

A Four-Gene Signature from Blood to Exclude Bacterial Etiology of Lower Respiratory Tract Infection in Adults

Corresponding Author: Dr Ann Falsey

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Dr Falsey et al report the discovery and validation of a 4-gene signature for diagnosis of ARI in hospitalized adults. In short, while the work seems largely technically valid (though I have comments below), it does not seem like a substantial advancement in a very crowded field with numerous similar studies published over the past decade.

Some strengths of the paper are:

1. The 4-gene set is parsimonious (though Michael Levin's group has published a 2-gene set)
2. There is a locked model that is applied into other RNAseq datasets with reasonable performance
3. A very large sample size was winnowed to a very well-defined group of patients

The weaknesses are:

1. Participant enrollment within 24h of admission only may not be a group with clinical equipoise; this is not an emergency dept group where antibiotics may actually be withheld.
2. It is not clear whether the NI 'controls' were suspected of infection and so represent a group with equipoise.
3. There is a substantial difference in illness severity between the groups, potentially leading to confounding.
4. AUCs are reported as 'bacterial vs viral' and do not appear to include the noninfected patients, which causes upward bias in the reported number.
5. Primary dataset accuracy is nested-LOBO-CV, which will be upwardly biased.
6. The authors report NIH funding of an RNAseq dataset but do not supply a public deposition of the data in GEO.
7. One of the two 'external' datasets was gathered by the same authors. And there are literally dozens of other datasets the authors could have used in validation, leading to a concern of cherry-picking validation.
8. No context is given in terms of the performance of standard biomarkers such as WBC, CRP, or PCT.
9. No context is given in terms of the performance of other reported bacterial-viral gene sets, of which there are many more than those cited by the authors. E.g. see PMC8858657, which reports 28 B-V host signatures (and is already 3 years out of date).
10. The authors correctly point out that several tests are either FDA cleared (TriVerity, MeMed BV) or under development (e.g. BioFire), and here the authors do not have a rapid platform, meaning the work is substantially less likely than other tests to make a clinical impact.

Overall, the work seems broadly technically valid and generally well-reported. In my opinion, it deserves to be published, but is likely to be of relatively low impact given the large number of similar gene sets, some of which are already through FDA clearance.

Reviewer #2

(Remarks to the Author)

This manuscript by Falsey et al. studied peripheral blood gene expression in a cohort of adults hospitalized for the treatment of acute respiratory infection. The study aimed to identify a host gene expression signature to discriminate bacterial from non-bacterial infection (viral or bacterial/viral). The study identified a 4-gene signature that was validated in two public gene expression datasets.

The manuscript is well written and easy to understand. A notable strength of the study is the inclusion of bacterial-viral co-detections. However, the study largely builds on an extensive body of research and prior findings without making significant advances. The authors cite many of these in their manuscript (e.g., references 13-25, 37-43, among many others). Please consider the following comments related to the study's findings and their presentation.

The study authors excluded subjects with non-infectious acute respiratory illness, with the exception of a small control group that was used to assess the model after it was trained. There needs to be a better justification for the exclusion of this non-infected (NI) group. Specifically, it is clinically difficult to determine whether an infection is present or absent in patients with acute respiratory illness. NI etiologies overlap considerably in clinical presentation. As such, requiring clinicians to first diagnose an infectious etiology significantly limits the value of any test. This point is underscored by emerging host response technologies including MeMed and Inflammatrix, both of which report results for bacterial vs. non-bacterial etiologies, the latter inclusive of NI etiologies. The exclusion of NI patients in this study, save for the results shown in Figure 5B, fails to build on these precedents. The authors indicate in the supplementary methods that "Noninfected controls (NI) were not included in the nonbacterial group for the primary analysis because our primary goal was to develop a classifier for patients with a respiratory illness when infection was considered in the differential diagnosis." More often than not, NI etiologies are considered alongside infectious etiologies of acute respiratory illness. Therefore, the justification for the exclusion is insufficient. The authors should consider reanalyzing the data inclusive of the NI group or offer a more scientifically and clinically valid reason not to.

As noted, the NI group was used in a validation experiment (depicted in figure 5B). This important control group should be described in the Cohort Description section of the Results.

The study focuses in on a 4-gene signature. However, the authors should add information about the performance of larger and smaller signatures. Could performance be improved with 5 genes, 10 genes, etc.? How much does it suffer with 3 genes, 2 genes, or just 1?

Is this specific combination of 4 genes unique or could other genes be substituted in with equivalent performance?

The authors selected two publicly available gene expression datasets to perform external validation when there are dozens of appropriate datasets. Moreover, they do not justify why these two were selected, raising the possibility that results are only shown for those experiments where validation was successful. While this is unlikely to be the case, there is no alternative explanation offered. Instead of justifying why these two, the authors should validate their 4-gene model in a much larger number of datasets. When additional datasets are chosen for validation, they should include subjects with NI etiologies.

When reporting results for these validation cohorts, they should include standard performance characteristics such as sensitivity (or PPA), specificity (or NPA).

Were there patterns or characteristics observed among subjects that were incorrectly classified, whether in the derivation cohort or the validation cohorts?

On lines 149-152, the authors write, "Based upon prior feedback from clinicians regarding the acceptability of a diagnostic test to exclude bacterial cause of LRTI, we aimed to achieve at least 90% sensitivity while also providing >90% negative predictive value (NPV) for the determination of bacterial infection etiology in LRTI across a range of likely prevalences." The authors should provide a reference or justification for this threshold that is more valid than anecdotal clinician feedback. It could easily be argued that a NPV of 95% or at least 98% is needed to drive clinical adoption. Without reference or justification, the use of >90% is arbitrary.

The study identifies a gene signature threshold of -0.886 to discriminate bacterial from non-bacterial etiologies. Please explain where this number comes from and what it means. For example, on what scale does it exist; does it represent anything such as probability or deviation from baseline; is it dependent on the methodology used to measure gene expression (sequencing, microarray, PCR); why does it have a negative value. There is no context offered to understand where this number comes from and how one should interpret it.

The study indicates that procalcitonin was measured and was significantly different between bacterial and non-bacterial groups. Since this biomarker is often used to discriminate between these two groups, the authors should report the performance characteristics of procalcitonin to discriminate bacterial from non-bacterial etiologies in this cohort.

The first two paragraphs are largely background material that reproduce what was in the Introduction. Consider significantly trimming down, though this is purely a suggestion.

The authors claim in the Discussion on lines 186-188 that many biomarkers have shown limited success to supplement clinical judgement in assessing patients with ARI. First, while there is variability in the literature on this topic, there is an abundance of evidence to show exactly the opposite – that each of these biomarkers adds diagnostic value in assessing patients with ARI. The authors also separately call out CRP, TRAIL and IL-10. Second, IL-10 is likely incorrect and should be IP-10. The former is not discussed in the referenced citations whereas IP-10 is. Third, the combination of CRP, TRAIL, and IP-10 comprises the MeMed BV test. There is ample evidence highlighting the test's performance and additive clinical value above and beyond standard of care. These points should all be mentioned.

On line 220, the authors describe the TriVerity test. Note that it was incorrectly described as a 'trial' whereas it should be a

'study'.

The authors should explain why they chose to focus on respiratory infection and not all types of infection. A bacterial vs. non-bacterial classifier has applications far beyond ARI. Even if they chose to do discovery in ARI, validation should include non-respiratory infections.

While including so many bacterial-viral co-detections is a strength, the authors should expand on the challenges in determining whether these patients truly had co-infection or just co-detection. How likely was each pathogen to be causing active infection at the time of enrollment? I concede this is difficult if not impossible to determine, but it does raise questions about the certainty of the microbiological etiologies for these co-detection cases.

Line 334 refers to NS PCR. What does NS mean?

Line 352 refers to a study by Kho et al. The author's name is Ko.

For Figures 4 and supplemental Figure 1, please add labels for the x-axis. Also, the plots should ideally be made on the same scale for comparison purposes, including both x and y axes.

In the Methods, please provide more details about the software and tools used to perform the analyses.

Reviewer #3

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #4

(Remarks to the Author)

I read with interest the work from Falsey et al. In their work, the authors proposed a transcriptomics-based methodology for discriminating bacterial from non-bacterial infections.

In general, the work is scientifically sound, well written, and most of the conclusions are warranted by the presented results. Nevertheless, I have a few points that could be addressed in a next iteration of the manuscript.

1. I understand the authors obtained a subset of important features, i.e., the genes ITGA7, IFI27, FAM20A, and ITGB4 during Lasso selection. The downstream results the authors derived are only based on the expression level of the four genes initially obtained. I wonder if the authors can, for control purposes, obtain (a) set(s) of genes with coefficient = 0 and use these in their predictions. In doing this, the authors could evaluate any potential bias of Lasso, confirm the importance of their set of genes, and check for any potential missed gene that might have been over penalized.

2. Once obtained their genes of interest, the authors used them in the relaxation step of Lasso to classify in any bacterial infection. The authors established a threshold based on clinician inputs. I understand Lasso-based models have less complex decision boundaries which is useful for clinical environments but Lasso assumes linear relationships between the genes and the classes. It also tends to rule out genes that are correlated (it chooses one and ignores the rest). Have the authors considered other models for classification that can handle nonlinear relationship and interaction between features?

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I appreciate the authors' responses to my points.

First, a quick clarification: re: 24h, I am not suggesting 'later' timepoints, I am suggesting 'earlier' timepoints, e.g. the first 60 min, first 3 hours, etc. 24h after admission is potentially quite a late timepoint here.

Overall, my concern about impact remains and is perhaps magnified by the response, for the following reasons:

- The authors acknowledge severity as a cofounder but make no attempt to quantify it
- Agree the LOBO-CV is not *very* biased, but anything reported from the discovery dataset is not the same as external

validation

-The authors continue to substantially minimize the number of datasets tested. See e.g. Chawla, Cell Systems 2022, or Bodkin, Genome Med 2022. There are dozens and dozens of possible datasets. The whole point of a locked classifier is to be externally validatable, meaning even a dataset with only bacterial samples can be tested, because it's not 'relative differences' but rather a fixed classifier with a fixed cutoff. The RNAseq limitation is only partially convincing because RNAseq cannot be made into a bedside tool. So either the signature is not translatable to practice, or the studies are not being fully utilized. Plus, the one 'older' study used (GSE60244) was used in the Bodkin paper and *most* of those gene signatures had a higher apparent performance than the present one.

-The other platforms mentioned (TriVerity and MeMed) have both had publications specifically addressing bacterial-viral coinfections. The fact is the current signature does not show any particular utility beyond currently fielded platforms, and is several years behind.

I reiterate that I find the work to be reasonably technically sound, but of only low-moderate impact.

Reviewer #2

(Remarks to the Author)

The authors have made many changes to address concerns raised in the initial review. However, some concerns remain.

The authors declined to report performance characteristics of PCT because it was available to adjudicators, and this raised concerns about circular reasoning. I disagree with this justification. It is possible that PCT performance could be overestimated due to incorporation bias (rather than circular reasoning). However, this analysis can still be performed with that limitation as is the case for many other publications using PCT, CRP, or WBC as comparators. Poor performance of PCT despite incorporation bias would strengthen the argument that it is insufficient as a biomarker. And if PCT performance was very good, then it may very well be due to incorporation bias and acknowledged as such in the discussion. Moreover, adjudicators were not bound to PCT results and incorporated all clinical information in making their diagnostic assignments. I therefore maintain concerns that not reporting PCT performance is a weakness in the manuscript.

Regarding the focus on ARI as opposed to suspected infection more broadly, the authors suggest that "While it would be ideal to have one test that fits all infections, it seems likely the host response to respiratory infection and non-respiratory infections may differ and a semi-focused approach can optimize assay accuracy." The authors do not offer any evidence to support this claim. In contrast, there is an abundance of evidence for the opposite - that the host response to bacterial or viral infection is conserved across anatomic sites of infection. I recommend the authors validate their signature in datasets that include all infectious syndromes to support external validity and generalization. Even if it was derived in ARI, it is critical to understand how the signature performs in other syndromes.

With respect to impact, the manuscript offers incremental advances in our scientific knowledge in this field. The authors cite the inclusion of coinfection as the major advance. In a cursory search of PubMed, I identified eight pre-existing publications that included coinfection. That is not to say the current study isn't a useful addition. But it remains unclear how the current discoveries advance the field, especially given the narrow focus on LRTI in adults (as noted above).

That list, if interested, includes these titles:

- Discriminating Bacterial and Viral Infection Using a Rapid Host Gene Expression Test
- Longitudinal Dynamics of a Blood Transcriptomic Signature of Tuberculosis
- Comparison of two rapid host-response tests for distinguishing bacterial and viral infection in adults with acute respiratory infection
- BV score differentiates viral from bacterial-viral co-infection in adenovirus PCR positive children
- Detection of Viral Infection and Bacterial Coinfection and Superinfection in Coronavirus Disease 2019 Patients Presenting to the Emergency Department Using the 29-mRNA Host Response Classifier IMX-BVN-3: A Multicenter Study
- Validation of a host response test to distinguish bacterial and viral respiratory infection
- Host gene expression classifiers diagnose acute respiratory illness etiology
- A distinct influenza infection signature in the blood transcriptome of patients with severe community-acquired pneumonia

Reviewer #3

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I remain consistent in my overall determination that the work is technically sound but of low impact. Here are some point-by-point thoughts:

-The authors write, "In addition, since patients have been usually symptomatic for a range of days prior to admission (anywhere from 1-14 days), the collection of samples within 24 hours of admission should not affect the interpretation of the gene expression results." But this is apples-to-oranges, patients aren't on IV antibiotics at home! We do see changes in gene expression within 24 hours of IV antibiotics. The study isn't changeable, and the choice was pragmatic, but it still lowers the impact and external validity of the findings.

-The authors write that "there is no such thing as confounding of a classifier" (!). This is an absurd thing to say. The authors are claiming that classifiers can't show spurious associations between dependent and predictor variables through the presence of extraneous variables? Classifiers can't be biased?

-The authors then write that 'Semantics aside, if the concern is that our gene signature might relate not only to bacterial infection but also to disease severity, we do not believe that it would pose any issue for the classifier of bacterial infection. Indeed, if bacterial infections tend to be present among those with severe disease (whether or not the relationship is causal), as appears to be the case, then any good classifier of bacterial infection should be expected to also be at least somewhat related to disease severity.'

The authors are surely aware of causes of severity other than bacterial infection, such that severe non-infectious illnesses will be incorrectly classified as bacterial? Or that mild bacterial infections are incorrectly classified as non-bacterial? It's one thing to say 'this is a weakness and here's how it affects interpretation' – fine. It's another thing to deny the weakness itself. Fact is, there are ways to build classifiers that correct for association with severity without explicitly including a severity variable. The authors' apparently limited knowledge in this arena does not make the concern incorrect.

-"...external validation is actually pessimistically biased" – Yes, it's harder and more representative of real-world validation. I believe what the authors mean is that between two sources of potential change, they are stating that technical platform difference outweighs clinical difference, so negative result should be ignored. I'd encourage them to study the MAQC – technical differences are often swamped by biological differences. Negative results aren't just 'microarrays are different'.

-“ In this second revised manuscript, we have included validation in 2 additional populations” (one of which they imputed a gene). The issue is not to include 2 more populations, it's to understand why those two more cohorts were chosen out of dozens available. Were they the ones with good scores? This is a top-tier journal. Validation datasets should be approached in a systematic fashion, and that systematic approach is critical to showing that the validation results are meaningful. This is the "File Drawer Problem".

-The abstract still lists an AUROC range lower end of 0.89 instead of 0.73.

-The authors write 'Therefore, the concept of a fixed classifier is somewhat premature'. They then cite the Lydon paper that proves exactly my point – without a fixed classifier or a POC platform, there has been NO further development from the Duke group 6 years out (Lydon et al), showing the relatively low impact of presenting a classifier that no one will actually ever use.

I am very sorry for my blunt language – I know personally how many hundreds of hours go into the construction of these cohorts. But, then again, the authors did very little to positively respond to the points raised, other than to add 2 datasets (out of dozens) with no explanation of how. In fact, to one of my points, they went the other way, REMOVING the discussion of weakness!

My summary is that the science here is sound but of low impact, and the rebuttal and edits do not change my conclusion. Since the authors show little appetite for revisions, it is up to the editor to choose a path forward.

Reviewer #2

(Remarks to the Author)

The authors have responded to all comments presented in my prior review. Despite those responses, I respectfully disagree that this study and accompanying manuscript advance the field. The study is technically sound and should be published. But it fails to be of significance to the field and largely represents a me-too study in the established literature in the opinion of this reviewer.

Reviewer #3

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Line references refer to the clean version of the manuscript.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Dr Falsey et al report the discovery and validation of a 4-gene signature for diagnosis of ARI in hospitalized adults. In short, while the work seems largely technically valid (though I have comments below), it does not seem like a substantial advancement in a very crowded field with numerous similar studies published over the past decade.

Some strengths of the paper are:

1. The 4-gene set is parsimonious (though Michael Levin's group has published a 2-gene set)
2. There is a locked model that is applied into other RNAseq datasets with reasonable performance
3. A very large sample size was winnowed to a very well-defined group of patients

Response: We appreciate the reviewer's positive comments

The weaknesses are:

1. Participant enrollment within 24h of admission only may not be a group with clinical equipoise; this is not an emergency dept group where antibiotics may actually be withheld.

Response: We are not entirely sure what the reviewer is suggesting regarding clinical equipoise. If they suggest it would have been better to enroll persons at later time points in their hospital course, we respectfully disagree. This type of diagnostic test to discriminate bacterial and non-bacterial respiratory infection is meant to be used early in the patient's course for maximal benefit. The incidence of side effects from antibiotics rise with days on treatment and the host response likely changes over time, related to time from illness onset and treatment. Applying the test later in the course might introduce more variation. Additionally, the hospital team managing the patient are often left with the decision whether to continue antibiotics already started in the ER or continue to hold antibiotics. We believe the maximal utility of a gene expression diagnostic test would be if it could be used as early as possible after presentation to the hospital. Inclusion of patients within 24 hours reflects the ability of the study team to identify and enroll participants.

2. It is not clear whether the NI 'controls' were suspected of infection and so represent a group with equipoise.

Response: NI controls were specifically selected as having cardiopulmonary symptom and signs (i.e., dyspnea, cough, chest pain) but the clinicians had no concern for infection as a clear alternative diagnosis was evident (i.e., pulmonary embolus, pneumothorax) and thus it does not represent a group with equipoise. This group (per the grant and the protocol) was included to evaluate the specificity of the gene predictors once developed using the bacterial and non-bacterial diagnosis groups. Supplementary Table 3 provides additional information on the NI group and Figure 1

defines the exclusions. Per reviewer 2 who also wanted additional data on the NI group, this information is now presented in the results on lines 106-111.

“The 63 NI controls included a variety of illnesses (28- congestive heart failure [CHF], 18 pulmonary emboli [PE], 7 atrial fibrillation with rapid ventricular response, 4 asthma exacerbations, 3 AECOPD, and 3 other) deemed non-infectious by the provider team and confirmed as such by the clinical adjudication committee. The reasons for exclusion of 18 NI are shown in Figure 1. Additional details of the NI cohort compared to the primary analysis group are shown in Supplemental Table 3.”

3. There is a substantial difference in illness severity between the groups, potentially leading to confounding.

Response: We agree that severity was different and may be a confounding factor. This is now included in the limitations section of the discussion on lines 344-345. It may be difficult to tease out the confounding effects since there may be a biologic basis that bacterial infections are generally more severe than viral alone (with the exception of COVID-19 in a non-immune patient).

“Finally, illnesses adjudicated as bacterial were more severe than those adjudicated as viral, thus severity may be a confounding factor in development of our gene signature.”

4. AUCs are reported as ‘bacterial vs viral’ and do not appear to include the noninfected patients, which causes upward bias in the reported number.

Response: Because of concerns here and from reviewer 2 we now report the performance and AUC with ROC for our 4 gene signature – including the NI group with the nonbacterial. This information is provided on lines 142-143 and Figure 3B. The AUC was essentially the same including NI as it was without this group.

5. Primary dataset accuracy is nested-LOBO-CV, which will be upwardly biased.

Response: We respectfully disagree with this statement. Upward bias can result when either (1) non-nested CV is used or (2) standard k-fold CV is used. By using LOBO-CV, we ensured that no subjects from any of the batches used to train the model would be included the left-out validation fold, thereby guaranteeing independence between each set of training and validation data. By further appropriately using nested CV, we eliminated the bias that can result when investigators optimize the CV criterion and then report that optimized criterion. Nested LOBO-CV is thus essentially unbiased, unlike standard k-fold CV or, worse, naïve estimates of AUC obtained by selecting a model and naively estimating AUC using all of the same data used to select the model and estimate its coefficients. Finally, we also performed external validation.

6. The authors report NIH funding of an RNAseq dataset but do not supply a public deposition of the data in GEO.

Response: The raw data has been deposited in dbGap and will be accessible using the identifier phs003931.v1.p1.

7. One of the two ‘external’ datasets was gathered by the same authors. And there are literally dozens of other datasets the authors could have used in validation, leading to a concern of cherry-picking validation.

Response: We can understand that the reviewer is concerned about cherry picking the external validation data sets. This was certainly not our intent, but the majority of studies available are not ideal to validate our study (please see the table below). Thus, we initially used 2 studies that used RNAseq in adult populations with respiratory infections. Most of the other available data sets used microarray assays rather than RNAseq, were applied to pediatric populations, or did not include the desired infectious phenotypes (bacterial and nonbacterial, viral infections). However, we have now performed 3 additional validations (2 adult studies and one pediatric study) using RNAseq data. We included the controls as requested by reviewer 2 although controls in 2 of the 3 studies with controls were healthy volunteers. The value of including “healthy” controls as opposed to non-infected controls seems minimal since it is likely a clinician can discriminate a healthy person from a sick one. Thus, we provide the validation results with and without the control groups. These results are presented in Table 2 which now replaces Supplemental Figure 2. Although we don’t feel it is best to validate using pediatric studies, given the limited RNAseq data sets we included a pediatric study that was recently published (iScience 28, 111747, February 21, 2025). Interestingly, in the pediatric study one of our genes (FAM20A) was also included in their 5 gene predictor.

Added to the 2 prior validation sets we now provide validation results from 5 independent RNAseq data sets. We have added this information on lines 156-162.

“We also evaluated our gene signature in 2 additional adult data sets (GSE 163151: 6 bacterial and 20 viral) AUC of 0.89 and (GSE161731: 24 bacterial and 78 viral) with AUC of 0.98. If healthy controls were included the AUCs dropped to 0.82 and 0.97, respectively. Finally, we evaluated our 4-gene set in a pediatric pneumonia study (GSE261482) that included 98 definite bacterial and 12 definite viral infections resulting in an AUC of 0.74. If the healthy controls were included the AUC rose to 0.92.”

The difference in platform characteristics make the micro array data sets less ideal for validating RNAseq studies. We could not use the largest and most appropriate data set (GSE63990) because it did not contain the FAM20A gene which was important in our gene predictor set. We did run one validation on the next most appropriate micro array data set that included adults with respiratory infection (GSE60244). Our 4 gene predictor did discriminate bacterial and nonbacterial illness in the micro array data set (AUC 0.73) without the healthy controls and somewhat lower when including healthy controls (AUC 0.67); which were clearly were not as closely aligned with validation using RNAseq data. This information is included on lines 162-166.

“Despite the different platforms we also performed a validation using one micro array data set that included 154 adults with viral and bacterial respiratory infection as well as healthy controls (GSE60244). Our 4 gene predictor did discriminate bacterial from non-bacterial infections (AUC 0.73), but results did not align as closely as validations with an RNAseq data sets”.

For the reviewers reference we have created a table of the additional publicly available datasets and limitations that we believe make them inappropriate I for our validation analyses.

<i>RNASeq (16)</i>
<i>COVID, viral vs Healthy (10) (no bacterial or viral-bacterial co-infection)</i>

CRA002390 GSE156063 GSE152418 GSE152641 GSE157103 GSE161777 GSE155454 GSE157859 GSE147507 GSE171110
<i>Viral vs Healthy Children (2) (Pediatric and no bacterial or viral-bacterial co-infection)</i> GSE99992 GSE113210
<i>Nasal samples only (1)</i> GSE152075
<i>Viral/Bacterial/Healthy - additional validations now presented (3)</i>
GSE165731 – <i>Adults (1)</i>
GSE163151 – <i>Adults (1)</i>
GSE261482 – <i>Children (1)</i>
<i>Micro Array (50)</i>
<i>Bacterial vs. Healthy (9) only (no viral or viral-bacterial co-infection)</i> GSE13015 GSE30119 GSE54514 GSE29161 GSE40586 GSE33341 GSE72946 E-MTAB-3423 GSE69528
<i>Bacterial vs. SIRS vs. Healthy (3)</i> GSE9960 GSE22098 GSE28750
<i>Viral vs. Healthy (18) (no bacterial or viral-bacterial co-infection)</i> GSE67059-1 GSE29366 GSE38900-2 GSE27131 GSE21802 GSE17156 GSE34205 GSE82050 GSE114466 GSE106475 GSE103842 GSE38900-1 GSE61754 GSE61821

GSE67059-2
GSE68310
GSE77087
GSE80179
<i>Bacterial vs. Viral (20)</i>
<i>Pediatric only (14)</i>
GSE6269-1
GSE6269-2
GSE25504-1
GSE42026
GSE72810
GSE9692
GSE119217
GSE25504-2
GSE103119
GSE40396
GSE64456
GSE73462
GSE68004
GSE73461
<i>Adults (6)</i>
GSE60244- <i>now presented</i>
GSE20346 – <i>only 10 subjects</i>
GSE185576 – <i>no publication associated with the dataset</i>
GSE63990- <i>data does not include FAM20A (one of our predictor genes)</i>
GSE40012- <i>longitudinal sampling</i>
E-MEXP-3589 <i>-whole data set not available</i>

8. No context is given in terms of the performance of standard biomarkers such as WBC, CRP, or PCT.

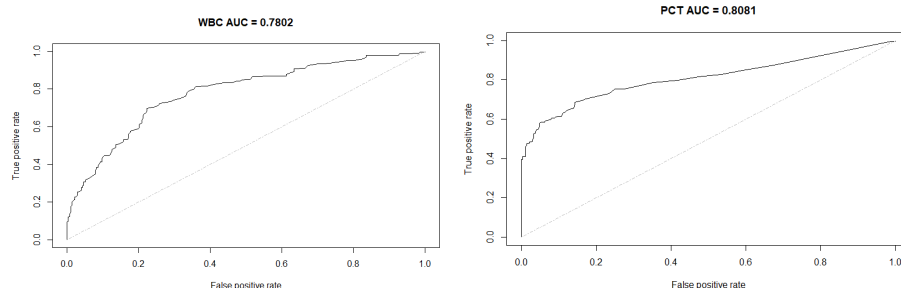
Response: Since all three of these measures (particularly WBC and PCT) were available to the clinicians to assist with adjudication of the cases, including these clinical variables would reflect circular reasoning and lead to upward bias. An AUC for each parameter (CRP not included due to many missing results) has been computed for the reviewer's interest. But we would prefer to not include them in the manuscript, given their upward bias.

B+BV (N=224) vs V (N=280)

WBC (502 non-missing values) AUC = 0.78

PCT (492 non-missing values) AUC = 0.81

Combining WBC and PCT in the same model yields AUC = 0.86 (still lower than our 4-gene signature, despite the upward bias of this clinical model)



9. No context is given in terms of the performance of other reported bacterial-viral gene sets, of which there are many more than those cited by the authors. E.g. see PMC8858657, which reports 28 B-V host signatures (and is already 3 years out of date).

Response: Please see the answer to question 7 (which outlines validation data in an additional 4 independent data sets), and we have substantially revised the discussion (lines 333-356) to provide context for our work in view of the published literature.

10. The authors correctly point out that several tests are either FDA cleared (TriVerity, MeMed BV) or under development (e.g. BioFire), and here the authors do not have a rapid platform, meaning the work is substantially less likely than other tests to make a clinical impact.

Response: While it is true that we have not initiated commercial development using our gene set, we respectfully disagree that it will have no impact. The most novel and very important aspect of the current study is that we have successfully dealt with mixed viral bacterial infections. This important subgroup is critically important to successful use with respiratory infections. Triverity has been well studied in sepsis which is a different population. Although it might be possible that one test might be valid for all types of infections and age groups, we believe that one geared to respiratory infections in adults may demonstrate better operating characteristics for specific subgroups, such as mixed viral-bacterial infections. Other tests mentioned do not provide substantial information about the test characteristics for these mixed infections. PCR testing is quite good for diagnosing viral respiratory infections yet has not led to changes in practice in hospitalized patient because concomitant bacterial infection is not ruled out. Thus, although we are behind in development of our gene predictor, we feel that we are filling a critical need by including the mixed viral-bacterial group, and that the parsimonious 4-gene set lends itself to commercial development in the future. Our discussion has been revised to raise these points.

Overall, the work seems broadly technically valid and generally well-reported. In my opinion, it deserves to be published but is likely to be of relatively low impact given the large number of similar gene sets, some of which are already through FDA clearance.

Response: Please see above regarding impact.

Reviewer #2 (Remarks to the Author):

This manuscript by Falsey et al. studied peripheral blood gene expression in a cohort of adults hospitalized for the treatment of acute respiratory infection. The study aimed to identify a host

gene expression signature to discriminate bacterial from non-bacterial infection (viral or bacterial/viral). The study identified a 4-gene signature that was validated in two public gene expression datasets.

The manuscript is well written and easy to understand. A notable strength of the study is the inclusion of bacterial-viral co-detections. However, the study largely builds on an extensive body of research and prior findings without making significant advances. The authors cite many of these in their manuscript (e.g., references 13-25, 37-43, among many others). Please consider the following comments related to the study's findings and their presentation.

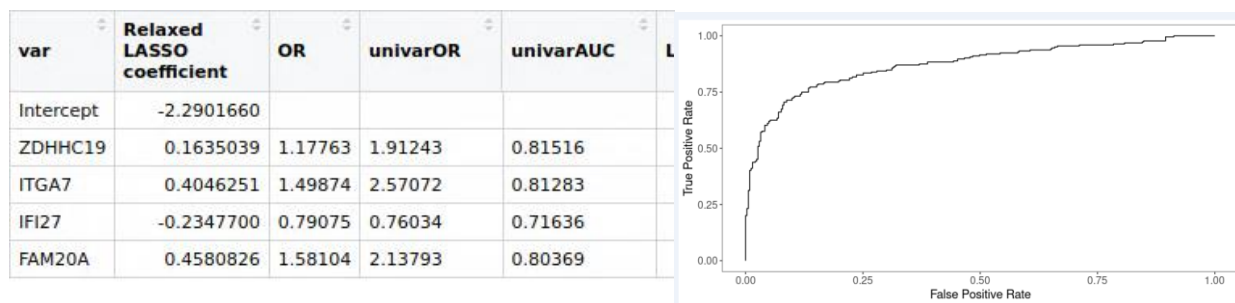
Response: We appreciate the reviewer's positive comments. While it is true that there is now a large body of literature using the host response to differentiate viral and bacterial illnesses, we believe that our findings do represent a significant advance, namely the inclusion of the large and critically important category of viral-bacterial co-infection. The reviewer notes this as a strength of our study which we appreciate but would like to emphasize this is a significant issue in respiratory infections and almost none of the prior studies have dealt with this very important group. The concern about bacterial co-infection is why the multiplex viral PCR tests have resulted in minimal change in clinical management.

The study authors excluded subjects with non-infectious acute respiratory illness, with the exception of a small control group that was used to assess the model after it was trained. There needs to be a better justification for the exclusion of this non-infected (NI) group. Specifically, it is clinically difficult to determine whether an infection is present or absent in patients with acute respiratory illness. NI etiologies overlap considerably in clinical presentation. As such, requiring clinicians to first diagnose an infectious etiology significantly limits the value of any test. This point is underscored by emerging host response technologies including MeMed and Inflammatrix, both of which report results for bacterial vs. non-bacterial etiologies, the latter inclusive of NI etiologies. The exclusion of NI patients in this study, save for the results shown in Figure 5B, fails to build on these precedents. The authors indicate in the supplementary methods that "Noninfected controls (NI) were not included in the nonbacterial group for the primary analysis because our primary goal was to develop a classifier for patients with a respiratory illness when infection was considered in the differential diagnosis." More often than not, NI etiologies are considered alongside infectious etiologies of acute respiratory illness. Therefore, the justification for the exclusion is insufficient. The authors should consider reanalyzing the data inclusive of the NI group or offer a more scientifically and clinically valid reason not to.

Response: We do not disagree with the reviewer that frequently infectious and non-infectious etiologies are considered when patients present to the ER with difficulty breathing. On the other hand, there are many clinical scenarios when infection is not high in the differential diagnosis and an alternative diagnosis rapidly can be identified (i.e., pulmonary embolus, pulmonary edema or pneumothorax). We believe that any diagnostic test is best employed in the proper clinical situation rather than used indiscriminately for all patients with variety of conditions. Our study population of infected participants did not require absolute diagnosis of infection but rather included possible or suspected infection. The investigators then performed comprehensive

diagnostic testing to identify the microbiologic etiology. In contrast, the NI controls were required to have a non-infectious clinical diagnosis by the treating provider, negative microbiologic diagnostic testing performed as either standard of care or by research testing and agreement by the adjudication committee that the etiology of illness truly was non-infectious. Consequently, we felt it best not to include this group with either the bacterial or non-bacterial groups in the primary analysis who were felt to have evidence of infection. Since the grant and the protocol were written as such, we don't feel it is appropriate to perform a post hoc analysis of our data for this publication. However, we do now present the performance of our 4-gene signature including the NI controls categorized with the nonbacterial group. This information is provided on lines 142-143 and Figure 3B. As shown, the AUC was essentially the same at 0.90.

Given the reviewers concern we have reanalyzed our raw data including the non-infected controls as nonbacterial and the results are very similar with one gene difference and an AUC of 0.89 as shown below. However, given the reasons as noted above we prefer not to present these data in the paper.



As noted, the NI group was used in a validation experiment (depicted in figure 5B). This important control group should be described in the Cohort Description section of the Results.

More detail has been presented on the NI group in the results on lines 106-111. We also refer the reviewer to supplemental table 3 and Figure 1.

The study focuses in on a 4-gene signature. However, the authors should add information about the performance of larger and smaller signatures. Could performance be improved with 5 genes, 10 genes, etc.? How much does it suffer with 3 genes, 2 genes, or just 1?

Response: We employed statistical methods designed to select a parsimonious set of genes; we did not fix the number of genes at 4 but rather our methods resulted in selecting 4 genes. We also reported univariate AUC for each of these genes, so that answers the 1-gene question; these univariate AUC (0.80 for ITGB4, e.g.) were not cross-validated and are thus upwardly biased, so a 1-gene model is clearly worse ($AUC \leq 0.80$) than our 4-gene model ($AUC = 0.90$). Similarly, 2-3 gene models fall in between, and would have been selected had they performed as well as 4 genes. Adding more genes (5+) would not substantially improve performance, or else more genes would have been selected; e.g. before we employed hard-thresholding of small LASSO-constrained coefficients to 0, the Relaxed LASSO procedure selected 9 genes and achieved the same $AUC = 0.90$, whereas standard (unrelaxed) LASSO would select many more genes without improving (and possibly degrading) performance.

Is this specific combination of 4 genes unique or could other genes be substituted in with equivalent performance?

Response: This is a great question. These 4 genes are not necessarily uniquely able to discriminate between bacterial and non-bacterial infection, as we identified 5,401 differentially expressed genes (see Figure 2). Critically, as a set, they demonstrate the best performance for discriminating between bacterial and non-bacterial infections. We tried excluding the 4 selected genes and re-running the procedure on the remaining genes; the resulting disjoint set of selected genes performed a little (but not a lot) worse, with AUC in the high 80's.

The authors selected two publicly available gene expression datasets to perform external validation when there are dozens of appropriate datasets. Moreover, they do not justify why these two were selected, raising the possibility that results are only shown for those experiments where validation was successful. While this is unlikely to be the case, there is no alternative explanation offered. Instead of justifying why these two, the authors should validate their 4-gene model in a much larger number of datasets. When additional datasets are chosen for validation, they should include subjects with NI etiologies.

Response: For full details including a table for reviewers, please see the response to reviewer one question 7.

We can understand that the reviewer is concerned about picking the external validation data sets that would be successful. This was certainly not our intent, but the majority of studies available are not ideal to validate our study (please see the table in the response to reviewer 1, #7). Initially, we initially used 2 studies that used RNAseq in adult populations with respiratory infections. Most of the other available data sets used microarray assays rather than RNAseq, were applied to pediatric populations, or did not include the desired infectious phenotypes (bacterial and nonbacterial viral infections). However in response to both reviewers, we have now performed 3 additional validations (2 adult studies and one pediatric study) using RNAseq data sets. We included the controls as requested although controls in 2 of the 3 studies with controls were healthy volunteers. The value of including “healthy” controls as opposed to non-infected controls seems minimal since it is likely a clinician can discriminate a healthy person from a sick one. Thus, we provide the validation results with and without the control groups. These results are presented in Table 2 which now replaces Supplemental Figure 2. Although we don't feel it is best to validate using pediatric studies, given the limited appropriate RNAseq data sets we included a pediatric study that was recently published (iScience 28, 111747, February 21, 2025). Interestingly, in the pediatric study one of our genes (FAM20A) was also included in their 5 gene predictor.

Added to the 2 prior validation sets we now provide validation results from 5 independent RNAseq data sets. We have added this information on lines 156-162.

“We also evaluated our gene signature in 2 additional adult data sets (GSE 163151: 6 bacterial and 20 viral) with AUC of 0.89 and (GSE161731: 24 bacterial and 78 viral) AUC of 0.98. If healthy controls were included the AUCs dropped to 0.82 and 0.97, respectively. Finally, we evaluated our 4-gene set in a pediatric pneumonia study (GSE261482) that included 98 definite bacterial and 12 definite viral infections resulting in an AUC of 0.74. If the healthy controls were included the AUC rose to 0.92.”

The difference in platform characteristics make the micro array data sets less ideal for validating RNAseq studies. We could not use the largest and most appropriate data set (GSE63990) because it did not contain the FAM20A gene which was important in our gene predictor set. We did run one validation on the next most appropriate micro array data set that included adults with respiratory infection (GSE60244). Our 4 gene predictor did discriminate bacterial and nonbacterial illness in the micro array data set (AUC 0.73) without the healthy controls and somewhat lower when including healthy controls (AUC 0.67); which were clearly were not as closely aligned with validation using RNAseq data. This information is included on lines 162-166.

“Despite the different platforms we also performed a validation using one micro array data set that included 154 adults with viral and bacterial respiratory infection as well as healthy controls (GSE60244). Our 4 gene predictor did discriminate bacterial from non-bacterial infections (AUC 0.73), but results did not align as closely as validations with an RNAseq data sets”.

When reporting results for these validation cohorts, they should include standard performance characteristics such as sensitivity (or PPA), specificity (or NPA).

Response: These data are now presented in Table 2.

Were there patterns or characteristics observed among subjects that were incorrectly classified, whether in the derivation cohort or the validation cohorts?

Response: The details of the bacterial cases misclassified as nonbacterial are provided in Supplemental Table 5 and the discussion of these results are provided on lines 182-190. Most were nonpneumonic mixed viral- bacterial cases. We were less concerned regarding those NB classified as B (specificity) since the default is usually to give antibiotics. There did not seem to be any striking features in these cases- although most were nonpneumonic (86%). Most were either rhinovirus (42%) or influenza A (22%) but these were the most commonly identified viruses. We added information describing the nonbacterial misclassified as bacteria on lines 190-192.

“Additionally, using this threshold only 1 of 10 mycoplasma pneumonia cases was misclassified as nonbacterial. The sensitivity of our 4- gene predictor was higher for more serious bacterial infection such as pneumonia and bacteremia at 93% and 100%, respectively. The majority of nonbacterial classified as bacterial using the 4-gene signature were either due to rhinovirus or influenza A and were nonpneumonic (86%) without distinguishing clinical features.”

On lines 149-152, the authors write, “Based upon prior feedback from clinicians regarding the acceptability of a diagnostic test to exclude bacterial cause of LRTI, we aimed to achieve at least 90% sensitivity while also providing >90% negative predictive value (NPV) for the determination of bacterial infection etiology in LRTI across a range of likely prevalences.” The authors should provide a reference or justification for this threshold that is more valid than anecdotal clinician feedback. It could easily be argued that a NPV of 95% or at least 98% is needed to drive clinical adoption. Without reference or justification, the use of >90% is arbitrary.

Response: We apologize for the lack of clarity regarding choosing this threshold since the reference was not included on lines 170-171. While it was not arbitrary and was based

on a survey of 97 hospital physicians, we agree with the reviewer that higher NPV may be needed to drive clinical adoption. But this would also likely depend on the patient population and severity of illness and will need to be evaluated in a future clinical trial. However, given the reviewer's comment, we now point out on lines 188-190 that the sensitivity of our 4- gene predictor was actually higher for more serious bacterial infection such as pneumonia and bacteremia for which it was 93% and 100% sensitive, respectively. The reference for the survey that we previously published (Falsey, A.R., Murata, Y. & Walsh, E.E. Impact of rapid diagnosis on management of adults hospitalized with influenza. Arch Intern Med 2007;167, 354-360) is now cited in the discussion. This survey result prompted our decision to use at least 90% sensitivity to guide our analysis on line 171.

The study identifies a gene signature threshold of -0.886 to discriminate bacterial from non-bacterial etiologies. Please explain where this number comes from and what it means. For example, on what scale does it exist; does it represent anything such as probability or deviation from baseline; is it dependent on the methodology used to measure gene expression (sequencing, microarray, PCR); why does it have a negative value. There is no context offered to understand where this number comes from and how one should interpret it.

Response: We now provide a hopefully better explanation of the threshold on lines 173-180. Our 4-gene signature is a linear combination of the expression levels of 4 selected genes and represents a risk score or estimated log(odds) of bacterial infection for a patient with the given expression levels. The higher the risk score, the more likely the patient is to have a bacterial infection. Classifying patients as bacterial if the risk score exceeds -0.886 (i.e. if the estimated probability of bacterial infection >29%) and nonbacterial if the risk score < -0.886 (i.e. if the estimated probability of bacterial infection <29%) resulted in 90% sensitivity and 71% specificity to detect any bacterial infection in our training data.

The study indicates that procalcitonin was measured and was significantly different between bacterial and non-bacterial groups. Since this biomarker is often used to discriminate between these two groups, the authors should report the performance characteristics of procalcitonin to discriminate bacterial from non-bacterial etiologies in this cohort.

Response: Thank you for this comment that is similar to the point made by reviewer 1 question 8. Since PCT was available in almost all cases and used by the clinicians to adjudicate cases we feel reporting the performance characteristics would be invalid due to circular reasoning.

The first two paragraphs are largely background material that reproduce what was in the Introduction. Consider significantly trimming down, though this is purely a suggestion.

Response: Thank you for this suggestion. We have trimmed the beginning of the discussion and substantially reorganized the discussion to accommodate the reviewers desire for a more thorough discussion of the literature and currently approved assays.

The authors claim in the Discussion on lines 186-188 that many biomarkers have shown limited success to supplement clinical judgement in assessing patients with ARI. First, while there is variability in the literature on this topic, there is an abundance of evidence to show exactly the opposite – that each of these biomarkers adds diagnostic value in assessing patients with ARI.

The authors also separately call out CRP, TRAIL and IL-10. Second, IL-10 is likely incorrect and should be IP-10. The former is not discussed in the referenced citations whereas IP-10 is. Third, the combination of CRP, TRAIL, and IP-10 comprises the MeMed BV test. There is ample evidence highlighting the test's performance and additive clinical value above and beyond standard of care. These points should all be mentioned.

Response: We apologize for the error and the reviewer is correct, IP-10 is the correct protein. This has been corrected. We agree with the reviewer that biomarkers such as PCT and the newer host response tests offer value when implemented and interpreted properly. We have provided more discussion on these points on lines 318-331. Yet, biomarkers and host response testing have not been widely adopted nor have they been recommended by the IDSA or ATS.

“Finally, two rapid host response-based tests are now commercially available. A rapid protein based assay which combines measurements of TNF-related apoptosis inducing ligand (TRAIL), IP-10 and CRP available as MeMed BV has demonstrated improved diagnostic performance than any protein measurement alone.³⁸ Studies of MeMed BV in children and adults presenting to emergency rooms or urgent care with fever or ARI suggests physicians will respond appropriately to the information provided.^{39,40} TriVerity®, a 29 gene-based predictor of bacterial sepsis (AUC 0.83), is FDA approved based on a large clinical study of suspected sepsis.^{45,46} These tests were compared in 169 adults with respiratory infection by Dedeoglu et al and both demonstrated similar overall accuracy.²⁷ However, more bacterial infections were missed by Triverity, whereas MeMed BV misclassified most viral infections as bacterial. Both tests provide a range of probability scores for the likelihood of bacterial or viral infection but with some subjects falling into an uncertainly range. The true utility of any host response test including ours will require prospective interventional trials adequately designed and powered to assess the impact on antibiotic use and clinical recovery.”

On line 220, the authors describe the TriVerity test. Note that it was incorrectly described as a ‘trial’ whereas it should be a ‘study’.

Response: Thank you for pointing out this error. We have corrected the term to be study as noted.

The authors should explain why they chose to focus on respiratory infection and not all types of infection. A bacterial vs. non-bacterial classifier has applications far beyond ARI. Even if they chose to do discovery in ARI, validation should include non-respiratory infections.

Response: We have refined our introduction to better explain our reasoning to focus on respiratory infections rather than all infections in the beginning of the discussion. Respiratory infections are a major cause of hospitalization in adults and range from asthma, acute and chronic bronchitis, and pneumonia. The majority of hospitalized adults receive antibiotics because of diagnostic uncertainty. Even with extensive testing the microbiologic etiology can be determined in only approximately 50%. While multiplex testing for viruses has shown nicely that many infections are due to viruses, their use has not effected significant change in antibiotic use due to provider’s concern about bacterial co-infections. We believe ARI is a very large group to focus efforts on, especially because even with viral pathogen specific testing diagnostic uncertainty persists. While it would be ideal to have one test that fits all infections, it seems likely the

host response to respiratory infection and non-respiratory infections may differ and a semi-focused approach can optimize assay accuracy.

While including so many bacterial-viral co-detections is a strength, the authors should expand on the challenges in determining whether these patients truly had co-infection or just co-detection. How likely was each pathogen to be causing active infection at the time of enrollment? I concede this is difficult if not impossible to determine, but it does raise questions about the certainty of the microbiological etiologies for these co-detection cases.

Response: The reviewer raises a legitimate concern regarding the accuracy of the diagnosis of co-infection since there is no gold standard and colonization with bacteria in the respiratory tract is well documented. Viruses are not known to colonize the airways but it is possible that detection reflected prolonged shedding from an antecedent viral infection and the primary pathogen at the time of hospitalization was a bacteria. Since viral bacterial (VB) infections were grouped with bacterial infection this should not have created any issue. In addition, we had rigorous adjudication of all cases that considered the microbiologic studies in the context of the clinical features of the illness. Of note, the VB cases more closely resembled B cases than V alone (Supplemental Table 2) Nonetheless, we concede some of the VB infections could have reflected viral infection with bacterial colonization. We mentioned the possibility of misclassification of VB on lines 275-277 in the discussion.

“Illnesses with both viral and bacterial detections may reflect bacterial colonization rather than true invasive infection.”

Line 334 refers to NS PCR. What does NS mean?

Response: NS refers to nasal swab. This is now defined on line 380.

Line 352 refers to a study by Kho et al. The author's name is Ko.

Response: Thank you for pointing out this error and we apologize for misspelling this author's name. It has been corrected.

For Figures 4 and supplemental Figure 1, please add labels for the x-axis. Also, the plots should ideally be made on the same scale for comparison purposes, including both x and y axes.

Response: The figures have been updated as suggested.

In the Methods, please provide more details about the software and tools used to perform the analyses.

Response: We appreciate the reviewer's concern and have incorporated information about the library packages and the platform used in analyzing the data in Methods section, subheadings Laboratory Methods and Statistical Methods on lines 382-415.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #4 (Remarks to the Author):

I read with interest the work from Falsey et al. In their work, the authors proposed a transcriptomics-based methodology for discriminating bacterial from non-bacterial infections.

In general, the work is scientifically sound, well written, and most of the conclusions are warranted by the presented results. Nevertheless, I have a few points that could be addressed in a next iteration of the manuscript.

Response: We appreciate the reviewer's positive comments

1. I understand the authors obtained a subset of important features, i.e., the genes ITGA7, IFI27, FAM20A, and ITGB4 during Lasso selection. The downstream results the authors derived are only based on the expression level of the four genes initially obtained. I wonder if the authors can, for control purposes, obtain (a) set(s) of genes with coefficient = 0 and use these in their predictions. In doing this, the authors could evaluate any potential bias of Lasso, confirm the importance of their set of genes, and check for any potential missed gene that might have been over penalized.

Response: Adding more genes (5+) would not substantially improve performance, or else more genes would have been selected; e.g. before we employed hard-thresholding of small LASSO-constrained coefficients to 0, the Relaxed LASSO procedure selected 9 genes and achieved the same AUC = 0.90, whereas standard (unrelaxed) LASSO would select many more genes without improving (and possibly degrading) performance. That said, our 4 selected genes are not necessarily uniquely able to discriminate between bacterial and non-bacterial infection, as we identified 5,401 differentially expressed genes (see Figure 2). Critically, as a set, they demonstrate the best performance for discriminating between bacterial and non-bacterial infections. We tried excluding the 4 selected genes and re-running the procedure on the remaining genes; the resulting disjoint set of selected genes performed a little (but not a lot) worse, with AUC in the high 80's.

2. Once obtained their genes of interest, the authors used them in the relaxation step of Lasso to classify in any bacterial infection. The authors established a threshold based on clinician inputs. I understand Lasso-based models have less complex decision boundaries which is useful for clinical environments but Lasso assumes linear relationships between the genes and the classes. It also tends to rule out genes that are correlated (it chooses one and ignores the rest).

Have the authors considered other models for classification that can handle nonlinear relationship and interaction between features?

Response: The short answer is, no, we did not apply multiple classification approaches, and we restricted ourselves to variants of the procedure reported here (with and without hard-thresholding, with and without restricting the range of the relaxation parameter, with and without standardizing predictors to have unit variance, etc.). While methods such as random forests implicitly allow for (and, in fact, force) interactions between features, we did not apply such methods here.

We appreciate the opportunity to address the reviewers' time and consideration in their review of our manuscript. Below is a point-by-point response to their specific comments and concerns.

Reviewer #1 (Remarks to the Author):

First, a quick clarification: re: 24h, I am not suggesting 'later' timepoints, I am suggesting 'earlier' timepoints, e.g. the first 60 min, first 3 hours, etc. 24h after admission is potentially quite a late timepoint here.

Thank you for your clarification. We cannot change the study design but would suggest that a test within 24 hours will still provide meaningful data for which providers can adjust their antibiotic management decisions. In addition, since patients have been usually symptomatic for a range of days prior to admission (anywhere from 1-14 days), the collection of samples within 24 hours of admission should not affect the interpretation of the gene expression results.

Overall, my concern about impact remains and is perhaps magnified by the response, for the following reasons:

-The authors acknowledge severity as a cofounder but make no attempt to quantify it

We would like to correct our error in stating in our previously revised manuscript that "severity may be a confounding factor in development of our gene signature." Our data show associations between bacterial infection and measures of severe illness, which is unsurprising, but that itself does not imply confounding. Indeed, the notion of confounding pertains to measures of association, such as odds ratios between predictors (e.g. gene expression levels) and a binary outcome (e.g. bacterial vs non-bacterial infection), not to a classifier. While measures of association such as odds ratios can be adjusted for a potential confounder by including it in the model, there is no such thing as confounding or adjustment of a classifier. In the case of disease severity, a classifier can either ignore it (as ours and others do) or utilize it (which could improve apparent performance, possibly via bias). As previously noted, we chose not to include any clinical predictors such as disease severity in our classifier since they were all considered by the adjudicators in defining the outcome, and thus their inclusion could result in optimistic bias in classification performance. This sort of bias could even impact the results of external validation, e.g. if any measures of disease severity were also considered in defining the outcome in the validation data. We have removed our incorrect reference to possible confounding in the limitations paragraph of the discussion. Semantics aside, if the concern is that our gene signature might relate not only to bacterial infection but also to disease severity, we do not believe that it would pose any issue for the classifier of bacterial infection. Indeed, if bacterial infections tend to be present among those with severe disease (whether or not the relationship is causal), as appears to be the case, then any good classifier of bacterial infection should be expected to also be at least somewhat related to disease severity.

-Agree the LOBO-CV is not *very* biased, but anything reported from the discovery dataset is not the same as external validation

We appreciate the reviewer's recognition of the value of both LOBO-CV-based internal validation and fully independent external validation, both of which we provide. We further

agree with the sentiment that performance in external validation should generally be expected to be somewhat lower than that in internal validation. However, this is not because LOBO-CV is optimistically biased (it is not) but rather because external validation is actually pessimistically biased due to the many potential differences in the data generating mechanisms.

The notion of generalizability is at the heart of these differences, and we agree it can be informative to investigate the degree of generalizability by attempting external validation using data sets with some known differences from the training data. For this reason, we expanded our set of external validation results to include some examples of microarray-based gene expression data (even one missing one of our selected genes), pediatric patients, and non-respiratory bacterial infections. Greater differences between the validation data and the training data lead to lower expected classification performance. When external validation results are favorable, despite marked differences from the training data, generalizability is high. However, when external validation results in substantially lower AUC, this can often be attributed to a lack of generalizability due to the test data being quite different from that used to develop the classifier.

-The authors continue to substantially minimize the number of datasets tested. See e.g. Chawla, Cell Systems 2022, or Bodkin, Genome Med 2022. There are dozens and dozens of possible datasets. The whole point of a locked classifier is to be externally validatable, meaning even a dataset with only bacterial samples can be tested, because it's not 'relative differences' but rather a fixed classifier with a fixed cutoff. The RNAseq limitation is only partially convincing because RNAseq cannot be made into a bedside tool. So either the signature is not translatable to practice, or the studies are not being fully utilized. Plus, the one 'older' study used (GSE60244) was used in the Bodkin paper and *most* of those gene signatures had a higher apparent performance than the present one.

In this second revised manuscript, we have included validation in 2 additional populations, bringing the total number of replication populations to 8. These two new replication populations include (1) a microarray data (GSE63990) set that we previously chose not to use because it lacked data for one of our signature genes (FAM20A), but here we imputed its expression to enable an analysis of this cohort (AUC of 0.90), and (2) a new RNAseq study (GSE282464), published this month. In this latter data set we performed validation without early pandemic COVID cases, which is more similar to our data, as well as validation including early pandemic, likely primary COVID cases; our 4-gene signature performed well with AUC of 0.89 and 0.84, respectively. These data are included in the revised Table 2.

In summary, we provide data indicating that this 4-gene set has excellent discriminative value in a variety of populations; in adult LRTI populations when gene expression is measured by RNAseq (Phs001248.v1.p11, GSE161731, GSE163151, and now also GSE282464), in pediatric LRTI populations when expression is measured by RNAseq (GSE261482), in adult populations with extra-pulmonary infection when measured with RNAseq (GSE211567), and in adult LRTI populations when measured by microarray (GSE60244 and GSE63990). The value of this 4-gene set appears somewhat reduced when using microarray data, as expected, likely because of differences in measurement approaches and their respective dynamic range.

We agree with the reviewer that RNAseq is not currently, and is unlikely in the near future to be, a bedside diagnostic tool. Therefore, the concept of a fixed classifier is somewhat premature. A reduction in performance across gene expression measurement modalities (e.g., microarray, RT-PCR, RNAseq, etc.) has been noted and acknowledged before by other investigators. Specifically, Lydon et al 2019 stated

...“Due to technical differences between microarray measurements and RT-PCR, we could not apply the microarray-based model to this new data. We therefore trained a new model on RT-PCR data”...

In consideration of our response to the reviewer’s prior comment (above), it is not surprising that our classifier performance is more variable in microarray data sets (GSE60244 and GSE63990). Even so, performance is excellent across cohorts with respect to age and measurement modality. It is our expectation, and current plan, that a more appropriate point-of-care assay such a qPCR will need to be recalibrated based upon the sensitivity and dynamic range of measurements for that assay.

As for the Bodkin paper, their definition of a “signature” was a list of genes, with no associated model, coefficients, or classification rule; to “validate” each list, they then went on to fit LASSO-penalized logistic regression models each to n-1 patients from each “validation” data set, thus estimating n sets of coefficients for selected subsets of the genes in the “signature,” and they employed nested Leave-One-Out Cross-Validation (LOO-CV) to estimate the resulting AUC. There were no fixed classifiers, and no truly external validation was performed. This partly explains the generally high apparent performance, as all validation was largely internal. They report the distribution of these LOO-CV-AUC over the 51 validation data sets; it’s impossible to determine which points correspond with GSE60244, but some $AUC \leq 0.5$ (corresponding with absolutely no discrimination power whatsoever).

The other platforms mentioned (TriVerity and MeMed) have both had publications specifically addressing bacterial-viral coinfections. The fact is the current signature does not show any particular utility beyond currently fielded platforms, and is several years behind.

We have reviewed the literature quite thoroughly and suggest to the reviewer that large, detailed studies including coinfection are limited. Many publications do not provide details on numbers or performance for co-infection, are relatively small or utilize large numbers of classifier genes.

I reiterate that I find the work to be reasonably technically sound, but of only low-moderate impact.

We appreciate the reviewer’s comment regarding the soundness of our work, however we continue to argue that the strength of this work and the advances it may offer to the field are the well-validated parsimonious gene set that will speed translation to a bedside tool, and the application to a very large and vexing common subgroup of patient situations (i.e., viral bacterial-coinfected patients). There remains great opportunity to increase the number of individuals with viral-bacterial co-infection and thereby improve diagnostic classifiers. Our study, developing a classifier to identify bacterial infection in both the absence (n=129) or presence (n=95) of viral co-infection, specifically helps to address this knowledge gap.

Reviewer #2 (Remarks to the Author):

The authors have made many changes to address concerns raised in the initial review. However, some concerns remain.

The authors declined to report performance characteristics of PCT because it was available to adjudicators, and this raised concerns about circular reasoning. I disagree with this justification. It is possible that PCT performance could be overestimated due to incorporation bias (rather than circular reasoning). However, this analysis can still be performed with that limitation as is the case for many other publications using PCT, CRP, or WBC as comparators. Poor performance of PCT despite incorporation bias would strengthen the argument that it is insufficient as a biomarker. And if PCT performance was very good, then it may very well be due to incorporation bias and acknowledged as such in the discussion. Moreover, adjudicators were not bound to PCT results and incorporated all clinical information in making their diagnostic assignments. I therefore maintain concerns that not reporting PCT performance is a weakness in the manuscript.

We maintain that incorporation bias is a serious flaw and that is why we previously provided the reviewer with the AUC of PCT and WBC in the last response but chose not to include this information in the manuscript. However, since the reviewer feels strongly about this topic, we now include this data in Supplementary Figure 2 and on lines 145-149 with a statement regarding the significant limitations interpreting these data due to incorporation bias.

Regarding the focus on ARI as opposed to suspected infection more broadly, the authors suggest that "While it would be ideal to have one test that fits all infections, it seems likely the host response to respiratory infection and non-respiratory infections may differ and a semi-focused approach can optimize assay accuracy." The authors do not offer any evidence to support this claim. In contrast, there is an abundance of evidence for the opposite - that the host response to bacterial or viral infection is conserved across anatomic sites of infection. I recommend the authors validate their signature in datasets that include all infectious syndromes to support external validity and generalization. Even if it was derived in ARI, it is critical to understand how the signature performs in other syndromes.

In the initial and 1st revision we included an RNAseq data set from a broad patient population (Ko et al Scientific Reports 2023; GSE211567) enrolled globally with a diverse set of conditions (not just ARI) and the gene set performed very well with AUC of 0.89, supporting the external validity and generalizability.

However, regarding the need for a test to be used for all infectious syndromes, we think a more focused approach is more likely to change practice. It could be argued that a clinician can identify certain types of infections quite well (urinary tract, abdominal and soft tissue infection, hemodynamically unstable sepsis) that invariably are caused primarily by bacterial pathogens and antibiotics should be prescribed. Testing these patients will not change management and thus the value of testing all suspected infections is unclear, especially in an era of antibiotic and diagnostic testing stewardship. Whereas in more than 50% of respiratory infections the diagnosis is uncertain since viral, bacterial, and co-infection are all common scenarios.

With respect to impact, the manuscript offers incremental advances in our scientific knowledge in this field. The authors cite the inclusion of coinfection as the major advance. In a cursory search of PubMed, I identified eight pre-existing publications that included coinfection. That is not to say the current study isn't a useful addition. But it remains unclear how the current discoveries advance the field, especially given the narrow focus on LRTI in adults (as noted above).

That list, if interested, includes these titles:

- 1-Discriminating Bacterial and Viral Infection Using a Rapid Host Gene Expression Test
- 2-Longitudinal Dynamics of a Blood Transcriptomic Signature of Tuberculosis
- 3-Comparison of two rapid host-response tests for distinguishing bacterial and viral infection in adults with acute respiratory infection
- 4-BV score differentiates viral from bacterial-viral co-infection in adenovirus PCR positive children
- 5-Detection of Viral Infection and Bacterial Coinfection and Superinfection in Coronavirus Disease 2019 Patients Presenting to the Emergency Department Using the 29-mRNA Host Response Classifier IMX-BVN-3: A Multicenter Study
- 6-Validation of a host response test to distinguish bacterial and viral respiratory infection
- 7-Host gene expression classifiers diagnose acute respiratory illness etiology
- 8-A distinct influenza infection signature in the blood transcriptome of patients with severe community-acquired pneumonia Bad example-

We thank the reviewer for providing this list of references. We would refer to our response to Reviewer 1 (comment 3) on this topic. We are aware of most of these studies and have referenced those that provide data that are useful for interpreting our results in our prior submission. When read in detail, studies 2, 3, 4, 5 and 8 either do not apply to our population, have problematic methodology to define coinfection, or indicate less than ideal discrimination. Papers 1, 6 and 7 provide details of co-infected cases 36 [12 confirmed], 53 [19 confirmed] and 26 confirmed, respectively, but require large numbers of probes for classification. In our study we used 95 confirmed co-infections and a parsimonious classifier that we feel addresses a significant knowledge gap.

Regarding the narrow focus, we must respectfully disagree. Health Care Cost Utilization Project (HCUP) data indicates pneumonia, COPD and bronchiectasis and respiratory insufficiency are among the top 15 principal diagnoses for nonmaternal, non-neonatal hospital stays resulting in substantial morbidity and mortality and billions of health care costs each year in the US. In addition, respiratory infection is one of the most common diseases where both viral and bacterial infection are common and diagnostic uncertainty remains a huge problem.

<https://hcup-us.ahrq.gov/reports/statbriefs/sb277-Top-Reasons-Hospital-Stays-2018.pdf>

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We thank the reviewers and appreciate them taking time to review our manuscript.

Response to Review

Reviewer #1 (Remarks to the Author):

-The authors write, “In addition, since patients have been usually symptomatic for a range of days prior to admission (anywhere from 1-14 days), the collection of samples within 24 hours of admission should not affect the interpretation of the gene expression results.” But this is apples-to-oranges, patients aren’t on IV antibiotics at home! We do see changes in gene expression within 24 hours of IV antibiotics. The study isn’t changeable, and the choice was pragmatic, but it still lowers the impact and external validity of the findings.

We do not disagree that gene expression may change over time and with antibiotic use and yet to be a useful clinical tool we maintain that sample collection within 24 hours of admission is reasonable and practical. We respectfully disagree that this lowers the impact or the external validity.

-The authors write that “there is no such thing as confounding of a classifier” (!). This is an absurd thing to say. The authors are claiming that classifiers can’t show spurious associations between dependent and predictor variables through the presence of extraneous variables? Classifiers can’t be biased?

-The authors then write that ‘Semantics aside, if the concern is that our gene signature might relate not only to bacterial infection but also to disease severity, we do not believe that it would pose any issue for the classifier of bacterial infection. Indeed, if bacterial infections tend to be present among those with severe disease (whether or not the relationship is causal), as appears to be the case, then any good classifier of bacterial infection should be expected to also be at least somewhat related to disease severity.’

The authors are surely aware of causes of severity other than bacterial infection, such that severe non-infectious illnesses will be incorrectly classified as bacterial? Or that mild bacterial infections are incorrectly classified as non-bacterial? It’s one thing to say ‘this is a weakness and here’s how it affects interpretation’ – fine. It’s another thing to deny the weakness itself. Fact is, there are ways to build classifiers that correct for association with severity without explicitly including a severity variable. The authors’ apparently limited knowledge in this arena does not make the concern incorrect.

The classifier accurately discriminates between bacterial and nonbacterial illness and we are reassured by the 8 external validations which provide illnesses with a range of severity. In addition, the 4 gene score discriminated bacterial and nonbacterial with the non-infected controls when were included as nonbacterial, some of whom were quite ill. Nonetheless, we agree that there is an association between surrogates of severe illness and bacterial infection. Since this information was available to adjudicators and likely contributed categorizing

some cases as either viral or bacterial, we excluded surrogates of disease severity from our model to avoid optimistically biasing performance. However, we now acknowledge in the limitations section that our gene predictor may relate not only to bacterial infection but also to disease severity and thus may be a potential confounder of the association between genes and bacterial infection.

-“...external validation is actually pessimistically biased” – Yes, it’s harder and more representative of real-world validation. I believe what the authors mean is that between two sources of potential change, they are stating that technical platform difference outweighs clinical difference, so negative result should be ignored. I’d encourage them to study the MAQC – technical differences are often swamped by biological differences. Negative results aren’t just ‘microarrays are different’.

-“ In this second revised manuscript, we have included validation in 2 additional populations” (one of which they imputed a gene). The issue is not to include 2 more populations, it’s to understand why those two more cohorts were chosen out of dozens available. Were they the ones with good scores? This is a top-tier journal. Validation datasets should be approached in a systematic fashion, and that systematic approach is critical to showing that the validation results are meaningful. This is the “File Drawer Problem”.

We appreciate the reviewers’ comment about the need to explain how the validation cohorts were selected and now include this information in the methods. We did not cherry pick those with good scores and we present all the validations that were performed. We had clear logic for our validation data set selections. Six RNAseq data sets were selected for validation which represented ALL RNAseq data sets that were appropriate since they included both viral and bacterial infections (5 adult and 1 pediatric). Of the 50 Microarray data sets available, only 20 included both viral and bacterial diagnoses (14 pediatric and 6 adult). Because we continue to feel that validation with a different platform such as microarray is suboptimal, we chose to focus on the 6 adult studies. Four studies were rejected because of sample size < 10, no publication associated with the data, full data set was unavailable, or longitudinal sampling, leaving two studies of adult ARI for validation.

-The abstract still lists an AUROC range lower end of 0.89 instead of 0.73.

As written, the abstract was technically correct and believe the reviewer missed that we refer to only adult RNAseq validations. We did this to stay within the word limit. We have separated the types of validation now and state “The 4-gene signature is validated in five independent adult RNAseq cohorts (AUC = 0.89 - 0.98), two adult microarray cohorts (AUC = 0.73-0.90), and one pediatric

pneumonia RNAseq cohort (AUC 0.74)”for improved clarity. The abstract is now slightly over the abstract word limit.

-The authors write ‘Therefore, the concept of a fixed classifier is somewhat premature’. They then cite the Lydon paper that proves exactly my point – without a fixed classifier or a POC platform, there has been NO further development from the Duke group 6 years out (Lydon et al), showing the relatively low impact of presenting a classifier that no one will actually ever use.

Reviewer #2 (Remarks to the Author):

The authors have responded to all comments presented in my prior review. Despite those responses, I respectfully disagree that this study and accompanying manuscript advance the field. The study is technically sound and should be published. But it fails to be of significance to the field and largely represents a me-too study in the established literature in the opinion of this reviewer.

We are sorry the reviewer feels this way but we respectfully disagree.