CALL FOR PAPERS | Computational Modeling of Physiological Systems

Discovery of the gene signature for acute lung injury in patients with sepsis

Judie A. Howrylak,1 Tamas Dolinay,2 Lorrie Lucht,1 Zhaoxi Wang,3 David C. Christiani,3,4 Jigme M. Sethi,1 Eric P. Xing,2 Michael P. Donahoe,1 and Augustine M. K. Choi2

1Division of Pulmonary, Allergy and Critical Care Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania; 2Division of Pulmonary and Critical Care Medicine, Brigham and Women’s Hospital; 3Department of Environmental Health, Harvard School of Public Health; 4Pulmonary and Critical Care Unit, Massachusetts General Hospital, Boston, Massachusetts; and 5Department of Machine Learning, Carnegie Mellon University, Pittsburgh, Pennsylvania

Submitted 18 June 2008; accepted in final form 16 January 2009

Howrylak JA, Dolinay T, Lucht L, Wang Z, Christiani DC, Sethi JM, Xing EP, Donahoe MP, Choi AM. Discovery of the gene signature for acute lung injury in patients with sepsis. Physiol Genomics 37: 133–139, 2009. First published January 27, 2009; doi:10.1152/physiolgenomics.90275.2008.—The acute respiratory distress syndrome (ARDS)/acute lung injury (ALI) was described 30 yr ago, yet making a definitive diagnosis remains difficult. The identification of biomarkers obtained from peripheral blood could provide additional noninvasive means for diagnosis. To identify gene expression profiles that may be used to classify patients with ALI, 13 patients with ALI + sepsis and 20 patients with sepsis alone were recruited from the Medical Intensive Care Unit of the University of Pittsburgh Medical Center, and microarrays were performed on peripheral blood samples. Several classification algorithms were used to develop a gene signature for ALI from gene expression profiles. This signature was validated in an independently obtained set of patients with ALI + sepsis (n = 8) and sepsis alone (n = 1). An eight-gene expression profile was found to be associated with ALI. Internal validation found that the gene signature was able to distinguish patients with ALI + sepsis from patients with sepsis alone with 100% accuracy, corresponding to a sensitivity of 100%, a specificity of 100%, a positive predictive value of 100%, and a negative predictive value of 100%. In the independently obtained external validation set, the gene signature was able to distinguish patients with ALI + sepsis from patients with sepsis alone with 88.9% accuracy. The use of classification models to develop a gene signature from gene expression profiles provides a novel and accurate approach for classifying patients with ALI.

ACUTE LUNG INJURY (ALI) and the acute respiratory distress syndrome (ARDS) were initially described in 1967 (3). However, accurate early diagnosis of patients with this syndrome remains difficult. In 1992, the American-European Consensus Conference (AECC) on ARDS was formed in an effort to develop well-defined clinical criteria for diagnosis (4). The clinical criteria developed at this conference are currently in use, and these syndromes are characterized by acute onset, radiographic evidence of diffuse alveolar damage in the absence of left atrial hypertension, and severe hypoxemia.

However, despite the development of clinical criteria for the diagnosis of ARDS, there remains both a discrepancy between clinical criteria and histological autopsy findings (8), and a high rate of intraobserver variability in diagnosis that makes it particularly difficult to select appropriate patients for ongoing clinical trials of therapies for ARDS/ALI (26). In addition, it is difficult to identify particular subgroups of patients who may benefit from more specialized treatment. Due to these difficulties with diagnosis, there have been many recent efforts to identify biologic markers for ARDS/ALI in critically ill patients, including studies of pulmonary edema fluid (1, 6, 7, 14, 18, 27), blood and urine (5, 11, 12, 15).

The development of microarray technology has also led to the identification of several gene expression markers of potential diagnostic and prognostic significance (13, 29). More recently, the emerging field of genomics has provided an additional methodology for developing classification models to identify biologic markers of diagnostic and prognostic significance in pulmonary and critical care medicine (17, 19, 21).

This study represents an attempt to use a genomic classification model to enhance the clinical criteria for the diagnosis of ARDS/ALI. Because sepsis is the condition with the highest risk of progression to ARDS/ALI (16, 28), we used gene expression analysis to identify molecular classifiers of ARDS/ALI in a cohort of patients with sepsis. We grouped patients based on their clinical presentation as sepsis alone or sepsis + ALI/ARDS and compared gene expression profiles in each group in an effort to identify unique molecular classifiers able to differentiate the two groups. We believe this is the first study to use a genomic classification strategy to attempt to distinguish patients with ALI + sepsis from patients with sepsis alone based upon plasma gene expression patterns.

MATERIALS AND METHODS

Patient Recruitment

Patients from two independently obtained cohorts were used for this analysis. For the training cohort, a retrospective observational study of sepsis and ARDS/ALI was reviewed and approved by the University of Pittsburgh Institutional Review Board. Patients were recruited from the Medical Intensive Care Unit (MICU) of the University of Pittsburgh Montefiore NW 628, 3459 5th Ave., Pittsburgh, PA 15213 (e-mail: howrylakj@upmc.edu).

The costs of publication of this article were defrayed in part by the payment of page charges. The article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
University of Pittsburgh Medical Center between February 2005 and June 2007. Patients admitted to the MICU for 48 h or less who were intubated and receiving mechanical ventilation were considered eligible for the study. Patients were classified as having both sepsis and ALI (n = 13) if they met the criteria for sepsis as defined by the Society of Critical Care Medicine Consensus statement (2), and the criteria for ALI as defined by the AECC on ARDS (4).

Specifically, patients were classified as having ALI if they had an acute decrease in the ratio of partial pressure of arterial oxygen to fraction of inspired oxygen to 300 or less (indicating the onset of hypoxemia), bilateral pulmonary infiltrates on chest radiograph consistent with the presence of edema, and no clinical evidence of left atrial hypertension. Patients were classified as having sepsis only (n = 20) if they met only the above criteria for sepsis. In addition, to eliminate ambiguity between septic patients with bilateral pneumonia and septic patients with ALI, patients with sepsis and bilateral infiltrates who did not meet the AECC criteria for ALI were not included in the study. The initial patient screening was conducted by the study coordinator, who subsequently confirmed whether a patient met the criteria of either ALI + sepsis or sepsis alone with both the primary author of this paper, and one of the senior authors of this paper, a board-certified Pulmonary/Critical Care Medicine physician. At the time of enrollment in the study, blood was collected from each patient into PAXgene Blood RNA tubes (PreAnalytiX, Valencia, CA) for RNA preservation and was stored at −70 degrees. None of the study patients received bronchoalveolar lavage prior to the collection of blood.

For the validation cohort, the results from microarray analysis were obtained from a recently published study involving patients with ALI + sepsis (n = 8), and sepsis alone (n = 1) (25). The patient population was similar to that in the training cohort, because the patients were clinically ill and were recruited within 24 h of admission to the ICU. In addition, all patients were intubated and receiving mechanical ventilation (9). In a similar fashion, blood samples were collected from each patient, and RNA was extracted and used for subsequent microarray analysis. Details regarding RNA extraction and microarray hybridization of this cohort may be obtained from the original article.

RNA Extraction and Microarray Hybridization

RNA extraction on the training cohort was performed using the PAXgene Blood RNA kit (Qiagen, Valencia, CA), and RNA quality was assessed so that only nondegraded RNA samples were used for microarray and reverse transcription-PCR (RT-PCR) experiments. No RNA amplification step was performed prior to hybridization. Microarray hybridization was performed using the GeneChip Human Genome U133A 2.0 Array (Affymetrix, Santa Clara, CA). One patient sample was placed on each chip.

RT-PCR Validation Studies

Expression of selected transcripts (cpl1, kip2) was confirmed by real-time RT-PCR using TaqMan Gene Expression Assays (Applied Biosystems, Foster City, CA). Detailed methods and ABI gene expression IDs are available as supplementary material.

RESULTS

Patient Characteristics

Table 1 lists the demographic and clinical characteristics of the patients used to develop and train the diagnostic classification model. The group of patients with ALI + sepsis is similar to the group of patients with sepsis alone in terms of sex, age, race, APACHE II score, and mortality. Table 2 lists the demographic and clinical characteristics of the patients in the independent cohort used to validate the classification model and is similar in terms of sex and age.

Classification of Gene-Expression Profiles to Improve Diagnosis

The goal of developing classification models is to develop multivariate predictors that will allow for the determination of whether patients with sepsis have ALI or not. The multivariate predictors developed by the classification models will consist of those genes that make up the “genetic signature” for ALI. The genetic signature is composed of those genes that allow the classification model to accurately partition the septic patients into two subgroups, those with ALI and those without.

The classification model and genetic signature were developed in a series of stages. First, a classification model was developed by identifying the gene signature, or genes that allow for the differentiation of patients with ALI + sepsis from patients with sepsis alone. Next, the gene signature was cross-validated within the training set. Finally, the gene signature was used to predict the disease category of an independent data set. Figure 1 depicts a flowchart of this procedure. In addition, each of the aforementioned steps are described in greater detail in the following paragraphs.

1 The online version of this article contains supplemental material.
Classifier training and gene selection. Classification was performed using both SVM and LDA described above on the normalized set of genes. Leave-one-out cross-validation served as a means of internal validation. To decrease the number of genes used to determine the genetic signature for ALI patients, the top 100 genes were selected by SAM score. Then, the selected genes were sorted by decreasing order of SAM score, and this result was used for subsequent greedy forward gene selection.

For the set of 100 genes that were ranked by SAM score, a series of iterations was performed to identify the optimal set of genes for disease classification. For each iteration, starting with the top gene by value of score, one additional gene was added by decreasing order of SAM score. Similarly, the classification error rate was calculated, and the optimal gene signature was identified as the set of genes that resulted in the lowest error rate.

**Table 1. Baseline characteristics for patients in training group**

<table>
<thead>
<tr>
<th>Table 2. Baseline characteristics for patients in external validation group</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ALI</strong></td>
</tr>
<tr>
<td>Patients, n</td>
</tr>
<tr>
<td>Sex (%)</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Age, yr</td>
</tr>
<tr>
<td>APACHE III</td>
</tr>
</tbody>
</table>

Plus/minus values are means ± SE. ALI, acute lung injury; APACHE, Acute Physiology and Chronic Health Evaluation score; COPD, chronic obstructive pulmonary disease; PEEP, positive end-expiratory pressure; BAL, bronchoalveolar lavage; CSF, cerebrospinal fluid.

**Classifier training and gene selection.** Classification was performed using both SVM and LDA described above on the normalized set of genes. Leave-one-out cross-validation served as a means of internal validation. To decrease the number of genes used to determine the genetic signature for ALI patients, the top 100 genes were selected by SAM score. Then, the selected genes were sorted by decreasing order of SAM score, and this result was used for subsequent greedy forward gene selection.

For the set of 100 genes that were ranked by SAM score, a series of iterations was performed to identify the optimal set of genes for disease classification. For each iteration, starting with the top gene by value of score, one additional gene was added by decreasing order of SAM score. Similarly, the classification error rate was calculated, and the optimal gene signature was identified as the set of genes that resulted in the lowest error rate.

**Fig. 1.** Graphical depiction of the algorithm used to develop a gene signature to distinguish patients with acute lung injury (ALI) + sepsis from patients with sepsis alone. SAM, Statistical Analysis of Microarrays; LOOCV, leave-one-out cross-validation.
classification error rate. Thus, starting out with just one gene as a variable for the classification model, with each iteration one additional gene was added as an additional variable to determine the optimal gene signature for each classification model. Figure 2 depicts the error rate of each classifier during the process of gene selection.

It may be observed from Fig. 2 how the process of gene selection improves the overall accuracy of each classifier. For example, for the LDA classifier, when all 100 genes are used for classification, the accuracy of the classifier is only 87.9%. However, after the process of gene selection, when only the top eight genes by SAM score were used for classification, the accuracy of LDA increased to 97.0%. Results for the classification models after the process of gene selection are reported in Table 3.

The highest classification accuracy was achieved using the SVM classifier with a set of eight genes. It can be seen in Fig. 2 that for the LDA classifier, during the process of gene selection, the error rate initially decreased with the addition of each gene. The error rate reached its lowest value with the addition of the top seven genes by SAM d-score. After this, the addition of genes did not cause the error rate to increase any further. In fact, often the addition of extra genes caused the error rate to decrease, presumably due to the fact that the addition of further genes did not contribute any more information allowing the classifier to discriminate between the two patient groups. Conversely, the error rate of the SVM classifier remained relatively constant, even after the addition of extra genes.

The eight genes used to achieve this optimal classification were, by decreasing order of SAM score, ferritin, heavy polypeptide 2, ADP-robosylation factor 3, BTG family, member 2, NAD(P)H dehydrogenase, quinone 2, cyclin-dependent kinase inhibitor 1A (p21, cip1), patatin-like phospholipase domain containing 2, aminopeptidase-like 1, and CREB/ATF bZIP transcription factor. A heat map depicting differential expression levels of each of the eight genes making up the gene signature is presented in Fig. 3.

External validation of classifier. Following optimization of the classification model and gene signature, the best performing classifier, which was SVM with a gene signature of the top eight genes by SAM score, was used to classify the patients in an independently recruited cohort of patients with either ALI + sepsis or sepsis alone (25). The demographic and clinical characteristics of both sets of independently recruited patients are similar, as may be observed by comparing Tables 1 and 2, which contain the training and validation cohorts of patients, respectively. Results of external validation are reported in Table 4. As reported in Table 4, the overall accuracy of the optimal classifier in distinguishing patients with ALI + sepsis from patients with sepsis alone decreased slightly from 100% accuracy in the training cohort to 88.89% accuracy in the validation cohort. This decrease in accuracy indicates that there was a small degree of overfitting between the optimal classifier and the training cohort, although the accuracy of the classifier in distinguishing between patients in the validation cohort is still quite high.

Confirmation by RT-PCR

Expression of selected transcripts (Il1b, cip1, kip2) was confirmed by RT-PCR. The Wilcoxon-Mann-Whitney test was used to compare expression values of RNA from patients with ALI + sepsis relative to patients with sepsis alone. The P values for these comparisons were 0.05 for cip1 and 0.039 for kip2. Results of RT-PCR are presented in Fig. 4.

**DISCUSSION**

This analysis represents the first attempt to use gene expression profiles along with classification models to develop a genetic signature for ALI. Our results found that when gene selection was performed to identify an optimal number of genes, the performance of both the SVM and the LDA classifiers were comparable. The classification accuracy of LDA was 97% using a gene signature composed of the top seven genes by SAM score, and the classification of SVM was 100% using a gene signature composed of the top eight genes by SAM score. Thus, it appears that the top seven to eight genes that are differentially expressed between patients with ALI + sepsis
and sepsis alone have the ability to successfully classify almost all of the study subjects, regardless of the specific type of classification algorithm used. This fact suggests that successful classification based on a gene signature containing the top seven to eight genes reflects the underlying pathophysiology and transcends the particular statistical method used for the analysis.

The gene with the highest SAM score that alone was able to classify the study subjects with 88.8% accuracy was ferritin, heavy polypeptide 1. This gene was found to be upregulated in patients with ALI/sepsis relative to patients with sepsis alone. It has been reported previously that this gene is induced by pro-oxidants as part of the cellular anti-oxidant response to protect cells from oxidative stress (23). Thus, the fact that this gene is upregulated in patients with ALI/sepsis relative to patients with sepsis alone is consistent with the function of this gene as an antioxidant, since patients with ALI + sepsis are likely to have an even higher degree of oxidative stress than patients with sepsis alone.

Although there have been other human studies that have used gene expression profiling to uncover candidate genes involved in ALI (13, 30), the search for a gene signature has been difficult for several reasons. One reason is due to the difficulties associated with obtaining RNA samples from critically ill patients with ALI. Previous gene expression profiling studies had a much smaller number of patient samples, which increased the difficulty in generalizing the study results to a larger patient population.

One potential limitation of this study is the relatively small sample size of patients with ALI + sepsis (n = 13) and patients with sepsis alone (n = 20). Although this sample size is larger than other gene expression profiling studies of patients with ALI, it is a small relative to other clinical studies that have developed classification models for pulmonary diseases (17, 19, 21). When using classification algorithms with smaller samples sizes to detect diagnostic markers there is a risk of overfitting, which occurs when a gene expression profile is a

---

### Table 4. Diagnostic performance of genetic signature on internal and external validation sets

<table>
<thead>
<tr>
<th>Validation</th>
<th>Internal, %</th>
<th>External, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Specificity</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>100</td>
<td>88.89</td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Accuracy*</td>
<td>100</td>
<td>88.89</td>
</tr>
</tbody>
</table>

*Accuracy refers to the rate of correct predictions and is equal to the sum of true positive and true negative observations divided by the total number of observations.

---

Fig. 3. Heat map for genes present in the gene signature associated with ALI as determined by the optimal classifier. The genes composing the gene signature are listed on the right edge of the heat map. The classes of each of the patients are listed along the bottom edge of the heat map. ARDS, acute respiratory distress syndrome.

Fig. 4. Results of RT-PCR validation of gene expression microarrays.
good diagnostic metric for a small sample of patients, but cannot be generalized to a larger patient population (20). To avoid this potential pitfall, we used the method of leave-one-out cross-validation whereby the classifier was trained on all but one patient sample and the remaining patient sample was used as the validation set. This method creates an independent relationship between training and validation sets. This study has shown that it is possible to achieve a gene signature of high sensitivity and specificity. A further level of validation included in this analysis involved using the developed classifier on an independently recruited validation cohort of patients. This level of validation showed the possibility that the classifier may be generalized to a wider population than that used for initial training. Such analysis introduces the potential for this gene signature to serve as an additional diagnostic metric for distinguishing patients with ALI + sepsis from patients with sepsis alone in the ICU.

Another limitation of this study is the fact that the initial set of subjects was very small in size, resulting in an even smaller training cohort for classification analysis. To improve the chances of the classification analysis finding an underlying pattern in the gene expression profiles of the training cohort, an initial screening of the subjects selected for microarray analysis was performed to render the two cohorts (i.e., subjects with ALI + sepsis and subjects with sepsis alone) more dichotomous. To achieve this goal, cases where there was uncertainty in the clinical diagnosis were excluded from the classification analysis. The drawback to this approach is that it results in the exclusion of the uncertain cases that it would be most beneficial to effectively classify. It is hoped that with the recruitment of larger cohorts of subjects in follow-up studies, more inclusive subject cohorts including borderline cases may be included in subsequent analysis.

A third problem that arose from this study was that the clinical criteria were used as a “gold standard” for classifying patients. However, our initial reason for undertaking the study was our belief that the gold standard might be improved upon. The initial assignment of each patient to either the ALI + sepsis cohort or sepsis alone cohort before gene expression profiling was performed using the current, clinical criteria. The potential difficulty of this approach is that, as was noted in the introduction, the clinical criteria are imperfect and subject to intraobserver variability. This variability could subsequently be a factor in patients being initially assigned to the incorrect patient cohort.

We attempted to minimize this problem by developing detailed phenotypic profiles for each patient included in the analysis. Furthermore, patients with potentially ambiguous clinical phenotypes, such as those with sepsis and bilateral infiltrates who did not meet the exact AECC criteria for ALI, were excluded from the study. In addition, much attention was paid to selecting the appropriate control group for this analysis. Because we were interested in detecting the signal for ALI in patients with sepsis, we chose patients with sepsis alone as a control group. It is well known that sepsis causes multiple changes in gene expression (21). Since most cases of ALI are due to sepsis, it was important to control for the presence of sepsis, since we did not want our analysis to simply detect the gene expression signal due to sepsis, which would interfere with our detection of the gene expression signature due to ALI. Our rationale was that in choosing patients with sepsis alone as our control group, differential gene expression we would detect would be due to the presence of ALI in our group of patients with ALI + sepsis. Thus, this analysis should allow us to discern the true ALI gene signature, which may often be masked by a strong sepsis gene signature.

The conclusions that may be drawn from this study are limited, due to the small sizes of the initial study population and the validation cohort. However, the fact that the same gene signature was capable of accurately classifying a high number of study subjects in both the initial study population and the validation cohort suggests that the gene signature identified by this study has biological relevance and should be further validated in a larger cohort of subjects.

**GRANTS**

Support for this study: National Heart, Lung, and Blood Institute Grants R01-HL-060-234, P01-HL-70807, R01-HL-55330, R01-HL-079904.

**REFERENCES**


14. Olman MA, White KE, Ware LB, Simmons WL, Benveniste EN, Zhu S, Pugin J, Matthey MA. Pulmonary edema fluid from patients with early...


