.ebi.ac.uk/ena under the study accession number ERP126851. The metadata is available in S7 Table. All data analysed during this study and used in the prediction models are included in this published article and its S3 Table, including all clinical and microbiota data. The R

Abstract

Background

The ability to accurately distinguish bacterial from viral infection would help clinicians better target antimicrobial therapy during suspected lower respiratory tract infections (LRTI). Although technological developments make it feasible to rapidly generate patient-specific microbiota profiles, evidence is required to show the clinical value of using microbiota data for infection diagnosis. In this study, we investigated whether adding nasal cavity microbiota profiles to readily available clinical information could improve machine learning classifiers to distinguish bacterial from viral infection in patients with LRTI.

script in knitr format for the prediction modelling is available at https://github.com/ErasmusMC-Bioinformatics/TTT.

Funding: J.P. Hays, D. Engelhard, E. Eden, E. González and L.J. Bont received funding for The TAILORED Treatment study from the European Union's Seventh Framework Programme (grant number 602860, https://cordis.europa.eu/project/id/602860). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: ANC, Absolute neutrophil count; ARI, Acute respiratory infection; AUC, Receiver operating characteristic curve; CRP, C-Reactive protein; CTSB, Cathepsin B; eCRF, Electronic case report form; FDR, False discovery rate; HK3, Hexokinase 3; HRV, Human rhinovirus; IFI27, Interferon alpha inducible protein 27; IP-10, Interferon gamma induced protein-10; Lcn2, Lipocalin-2; LRTI, Lower respiratory tract infections; MxA1, Myxoma resistance protein; OTU, Operational taxonomic units; PCT, Procalicitonin; POC, Point-of-Care; SEM, Standard error of mean; TRAIL, TNF-related apoptosis-inducing ligand.

Results

Various multi-parametric Random Forests classifiers were evaluated on the clinical and microbiota data of 293 LRTI patients for their prediction accuracies to differentiate bacterial from viral infection. The most predictive variable was C-reactive protein (CRP). We observed a marginal prediction improvement when 7 most prevalent nasal microbiota genera were added to the CRP model. In contrast, adding three clinical variables, absolute neutrophil count, consolidation on X-ray, and age group to the CRP model significantly improved the prediction. The best model correctly predicted 85% of the 'bacterial' patients and 82% of the 'viral' patients using 13 clinical and 3 nasal cavity microbiota genera (*Staphylococcus*, *Moraxella*, and *Streptococcus*).

Conclusions

We developed high-accuracy multi-parametric machine learning classifiers to differentiate bacterial from viral infections in LRTI patients of various ages. We demonstrated the predictive value of four easy-to-collect clinical variables which facilitate personalized and accurate clinical decision-making. We observed that nasal cavity microbiota correlate with the clinical variables and thus may not add significant value to diagnostic algorithms that aim to differentiate bacterial from viral infections.

Introduction

Lower respiratory tract infections (LRTI) are a leading global cause of mortality in all age groups [1, 2]. For example, in the Unites States of America, acute LRTI are associated with a greater morbidity and mortality than any other infection (https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm). Further, in Europe, pneumonia is associated with approximately 230,000 deaths per year [3]. However, the accurate differentiation of a bacterial infection, viral infection or no infection in cases of lower respiratory tract complaints (useful in deciding if a clinician should prescribe antibiotics or not to patients), is still difficult. Expert guidelines to direct this decision-making process have been available for several years [4, 5] with many algorithm-based guidelines incorporating the popular biomarker C-reactive protein (CRP) and/or procalcitonin (PCT) for the presumed diagnosis of an infection [6–9]. Further, biomarkers continue to be described as being potentially useful in differentiating between bacterial versus viral respiratory infections/sepsis. These include for example: TNF-related apoptosis-inducing ligand (TRAIL) [10], Interferon gamma induced protein-10 (IP-10) [11], Myxoma resistance protein (MxA1) [12, 13] and Lipocalin-2 (Lcn2) [14]. Diagnostic algorithms incorporating such biomarkers may help guide targeted antibiotic prescribing for bacterial LRTI, helping reduce (long term) costs and possible unnecessary side-effects in cases of viral LRTI, while helping reduce the increasing global epidemic of antibiotic resistance [15, 16].

Interestingly, research has also implicated the respiratory tract microbiota in the etiology of LRTI, with a particular emphasis on the nasopharyngeal microbiota [17, 18], although the actual site of sampling may be instrumental in the success of such approaches [19]. However, with respect to practical considerations (including the performance of large clinical trials), taking a nasal swab tends to be more convenient for nurses and doctors, and more comfortable/acceptable for patients than taking a nasopharyngeal swab, nose/nasopharyngeal washing, tracheal aspirate or bronchoalveolar lavage [20].

The development of rapid sequencing techniques has facilitated the prospect of (rapid) microbiota profiling as a diagnostic parameter in infectious diseases to enhance diagnostic algorithms based on a patient's personalized clinical data. Therefore, in this publication, we investigated the added value of nasal cavity microbiota profiling in increasing the predictive power of diagnostic algorithms for classifying bacterial versus viral infection in children and adults with LRTI. The basis for this research was the EU-funded FP7 TAILORED-Treatment study (Grant ID 602860), which was charged with establishing a multi-omics approach to aid effective antibiotic prescribing in lower respiratory tract infections. After rigorous testing, we observed that nasal cavity microbiota correlate with the clinical variables and thus do not add significant value in the classifiers to differentiate bacterial from viral infections. Our prediction model with four easy-to-collect clinical variables greatly improves current diagnosis practice and will further facilitate personalized and accurate clinical decision-making for patients with LRTI.

Materials and methods

Ethics statement

The TAILORED-Treatment study is registered on ClinicalTrials.gov in January 2014 under the registration number NCT02025699 and was approved by the ethics committees in the participating countries. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Patients or their legal representatives (parents or guardians) received explanations about the study protocol and signed the study informed consent forms. Informed consent was obtained from all individual participants included in the study.

Specifically, ethical approval was obtained from the Medical Ethical Committee of UMCU (14–104) and the Institutional Review Boards of Hillel Yaffe Medical Centre (HYMC-0108-13 and HYMC-0107-13), Bnai Zion Medical Centre (BNZ-0107-14 and BNZ-0011-14) and Hadassah University Medical Centre (HMO-0007-14 and HMO-0006-14).

Patient cohort and study design

Patients were recruited for the EU FP7-funded 'TAILORED-Treatment' study, which ran from April 2014 to September 2016 [21]. During this period, a total of 1,261 pediatric and adult patients aged 1 month and older with suspected LRTI and/or sepsis were recruited at 7 participating emergency departments and hospital wards of Dutch and Israeli medical centers. From these patients a total of 516 nasal cavity microbiota profiles met the quality criteria of the study, using 16S rRNA gene sequencing quality criteria of >950 16S rRNA molecules DNA/μl and a sequencing depth of >1000 reads per sample. One hundred and sixty four of these 516 profiles were subsequently removed from the study as the relevant patients did not match our expert panel criteria for the clinical definition of bacterial or viral infection, i.e. the patients were classified as 'non-infectious' or 'undetermined'. Suspected mixed viral and bacterial infections were labelled as bacterial infections, as they often elicit similar patient management —in LRTI most therapy of patients is prescribed based on infection with bacterial not viral pathogens. An additional 59 patients were also excluded from further analysis due to missing CRP data (CRP is already used in clinical practice as an important predictive variable of inflammation/infection).

Ultimately, this meant that machine learning classification analysis was performed on a total cohort of 293 patients (**Fig 1** and **S1 Table** for patient characteristics). By following the

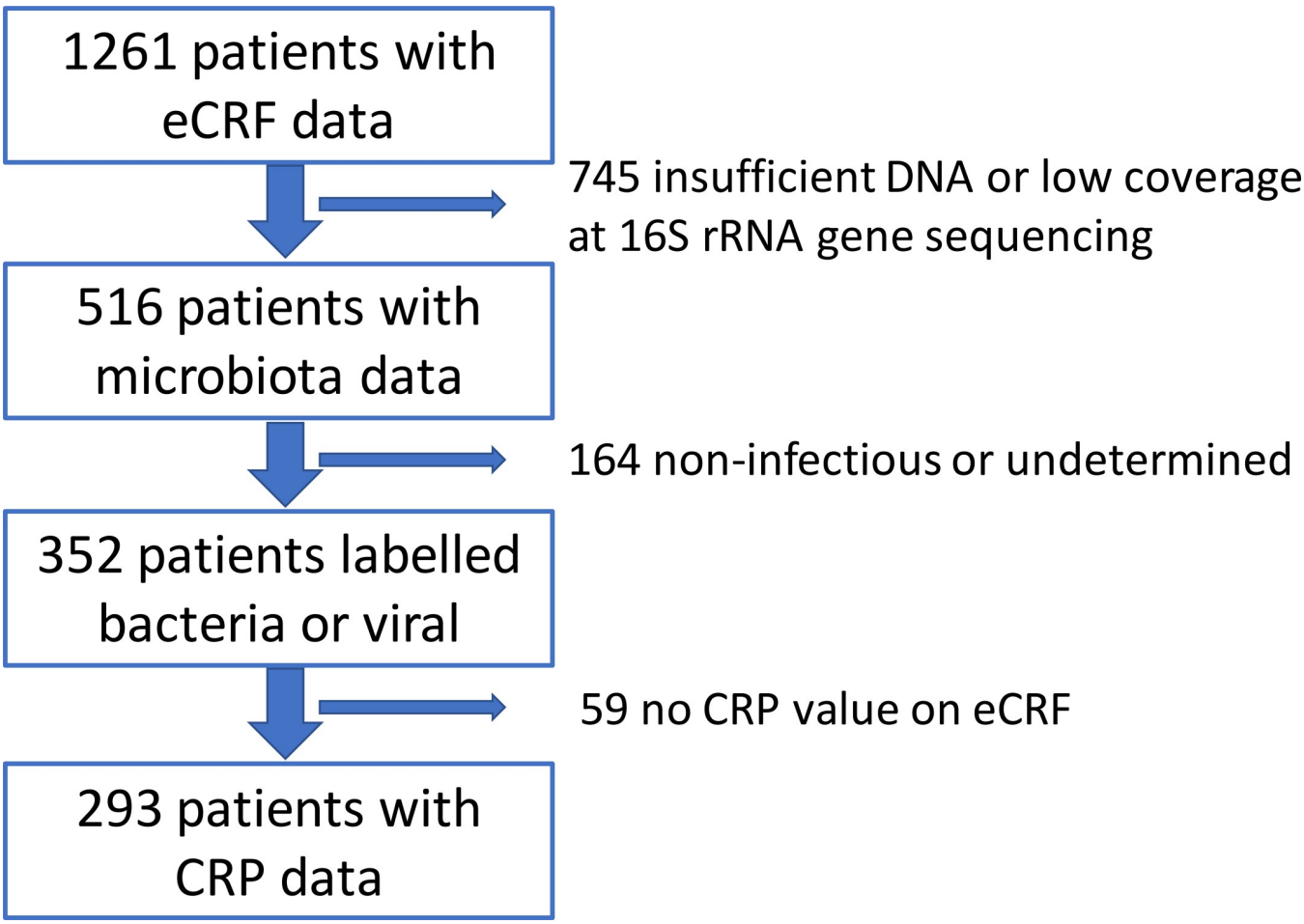

Fig 1. Patient filtering. LRTI patients from TAILORED-Treatment study underwent several filtering steps before they entered the classifier development stage. The eCRF data utilized in this publication was obtained from the following patient cohort [21]. eCRF: electronic Case Report Forms. CRP: C-reactive protein.

https://doi.org/10.1371/journal.pone.0267140.g001

TAILORED-Treatment study protocol [21], a subset of 242 patients that were recruited during the discovery phase were included in the initial cohort. They were used to reduce uninformative variables and investigate whether adding microbiota data on top of clinical variables on the electronic Case Report Forms (eCRFs) helped to achieve better prediction of 'bacterial infection' versus 'viral infection' at patient presentation. Subsequently, the other 51 patients that were recruited later were added to the initial cohort in order to evaluate the two types of variables simultaneously (i.e. clinical and microbiota) in 5-fold cross-validation setting (**Fig 2**) to evaluate the generalization of the classification performance and to build the final classifier.

Expert panel

A recently published reference standard using an expert panel protocol was used to assign TAILORED-Treatment study patients to either 'bacterial infection' or 'viral infection' class labels [16, 21]. The TAILORED-Treatment expert panel comprised 3 experienced physicians who were provided with all available clinical and laboratory information as listed in the electronic case record forms, including 28-day follow-up evaluation data. Each expert was blinded to the research results and to the labels of his/her peers on the expert panel. A final diagnosis

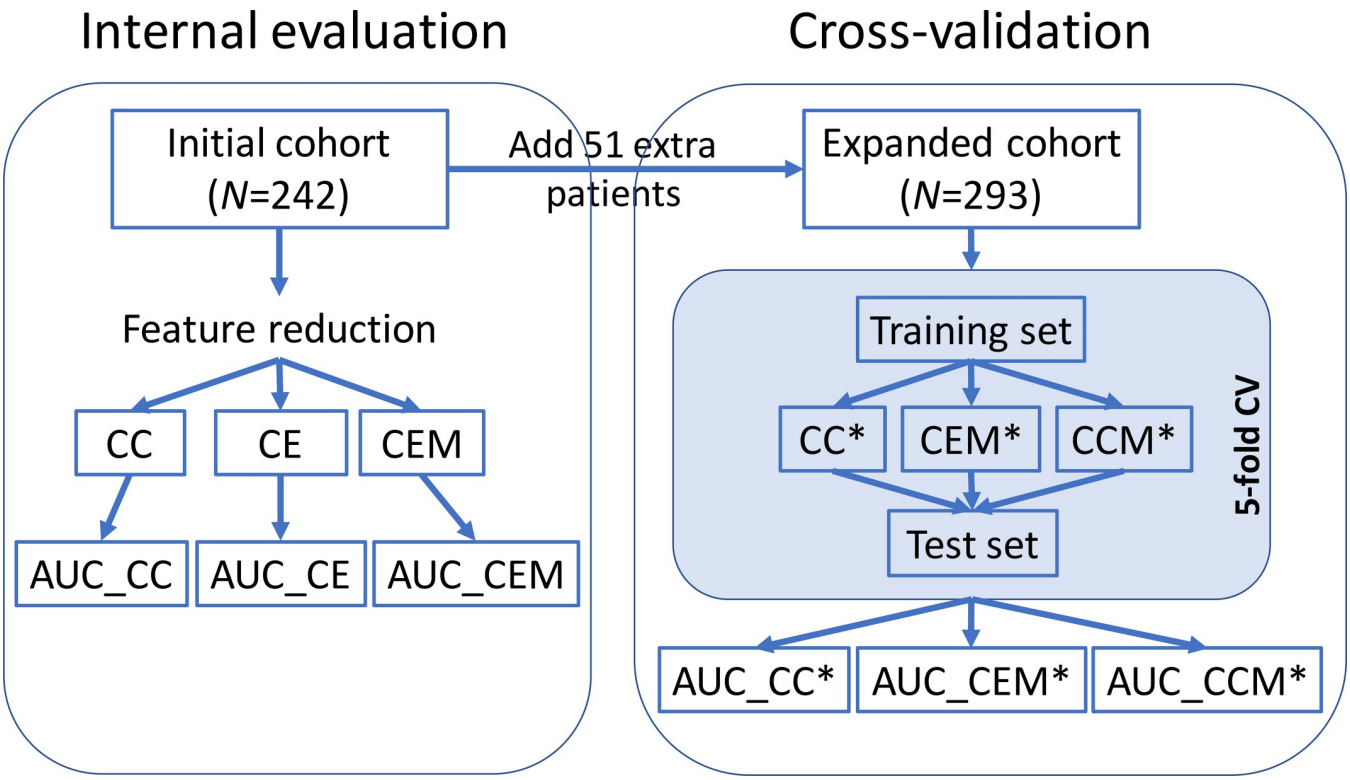

Fig 2. Study overview. A cohort of 242 patients were included in the internal evaluation phase according to the date of recruitment to the TAILORED-Treatment study. This cohort was used to compare the prediction performances of the classifiers using eCRF variables alone, as well as classifiers using both eCRF and microbiota variables (Internal evaluation phase). In the expanded cohort (51 extra patients), 5-fold cross-validation (CV) analysis was conducted to evaluate the contribution of eCRF and microbiota variables to prediction performance (Cross-validation phase). **CC**: **C**lassifier using **C**RP only in the initial cohort. **CE**: **C**lassifiers using two or more **e**CRF variables (incl. CRP) in the initial cohort. **CEM**: **C**lassifiers using all input **e**CRF variables (incl. CRP) and at least one **m**icrobiota in the initial cohort. **CC***: **C**lassifier using **C**RP only in the 5-fold CV of the expanded cohort. **CEM***: **C**lassifiers using two or more variables (regardless **e**CRF or **m**icrobiota) in the 5-fold CV of the expanded cohort. **CCM***: **C**lassifiers using **C**RP and all input **m**icrobiota variables in the 5-fold CV of the expanded cohort. AUC: Area Under the ROC Curve.

https://doi.org/10.1371/journal.pone.0267140.g002

was determined based on the consensus agreement among all three experts. Suspected mixed viral and bacterial infections were labelled as bacterial infections, as they often elicit similar patient management—in LRTI most therapy of patients is prescribed based on infection with bacterial not viral pathogens. Therefore, including 'mixed' infections in 'bacterial infection' will include all patients that may eventually require antibiotic therapy. Further, mixed viral and bacterial infections may actually lead to more serious disease and consequently bear more chance of antibiotic prescribing [22–24].

Clinical eCRF data

Clinical eCRF data were generated using standard laboratory methods according to the published TAILORED-Treatment clinical trial protocol [21]. All clinical information as listed in the eCRFs was used by the expert panel to determine the etiology of infection. However, not all eCRF variables were actually included in the current prediction modelling as some of these variables were not available at patient presentation, and thereby not applicable for clinical diagnosis at patient presentation. In this respect, 195 eCRF variables were selected to be used in the analysis (**S2 Table**). After conversion of the categorical variables into Boolean data using one hot encoding, a total of 1,624 eCRF variables remained. This dataset was then cleaned to correct for input errors, to unify metrics and to merge variables within the same category.

Variables only available for ≤2 patients were also removed, resulting in 58 numeric variables and 235 categorical variables available for further analysis.

In the internal evaluation phase, we performed feature reduction based on 242 patients in the initial TAILORED-Treatment cohort by testing the correlation between these 293 variables and the class labels (i.e. 'bacterial infection' or 'viral infection'). More specifically, we tested the categorical variables by Fisher's Exact test using R [25] function *fisher.test*. The resulting *p*-values were subject to Benjamini–Hochberg procedure to control for the False Discovery Rate (FDR). Variables having an adjusted *p*-value lower than 0.05 were retained for further classification modelling. For the numeric variables, we applied both parametric *t*-test and nonparametric Mann-Whitney *U* test. Twelve numeric variables generated significant results with $p<0.05$ in both tests. Among these, 4 redundant variables with a Pearson's $r>0.7$ and $p<0.05$ were removed.

After feature reduction, 21 out of 235 categorical eCRF variables and 8 out of 58 numeric eCRF variables were identified that differed significantly between the 'bacterial infection' and 'viral infection' patients in the initial cohort. They served as the clinical input variables for the classifiers. All clinical data used for the prediction model can be found in **S3 Table**.

16S rRNA gene sequencing and pre-processing

Nasal cavity swab samples were collected from 1261 patients using e-swabs (Copan, USA) containing 1mL of liquid Amies medium and stored at -80˚C for subsequent 16S rRNA gene sequencing analysis. DNA was extracted from all nasal cavity swab samples, using a phenol/bead-beating protocol combined with the AGOWA mag Mini DNA Isolation Kit (LGC) as described previously [26]. In addition, DNA from elution buffer BL (LGC) was extracted as negative extraction control samples at the same time to assess the composition of contaminating bacterial DNA in the experimental methodologies. 16S rRNA gene amplicon library preparation was performed as previously published [27, 28]. The hypervariable V5 and V6 regions (276 bp) of the 16S rRNA gene were amplified using the 785F (5′-GGA TTA GAT ACC CBR GTA GTC-3′) and 1061R (5′-TCA CGR CAC 20 GAG CTG ACG AC-3′) primers [27] and dual indexing [28]. Amplicons were generated in 30 cycli using the FastStart High Fidelity System (Roche), normalized using the SequalPrep Normalization Plate kit (Thermo Fischer Scientific) and pooled in batches of approximately 250 samples. Pools were purified prior to sequencing using the Agencourt AMPure XP (Beckman Coulter Life Science, Indianapolis, IN) and the amplicon size and quantity of the pools were assessed on the LabChip GX (PerkinElmer Inc., Groningen, The Netherlands). The PhiX Control v3 library (Illumina Inc., San Diego, CA) was combined (~10%) with the pooled amplicon libraries and each pool was sequenced on an Illumina MiSeq sequencer 2 (MiSeq Reagent Kit v3, 2 x 300 bp).

Bidirectional sequencing of the 16S rRNA gene amplicon libraries was performed using the Illumina MiSeq platform, with FASTQ-formatted sequences being extracted from the machine and further processed using a Galaxy mothur toolset [29] based on the standard mothur bioinformatics pipeline [30]. For those samples with more than 5000 reads, we randomly subsampled 5000 reads per sample which went further into the pipeline. Forward and reverse FASTQ-formatted sequence files were merged using the *make.contigs* command. Unique sequences were aligned against the SILVA [31] reference alignment release 123, where the reference sequences were trimmed to only include the V5-V6 region of the 16S rRNA gene using the *pcr.seqs* command. To filter out sequencing errors, we applied pre-clustering using the *pre.cluster* command allowing for up to two differences between sequences. Potential chimeric sequences were removed using UCHIME [32]. The remaining sequences were assigned to taxonomy using the *classify.seqs* command based on the customized SILVA alignment release

123. Sequences were then clustered into operational taxonomic units (OTUs) at 97% similarity using the *dist.seq* and the *cluster* commands. Finally, each OTU was assigned to a consensus taxonomy using the *classify.otu* command. For quality reasons, only samples generating >950 16S rRNA gene molecules/μl and >1000 reads were included in further analysis. Additional filtering to reduce noise involved removing OTUs with 2 or fewer reads, since they were very likely to be sequence artifacts e.g. chimeras. We calculated the relative abundance for each OTU, by dividing the OTU count number by the total count per sample. OTUs were then summarized into 240 genera for further analysis. Next, to avoid overfitting the model of genera that may not be representative in the general LRTI patient population, the relative genus-level abundance data was transformed and filtered according to recommendations in Rhea [33] in two steps: 1) All relative abundances in any sample below 0.5% were considered as absent (i.e. missing value) and 2) for prevalence-based filtering—if a genus was present in <30% of samples in both 'bacterial infection' and 'viral infection' classes, then the genus was considered 'too sparse' and was not included for further analysis.

Classification modelling

Several classifiers were built in the initial cohort as well as in the expanded cohort and their performances were evaluated. Since CRP is widely used in current clinical practice to detect inflammation/infection, the classification performance using only CRP (i.e. CC and CC* in **Fig 2**) serves as the benchmark. Besides CRP, we would like to investigate the prediction value of readily available clinical variables (i.e. CE in **Fig 2**) and the added value of microbiota (i.e. CEM, CEM* and CCM* in **Fig 2**). The initial cohort was dedicated to reduce uninformative variables and investigate whether adding microbiota data on top of clinical variables (i.e. CEM vs. CE) helped to achieve better prediction of 'bacterial infection' versus 'viral infection' at patient presentation. The expanded cohort not only contained more patient samples, but also was used to evaluate clinical and microbiota variables simultaneously for a fair comparison. That is, instead of adding the microbiota variables after the clinical ones as in the initial cohort (i.e. CEM), here all input variables entered the modelling without imposed order to rank their importance for prediction (i.e. CEM*).

Given the nature of our dataset (a mix of categorical and numeric variables with missing data), Breiman's Random Forests classification method [34] was chosen and implemented using R package *randomForestSRC* version 2.11.0. The advantage of using Random Forest as the classification model in this case is three-fold: 1) It can handle both categorical data (some of the clinical variables) and numeric data (microbiota and some of the clinical variables); 2) It is a tree-based model and hence does not require feature scaling; 3) It can handle missing values by using inference or imputation. Random Forests are ensembles of decision trees which vote for the most popular class. When growing these ensembles, bagging (i.e. bootstrap aggregating) [35], is used in tandem with random feature selection. To grow each decision tree in the training phase, a new sub-training set is drawn with replacement from the original training dataset while about one-third of the samples are left out (i.e. out-of-bag). Then a tree is grown on the new sub-training set using random feature selection. The out-of-bag samples are used to get a running unbiased estimate of the classification error as they are excluded in the sub-training set for that particular tree. Therefore, this method can provide an unbiased error rate estimate from the out-of-bag error rate even internally without the need for cross-validation or a separate test set. In our implementation, R package *randomForestSRC* was deployed using *mtry* = NULL, *ntree* = 500, and *na.action* = "na.impute". Missing data were imputed using in-bag non-missing data only, via a modification of the missing data algorithm of Ishwaran *et al.* [36].

Results

Nasal cavity microbiota

The average sequence coverage per sample was 88,461. Pre-clustering, chimera removal and operational taxonomic units (OTU) clustering at a similarity of 97% identified a total of 2,838 OTUs. After filtering out OTUs with 2 or fewer reads, the relative abundance were calculated for the remaining OTUs which were then summarized into 240 genera for further analysis. The genus-level relative abundance data was transformed and filtered according to recommendations in Rhea [33] (see Materials and Methods). Seven genera remained after Rhea filtering and served as the input for the classification modelling of the microbiota variables. These genera were: *Staphylococcus*, *Moraxella*, *Dolosigranulum*, *Corynebacterium*, *Streptococcus*, *Haemophilus*, and *Anaerococcus*, which have already been associated with the nasal microbiota in many publications, including in infants, children and adults [37–40]. **Fig 3** shows the relative abundance of these seven genera in relation to age and infection origin of the TAILORED-Treatment cohort. Although suspected mixed viral and bacterial infections were labelled as bacterial infections, we list them separately in this figure in order to highlight the differences between "mixed" (bacterial and viral) and "bacterial" infections (see Discussion). All microbiota data used for the prediction model can be found in **S3 Table**.

Classification performance & variable importance

In total, 29 clinical variables and 7 microbiota variables entered the classifier modelling phase (**S1 Fig**). Notably, more than 50% of data was missing for 6 of these 36 variables. In both phases, the performance of the models was assessed by calculating the area under the receiver operating characteristic curve (AUC) for overall performance and the percentage of correctly predicted cases.

In the internal evaluation phase, 242 patients were included to compare the prediction performances of classifiers using eCRF variables alone and classifiers using both eCRF and microbiota variables to assess whether adding microbiota on top of clinical data helped to distinguish bacterial infection from viral infection (**Fig 2**). More specifically, we first ranked the 29 input eCRF variables and 7 nasal cavity microbiota variables separately by function *vimp* in the initial cohort [34] (**S1 Fig**), then the eCRF variables were included into the Random Forests model incrementally according to their importance ranking and subsequently the ranked nasal cavity microbiota variables. Each time 1000 forests were grown and the best forest was kept. This was iterated 10 times to evaluate the robustness of the model.

The resulting AUC and accuracy per class are shown in **Fig 4A** and **S4 Table**, which are out-of-bag error rates calculated by *randomForestSRC*. More specifically, the classifier using the best eCRF variable CRP alone ("CC"), generated an average AUC of 0.75 in the 10 times repetitions and accuracies of 68% for bacterial infection and 70% for viral infection cases. Further, the AUC and accuracies for the prediction of both classes were greatly improved, and remained stable, after adding the first few (non-microbiota) clinical variables into classifiers "CE". The average AUC of CE ranges from 0.78 to 0.91 when using 2 to 29 eCRF variables, with the maximum AUC being reached when using the top ranked 15 eCRF variables. However, adding nasal cavity microbiota variables on top of these 29 eCRF variables into classifiers "CEM" did not increase the AUC or the accuracy of the classifier. The average AUC of CEM was between 0.89 and 0.90 using all 29 eCRF variables plus 1 to 7 microbiota genera (**S4 Table**).

In the expanded cohort with 51 extra patients, 5-fold cross-validation was conducted to evaluate the contribution of the eCRF and microbiota variables to the prediction performance.

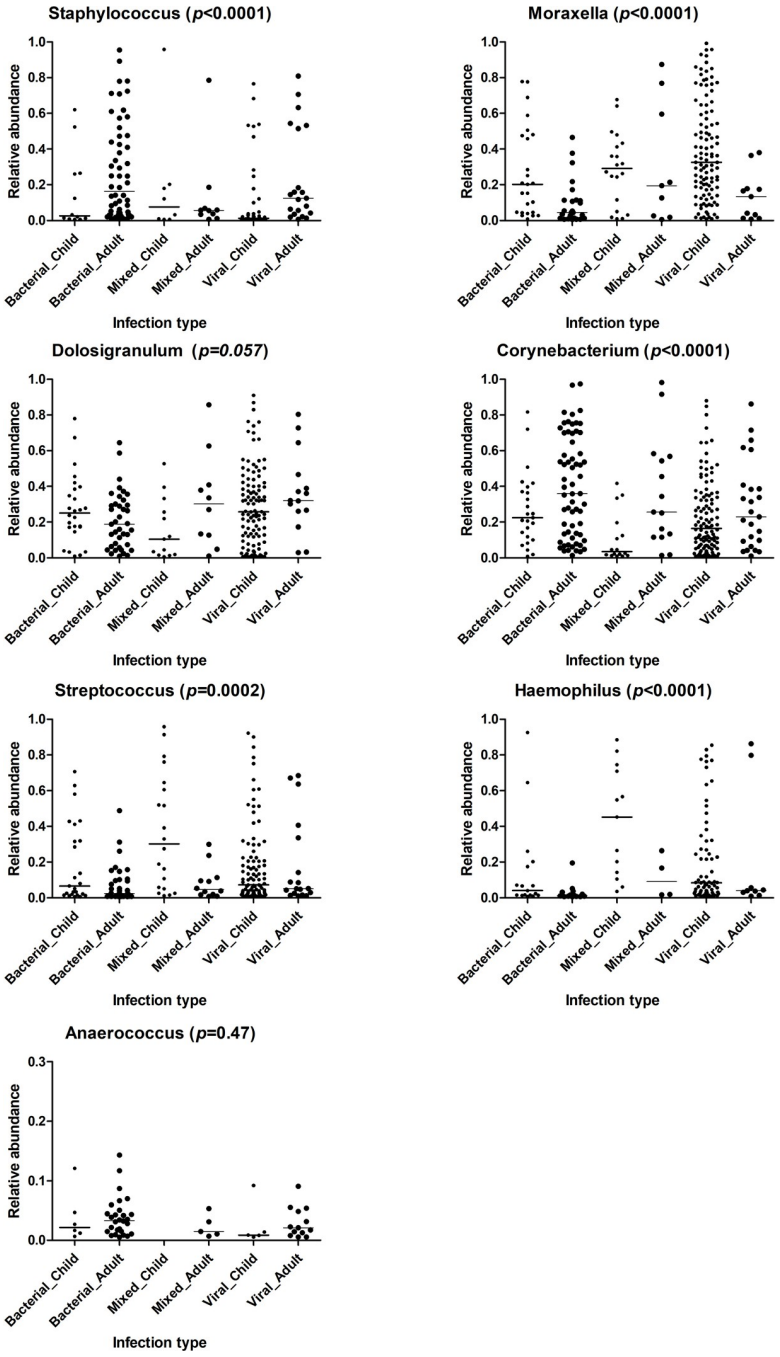

Fig 3. Relative abundance of seven most common bacterial genera related to age and infection origin of TAILORED-Treatment cohort. Kruskal-Wallis test was performed to calculate the *p*-values. Horizontal bars represent the median values.

https://doi.org/10.1371/journal.pone.0267140.g003

The total cohort of 293 patients was randomly split into 5 subsets, or so called 'folds'. Each time one fold containing 58 or 59 patients was taken as the test set and the remaining 4 folders formed the training set. In each round, the test fold was withheld in the training step and used to assess the prediction accuracy on "unseen" data in the test fold. Eventually all samples in the dataset were evaluated exactly once to obtain the overall performance of the models. Since the

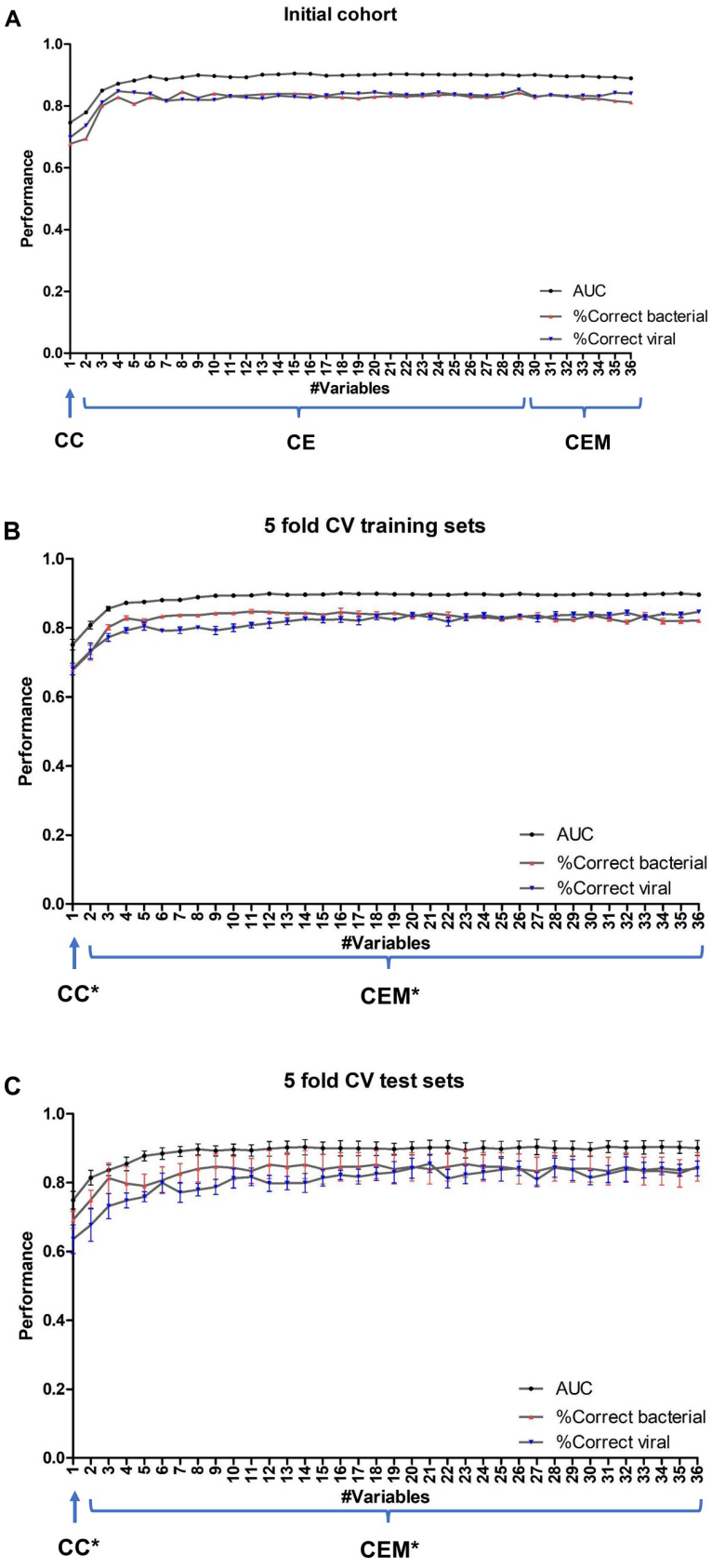
A
Initial cohort
Performance
AUC
%Correct bacterial
%Correct viral
#Variables
CC
CE
CEM
B
5 fold CV training sets
Performance
AUC
%Correct bacterial
%Correct viral
#Variables
CC*
CEM*
C
5 fold CV test sets
Performance
AUC
%Correct bacterial
%Correct viral
#Variables
CC*
CEM*

Fig 4. Performance of the classifiers. Classifier performance in **A**) the initial cohort, **B**) 5-fold cross-validation training sets in the expanded cohort, and **C**) 5-fold cross-validation test sets in the expanded cohort. X-axis shows the number of variables included in the classifier. The lines represent the mean of AUC, the accuracy of class 'bacterial infection', and the accuracy of class 'viral infection', respectively. The bars represent the standard error of the mean (SEM). In the initial cohort (Panel A), first eCRF variables were ranked separately and included in the classifier incrementally, followed by ranked microbiota variables. The ranking was based on their variable importance calculated by function *vimp* in the initial cohort. In the cross-validation (Panels B-C), the ranking of all eCRF and microbiota variables was calculated simultaneously based on the training set in the particular split and averaged across five splits. **CC**: Classifier using CRP only in the initial cohort. **CE**: Classifiers using two or more eCRF variables (incl. CRP) in the initial cohort. **CEM**: Classifiers using all input eCRF variables (incl. CRP) and at least one nasal cavity microbiota variable in the initial cohort. **CC***: Classifier using only CRP in the 5-fold cross-validation of the expanded cohort. **CEM***: Classifiers using two or more variables (regardless of eCRF or nasal cavity microbiota origin) in the 5-fold cross-validation of the expanded cohort. AUC: Area Under the ROC Curve. CV: cross-validation. SEM: standard error of the mean.

https://doi.org/10.1371/journal.pone.0267140.g004

two classes were unbalanced, the majority class label 'viral infection' was randomly subsampled to reach equal prior probability as class 'bacterial infection' in each training set. Unlike the internal evaluation phase where the eCRF and microbiota variables were ranked separately and added sequentially (i.e. first the eCRF variables and then the microbiota ones), here all variables were ranked simultaneously based on the training set at hand using function *vimp* and added into the model accordingly.

In the 5-fold CV, classification performance was assessed on both training set (using 'out-of-bag' predictions) and test set at hand. We then summarized the performance metrics by taking the mean and standard error of the mean. The resulting AUC graphs and accuracy per class of the classifier using CRP alone ("CC*") and classifiers using two or more variables ("CEM*") are shown in **Fig 4B and 4C** and **S4 Table**. In both the training sets and test sets of 5-fold cross-validation, the classifier using CRP alone generated an average AUC of 0.75 which is the same as in the internal evaluation phase. The average accuracies of bacterial and viral classes in the test sets were 69% and 64%, respectively. Both the AUC and accuracy per class were consistently improved after adding the first few variables and remained stable upon the addition of subsequent variables. More specifically, adding 3 more clinical variables, including 'absolute neutrophil count' (ANC), 'consolidation on X-ray', and 'age group' on top of CRP significantly improved the prediction towards an average AUC of 0.85 in the 5-fold cross-validation test sets. The distribution of the four most predictive variables can be found in **S2 Fig**. The highest average AUC in the 5 training sets was reached using the top 16 variables (**S5 Table**), and the average performance using these variables in the 5 test sets was an AUC of 0.90, with 85% of bacterial infection and 82% of viral infection being correctly predicted. Of note, 3 genera were present in the top 16 variables: *Staphylococcus* (rank #11), *Moraxella* (rank #13) and *Streptococcus* (rank #14).

We also conducted 5-fold cross-validation to investigate the added value of adding nasal cavity microbiota variables to the best eCRF predictor, i.e. CRP. That is, classifier CCM* was built using CRP plus all 7 microbiota genera variables as the input variables. Compared to CRP alone, adding nasal cavity microbiota variables showed a slight improvement in AUC (from 0.75 to 0.81), which held for both training sets and test sets (**S4 Table**). The percentage of correctly predicted 'viral infection' patients in the test sets was pronouncedly improved (from 63% to 80%) after adding 7 nasal cavity microbiota variables to CRP, whereas the percentage of correctly predicted cases of 'bacterial infection' patients was only slightly improved (from 69% to 74%).

The use of the Random Forests model meant that unbiased error rate estimate could be derived internally from the out-of-bag samples without the need for cross-validation or a separate test set, because the out-of-bag samples are excluded in the training set for the particular

tree to obtain a running unbiased estimate of the classification error. Indeed, we observed very similar performance in our training set (using the out-of-bag samples) and test sets (using independent samples) as shown in **Fig 4B and 4C**. This demonstrates the robustness of our prediction model. Further, the standard error of mean (SEM) in the test sets was modest, suggesting that the sample size of the cohort used in this study was sufficient.

Correlations between nasal cavity microbiota and clinical variables

Since we observed no improvement after adding nasal cavity microbiota variables into the classifiers when using all eCRF variables, we hypothesized that nasal cavity microbiota variables were probably correlated with the eCRF variables and thereby provided limited extra predictive value. Based on this hypothesis, we investigated the correlations between the 7 genera (nasal cavity microbiota) variables and 29 clinical (eCRF) variables in the initial cohort for each class separately. Spearman's *rho* was calculated for numeric variables, and Mann-Whitney *U* test or Kruskal–Wallis test was performed to calculate the correlation between numeric variables and categorical variables. The results are shown in **S6 Table**, together with the correlations among the 7 genera. As we expected, all 7 nasal cavity microbiota variables were correlated with one or more clinical variables. In particular, *Corynebacterium* and *Streptococcus* genera were significantly correlated with the highest number of clinical variables (both 11), while *Dolosigranulum* and *Anaerococcus* were significantly correlated with the least number of clinical variables (1 and 2, respectively).

Discussion

In this publication, we describe the development of a comprehensive machine learning algorithm for distinguishing 'bacterial infection' from 'viral infection' in a cohort comprising child and adult patients presenting with LRTI. The TAILORED-Treatment study methodology and cohort recruitment allowed an assessment to be made of easy-to-collect clinical variables together with nasal cavity microbiota variables for predicting 'bacterial infection' versus 'viral infection'. After rigorous testing, we observed that nasal cavity microbiota correlate with the clinical variables and thus do not add significant value besides the clinical variables in the classifiers to differentiate bacterial from viral infections. Our findings indicate that 4 clinical variables, i.e. CRP, absolute neutrophil count (ANC), consolidation on X-ray and age group, gave an average AUC of 0.85 and that the addition of nasal cavity microbiota variables leads to a marginal improvement in the accuracy of the algorithm.

Since this marginal improvement may not be sufficient to affect the antibiotic prescribing decision of clinicians in patients presenting with LRTI, our study provides evidence of limited clinical value of using nasal microbiota data for infection diagnosis. Furthermore, we propose an accurate multi-parametric prediction model which takes the interactions between multifactors into account and is applicable for both child and adult patients. Our prediction model with four easy-to-collect clinical variables greatly improves current diagnosis practice and will further facilitate personalized and accurate clinical decision-making for patients with LRTI.

The data presented in this publication was collected as part of a European-wide approach to combat ever increasing global antibiotic resistance. We expect to demonstrate our prediction model in a future cohort should it be available. Targeted antibiotic prescribing to patients suffering from bacterial, as opposed to viral, infections could potentially limit the amount of unnecessary antibiotics prescribed by clinicians (TAILORED-Treatment) [16]. Better targeted antibiotic treatment involving a reduction in the unnecessary prescription of antibiotics at local, national and international levels, will help limit the increasing prevalence of antibiotic resistance, as well as inhibit the development of new antibiotic resistances.

The nasal cavity microbiota

The use of established 16S rRNA gene sequencing techniques has shown that the respiratory tract microbiota varies with age, with Proteobacteria (*Haemophilus*, *Neisseria* and *Moraxella*), and Firmicutes (*Strepotcoccus*, *Gemella*, *Dolosogranulum* and *Granulicatella*) being overrepresented in prepubertal children and Actinobacteria (*Corynebacteium*, *Propionibacterium* and *Turicella*) being overrepresented in adults [41]. The presence of potential bacterial pathogens in the respiratory microbiota have also been associated with LRTI. This variation usually involves colonization with *Streptococcus pneumoniae*, *Haemophilus influenzae* and *Moraxella catarrhalis* (three bacteria often associated with the development of upper respiratory tract infections in infants, including otitis media in young children) and *Staphylococcus aureus*, *Dolosigranulum* spp. or *Corynebacterium* spp. [42–44]. However, published results tend to be presented for either a mixture of both nasal and nasopharynx microbiota or only the nasopharyngeal microbiota [22, 45], while differences in the microbiota of nasal, nasopharyngeal and oropharyngeal microbiota and over time have been reported [41, 46]. Importantly, there are several physical reasons why the nasal cavity microbiota may vary from the nasopharyngeal microbiota including: 1) the presence of stiff coarse hairs (vibrissae) in the anterior nares; 2) epithelium type and 3) the cooler environment present compared to the nasopharynx [47]. Also, the sampling methodology used may affect the microbiota composition found from nasal/nasopharyngeal sites, with nasal washings not being 'spatially specific' [48]. In this respect, de Boeck *et al*. compared the healthy nose and nasopharynx microbiota, with the observation that both niches possessed a low overall species richness and uneven distribution and the nasopharynx was found to possess more pathogenic bacterial species than the nose [49]. A previous publication by Toivonen *et al*. indicated a role for the nasal microbiota in an increased rate of acute respiratory infection (ARI) in children up to 2 years of age [50]—note that the TAILORED-Treatment cohort involved child and adult participants combined. Interestingly, we also identified 3 out of the 5 microbiota genera (i.e. *Moraxella*, *Streptococcus*, *Dolosigranulum*, *Staphylococcus* and *Corynebacteriacea*) described by Toivonen as potential contributors to an enhanced 'bacterial infection' versus 'viral infection' diagnostic algorithm. However, our findings indicated that the addition of such microbiota variables only marginally increased the predictive power of our combined child and adult diagnostic algorithm. Luna *et al*. indicated that although the microbiota and the anterior nares and nasopharynx are distinct, there may be "considerable" overlap between microbiota genera from these two sites (*Haemophilus* and *Moraxella*, but not *Staphylococcus* genera), at least in infants hospitalized with bronchiolitis [20]. The authors suggested that nasal swabs are "effective sample types" and can be used "to detect microbial risk markers". Again, the cohort characteristics of this publication were somewhat different (infants <1 year of age) as compared to the current publication and our study was not established to investigate bronchiolitis alone.

Finally, the nasal microbiota data used in our multi-parametric modelling was available at genus level due to limitations in the short-read sequencing technology used [51, 52]. As some microbiota genera found in the nose could contain pure commensal and potentially pathogenic species of bacteria, the utilisation of species level data, e.g. via long-read sequencing technologies, may generate increased accuracy in the development of nasal microbiota-based algorithms [53].

CRP as biomarker

The most widely used biomarker for the detection of inflammation and guidance of clinical decision-making in primary, secondary and tertiary care is C-reactive protein (CRP). A recent meta-analysis and review concluded that the use of CRP-based algorithms "seems to reduce

antibiotic treatment duration in neonates, as well as to decrease antibiotic treatment initiation in adult outpatients" [54]. This finding was also verified by Verbakel *et al.*, who in a 2019 systemic review and analysis concluded that "performing a point-of-care CRP test in ambulatory care accompanied by clinical guidance on interpretation reduces the immediate antibiotic prescribing in both adults and children" [55]. It is therefore not surprising, and indeed reassuring, that CRP is one of the major clinical factors identified in the TAILORED-Treatment algorithm. However, that said, the use of CRP in helping antibiotic prescribing decision-making processes may not always be obvious and may depend on factors not directly associated with CRP measurement per se [56, 57]. Point-of-Care (POC) diagnostic devices now exist for the measurement of CRP in different clinical environments e.g. QuikRead go (Orion diagnostics) [58, 59]. Further, novel diagnostic algorithms continue to be developed in spite of the relative success of CRP as a biomarker in clinical decision-making, largely due to the different circumstances in which CRP may be utilised and the 'grey zone' that exists in the 20–100 mg/l range [42]. In this respect, rapid biomarker-based diagnostics are currently available, or are being developed, that may more accurately distinguish between bacterial and viral infections in cases of infection, by using multiple biomarkers. Some examples of these diagnostics include: 1) CRP and Myxovirus Resistance Protein 1 (MxA1)—FebriDx® (RPS Inc., USA) [60]; 2) tumour necrosis factor-related apoptosis-inducing ligand (TRAIL), interferon gamma induced protein-10 (IP-10) and CRP—MeMed Key™ (MeMed BV, Israel) [61] and 3) Cathepsin B (CTSB), Hexokinase 3 (HK3), Interferon alpha inducible protein 27 (IFI27)—HostDx Tests (Inflammatix Inc., USA) [62].

Mixed infections

After 5-fold cross-validation, all 293 patients obtained their predicted class labels in the test sets. Subsequently, we investigated the prediction of the classifier using all 36 variables. More specifically, 22 out of 161 'viral' patients were misclassified as 'bacterial', while 24 out of 132 'bacterial' patients were misclassified as 'viral'. When we looked into these 24 misclassified 'bacterial' cases, we found that 15 of misclassified patients were actually associated with 'mixed' infection i.e. the expert panel had indicated that these patients were likely suffering from a viral and bacterial co-infection. Further, Fisher's exact test showed that 'mixed' patients were significantly more likely to be misclassified ($p = 0.0001$). We can also see the differences of the microbiota values between "mixed" and "bacterial" patients in **Fig 3**. In this study we treat these co-infected patients as 'bacterial infection' since they also are likely to require antibiotic treatment. However, inclusion of 'mixed' infections in the 'bacterial' infection category inevitably decreased the accuracy of predicting the 'bacterial infection' class. We tested this using the classifier with all 36 variables: when the 'mixed' patients were removed from the cohort, the classifier's AUC was increased from originally 0.90 to 0.94, and the accuracy of predicting the 'bacterial infection' class was improved to 87% while the accuracy of 'viral infection' class remained almost unchanged.

The algorithm developed in this study was designed to help clinicians decide on a bacterial versus viral infection at patient presentation. However, in real life situations, we cannot always be sure that these 'mixed' infections will be rapidly identified since clinicians do not have information immediately available on possible combinations of viral and bacterial infection. Consequently, the 'bacterial infection' class may include 'mixed' patients not identified as such at patient presentation. In this sense, the classification performance we report in this publication is an underestimation, with better algorithm performance being achieved if patients can be divided into strict 'bacterial' class at presentation where no viral infection is present. Of course, viral and bacterial infections often go hand in hand, so defining a strictly 'bacterial' class patient population may not be realistic.

Supporting information

S1 Fig. Input variables and missing data. The selected input variables for the classifiers and their respective missing data percentages. (A) 29 clinical eCRF variables and (B) 7 microbiota variables at genus level after Rhea transformation, i.e. all relative abundances in any sample below 0.5% were considered as absent. The variables are sorted by the importance calculated based on the initial cohort (see Methods). *: Numeric variables.
(TIF)

S2 Fig. Distribution of the four most predictive variables in the expanded cohort of 293 patients. (A) C-reactive protein (CRP). (B) Absolute neutrophil count (ANC). (C) X-ray signs Consolidation. (D) Age group. In (A) and (B), the mean and standard error of mean are shown. In (C) and (D), the number of patients per category is shown. The patients with mixed infection are indicated separately although they were calculated as possessing 'bacterial infection' class labels in this study.
(ZIP)

S1 Table. Clinical characteristics for the cohort of 293 patients, including (A) adult and (B) child age group. (A) Age group "Adult". (B) Age group "Child" (<18 years old). For each characteristic, the (mean) value, percentage, and range are shown, together with a *p*-value by two-tailed *t*-test for numeric data or chi-square test for categorical data. *Excluding the missing values. #By Fisher's Exact test.
(XLSX)

S2 Table. 195 clinical eCRF variables as the input to the analysis.
(XLSX)

S3 Table. All clinical and microbiota data used for the prediction model.
(XLSX)

S4 Table. Summary of the performance of classifiers. CC: Classifier using CRP only in the initial cohort. CE: Classifiers using two or more eCRF variables (incl. CRP) in the initial cohort. CEM: Classifiers using all input eCRF variables (incl. CRP) and at least one nasal cavity microbiota variable in the initial cohort. CC*: Classifier using only CRP in the 5-fold cross-validation of the expanded cohort. CEM*: Classifiers using two or more variables (regardless of eCRF or nasal cavity microbiota origin) in the 5-fold cross-validation of the expanded cohort. CCM*: Classifier using CRP plus all 7 microbiota genera variables in the 5-fold cross-validation of the expanded cohort. AUC: Area Under the ROC Curve. SEM: standard error of the mean.
(XLSX)

S5 Table. Top sixteen predictive variables in the 5-fold cross-validation training sets. Given each training set in the 5-fold cross-validation of the expanded cohort, all input variables were ranked by their individual variable importance [34]. The overall ranking in the five training sets is shown here.
(XLSX)

S6 Table. Correlations between microbiota and eCRF variables in the initial cohort. For the significant correlations between the 7 microbiota and 21 categorical eCRF variables (upper part of the table), Mann-Whitney *U* test or Kruskal–Wallis two-tailed test *p*-values are shown. For the significant correlations between the 7 microbiota and numeric variables (incl. 8 numeric eCRF variables and the other microbiota variables, lower part of the table), Spearman's *rho* values are shown. These results were generated per class: those in class 'bacterial

infection' have a superscript 'b' and those in class 'viral infection' have a superscript 'v'. Results with p-val<0.05 were considered significant.
(XLSX)

S7 Table. Metadata for the 16S RNA sequencing dataset.
(XLSX)

Acknowledgments

We thank the study team from the University Medical Centre Utrecht, The Netherlands (Brigitte Buiteman, Maaike van der Lee, Wouter van der Valk), Department of Internal Medicine, Gelderse Vallei Hospital, Ede, The Netherlands (Rik Heijligenberg), Department of Paediatrics, Hillel Yaffe Medical Centre, Hadera, Israel (Sharon Reisfeld, Ronit Rachmilevitch, Itzhak Braverman, Valery Karton), Department of Paediatrics, Bnai Zion Medical Centre, Haifa, Israel (Irena Chistyakov), and from MeMed Diagnostics, Tirat Carmel, Israel (Liran Shani, Omer Sadeh, Stav Rakedzon, Tzah Feldman) for patient recruitment and data collection.

Author Contributions

Conceptualization: Yunlei Li, Andrew P. Stubbs, John P. Hays.

Data curation: Yunlei Li, Chantal B. van Houten, Stefan A. Boers, Ruud Jansen, Asi Cohen, Dan Engelhard, Robert Kraaij, Wouter J. de Waal, Karin M. de Winter-de Groot, Tom F. W. Wolfs, Pieter Meijers, Bart Luijk, Jan Jelrik Oosterheert, Sanjay U. C. Sankatsing, Aik W. J. Bossink, Michal Stein, Adi Klein, Jalal Ashkar, Ellen Bamberger, Isaac Srugo, Majed Odeh, Yaniv Dotan, Olga Boico, Liat Etshtein, Meital Paz, Roy Navon, Tom Friedman, Einav Simon, Tanya M. Gottlieb, Ester Pri-Or, Gali Kronenfeld, Kfir Oved, Eran Eden, Louis J. Bont.

Formal analysis: Yunlei Li, Stefan A. Boers, Robert Kraaij, Saskia D. Hiltemann, Jie Ju, David Fernández, Cristian Mankoc, Eva González.

Funding acquisition: John P. Hays.

Investigation: Yunlei Li, Robert Kraaij, Andrew P. Stubbs, Louis J. Bont, John P. Hays.

Methodology: Yunlei Li.

Software: Yunlei Li.

Validation: Yunlei Li.

Visualization: Yunlei Li.

Writing – original draft: Yunlei Li, John P. Hays.

Writing – review & editing: Yunlei Li, Chantal B. van Houten, Stefan A. Boers, Andrew P. Stubbs, Louis J. Bont, John P. Hays.

References

1. Feldman C, Shaddock E. Epidemiology of lower respiratory tract infections in adults. Expert Rev Respir Med. 2019; 13(1):63–77. https://doi.org/10.1080/17476348.2019.1555040 WOS:000455021200005. PMID: 30518278
2. Hossain MZ, Bambrick H, Wraith D, Tong SL, Khan A, Hore SK, et al. Sociodemographic, climatic variability and lower respiratory tract infections: a systematic literature review. Int J Biometeorol. 2019; 63 (2):209–19. https://doi.org/10.1007/s00484-018-01654-1 WOS:000459790300011. PMID: 30680618

3. Marshall DC, Goodson RJ, Xu YW, Komorowski M, Shalhoub J, Maruthappu M, et al. Trends in mortality from pneumonia in the Europe union: a temporal analysis of the European detailed mortality database between 2001 and 2014. Resp Res. 2018; 19. https://doi.org/10.1186/s12931-018-0781-4 WOS:000431832800001. PMID: 29728122

4. Moore A, Harnden A, Grant CC, Patel S, Irwin RS, Altman KW, et al. Clinically Diagnosing Pertussis-associated Cough in Adults and Children CHEST Guideline and Expert Panel Report. Chest. 2019; 155 (1):147–54. https://doi.org/10.1016/j.chest.2018.09.027 WOS:000454907500028. PMID: 30321509

5. Hill AT, Gold PM, El Solh AA, Metlay JP, Ireland B, Irwin RS, et al. Adult Outpatients With Acute Cough Due to Suspected Pneumonia or Influenza CHEST Guideline and Expert Panel Report. Chest. 2019; 155(1):155–67. https://doi.org/10.1016/j.chest.2018.09.016 WOS:000454907500029. PMID: 30296418

6. Keitel K, Samaka J, Masimba J, Temba H, Said Z, Kagoro F, et al. Safety and Efficacy of C-reactive Protein-guided Antibiotic Use to Treat Acute Respiratory Infections in Tanzanian Children: A Planned Subgroup Analysis of a Randomized Controlled Noninferiority Trial Evaluating a Novel Electronic Clinical Decision Algorithm (ePOCT). Clin Infect Dis. 2019; 69(11):1926–34. https://doi.org/10.1093/cid/ciz080 WOS:000501730000013. PMID: 30715250

7. Çetinkaya A, Uysal MA, Niksarlıoglu EY, Durna AS, Ozer NO, Camsarı G. Can Neutrophil/lymphocyte Ratio, C-reactive protein (CRP) and Procalcitonin predict the hospitalization time in patients with lower tract respiratory infections? European Respiratory Journal. 2019; 54. https://doi.org/10.1183/13993003.congress-2019.PA3849

8. Jun-Hua T, Dong-Ping G, Zou P. Comparison of serum PCT and CRP levels in patients infected by different pathogenic microorganisms: a systematic review and meta-analysis. Brazilian Journal of Medical and Biological Research. 2018; 51(e6783). https://doi.org/10.1590/1414-431x20176783 PMID: 29846409

9. Póvoa P, Coelho L, Bos L. Biomarkers in Pulmonary Infections. Clinical Pulmonary medicine. 2019; 26 (4):118–25(8). https://doi.org/10.1097/CPM.0000000000000322

10. Oved K, Cohen A, Boico O, Navon R, Friedman T, Etshtein L, et al. A novel host-proteome signature for distinguishing between acute bacterial and viral infections. PLoS One. 2015; 10(3):e0120012. Epub 2015/03/19. https://doi.org/10.1371/journal.pone.0120012 PONE-D-14-53399 [pii]. PMID: 25785720; PubMed Central PMCID: PMC4364938.

11. Stein M, Lipman-Arens S, Oved K, Cohen A, Bamberger E, Navon R, et al. A novel host-protein assay outperforms routine parameters for distinguishing between bacterial and viral lower respiratory tract infections. Diagn Microbiol Infect Dis. 2018; 90(3):206–13. Epub 2017/12/24. https://doi.org/10.1016/j.diagmicrobio.2017.11.011 PMID: 29273482.

12. Engelmann I, Dubos F, Lobert PE, Houssin C, Degas V, Sardet A, et al. Diagnosis of viral infections using myxovirus resistance protein A (MxA). Pediatrics. 2015; 135(4):e985–93. Epub 2015/03/25. https://doi.org/10.1542/peds.2014-1946 PMID: 25802344.

13. Rhedin SA, Eklundh A, Ryd-Rinder M, Naucler P, Martensson A, Gantelius J, et al. Introducing a New Algorithm for Classification of Etiology in Studies on Pediatric Pneumonia: Protocol for the Trial of Respiratory Infections in Children for Enhanced Diagnostics Study. JMIR Res Protoc. 2019; 8(4): e12705. Epub 2019/04/27. https://doi.org/10.2196/12705 PMID: 31025954; PubMed Central PMCID: PMC6658235.

14. Venge P, Xu S. Diagnosis and Monitoring of Acute Infections with Emphasis on the Novel Biomarker Human Neutrophil Lipocalin. The Journal of Applied Laboratory Medicine. 2019; 3(4):664–74. https://doi.org/10.1373/jalm.2018.026369 PMID: 31639734

15. Mewes JC, Pulia MS, Mansour MK, Broyles MR, Nguyen HB, Steuten LM. The cost impact of PCT-guided antibiotic stewardship versus usual care for hospitalised patients with suspected sepsis or lower respiratory tract infections in the US: A health economic model analysis. Plos One. 2019; 14(4). https://doi.org/10.1371/journal.pone.0214222 WOS:000465223900005. PMID: 31013271

16. van Houten CB, Cohen A, Engelhard D, Hays JP, Karlsson R, Moore E, et al. Antibiotic misuse in respiratory tract infections in children and adults-a prospective, multicentre study (TAILORED Treatment). Eur J Clin Microbiol Infect Dis. 2019; 38(3):505–14. Epub 2019/02/02. https://doi.org/10.1007/s10096-018-03454-2 [pii]. PMID: 30707378; PubMed Central PMCID: PMC6394715.

17. Man WH, Houten MA, Mérelle ME, Vlieger AM, Chu MLJN, Jansen NJG, et al. Bacterial and viral respiratory tract microbiota and host characteristics in children with lower respiratory tract infections: a matched case-control study. The Lancet Respiratory Medicine. 2019; 7(5):417–26. https://doi.org/10.1016/S2213-2600(18)30449-1 PMID: 30885620

18. Dubourg G, Edouard S, Raoult D. Relationship between nasopharyngeal microbiota and patient's susceptibility to viral infection. Expert Review of Anti-infective Therapy 2019; 17(6). https://doi.org/10.1080/14787210.2019.1621168 PMID: 31106653

19. Langelier C, Kalantar KL, Moazed F, Wilson MR, Crawford ED, Deiss T, et al. Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. P Natl Acad Sci USA. 2018; 115(52):E12353–E62. https://doi.org/10.1073/pnas.1809700115 WOS:000454302600029. PMID: 30482864

20. Luna PN, Hasegawa K, Ajami NJ, Espinola JA, Henke DM, Petrosino JF, et al. The association between anterior nares and nasopharyngeal microbiota in infants hospitalized for bronchiolitis. Microbiome. 2018; 6. https://doi.org/10.1186/s40168-017-0385-0 WOS:000419158300001. PMID: 29298732

21. van Houten CB, Oved K, Eden E, Cohen A, Engelhard D, Boers S, et al. Observational multi-centre, prospective study to characterize novel pathogen-and host-related factors in hospitalized patients with lower respiratory tract infections and/or sepsis—the "TAILORED-Treatment" study. BMC Infect Dis. 2018; 18(1):377. Epub 2018/08/09. https://doi.org/10.1186/s12879-018-3300-9 PMID: 30086729; PubMed Central PMCID: PMC6081806.

22. Edouard S, Million M, Bachar D, Dubourg G, Michelle C, Ninove L, et al. The nasopharyngeal microbiota in patients with viral respiratory tract infections is enriched in bacterial pathogens. Eur J Clin Microbiol. 2018; 37(9):1725–33. https://doi.org/10.1007/s10096-018-3305-8 WOS:000442571200017. PMID: 30033505

23. Brealey JC, Chappell KJ, Galbraith S, Fantino E, Gaydon J, Tozer S, et al. Streptococcus pneumoniae colonization of the nasopharynx is associated with increased severity during respiratory syncytial virus infection in young children. Respirology. 2018; 23(2):220–7. https://doi.org/10.1111/resp.13179 WOS:000422787200015. PMID: 28913912

24. Mancini DAP, Alves RCB, Mendonca RMZ, Bellei NJ, Carraro E, Machado AMO, et al. Influenza virus and proteolytic bacteria co-infection in respiratory tract from individuals presenting respiratory manifestations. Rev Inst Med Trop Sp. 2008; 50(1):41–6. https://doi.org/10.1590/s0036-46652008000100009 WOS:000256274400009. PMID: 18327486

25. www.r-project.org. The R Project for Statistical Computing [12th May 2020]. Available from: www.r-project.org.

26. Biesbroek G, Sanders EAM, Roeselers G, Wang XH, Caspers MPM, Trzcinski K, et al. Deep Sequencing Analyses of Low Density Microbial Communities: Working at the Boundary of Accurate Microbiota Detection. Plos One. 2012; 7(3). https://doi.org/10.1371/journal.pone.0032942 WOS:000303021100043. PMID: 22412957

27. Bogaert D, Keijser B, Huse S, Rossen J, Veenhoven R, van Gils E, et al. Variability and Diversity of Nasopharyngeal Microbiota in Children: A Metagenomic Analysis. Plos One. 2011; 6(2). https://doi.org/10.1371/journal.pone.0017035 WOS:000287931400025. PMID: 21386965

28. Fadrosh DW, Ma B, Gajer P, Sengamalay N, Ott S, Brotman RM, et al. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. Microbiome. 2014; 2. https://doi.org/10.1186/2049-2618-2-6 WOS:000363189500001. PMID: 24558975

29. Hiltemann SD, Boers SA, van der Spek PJ, Jansen R, Hays JP, Stubbs AP. Galaxy mothur Toolset (GmT): a user-friendly application for 16S rRNA gene sequencing analysis using mothur. Gigascience. 2019; 8(2). Epub 2019/01/01. https://doi.org/10.1093/gigascience/giy166 PMID: 30597007; PubMed Central PMCID: PMC6377400.

30. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol. 2013; 79(17):5112–20. Epub 2013/06/25. https://doi.org/10.1128/AEM.01043-13 PMID: 23793624; PubMed Central PMCID: PMC3753973.

31. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig WG, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 2007; 35(21):7188–96. https://doi.org/10.1093/nar/gkm864 ISI:000251868800024. PMID: 17947321

32. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011; 27(16):2194–200. Epub 2011/06/28. https://doi.org/10.1093/bioinformatics/btr381 PMID: 21700674; PubMed Central PMCID: PMC3150044.

33. Lagkouvardos I, Fischer S, Kumar N, Clavel T. Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. PeerJ. 2017; 5:e2836. Epub 2017/01/18. https://doi.org/10.7717/peerj.2836 PMID: 28097056; PubMed Central PMCID: PMC5234437.

34. Breiman L. Random forests. Machine Learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324 WOS:000170489900001.

35. Breiman L. Bagging predictors. Machine Learning. 1996; 24(2):123–40. https://doi.org/10.1023/A:1018054314350 WOS:A1996UZ38000003.

36. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random Survival Forests. Ann Appl Stat. 2008; 2 (3):841–60. https://doi.org/10.1214/08-Aoas169 WOS:000261057900003.

37. Brugger SD, Eslami SM, Pettigrew MM, Escapa IF, Henke MT, Kong Y, et al. Dolosigranulum pigrum Cooperation and Competition in Human Nasal Microbiota. Msphere. 2020; 5(5). https://doi.org/10.1128/mSphere.00852-20 WOS:000579365500007. PMID: 32907957

38. van den Munckhof EHA, Hafkamp HC, de Kluijver J, Kuijper EJ, de Koning MNC, Quint WGV, et al. Nasal microbiota dominated by Moraxella spp. is associated with respiratory health in the elderly population: a case control study. Respir Res. 2020; 21(1):181. Epub 2020/07/16. https://doi.org/10.1186/s12931-020-01443-8 10.1186/s12931-020-01443-8 [pii]. PMID: 32664929; PubMed Central PMCID: PMC7362441.

39. Thibeault C, Suttorp N, Opitz B. The microbiota in pneumonia: From protection to predisposition. Sci Transl Med. 2021; 13(576). Epub 2021/01/15. https://doi.org/10.1126/scitranslmed.aba0501 PMID: 33441423.

40. Lemon KP. Human nasal microbiota. Current Biology. 2020; 30(19):R1118–R9. WOS:000579845200024. https://doi.org/10.1016/j.cub.2020.08.010 PMID: 33022252

41. Bomar L, Brugger SD, Lemon KP. Bacterial microbiota of the nasal passages across the span of human life. Current Opinion in Microbiology. 2018; 41:8–14. https://doi.org/10.1016/j.mib.2017.10.023 WOS:000429514300003. PMID: 29156371

42. Frank DN, Feazel LM, Bessesen MT, Price CS, Janoff EN, Pace NR. The Human Nasal Microbiota and Staphylococcus aureus Carriage. Plos One. 2010; 5(5). https://doi.org/10.1371/journal.pone.0010598 WOS:000277773700004. PMID: 20498722

43. Unger SA, Bogaert D. The respiratory microbiome and respiratory infections. J Infection. 2017; 74:S84–S8. WOS:000415365900014. https://doi.org/10.1016/S0163-4453(17)30196-2 PMID: 28646967

44. Cleary DW, Clarke SC. The nasopharyngeal microbiome Emerging Topics in Life Sciences. 2017; 1 (4):297–312. https://doi.org/10.1042/ETLS20170041 PMID: 33525776

45. Esposito S, Principi N. Impact of nasopharyngeal microbiota on the development of respiratory tract diseases. Eur J Clin Microbiol. 2018; 37(1):1–7. https://doi.org/10.1007/s10096-017-3076-7 WOS:000419149100001. PMID: 28795339

46. Bassis CM, Tang AL, Young VB, Pynnonen MA. The nasal cavity microbiota of healthy adults. Microbiome. 2014; 2. https://doi.org/10.1186/2049-2618-2-27 WOS:000363194700001. PMID: 25143824

47. Proctor DM, Relman DA. The Landscape Ecology and Microbiota of the Human Nose, Mouth, and Throat. Cell Host Microbe. 2017; 21(4):421–32. https://doi.org/10.1016/j.chom.2017.03.011 PMID: 28407480.

48. Perez-Losada M, Crandall KA, Freishtat RJ. Two sampling methods yield distinct microbial signatures in the nasopharynges of asthmatic children. Microbiome. 2016; 4(1):25. https://doi.org/10.1186/s40168-016-0170-5 PMID: 27306800.

49. De Boeck I, Wittouck S, Wuyts S, Oerlemans EFM, van den Broek MFL, Vandenheuvel D, et al. Comparing the Healthy Nose and Nasopharynx Microbiota Reveals Continuity As Well As Niche-Specificity. Front Microbiol. 2017; 8:2372. Epub 2017/12/15. https://doi.org/10.3389/fmicb.2017.02372 PMID: 29238339; PubMed Central PMCID: PMC5712567.

50. Toivonen L, Hasegawa K, Waris M, Ajami NJ, Petrosino JF, Camargo CA, et al. Early nasal microbiota and acute respiratory infections during the first years of life. Thorax. 2019; 74(6):592–9. https://doi.org/10.1136/thoraxjnl-2018-212629 WOS:000471125000010. PMID: 31076501

51. Kawai Y, Ozawa N, Fukuda T, Suzuki N, Mikata K. Development of an efficient antimicrobial susceptibility testing method with species identification by Nanopore sequencing of 16S rRNA amplicons. PLoS One. 2022; 17(2):e0262912. Epub 2022/02/04. https://doi.org/10.1371/journal.pone.0262912 PMID: 35113894; PubMed Central PMCID: PMC8812843 authors have read the journal's policy and have the following conflicts: Y. Kawai, N. Ozawa, T. Fukuda, N. Suzuki, K. Mikata are employees of Sumitomo Chemical Co., Ltd. The Sumitomo Chemical Co., Ltd. do not have any patents related to this work. These conflicts do not alter our adherence to all the PLOS ONE policies on sharing data and materials.

52. Church DL, Cerutti L, Gurtler A, Griener T, Zelazny A, Emler S. Performance and Application of 16S rRNA Gene Cycle Sequencing for Routine Identification of Bacteria in the Clinical Microbiology Laboratory. Clin Microbiol Rev. 2020; 33(4). Epub 2020/09/11. https://doi.org/10.1128/CMR.00053-19 PMID: 32907806; PubMed Central PMCID: PMC7484979.

53. Heikema AP, Horst-Kreft D, Boers SA, Jansen R, Hiltemann SD, de Koning W, et al. Comparison of Illumina versus Nanopore 16S rRNA Gene Sequencing of the Human Nasal Microbiota. Genes-Basel. 2020; 11(9). https://doi.org/10.3390/genes11091105 WOS:000581862900001. PMID: 32967250

54. Petel D, Winters N, Gore GC, Papenburg J, Beltempo M, Lacroix J, et al. Use of C-reactive protein to tailor antibiotic use: a systematic review and meta-analysis. Bmj Open. 2018; 8(12). https://doi.org/10.1136/bmjopen-2018-022133 WOS:000455309300034. PMID: 30580258

55. Verbakel JY, Lee JJ, Goyder C, Tan PS, Ananthakumar T, Turner PJ, et al. Impact of point-of-care C reactive protein in ambulatory care: a systematic review and meta-analysis. Bmj Open. 2019; 9(1). https://doi.org/10.1136/bmjopen-2018-025036 WOS:000471116800223. PMID: 30782747

56. Lemiengre MB, Verbakel JY, Colman R, De Burghgraeve T, Buntinx F, Aertgeerts B, et al. Reducing inappropriate antibiotic prescribing for children in primary care: a cluster randomised controlled trial of two interventions. Brit J Gen Pract. 2018; 68(668):E204–E10. https://doi.org/10.3399/bjgp18X695033 WOS:000425964500007. PMID: 29440016

57. Lemiengre MB, Verbakel JY, Colman R, Van Roy K, De Burghgraeve T, Buntinx F, et al. Point-of-care CRP matters: normal CRP levels reduce immediate antibiotic prescribing for acutely ill children in primary care: a cluster randomized controlled trial. Scand J Prim Health. 2018; 36(4):423–36. https://doi.org/10.1080/02813432.2018.1529900 WOS:000450634500010. PMID: 30354904

58. Matheeussen V, Van Hoof V, Loens K, Lammens C, Vanderstraeten A, Coenen S, et al. Analytical performance of a platform for point-of-care CRP testing in adults consulting for lower respiratory tract infection in primary care. Eur J Clin Microbiol. 2018; 37(7):1319–23. https://doi.org/10.1007/s10096-018-3253-3 WOS:000435950400016. PMID: 29744764

59. https://www.aidian.eu/us/products-available-in-the-usa/quikread-go-crp-for-the-usa#generally. QuikRead go [updated 12th May 2020]. Available from: https://www.aidian.eu/us/products-available-in-the-usa/quikread-go-crp-for-the-usa#generally.

60. https://www.febridx.com/. FebriDx [12th May 2020]. Available from: https://www.febridx.com/.

61. https://www.me-med.com/memed-key. MeMed Key [updated 12th May 2020]. Available from: https://www.me-med.com/memed-key.

62. https://inflammatix.com/hostdx-tests/. HostDx Tests. Available from: https://inflammatix.com/hostdx-tests/.