nature medicine

Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study

Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma:*[1]
Kerby Shedden[2,3,17], Jeremy M G Taylor[3,4,17], Steven A Enkemann[5,17], Ming-Sound Tsao[6,17], Timothy J Yeatman[5,17], William L Gerald[7,17], Steven Eschrich[5,17], Igor Jurisica[6,17], Thomas J Giordano[8], David E Misek[3,9], Andrew C Chang[3,9], Chang Qi Zhu[6], Daniel Strumpf[6], Samir Hanash[3], Frances A Shepherd[6], Keyue Ding[10], Lesley Seymour[10], Katsuhiko Naoki[11], Nathan Pennell[11], Barbara Weir[11], Roel Verhaak[11], Christine Ladd-Acosta[12], Todd Golub[12], Michael Gruidl[5], Anupama Sharma[5], Janos Szoke[7], Maureen Zakowski[7], Valerie Rusch[7], Mark Kris[7], Agnes Viale[7], Noriko Motoi[7], William Travis[7], Barbara Conley[13], Venkatraman E Seshan[14,17], Matthew Meyerson[11,12,17], Rork Kuick[3,17], Kevin K Dobbin[15,17], Tracy Lively[16,17], James W Jacobson[16,17] & David G Beer[3,9,17]

Although prognostic gene expression signatures for survival in early-stage lung cancer have been proposed, for clinical application, it is critical to establish their performance across different subject populations and in different laboratories. Here we report a large, training–testing, multi-site, blinded validation study to characterize the performance of several prognostic models based on gene expression for 442 lung adenocarcinomas. The hypotheses proposed examined whether microarray measurements of gene expression either alone or combined with basic clinical covariates (stage, age, sex) could be used to predict overall survival in lung cancer subjects. Several models examined produced risk scores that substantially correlated with actual subject outcome. Most methods performed better with clinical data, supporting the combined use of clinical and molecular information when building prognostic models for early-stage lung cancer. This study also provides the largest available set of microarray data with extensive pathological and clinical annotation for lung adenocarcinomas.

In the United States and in many Western countries, lung cancer represents the leading cause of cancer-related death[1]. The 5-year, overall survival rate is 15% and has not improved over many decades. This is mainly because approximately two-thirds of lung cancers are discovered at advanced stages, for which cure by surgical resection is no longer an option. Furthermore, even among early-stage patients who are treated primarily by surgery with curative intent, 30–55% will develop and die of metastatic recurrence. Recent multinational clinical trials (IALT, JBR10, ANITA, UFT, LACE) conducted in several continents have demonstrated that adjuvant chemotherapy significantly improves the survival of patients with early-stage (IB–II) disease[2]. Nevertheless, it is clear that a proportion of patients with stage I disease have poorer prognosis and may benefit significantly from adjuvant chemotherapy, whereas some with stage II disease with relatively good prognoses may not benefit significantly from adjuvant chemotherapies. It remains possible, however, that the latter patients could derive additional benefit from adjuvant targeted therapies[2–4]. Therefore, there is an urgent need to establish new diagnostic paradigms and validate in clinical trials methods for improving the selection of stage I–II patients who are most likely to benefit from adjuvant chemotherapy.

Global gene-expression profiling using microarray technologies has helped to improve our understanding of the histological heterogeneity of non–small cell lung cancer (NSCLC) and has identified potential biomarkers and gene signatures for classifying patients with significantly different survival outcomes[5–11]. However, the performance and general applicability of published classifiers has not been easy to establish because of small numbers of subjects examined and inclusion of heterogeneous tumor types. Furthermore, there have not been uniform criteria for sample inclusion, annotation, sample processing and data analyses. To address these concerns and to generate a large microarray database of NSCLC samples that have been collected and studied using a common protocol[12], we conducted a large retrospective, multi-site, blinded study. The study included a blinded validation step to characterize the performance of several newly developed prognostic models using a total of 442 lung adenocarcinomas, the specific type of lung cancer that is increasing in incidence[13].

To ensure scientific validity of the results, subject samples along with all relevant clinical, pathological and outcome data were collected by investigators at four institutions using data from six lung-cancer treatment sites with subject inclusion criteria defined a priori. Gene

*The complete list of affiliations appears at the end of the article. Correspondence should be addressed to J.W.J. (jacobsonj@ctep.nci.nih.gov) or D.G.B. (dgbeer@umich.edu).

Received 8 January; accepted 12 June; published online 20 July 2008; doi:10.1038/nm.1790

Table 1 Summary statistics of data

	UM	HLM	CAN/DF	MSK
Sample size	177	79	82	104
Age (mean, s.d.)	64 (10)	67 (10)	61 (10)	65 (10)
Sex (% male)	56%	51%	56%	36%
Stage I	66%	54%	68%	61%
Stage II	16%	26%	32%	19%
Stage III	18%	19%	0%	20%
Median follow-up (months)	54	39	40	43
Number of deaths	75	50	28	34

expression data on subsets of lung adenocarcinomas were generated by each of four different laboratories using a common platform and following a protocol previously demonstrated to be robust and reproducible[12]. We considered four separate hypotheses: (i) gene expression alone can predict outcomes for all samples; (ii) gene expression and basic clinical covariates (stage, age, sex) can predict outcomes for all samples; (iii) gene expression alone can predict outcomes for stage 1 samples; and (iv) gene expression and basic clinical covariates can predict outcomes for stage 1 samples. Note that prediction on stage 1 samples is more difficult than on the full study set as these samples are relatively homogeneous. The consideration of clinical covariates is highly relevant as the basic variables considered here will always be available in practice, and gene expression–based prediction is relevant in practice only if it provides more information than these measures. We followed a strict protocol for the data collection, data analysis and performance evaluation phases of our study. Data generated at two sites were used as a training set and the results were validated using the independent data sets from the other two participating sites following a blinded protocol. The results from this study provide not only valid assessment of outcome prediction in the multi-institutional setting but also a rich data set for future analysis and provide an example of how large data sets can be generated and tested by cooperation and pooling of resources among many investigators.

RESULTS

Consortium and classifier development

We collected a total of 442 lung adenocarcinomas with high-quality gene-expression data, pathological data and clinical information describing the severity of the disease at surgery and the clinical course of the disease after sampling. These samples, collected from six contributing treatment institutions, were grouped into four sets of data on the basis of the laboratory where samples were processed for microarray analysis. The distribution of several clinical variables for these four data sets is shown in **Table 1**. The first two data sets, UM and HLM, were released to members of the consortium for the development of classifiers appropriate for our four hypotheses. Details of our protocol for developing and evaluating classifiers are provided in the **Supplementary Methods** online.

Eight classifiers producing either categorical or continuous risk scores were developed by investigators using the training data and were tested for effectiveness on the two remaining data sets (MSK and CAN/DF). Most of these classifiers incorporated techniques that have repeatedly been applied in gene expression–based prognosis and found to work well in at least some instances. As an overview, data reduction was carried out using gene clustering (method A), univariate testing (methods B, C, D, E, F, G) or on a mechanistic basis (method H). Final scoring and classification was based on penalized Cox regression modeling with gene cluster expression summaries as the features (method A), on expression of individual genes (method B), on principal components of gene expression (methods F, G), on membership in clusters defined by gene expression (methods C, D) or on a vote of single-gene classifiers (method H). A number of other factors, such as subselection of the training samples, gene filtering and data transformation, were handled in various ways as described in detail in **Supplementary Methods**. We note that all classifiers started with the same set of expression summaries processed using the DChip algorithm[14], so handling of the raw data was uniform across the methods.

Classifier performance without and with clinical covariates

The estimated hazard ratios for the risk scores produced by the eight prognosis methods, with 95% confidence intervals, are shown for the two validation sets in **Figure 1**. Hazard ratios substantially greater than 1.0 indicate that subjects in the validation set with high predicted risk had poor outcomes. Confidence intervals in **Figure 1** and the corresponding *P*-values (**Supplementary Results** online) indicate which of the methods performed significantly better than expected by chance. As another performance measure, we calculated the concordance probability estimate (CPE), which measures how well the subject outcomes agree with the predicted risk scores. CPE values close to 0.5 indicate no concordance (poor predictivity); CPE values approaching 1.0 indicate strong concordance (good predictivity). On the basis of these measures, most of the classifiers performed well in at least some situations. Finally, for 3-year survival, we constructed receiver operating characteristics (ROC) curves for continuous predictors and tables of sensitivity and specificity estimates for categorical predictors (**Supplementary Results**).

There are some notable observations about the classifiers as a group. Most methods performed much better on sample sets containing all stages compared to sample sets containing just stage 1. This reflects an ability to stratify by stage even when stage is not explicitly included in the model. Including clinical covariates improved the performance of most of the models. In fact, without clinical covariates, no model achieved a hazard ratio significantly greater than 1 in both validation sets for the stage 1 samples. An important criterion was that a model should perform well in both validation sets as an indication of robust performance in routine clinical testing. For prediction on all stages

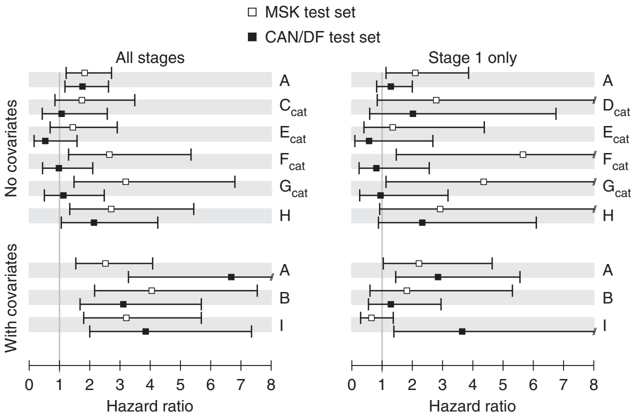

Figure 1 Classifier performance. Hazard ratios of methods A–I (see Introduction and **Supplementary Methods**) on validation data sets for the four hypotheses, along with 95% confidence intervals. Methods that placed patients into ordered categories rather than providing a continuous risk score are denoted "cat".

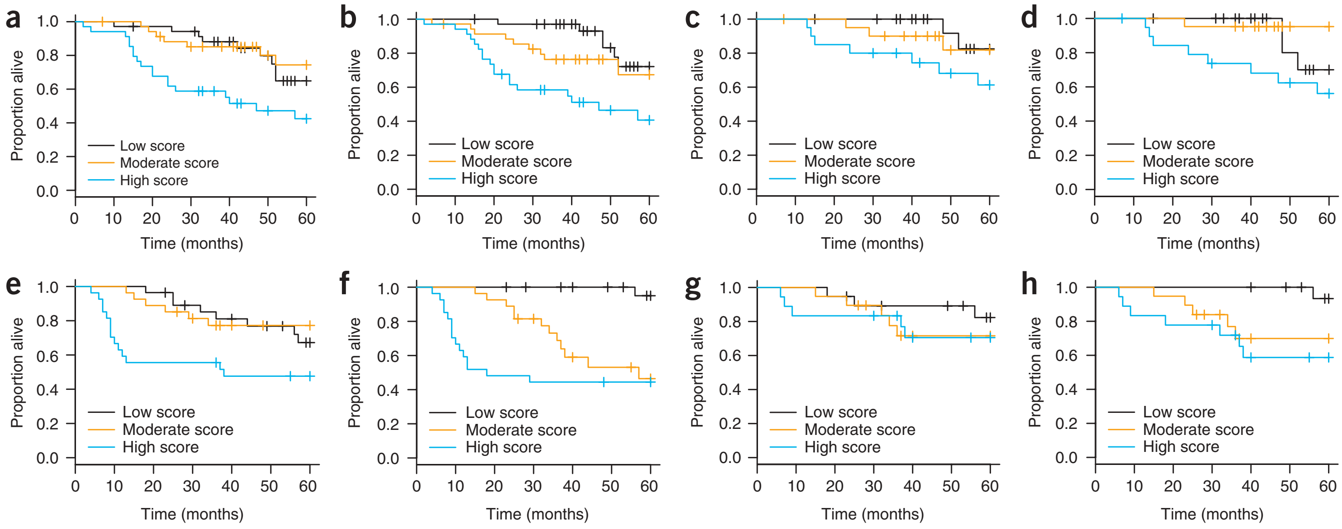

Figure 2 Kaplan-Meier estimates of the survivor function for method A on each validation data set for the four hypotheses. (**a**) MSK test set, all stages. (**b**) MSK test set with covariates, all stages. (**c**) MSK test set, stage 1 only. (**d**) MSK test set with covariates, stage 1 only. (**e**) CAN/DF test set, all stages. (**f**) CAN/DF test set with covariates, all stages. (**g**) CAN/DF test set, stage 1 only. (**h**) CAN/DF test set with covariates, stage 1 only. Low scores correspond to the lowest predicted risk and high scores correspond to the greatest predicted risk.

using gene expression data, only methods A and H performed with consistent statistical significance. For prediction on all stages using both gene expression and clinical covariates, methods A and B produced hazard ratios exceeding 2 for both validation sets. For prediction on subjects with stage 1 disease using gene expression data only, three of the methods (A, D and H) gave hazard ratios exceeding 1 for both validation sets. Of these, only method A had a hazard ratio significantly greater than 1 for one of the data sets. For prediction on subjects with stage 1 disease using gene expression data and clinical covariates, method A gave hazard ratios that exceeded 2 and were statistically significant for both data sets. For many of the classifiers, good performance in one setting was offset by poor performance in a different setting. Thus, method A seemed to have the best overall performance across the four hypotheses.

Using method A to stratify subjects into three groups, we generated Kaplan-Meier plots to illustrate the survival differences among the groups determined by this classification scheme for both the validation (**Fig. 2**) and the training (**Fig. 3**) data sets. This illustrates that lung adenocarcinomas can be divided into groups with different survival rates. Kaplan-Meier plots showing the performance of the other classifiers on the validation data sets are available in the **Supplementary Results**. The plots developed from method A again illustrate that risk predictors evaluated on all subjects performed better than those evaluated on subjects with stage 1 disease. Furthermore, using clinical covariates together with the gene expression data improved outcome prediction compared to using gene expression data alone. Method A included the null value 1 in its 95% hazard ratio confidence interval in only one of eight situations considered (**Fig. 1**). The one hypothesis wherein method A did not give significant prediction was stage 1; subjects scored using only gene-expression measures. As noted above, no method gave significant results for both validation sets in this setting. This suggests that stage 1 tumors may be classified more efficiently using clinical parameters along with gene expression data.

Analyses of additional classifiers

The additional classifiers shown in the **Supplementary Results** (J, K, L, M and N) were derived from the probe sets listed in refs. 9 and 10. Although we were unable to reconstruct the classifiers reported in the original papers, we used the reported probe sets to construct classifiers, and we tested them on our validation data. The performances of these classifiers were generally comparable to, although slightly poorer than, those for methods A–H developed for this paper. The hazard ratios were in most cases larger than 1, but they did not give statistically significant hazard ratios consistently for both validation data sets (**Supplementary Results**). For these classifiers, the addition of clinical covariates improved the predictive ability.

We considered two other ways to compare the classifiers developed for this study. The **Supplementary Results** show how each tumor sample was classified by each of the methods. Many subjects could be correctly classified by many different methods. These may represent extreme cases that can be easily recognized. There were a number of

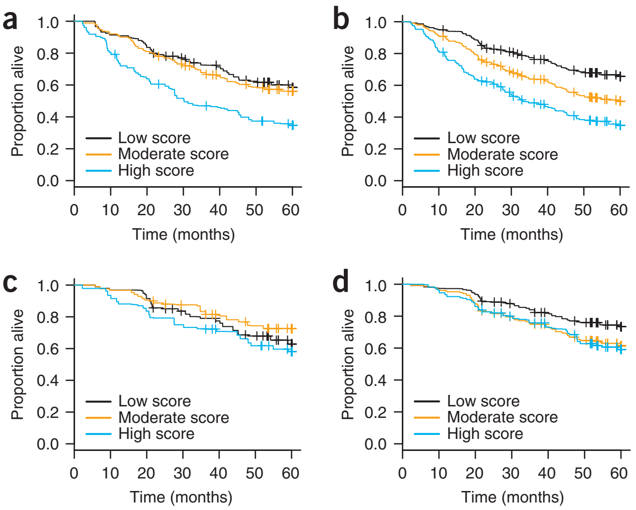

Figure 3 Kaplan-Meier estimates of the survivor function for method A (cross-validated) on training sets UM and MSK. (**a**) All stages. (**b**) All stages with covariates. (**c**) Stage 1 only. (**d**) Stage 1 only, with covariates.

tumors for which the classifiers disagreed, which could reflect classifier quality or tumors that show conflicting expression patterns in genes that typically correlate with greater or lesser risk. This highlights the greatest problem facing expression-based classification of tumors: are misclassifications due to inaccurate clinical information, tissue sampling problems or bad classifiers, or do they simply reflect the heterogeneity of tumor types that can arise? The overlap in predictivity is not explained by a high overlap in the probe sets used for classification (**Supplementary Data** online). There was overlap between the genes used in method H and those in one of the clusters observed to be important in method A. Many of these genes were associated with proliferation, which is consistent with more aggressive lung adenocarcinomas showing increased proliferative potential. For all the other newly developed classifiers, the overlap reflected similarities in the methods used to select genes. The variety of probe sets showing some predictive capacity suggests that information about lung adenocarcinoma outcomes may not be concentrated in just a few exceptional genes.

DISCUSSION

Several studies of primary lung adenocarcinoma or NSCLC have reported the ability to generate expression signatures capable of grouping subjects according to their survival outcomes. However, most studies are small (approximately 100 subjects or fewer) and typically draw data from a single treatment institution. Gene expression profiles with real clinical applicability must be recognizable despite variability that might occur in the processing of samples at different institutions. So far, little has been published on the ability of prognostic methods for lung cancer to perform in larger data sets or with independent validation samples. Often, the published signatures show little overlap in the genes identified as significant predictors of outcome. Thus, there is a strong possibility that sample collection methods, processing protocols, single-institution subject cohorts, small sample sizes and peculiarities of the different microarray platforms are contributing significantly to the results. To address these issues, we conducted a multi-institutional collaborative study to generate gene expression profiles from a large number of samples with a priori-determined clinical features that could be used to fully evaluate proposed prognostic models for potential clinical implementation.

The design and execution of the present study recognized the specific issues discussed above. Significant emphasis was placed on reducing technical variability by using similar protocols, reagents and platforms[12], so that the main uncontrolled variables represented the biology of the lung cancers and associated clinical data. We used two blinded, external validation sets to provide a realistic assessment of the performance of each prognostic method. This is in contrast to the more common approach of obtaining all the data from a single source and randomly assigning samples to training and validation sets for the development and assessment of classifiers. Furthermore, we took great care to standardize the pathological assessment of each tumor sample and the collection of clinical information across all institutions involved in this study. The lessons learned from this coordinated effort will likely influence the research practice for future profiling efforts in lung cancer.

Several classifiers were developed from the training data and tested on the independent data sets. These classifiers represent many of the established techniques for classifier development, with new approaches also represented. The classifiers varied in their success in stratifying subjects according to risk. Two of the methods (C and E) showed little predictive capacity. The poor performance of method E was expected, as one individual gene parameter is too sensitive to noise to perform well in gene expression data collected from more than one institution. More complex classifiers showed better success, with a few classifiers demonstrating the ability to classify across different institutional data sets as well as within the stage 1 tumors. The most successful classifiers at stratifying stage 1 samples were trained on samples from all stages. This suggests that heterogeneity of aggressiveness exists in stage 1 tumors and that the pattern of gene expression in higher stage tumors is informative for predicting the risk of stage 1 tumors. We note that the power for comparing classifiers tended to be lower than the power for identifying differentially expressed genes. This study was not adequately powered to draw sharp lines between the performances of different classifiers.

Method A, which worked with all tumor samples or with stage I samples alone, both with and without clinical covariates, showed the best overall predictive ability. Method H also had good performance without clinical covariates. The genes in these classifiers may provide insight into the biology of aggressive tumors. Method A relied on the correlated expression of 100 gene clusters to predict subject outcome. Relatively higher expression of genes in cluster 6 of method A (545 genes) was associated with poor subject outcome. This cluster included cell proliferation–related genes, including cyclin A (*CCNA2*) and other cyclins, *BUB1B*, topoisomerases, checkpoint genes (*CHEK1*), and chromosomal and spindle protein genes. Method H also relied heavily on these genes for classification. This is consistent with elevated cell proliferation and loss of cell cycle control being associated with poor outcomes[7]. Greater expression of genes in cluster 4 of method A (262 genes), cluster 5 (82 genes) and cluster 12 (427 genes) was associated with better survival. Cluster 4 included several differentiation related genes such as thyroid transcription factor 1 (*NKX2-1*), pulmonary-associated surfactant protein B (*SFTPB*), as well as G protein–coupled receptor 116 (*GPR116*) and MAP3K12 binding inhibitory protein 1 (*MBIP*), whereas cluster 12 contained many genes related to immunological functions. This is consistent with tumors showing some aspects of recognition by the immune system having better outcomes[15]. The variety of genes found useful for classification suggests that several mechanisms contribute to the clinical progression of lung adenocarcinomas and that several classifiers may be equally effective.

This study provides a realistic assessment of the challenges in developing prognostic models for early-stage lung cancer. A significant degree of outcome prediction accuracy was observed using gene expression data alone, yet the hazard ratios for most of our models increased with the inclusion of clinical data (**Fig. 1**). Conversely, gene expression data improved the predictive performance of clinical parameters alone (method I; compared to method A, which used gene expression and clinical variables). We note that even this large study was not adequately powered to make comparisons between classification methods with high statistical confidence. Nevertheless, some interesting trends emerged. For the all-stage analysis, method I (clinical variables only) was competitive with most of the procedures using gene expression data without clinical variables, consistent with gene expression largely recapitulating stage. However, it is notable that method A with covariates performed substantially better on the CAN/DF samples than either method A without covariates or method I. In the stage 1 analysis, the clinical variables reduce to age and sex. In the MSK test set, these variables were uninformative about disease risk, so it is noteworthy that gene expression seemed to stratify subjects by risk in method A. The predictive performance of method I in the stage 1 CAN/DF test set was driven by a strong association with age. However, it is unclear how far this relationship will generalize. Therefore, an integrated approach using gene expression together with associated clinical, pathological and other information may be more promising for future work, as has previously been pointed out in studies examining prostate and breast cancer[16,17]. Although it is not possible to attribute to specific classifier properties the slightly better results across the hypotheses and test sets with method A compared to the other methods, we do note that method A did use substantially more genes than the other approaches and incorporated an initial gene-clustering procedure. These properties may have contributed to its more consistent performance. We have provided a detailed discussion of the challenges in using gene expression profiling for lung cancer prognosis in practice in the **Supplementary Results**. Our findings suggest that clinical covariates should be collected with the same care as used for obtaining gene expression signatures.

The present study was designed to address three key issues in the field of gene expression–based outcome prediction. First, this study provides, to our knowledge, the largest gene-expression data set with pathological and clinical annotation for lung adenocarcinomas so far. Because of the large sample size, further analyses of prognostic genes associated with specific histological subtypes, such as bronchioalveolar carcinomas, can now be undertaken. Further pathological and mutational annotation of each specimen is ongoing: https://caarraydb.nci.nih.gov/caarray/publicExperimentDetailAction.do?expId=1015945236141280 and this careful assessment should provide a valuable resource for hypothesis generation. Second, we used these data to test rigorously the current methodologies that predict tumor biology and, by inference, patient prognosis from gene expression signatures. Finally, we identified issues relevant to the use of gene expression profiles that should be taken into consideration in designing future studies. We had observed previously[12] that the biological variation between tumors exceeds the technical variation introduced by microarray analysis. We observed in this study that clinical covariates improved upon gene expression alone as a mechanism for stratifying tumor samples. We have also learned that coordinating the collection of clinical and pathological data across several institutions is an important task for prospective studies designed to further refine prognostic signatures. There are also limitations in using subject survival as an endpoint that may be overcome by using time to tumor recurrence as the primary endpoint instead. Although there still remain significant challenges to the use of gene expression–based classifiers in the clinical setting, the potential that these tools can improve patient care and increase survival provides a strong impetus to continue to refine these approaches for eventual clinical use.

METHODS

Investigator consortium. Four institutions (University of Michigan Cancer Center (UM), Moffitt Cancer Center (HLM), Memorial Sloan-Kettering Cancer Center (MSK) and the Dana-Farber Cancer Institute (CAN/DF)) formed a consortium with support and collaboration of US National Cancer Institute investigators to develop and validate gene expression signatures of lung adenocarcinomas. Details of the specimens, criteria for inclusion, clinical covariates collected, mRNA processing and hybridization are described in the **Supplementary Methods**. Consent was obtained for all subjects and the protocols approved by the respective Institutional Review Board of each institution. The CEL files for the study are available at https://caarraydb.nci.nih.gov/caarray/publicExperimentDetailAction.do?expId=1015945236141280. Links to the pathological and clinical data are also available at this site.

Training and validation sets. Initial evaluation of the gene expression data suggested that the data from the UM, HLM and MSK were broadly similar, although distinguishable, but the data from CAN/DF showed some systematic differences from the other three sites due mainly to reduced signal intensity. The CAN/DF set was also different in that it lacked stage 3 samples. To give a realistic evaluation of how a prognostic method might be used in practice, we decided to use the combined data from UM and HLM as the training set, with MSK held out as a similar but external validation set and the CAN/DF data held out as a second and more challenging external validation set.

Analysis protocol. We followed a strict protocol for analysis, with data for the two validation sets held by a third-party 'honest broker' during analysis of the training data. Risk scoring procedures were developed on the training data for four distinct hypotheses described in the paper. The available clinical variables were age, sex and American Joint Committee on Cancer stage. Prognostic models were developed on the training data by research groups at each of the four institutions, with each group submitting one or more candidate models for some or all of the four hypotheses defined above. After the models were defined and documented, the honest broker released the gene expression and clinical data (but not the outcome data) for the two validation data sets to the four groups, who used each candidate prognostic model to predict outcomes for each subject. A method was permitted to standardize gene expression levels within each test set or to refer to percentile points of summary features in a test set, but otherwise predictions were made for each test sample in isolation. Some models produced a continuous risk score for each subject, whereas others grouped the subjects into a finite number of ordered risk categories. These predictions were then passed back to the honest broker, allowing evaluation of the performance of the prognostic models. Results for all methods we considered are presented in this paper.

For performance evaluation, we used each predicted risk score as the covariate in a univariate Cox proportional hazards model, with overall survival (censored at 60 months) as the outcome variable. The continuous risk scores were standardized to have unit interquartile range to make the hazard ratios from the proportional hazards model comparable to each other and approximately comparable to those from binary predictors. The estimated hazard ratio and its 95% confidence interval and *P*-value (**Supplementary Results**) allowed us to directly compare the performances of different procedures on a unidimensional scale. For graphical representation, continuous risk scores were binned into tertiles and Kaplan-Meier estimates of the survivor function were plotted for each subgroup. This allowed for assessment of any 'dose-response' relationship and also facilitated graphical comparison between different predictors.

An alternative measure of performance was provided by the concordance probability estimate (CPE)[18]. The CPE estimates the concordance probability, which is the probability that, for a given pair of subjects selected at random from the study population, the subject with better prognosis has a better outcome. CPE values close to 0.5 indicate poor predictive accuracy and values approaching 1.0 indicate increasingly good predictive accuracy.

Finally, we constructed ROC curves for the continuous predictors and tables of sensitivity and specificity values for the categorical predictors. Sensitivity and specificity were calculated using Bayes' theorem and Kaplan-Meier estimates of the survivor function to appropriately handle censoring. Details are provided in **Supplementary Methods** and **Supplementary Results**.

Risk scoring procedures. A variety of strategies were used to construct prognostic models. All of the methods used an initial step to reduce the amount of data for final modeling of the outcomes. Detailed descriptions of each method are provided in **Supplementary Methods**.

Note: Supplementary information is available on the Nature Medicine website.

ACKNOWLEDGMENTS

We thank M. Orringer, A. Pickens, F. Taylor, N. Liu, D. Lau, M. Whitehead, L. Chen, L. Vargas, Y. Xiao, M. Maddaus and C. Hoang. We thank M. Heiskanen, L. Liu, D. Reeves and S. Whitley from the US National Cancer Institute Center for Bioinformatics and W. Ricker from Information Management Services for assistance with development of the lung study database and data management. We thank D. Sawyer, J.M. Askew and A. Vaughn of the Cancer and Leukemia Group B Statistical Center, Duke University for quality control of the clinical data. We thank Affymetrix for technical support. This work was supported by US National Cancer Institute grants CA84953, CA84999, CA84995, CA85052 and CA46592 and contracts 263-MQ-319735, 263-MQ-319740, 263-MQ-319746 and 263-MQ-510430 and support from the Canadian Cancer Society.

AUTHOR CONTRIBUTIONS

Writing Committee: K.S., J.M.G.T., S.A.E., M.S.T., T.J.Y., W.L.G., S.E., I.J., V.E.S., M.M., R.K., K.K.D., T.L., J.W.J. and D.G.B. Members of the Writing Committee participated in the planning, initiation, data generation, data analysis and manuscript preparation for the project.

Additional participants: T.J.G., D.E.M., A.C.C. and S.H. participated in aspects of sample collection and preparation, data generation and data analysis at the University of Michigan. C.Q.Z., D.S., F.A.S., K.D. and L.S. participated in aspects of sample collection and preparation, data generation and data analysis at the Ontario Cancer Institute. K.N., N.P., B.W., R.V., C.L.-A and T.G. participated in aspects of sample collection and preparation, data generation and data analysis at the Dana-Farber Cancer Institute and Broad Institute. M.G. assembled the clinical data at the H. Lee Moffitt Cancer Center. J.S., M.Z., V.R., M.K., A.V., N.M., W.T. and A.S. participated in aspects of sample collection and preparation, data generation and data analysis at Memorial Sloan-Kettering Cancer Center. B.C. participated in the planning and initiation of the study.

Published online at http://www.nature.com/naturemedicine/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Jemal, A. *et al.* Cancer Statistics 2006. *CA Cancer J. Clin.* **56**, 106–130 (2006).
2. Booth, C.M. & Shepherd, F.A. Adjuvant chemotherapy for resected non-small cell lung cancer. *J. Thorac. Oncol.* **2**, 180–187 (2006).
3. Gandara, D.R., Wakelee, H., Calhoun, R. & Jablons, D. Adjuvant chemotherapy of stage I non-small cell lung cancer in North America. *J. Thorac. Oncol.* **7**(suppl. 3), S125–S127 (2007).
4. Shepherd, F.A. *et al.* Erlotinib in previously treated non-small-cell lung cancer. *N. Engl. J. Med.* **353**, 123–132 (2005).
5. Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* **98**, 13790–13795 (2001).
6. Garber, M.E. *et al.* Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl. Acad. Sci. USA* **98**, 13784–13789 (2001).
7. Beer, D.G. *et al.* Gene-expression profiles predict survival of subjects with lung adenocarcinoma. *Nat. Med.* **8**, 816–824 (2002).
8. Wigle, D.A. *et al.* Molecular profiling of non–small cell lung cancer and correlation with disease-free survival. *Cancer Res.* **62**, 3005–3008 (2002).
9. Potti, A. *et al.* A genomic strategy to refine prognosis in early-stage non–small-cell lung cancer. *N. Engl. J. Med.* **355**, 570–580 (2006).
10. Chen, H.Y. *et al.* A five-gene signature and clinical outcome in non-small-cell lung cancer. *N. Engl. J. Med.* **356**, 11–20 (2007).
11. Lu, Y. *et al.* A gene expression signature predicts survival of subjects with stage I non-small cell lung cancer. *PLoS Med.* **12**, e467 (2006).
12. Dobbin, K.K. *et al.* Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clin. Cancer Res.* **11**, 565–572 (2005).
13. Fry, W.A., Phillips, J.L. & Menck, H.R. Ten-year survey of lung cancer treatments and survival in hospitals in the United States. *Cancer* **86**, 1867–1876 (1999).
14. Li, C. & Wong, W.H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **98**, 31–36 (2001).
15. Moran, C.J. *et al.* Rantes expression by lung adenocarcinomas is a predictor of survival in stage I subjects. *Clin. Cancer Res.* **8**, 3803–3812 (2002).
16. Stephenson, A.J. *et al.* Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer* **104**, 290–298 (2005).
17. Sotiriou, C. & Piccart, M.J. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to subject care? *Nat. Rev. Cancer* **7**, 545–553 (2007).
18. Gonen, M. & Heller, G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* **92**, 965–970 (2005).

[1]The consortium consists of the Writing Committee plus additional participants as detailed in the Author Contributions section. [2]Department of Statistics, 1085 South University, University of Michigan, Ann Arbor, Michigan 48109, USA. [3]Cancer Center, 1500 East Medical Center Drive, University of Michigan, Ann Arbor, Michigan 48109, USA. [4]Department of Biostatistics, 1420 Washington Heights, University of Michigan, Ann Arbor, Michigan 48109, USA. [5]Department of Surgery, H. Lee Moffitt Cancer Center and Research Institute, University of South Florida, 12902 Magnolia Avenue, Tampa, Florida 33612, USA. [6]University Health Network, Ontario Cancer Institute and Princess Margaret Hospital, 610 University Avenue, Toronto, Ontario M5G 2M9, Canada. [7]Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA. [8]Department of Pathology, University Hospital 2G332/0054, University of Michigan, Ann Arbor, Michigan 48109, USA. [9]Department of Surgery, 1150 West Medical Center Drive, University of Michigan, Ann Arbor, Michigan 48109, USA. [10]National Cancer Institute of Canada Clinical Trials Group and Queen's University, 10 Stuart Street, Kingston, Ontario K7L 3N6, Canada. [11]Department of Medical Oncology, Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts, 02115, USA. [12]Broad Institute, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [13]Department of Medicine, B-414 Clinical Center, Michigan State University, East Lansing, Michigan 48824, USA. [14]Columbia University, 722 West 168th Street, New York, New York 10032, USA. [15]Biometric Research Branch National Cancer Institute, EPN 8121A, 6130 Executive Boulevard, Rockville, Maryland 20852, USA. [16]Cancer Diagnosis Program, National Cancer Institute, EPN 6035A, 6130 Executive Boulevard, Rockville, Maryland 20852, USA. [17]Writing Committee members.