Age- and Sex-Specific Genomic Profiles in Non–Small Cell Lung Cancer

Context Gene expression profiling may be useful in examining differences underlying age- and sex-specific outcomes in non–small cell lung cancer (NSCLC).

Objective To describe clinically relevant differences in the underlying biology of NSCLC based on patient age and sex.

Design, Setting, and Patients Retrospective analysis of 787 patients with predominantly early stage NSCLC performed at Duke University, Durham, North Carolina, from July 2008 to June 2009. Lung tumor samples with corresponding microarray and clinical data were used. All patients were divided into subgroups based on age (<70 vs ≥70 years old) or sex. Gene expression signatures representing oncogenic pathway activation and tumor biology/microenvironment status were applied to these samples to obtain patterns of activation/deregulation.

Main Outcome Measures Patterns of oncogenic and molecular signaling pathway activation that are reproductible and correlate with 5-year recurrence-free patient survival.

Results Low- and high-risk patient clusters/cohorts were identified with the longest and shortest 5-year recurrence-free survival, respectively, within the age and sex NSCLC subgroups. These cohorts of NSCLC demonstrate similar patterns of pathway activation. In patients younger than 70 years, high-risk patients, with the shortest recurrence-free survival, demonstrated increased activation of the Src (25% vs 6%; \( P < .001 \)) and tumor necrosis factor (76% vs 42%; \( P < .001 \)) pathways compared with low-risk patients. High-risk patients aged 70 years or older demonstrated increased activation of the wound healing (40% vs 24%; \( P = .02 \)) and invasiveness (64% vs 20%; \( P < .001 \)) pathways compared with low-risk patients. In women, high-risk patients demonstrated increased activation of the invasiveness (99% vs 2%; \( P < .001 \)) and STAT3 (72% vs 35%; \( P < .001 \)) pathways while high-risk men demonstrated increased activation of the STAT3 (87% vs 18%; \( P < .001 \)), tumor necrosis factor (90% vs 46%; \( P < .001 \)), EGFR (13% vs 2%; \( P = .003 \)), and wound healing (50% vs 22%; \( P < .001 \)) pathways. Multivariate analyses confirmed the independent clinical relevance of the pathway-based subphenotypes in women (hazard ratio [HR], 2.02; 95% confidence interval [CI], 1.34-3.03; \( P < .001 \)) and patients younger than 70 years (HR, 1.83; 95% CI, 1.24-2.71; \( P = .003 \)). All observations were reproducible in split sample analyses.

Conclusions Among a cohort of patients with NSCLC, subgroups defined by oncogenic pathway activation profiles were associated with recurrence-free survival. These findings require validation in independent patient data sets.

©2010 American Medical Association. All rights reserved.
GENOMIC PROFILES IN NON–SMALL CELL LUNG CANCER

Table 1. Clinical Characteristics of Data Sets Used in Analyses (N = 787)4

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Shedden et al,4 2008 (n = 442)</th>
<th>Bhattacharjee et al,2001 (n = 125)</th>
<th>Raponi et al,5 2006 (n = 129)</th>
<th>Bild et al,6 2007 (n = 91)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, y</td>
<td>Median (range) 65 (33-87)</td>
<td>64 (33-88)</td>
<td>68 (42-91)</td>
<td>67 (32-83)</td>
</tr>
<tr>
<td></td>
<td>Mean (SD) 64 (10)</td>
<td>63 (10)</td>
<td>67 (10)</td>
<td>65 (9)</td>
</tr>
<tr>
<td>Sex</td>
<td>Male 223 (50)</td>
<td>53 (42)</td>
<td>82 (64)</td>
<td>56 (62)</td>
</tr>
<tr>
<td></td>
<td>Female 219 (50)</td>
<td>72 (58)</td>
<td>47 (36)</td>
<td>35 (38)</td>
</tr>
<tr>
<td>Tobacco use</td>
<td>Smokers 300 (68)</td>
<td>107 (86)</td>
<td>119 (92)</td>
<td>80 (88)</td>
</tr>
<tr>
<td></td>
<td>Never smokers 49 (11)</td>
<td>17 (14)</td>
<td>4 (3)</td>
<td>4 (4)</td>
</tr>
<tr>
<td></td>
<td>Unknown 93 (21)</td>
<td>1 (0)</td>
<td>6 (5)</td>
<td>7 (8)</td>
</tr>
<tr>
<td>Histology</td>
<td>Adenocarcinoma 442 (100)</td>
<td>125 (100)</td>
<td>0</td>
<td>45 (49)</td>
</tr>
<tr>
<td></td>
<td>Squamous cell carcinoma 0</td>
<td>0</td>
<td>129 (100)</td>
<td>46 (51)</td>
</tr>
<tr>
<td>Disease stage</td>
<td>Early (IA, IB, IIA, IIB, IIIA)</td>
<td>428 (97)</td>
<td>107 (86)</td>
<td>123 (95)</td>
</tr>
<tr>
<td></td>
<td>Late (IIIB, IV)</td>
<td>11 (2)</td>
<td>18 (14)</td>
<td>6 (5)</td>
</tr>
<tr>
<td></td>
<td>Unknown 3 (1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Adjuvant chemotherapy, %</td>
<td>0</td>
<td>&lt;1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

*Data are presented as No. (%) unless otherwise specified.

Figure 1. Kaplan-Meier Survival Comparison in Patients Aged <70 vs ≥70 Years

METHODS

Complete details of the statistical methods and the Cox proportional hazard models are available in the eAppendix (available online at http://www.jama.com). All tumor samples were obtained from patients enrolled in institutional review board–approved clinical trials after written informed consent was obtained.

NSCLC Samples

Gene expression data from 787 NSCLC patient tumor samples (with corresponding clinical data) were obtained from 4 independent data sets.4-7 Data sets were selected based on the availability of microarray gene expression and clinical data (with follow-up data of ≥60 months) from non–small cell lung tumors. The samples were obtained from patients with mostly early stage disease (stages I-IIIA) at the time of diagnosis and only 1% of whom received adjuvant chemotherapy or radiation. All gene expression data were arrayed using Affymetrix Human Genome U133A GeneChips (http://www.affymetrix.com/products_services/arrays/index.affx) except for the data set of Bhattacharjee et al.5 These data were converted from U95A GeneChips to U133A using Chip Comparer (http://chipcomparer.genome.duke.edu).

When combining data sets from different platforms and different experiments, nonbiological experimental variation, or batch effects, can occur. To reduce the likelihood of batch effects, a normalizing algorithm, ComBat6 (http://statistics.byu.edu/johnson/ComBat/), was applied to the patient samples before performing any analysis.

To study the biology of NSCLC as a function of age and sex, data from the 787 NSCLC samples were divided into 4 different groups: patients younger than 70 years (n = 520), patients aged 70 years or older (n = 267), men (n = 414), and women (n = 373). Patient samples from each clinical group were also randomly divided into 2 equal cohorts to create training and validation subsets that would be used for further validation of the findings. Five-year recurrence-free survival (RFS) was used as the measure of a clinically relevant end point in all survival analyses.

Oncogenic Pathway and Tumor Microenvironment Analysis

Previously validated gene expression signatures of oncogenic pathway activation (β-catenin, E2F1, Myc, Ras, Src, epidermal growth factor receptor [EGFR], and signal transducer and activator of transcription 3 [STAT3]), as well as tumor biology/microenvironment status (chromosomal instability, epigenetic stem cell, invasiveness, tumor necrosis factor [TNF], and wound healing), were applied to all data sets and subsets to dissect the heterogeneity of NSCLC and identify relevant subphenotypes. Each signature is able to identify a pattern of gene up-regulation or overexpression associated with the specific pathway being “turned on” or activated. Briefly, using previously described Bayesian binary regression methods,7,8,10-20 predictions of oncogenic pathway and tumor biology/microenvironment activation for each of the aforementioned signatures were developed. These predictions are expressed as a relative probability (ranging from 0-1) and represent the degree of pathway activation found within each tumor sample. A probability greater than 0.5 is considered positive for pathway activation.

Hierarchical clustering of tumor predictions of pathway activation was performed using R/Bioconductor statistical packages, version 2.9.0 (http://www.r-project.org)21-23 after they had been grouped by age or sex variables. Heat maps were generated using R/Bioconductor statistical packages based on the clustered order of the patient samples from the 4 individual groups. Clusters are identified based on branching points of the corresponding dendrogram following hierarchical clustering of patient samples and represent specific patterns of pathway activation. Furthermore, cluster stability was assessed and confirmed by agglomeration bootstrap clustering. Low- and high-risk groups were defined as clusters, or cohorts, of patients that have the best and worst 5-year RFS, respectively.
Statistical Analysis

Standard Kaplan-Meier survival curves, using 5-year RFS end points, were generated for patient clusters with similar patterns of oncogenic pathway or tumor microenvironment deregulation using GraphPad software, version 5.02 (http://www.graphpad.com). A prognostically significant result is defined by log-rank $P < .05$. GraphPad software was also used to generate unpaired, 1-tailed $t$ tests to compare degree of pathway activation between sample groups, as well as 2-tailed $\chi^2$ tests to compare percentages of patients with pathway activation between high- and low-risk groups. Adjustments were made for multiple testing where necessary using Bonferroni correction.

Cox proportional hazards analyses were performed using data from patients ($N = 787$) and were repeated for the following data subsets: patients younger than 70 years ($n = 520$), patients aged 70 years or older ($n = 267$), women ($n = 373$), men ($n = 414$), adenocarcinoma ($n = 612$), squamous cell carcinoma ($n = 175$), never smokers ($n = 74$), and smokers ($n = 606$). Samples with missing data were excluded from the analyses.

For each subset, a multivariate Cox proportional hazards model was computed to determine if the clustering prognosis (high- vs low-risk groups) variable was significant in the presence of age ($<70$ vs $\geq 70$ years), sex, histology, smoking status, and disease stage (I-IIIA vs IIIB-IV). Estimated hazard ratios (HRs), confidence intervals (CIs), and $P$ values are provided for the pathway-based prognostic clusters and clinical variables.

RESULTS

The clinical and demographic features of all patients included in this analysis are described in Table 1. Among the NSCLC patient samples used for this analysis, the median age of patients ranged from 64 to 68 years.
The majority of patients were male and smokers (53% and 77%, respectively). Adenocarcinoma was the most common histology (78%), with patients predominantly having early stage (stages I-IIa) disease (95%).

**Age-Specific Differences in the Biology of NSCLC**

The association of age with clinical outcome was analyzed by comparing patients younger than 70 years old (n=520) with patients aged 70 years or older (n=267). Younger patients had significantly better RFS (P=.006; Figure 1).

Activation of individual oncogenic pathways and tumor microenvironment status was then examined in patients stratified by age (<70 or ≥70 years) (Figure 2). In this analysis, the β-catenin (P=.03), invasiveness (P=.04), and Src (P=.008) pathways were more likely to be activated in patients younger than 70 years.

Four distinct clusters of pathway activation in each age group (Figure 3) were identified. Kaplan-Meier survival analysis for younger patients (<70 years) revealed clusters with unique prognostic significance (P=.03; Figure 4). Using RFS as a clinically relevant phenotype, we were able to characterize distinct gene expression patterns associated with high-risk patients (cluster 4) with shorter RFS compared with patients with a better prognosis (cluster 3) with longer RFS. Clusters 1 and 2 represent patients with an intermediate risk compared with clusters 3 and 4, with RFS between that of the low- and high-risk patients. Analysis of clusters 3 (low-risk) and 4 (high-risk) in younger lung cancer patients demonstrated that the Src (25% vs 6%; P<.001) and TNF (76% vs 42%; P<.001) gene signatures had a significantly higher probability of activation in the high-risk cohort (eFigure 1A).

Likewise, Kaplan-Meier survival analysis for older patients showed clinically meaningful differences in median RFS (68 months vs 33.9 months between high-risk cluster 4 and low-risk cluster 3, respectively), although these differences were not statistically significant (P=.10; Figure 4). Biologically, in older patients, gene signatures for wound healing (40% vs 24%; P=.02) and invasiveness (64% vs 20%; P<.001) were significantly more activated in the high-risk (cluster 4) cohort vs the low-risk cohort (cluster 3) (eFigure 1A).

**Sex-Specific Variations in NSCLC**

The biology and clinical course of NSCLC are sex-specific. To further elucidate sex-specific differences at a pathway level, 373 women and 414 men with NSCLC were compared. Consistent with previous findings, a Kaplan-Meier analysis revealed that women in general had significantly better RFS than men (P=.008; Figure 5).

The probability of oncogenic pathway activation in patients stratified by sex is depicted in Figure 6. Men demonstrated a higher probability of activation of chromosomal instability (P=.001), epigenetic stem cell (P=.03), invasiveness (P=.005), Myc (P=.02), and others.
and wound healing ($P=.004$) pathways, while women demonstrated a higher probability of activation of the $E2F1$ pathway ($P=.04$).

Hierarchical clustering of the oncogenic pathway activity in women and men revealed 4 distinct clusters for each group (Figure 7). In women, the 4 clusters had distinct prognostic significance ($P=.002$; Figure 8). In addition, when comparing clusters 1 (high-risk) and 2 (low-risk), which had the largest difference in RFS, the prognostic significance was maintained ($P=.001$). Among men, the 4 individual clusters of pathway activation did not demonstrate prognostic significance (Figure 7 and Figure 8). However, when only the clusters with the best (cluster 2) and worst (cluster 1) RFS were compared, cluster 1 demonstrated a significantly lower RFS ($P=.02$). Clusters 3 and 4 for men and women represent patients with an intermediate risk compared with clusters 1 and 2, with RFS between that of the high- and low-risk patients.

Further analysis compared the activation of individual oncogenic and tumor microenvironment pathways in relevant prognostic clusters (eFigure 2A). In women, cluster 1 (high-risk) demonstrated increased activation of the invasiveness (99% vs 2%; $P<.001$) and $STAT3$ (72% vs 35%; $P<.001$) pathways compared with cluster 2. In men, the high-risk cohort (cluster 1) demonstrated increased activation of $STAT3$ (87% vs 18%; $P<.001$), TNF (90% vs 46%; $P<.001$), wound healing (50% vs 22%; $P<.001$), and $EGFR$ (13% vs 2%; $P<.003$) compared with the better-prognosis cohort (cluster 2). Thus, while the $STAT3$ pathway was uniformly activated in patients with poor prognosis irrespective of sex, the invasiveness, $EGFR$, TNF, and wound healing pathways probably underlie the sex-specific differences observed in high-risk NSCLC patients.

Error bars indicate 95% confidence intervals. WH indicates wound healing; TNF, tumor necrosis factor; $STAT3$, signal transducer and activator of transcription 3; CIN, chromosomal instability; EPI, epigenetic stem cell; IGS, invasiveness; $EGFR$, epidermal growth factor receptor. By unpaired, 1-tailed $t$ tests, $P<.05$ for $E2F1$, EPI, and $Myc$; $P<.01$ for CIN, IGS, and WH.
Validity and Reproducibility of the Biologically Distinct Subphenotypes
The true power and relevance of specific biological patterns of pathway activity lies in their reproducibility. Thus, to assess the stability and validity of biologically distinct clusters at a pathway level, patient samples from each clinical subgroup were randomly divided into 2 equal cohorts (training and validation subsets) using a split-sample approach. Similar patterns of pathway activation between training and validation subsets were observed within each clinical variable subgroup (Figure 9; eFigure 1, B and C, and eFigure 2, B and C), demonstrating the reproducibility of the pathway activation patterns.

To further elucidate the similarities between the training and validation subsets in an objective manner, correlation analyses were performed. Individual pathway activations were compared with each other within each cohort, and Pearson correlation coefficients ($-1 < r < 1$) were extracted and plotted as a grid. Near identical patterns of pathway activations and correlation were observed (as depicted in eFigure 1D and eFigure 2D), confirming the inherent biological similarities between the high- and low-risk phenotypes identified by pathway activation patterns in both the training and validation cohorts.

In addition, RFS analyses (Figure 10) of the training and validation subsets indicate consistent survival differences between high- and low-risk clusters for the training and validation subsets. These analyses were repeated using several random splits of subset cohorts, all with identical results. Taken together, while a completely independent validation of our findings would have been ideal, these split-sample analyses suggest that the molecular traits between high- and low-risk clusters within the current clinical classification scheme (age and sex) may represent biologically distinct, reproducible subphenotypes.

Prognostic Oncogenic Pathway Patterns
Cox proportional hazards analysis was performed to determine if the prognostic clusters identified through oncogenic pathway patterns in the above analyses were independent of known clinical variables (age, sex, histology, disease stage, and smoking status) (Table 2, eTable 1, and eTable 2). Multivariate Cox proportional hazard analysis of the pathway-based prognostic clusters (low- and high-risk clusters determined by pathway patterns shown in Figure 3 and Figure 7) was found to be statistically significant (Table 2 and eTable 1) in the following subsets: women (HR, 2.02; 95% CI, 1.34-3.03; $P < .001$) and patients younger than 70 years (HR, 1.83; 95% CI, 1.24-2.71; $P = .003$). It was not significant in men or in patients aged 70 years or older. When testing for interactions between prognostic clusters and clinical variables, $P > .05$ meant a failure to conclude dependence (eTable 2). Thus, we demonstrate that patterns are not only unique in characterizing the biology of NSCLC but are also prognostically independent of other relevant clinical variables, most notably stage of disease and histology (adenocarcinoma vs squamous).

COMMENT
The oncogenic process typically involving the somatic acquisition of large numbers of mutations, coupled with varied host genetic constitution, produces a disease of enormous complexity that is difficult to characterize and treat effectively. There is perhaps no better example of this challenge than
that seen in cancer, particularly lung cancer.12,24-26 Numerous previous studies report the value of dissecting NSCLC biology using gene expression profile5,27-31; however, none have described biologic differences in clinically relevant phenotypes of NSCLC. We examined the biology behind age- and sex-specific differences in NSCLC through the use of gene expression signatures to better understand this disease process and potentially identify molecular targets for therapy. This analysis represents one of the first large-scale attempts to comprehensively characterize the biology of early stage NSCLC at a molecular pathway level and demonstrates a clear distinction in gene expression profiles within relevant age and sex categories of NSCLC.

Our characterization of a large collection of patients with NSCLC has identified biologically distinct, clinically relevant (using survival as a phenotype) subgroups that are independent of disease stage, histology, and other known clinical variables (Table 2, eTable 1, and eTable 2), based on profiles depicting oncogenic pathway status and relevant tumor biology. While differences in clinical outcomes and the biology of NSCLC based on age and sex have been previously noted, we were able to describe the molecular networks contributing to these differences. As an example, younger patients and women have better cancer-related survival, but the underlying biology has been unclear and often debated.32-36 We were able to confirm superior RFS in women and younger patients (<70 years). Further analysis revealed that female patients with relatively poorer RFS demonstrated increased deregulation of the invasiveness and STAT3 pathways. Likewise, older patients with poor prognosis had increased activation of the invasiveness and wound healing pathways. Although this is a novel finding, biologically this is not entirely unexpected. The invasiveness and wound healing gene signatures likely identify tumors at high risk of metastasis,11,15 along with the wound healing signature identifying activation of angiogenesis pathways.15,37

We believe our findings represent a novel approach to defining clinically relevant cohorts of NSCLC stratified by age and sex that are enriched for spe-

Figure 9. Comparisons of Significantly Activated Pathways by Patient Percentages Between High- and Low-Risk Clusters for the Training and Validation Subsets in Each Age and Sex Group

A. Age

Age <70 y

Age ≥70 y

Cluster  Pathway  Patients, %

High-risk  Low-risk

Cluster  Pathway  Patients, %

High-risk  Low-risk

src  TNF  src  TNF

Training Subset  Validation Subset

Training Subset  Validation Subset

B. Sex

Women

Men

Cluster  Pathway  Patients, %

High-risk  Low-risk

Cluster  Pathway  Patients, %

High-risk  Low-risk

EGFR  Myc  STAT3  TNF  WH  EGFR  Myc  STAT3  TNF  WH

Training Subset  Validation Subset

Training Subset  Validation Subset

Training and validation subsets were created by randomly splitting the samples in each age or sex group. By 2-tailed χ² tests comparing high-risk vs low-risk clusters, among younger patients, \( P < .01 \) for TNF and \( P < .001 \) for src in the training subset and \( P < .01 \) for both comparisons in the validation subset. Among older patients, \( P < .05 \) for IGS and \( P < .001 \) for WH in the training subset and \( P < .001 \) for both comparisons in the validation subset. Among women, \( P < .001 \) for IGS and STAT3 in the training and validation subsets. Among men, \( P < .001 \) for all comparisons in both subsets except WH in the validation subset (\( P < .01 \)). Sample sizes are shown below the graphs in Figure 10 at time = 0.
Figure 10. Kaplan-Meier Survival Curves of Identified High- and Low-Risk Clusters From the Training and Validation Subsets

For younger patients, median recurrence-free survival in the training set was 39.27 months in cluster 2 and 79.5 months in cluster 4 and in the validation set it was 47 months in cluster 2 and 79.54 months in cluster 4. For older patients, in the training set it was 33.8 months in cluster 4 and 86.7 months in cluster 2 and in the validation set it was 28.25 months in cluster 3 and 54.2 months in cluster 2. For women, median recurrence-free survival in the training set was 44.94 months in cluster 3 and 119.5 months in cluster 1 and in the validation set it was 25 months in cluster 2 and 95.5 months in cluster 3. For men, in the training set it was 27.6 months in cluster 4 and 68 months in cluster 1 and in the validation set it was 31 months in cluster 2 and 70.3 months in cluster 1. Log-rank comparisons are between high- and low-risk cohorts within each subset.

Table 2. Univariate and Multivariate Cox Proportional Hazard Analyses for Prognostic High- and Low-Risk Patient Clusters

<table>
<thead>
<tr>
<th>Clinical Variable</th>
<th>Univariate Analysis</th>
<th>Multivariate Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hazard Ratio (95% Confidence Interval)</td>
<td>P Value</td>
</tr>
<tr>
<td>Female</td>
<td>2.00 (1.36-2.95)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Male</td>
<td>1.56 (1.01-2.41)</td>
<td>.04</td>
</tr>
<tr>
<td>Age, y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;70</td>
<td>1.54 (1.09-2.17)</td>
<td>.01</td>
</tr>
<tr>
<td>≥70</td>
<td>1.59 (1.02-2.49)</td>
<td>.04</td>
</tr>
</tbody>
</table>

a Based on the comparison of the 5-year median recurrence-free survival for low- and high-risk patient clusters with the best and worst recurrence-free survival, respectively.
ward in defining pathway-driven cohorts of NSCLC that likely explain the age- and sex-specific differences seen in NSCLC.

Author Contributions: Mr Mostertz, Dr Stevenson, and Dr Potti had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Mr Mostertz and Dr Stevenson contributed equally to this work.

Study concept and design: Mostertz, Stevenson, Potti. Acquisition of data: Mostertz, Chan, Potti. Analysis and interpretation of data: Mostertz, Stevenson, Acharya, Chan, Walters, Lamlethron, Barry, Crawford, Nevinns, Potti.

Drafting of the manuscript: Mostertz, Stevenson, Chan, Acharya, Potti.

Critical revision of the manuscript for important intellectual content: Mostertz, Stevenson, Acharya, Chan, Walters, Lamlethron, Barry, Crawford, Nevinns, Potti.

Statistical analysis: Acharya, Mostertz, Walters, Barry, Lamlethron, Chan, Potti.

Obtained funding: Potti.

Administrative, technical, or material support: Stevenson, Crawford, Nevinns, Potti.

Study supervision: Mostertz, Stevenson, Potti.

Financial Disclosures: None reported.

Funding/Support: This study was supported by research grants from the Emeline Brown Cancer Research Fund, the Harold and Linda Chapman Lung Cancer Fund, the Jimmy V Foundation, the American Cancer Society, and the National Cancer Institute.

Role of the Sponsors: The funding organizations had no role in the design and conduct of the study, in the collection, analysis, and interpretation of the data, or in the preparation, review, or approval of the manuscript.

Online-Only Material: The eAppendix, efigures 1 and 2, and eTables 1 and 2 are available online at http://www.jama.com.

REFERENCES


