Prediction of Recurrence-Free Survival in Postoperative Non–Small Cell Lung Cancer Patients by Using an Integrated Model of Clinical Information and Gene Expression

Eung-Sirk Lee, Dae-Soon Son, Sung-Hyun Kim, et al.

Clin Cancer Res 2008;14:7397-7404.

Prediction of Recurrence-Free Survival in Postoperative Non–Small Cell Lung Cancer Patients by Using an Integrated Model of Clinical Information and Gene Expression

Eung-Sirk Lee,[1,3,7] Dae-Soon Son,[1,6] Sung-Hyun Kim,[1] Jinseon Lee,[1] Jisuk Jo,[1] Joungho Han,[2] Heesue Kim,[1] Hyun Joo Lee,[1] Hye Young Choi,[1] Youngja Jung,[1] Miyeon Park,[1] Yu Sung Lim,[1] Kwhanmien Kim,[3] Young Mog Shim,[3] Byung Chul Kim,[6] Kyusang Lee,[6] Nam Huh,[6] Christopher Ko,[5] Kyunghee Park,[6] Jae Won Lee,[5] Yong Soo Choi,[1,3] and Jhingook Kim[1,3]

Abstract

Purpose: One of the main challenges of lung cancer research is identifying patients at high risk for recurrence after surgical resection. Simple, accurate, and reproducible methods of evaluating individual risks of recurrence are needed.

Experimental Design: Based on a combined analysis of time-to-recurrence data, censoring information, and microarray data from a set of 138 patients, we selected statistically significant genes thought to be predictive of disease recurrence. The number of genes was further reduced by eliminating those whose expression levels were not reproducible by real-time quantitative PCR. Within these variables, a recurrence prediction model was constructed using Cox proportional hazard regression and validated via two independent cohorts ($n = 56$ and $n = 59$).

Results: After performing a log-rank test of the microarray data and successively selecting genes based on real-time quantitative PCR analysis, the most significant 18 genes had P values of <0.05. After subsequent stepwise variable selection based on gene expression information and clinical variables, the recurrence prediction model consisted of six genes (*CALB1, MMP7, SLC1A7, GSTA1, CCL19,* and *IFI44*). Two pathologic variables, pStage and cellular differentiation, were developed. Validation by two independent cohorts confirmed that the proposed model is significantly accurate ($P = 0.0314$ and 0.0305, respectively). The predicted median recurrence-free survival times for each patient correlated well with the actual data.

Conclusions: We have developed an accurate, technically simple, and reproducible method for predicting individual recurrence risks. This model would potentially be useful in developing customized strategies for managing lung cancer.

Non–small cell lung cancer (NSCLC) is an increasingly common and lethal disease, accounting for 25% of all cancer deaths (1). Even after surgical resection of early stage primary tumors, ~50% of patients die from tumor recurrence (2).

A principal challenge in lung cancer research is identifying patients at high risk for recurrence after surgical resection, as well as patients who would benefit from adjuvant treatment (3–5). The current tumor-node metastasis staging system, which is based on clinical and pathologic findings, may have reached the limit of its usefulness (3–8). Individualized risk evaluation, rather than prediction based on broad population cohorts, would be helpful in the quest to customize treatment strategies for specific patients.

Recent molecular studies, such as those involving gene expression profiling, have shown that various inherent and acquired genetic alterations influence prognosis (9). However, genetic and epigenetic factors have not been considered for use in general clinical practice, much less for use in formulating treatment modalities. Thus, there is a need for new methodologies that are applicable to clinical practice (3–5, 7, 8).

When developing clinically applicable methods that rely on genetic factors to predict prognosis, several points should be considered. First, the outcomes should be expressed as time-dependent continuous variables [e.g., time-to-recurrence (TTR) or survival duration], not as mere binary events (e.g., alive/dead or recurrence/nonrecurrence). Many epidemiologic studies

Authors' Affiliations: [1]Cancer Research Center, Center for Clinical Research, Samsung Biomedical Research Institute and [2]Department of Pathology and [3]Department of Thoracic Surgery, Samsung Medical Center, College of Medicine, Sungkyunkwan University; [4]Department of Statistics, Korea University, Seoul, South Korea and [5]Bio & Health Lab and [6]Emerging Center, Samsung Advanced Institute of Technology; and [7]Department of Thoracic Surgery, Hando General Hospital, Gyunggi-do, South Korea

Received 11/21/07; revised 6/16/08; accepted 6/18/08.

Grant support: Bio & Health Lab, Samsung Advanced Institute of Technology.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: E-S. Lee and D-S. Son contributed equally to the study.

Requests for reprints: Jhingook Kim, Department of Thoracic Surgery, Samsung Medical Center, College of Medicine, Sungkyunkwan University, Seoul 135-230, South Korea. Phone: 82-2-3410-3489; Fax: 82-2-3410-0089; E-mail: jkimsmc@skku.edu.

doi:10.1158/1078-0432.CCR-07-4937

Translational Relevance

One of the most important aspects in lung cancer research is identifying patients who are at risk for recurrence. Clinical predictors for the prognosis of patients, such as age, gender, histologic features, and current tumor-node metastasis staging system may have reached the limit of their usefulness. So, we developed a practical gene expression analysis method for the risk prediction, using time-to-recurrence for gene selection and model development, to complement the current clinical practice. By stepwise variable selection of the gene expression information obtained from microarray data with significant level of 5% and the conventional clinical variables, an integrative recurrence prediction model was formulated consisting of pathogenetic variables, such as the selected six genes, pStage, and cell differentiation. In effect, it has been shown that the sensitivity of recurrence within 2 years is 72.1% and the specificity of nonrecurrence for 5 years is 90.0%, supporting that this prediction model for the recurrence risk provides an improvement in personalized prediction compared with the currently available ones and can help to identify those at high risk of developing postresection recurrences, as well as those beneficial from adjuvant treatments. This model needs to be improved further, but lays the basic foundation of a customized lung cancer management strategy to be used in day-to-day clinical practice.

have relied on such binary events because of the difficulties inherent in processing numerous variables.

Second, gene expression data should be "assayable" or, in other words, measurable in the range detected by real-time quantitative PCR (QPCR). Microarray analysis has made it possible to simultaneously measure the expression of thousands of genes (3, 4, 10–12). However, analysis of cumulative microarray data is impractical in the clinical setting. Real-time QPCR is an alternative and supplementary method that consists of a simple, efficient, and effective assay (13, 14).

Third, clinical and pathologic findings remain important variables in the development of clinically applicable methods for prognosis prediction. Traditional high-risk factors, which are based on clinical and pathologic findings (e.g., advanced stage of illness, high-level involvement of lymph nodes, tumor mass size, and presence of nonsquamous cancer), certainly have prognostic value and play significant roles in treatment planning. Therefore, gene expression profiling should be evaluated in light of any possible synergies with recognized clinical prognostic factors (4, 8).

In this study, we developed a gene expression analysis method that is appropriate for the clinical setting. Data used to develop the model were gathered by analyzing microarray studies of 138 patients and selecting genes that were expressed at statistically significant levels according to TTR. We next eliminated genes for which expression levels were not reproducible by real-time QPCR. The resulting genes were used to construct a recurrence prediction model for each patient, whereby the usefulness of this model in clinical practice was shown in two independent validation cohorts.

Materials and Methods

Patient population and sample selection. A total of 818 frozen tumor samples were collected from patients who underwent curative resection for NSCLC at Samsung Medical Center between January 1995 and December 2005. One or two pieces from the periphery of the tumor masses—avoiding necrotic regions—were immediately frozen at -80°C until retrieval. Frozen tissues were selected for RNA extraction after the application of exclusion criteria (Materials and Methods in Appendix A).

Tissue preparation and RNA extraction. Selected frozen tissues (n = 406) were stained with H&E to improve visualization. Necrotic tumor tissues and intervening normal tissues were removed as previously described (15). Extraction and determination of mRNA were done according to the manufacturers' protocols (easy-spinTM total RNA extraction kit, iNtRon Biotechnology; Nanodrop Technologies; Agilent RNA 6000 nano kit, Agilent Technologies, Inc.).

A total of 253 frozen tissues with acceptable RNA quality (i.e., RNA integrity number values of over 6.5) were used for the experiment. Oligonucleotide microarray experiments were done using higher quality RNA (i.e., RNA integrity number value of over 8.0, n = 138) with sufficient follow-up duration (i.e., frozen tissues collected from 1995 to 2003, with a median of 44 months of follow-up) as the "modeling cohort." Clinical information is shown in Table 1.

To validate our proposed model, RNAs of acceptable quality (i.e., RNA integrity number values = 8.0-6.5, n = 56) of the same duration described above were assigned to validation cohort I, whereas RNA from patients with relatively short follow-up periods (i.e., frozen tissues collected from 2004 to 2005 with a median of 33 months of follow-up, n = 59) were assigned to validation cohort II. Clinical information from the two validation cohorts is shown in Table 1. An outline of the study is shown in Fig. 1.

Microarray experiments and real-time QPCR. Total RNAs from 138 patients were used to analyze mRNA expression profiles on a DNA microarray platform, according to the standard Affymetrix U133 Plus 2.0 protocol with 17 RNAs as duplicate for quality control (correlation coefficient, 0.988). Our entire microarray data set is available online[8] under the data series accession number GSE8894.

To ensure consistency of the mRNA expression levels observed in the microarray experiments, the same set of RNA samples (n = 116, because 22 RNAs were present in insufficient amounts) was used in the development of the real-time QPCR-based model (Materials and Methods in Appendix A).

Patient follow-up data, including recurrence information, are shown in Table 1. The routine follow-up schedule for postoperative outpatients is described in the Materials and Methods in Appendix A. The last month for follow-up was June 2007, and follow-up for patients who died was complete at the time of death. Patient participation was consistent with protocols approved by the institutional review board of Samsung Medical Center, and written informed consent was obtained (IRB 2005-12-034).

Statistical analyses. Microarray data were preprocessed using GCRMA normalization with PM-MM data. Genes were selected using a log-rank test as a robust nonparametric method. The pool of genes was then reduced via real-time QPCR analysis by screening out "assayable" genes that showed ~100-fold differences, on average, between the percentiles of 0 to 3 and 97 to 100 in the microarray analysis. Among these genes, we selected 87 with unique cDNA sequences (based on a search of the public database) and with Universal Probe Library (UPL) and primers sequences that were identical to the 3′-mRNA probe used in the microarray experiments

[8] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8894

Table 1. Clinical characteristics of the modeling cohort and validation cohorts I and II

		Modeling cohort (n = 138)		Validation cohort I (n = 56)		Validation cohort II (n = 59)	
Age (y)		60.68 ± 9.67	(13~82)	57.89 ± 9.68	(34~74)	62.44 ± 10.36	(42~90)
Sex	Male	103	74.6%	45	80.4%	50	84.7%
	Female	35	25.4%	11	19.6%	9	15.3%
Smoking Hx (PY)	None	38	27.5%	11	19.6%	18	30.5%
	~20	13	9.4%	5	8.9%	3	5.1%
	20~50	53	38.4%	22	39.3%	27	45.8%
	50~	25	18.1%	15	26.8%	4	6.8%
	Unknown	9	6.5%	3	5.4%	7	11.9%
Cell type	ADC	63	45.7%	24	42.9%	25	42.4%
	SQC	75	54.3%	32	57.1%	34	57.6%
Differentiation	WD	20	14.5%	9	16.1%	3	5.1%
	MD	80	58.0%	25	44.6%	42	71.2%
	PD	29	21.0%	14	25.0%	12	20.3%
	unknown	9	6.5%	8	14.3%	2	3.4%
Size (cm)		4.38 ± 1.95	(1.5~14)	4.82 ± 2.41	(1.3~13)	4.03 ± 1.69	(1.7~10)
Stage	I a	16	11.6%	3	5.4%	7	11.9%
	I b	72	52.2%	18	32.1%	28	47.5%
	II a	6	4.3%	1	1.8%	1	1.7%
	II b	18	13.0%	17	30.4%	9	15.3%
	III a	16	11.6%	11	19.6%	11	18.6%
	III b	10	7.2%	6	10.7%	3	5.1%
T	1	24	17.4%	5	8.9%	9	15.3%
	2	94	68.1%	32	57.1%	43	72.9%
	3	10	7.2%	13	23.2%	4	6.8%
	4	10	7.2%	6	10.7%	3	5.1%
N	0	98	71.0%	33	58.9%	39	66.1%
	1	28	20.3%	13	23.2%	9	15.3%
	2	12	8.7%	10	17.9%	11	18.6%
Median F/U (mo)		44	(3~138)	44	(3~146)	33	(3~43)
Median RFS time		49.38	(0.93~138.48)	33.93	(2.03~145.80)	.	(1.8~43.20)
Recurrence	Local	13	9.4%	9	16.1%	8	13.6%
	Distant	33	23.9%	13	23.2%	11	18.6%
	Local + Distant	15	10.9%	8	14.3%	3	5.1%
	Lung and pleura	32	45.7%	23	50.0%	8	34.8%
	Bone	20	28.6%	13	28.3%	6	26.1%
	Abdominal organ	10	14.3%	7	15.2%	4	17.4%
	Brain	6	8.6%	3	6.5%	3	13.0%
	Neck lymph node	2	2.9%	0	0.0%	2	8.7%
Nonrecurrence		77	55.8%	26	46.4%	37	62.7%

Abbreviations: PY, pack year; F/U, follow-up; RFS, recurrence-free survival. Age values are mean ± SD.

(2). Construction of models using either 10, 20, or 30 genes with the lowest q values revealed that the 20-gene model was the most accurate with less number of genes. The mRNA expression of these genes, as determined via microarray and real-time QPCR, correlated well (correlation coefficients, 0.55-0.90; Table 2). Details of the real-time QPCR analysis are described in the Materials and Methods in Appendix A.

To construct a Cox proportional hazard regression model for prognosis prediction, we used seven clinical variables (i.e., pStage, cell type, differentiation, smoking history, tumor size, gender, and age) and 18 real-time QPCR-assayable genes (Table 2). The number of combined variables was reduced to eight by stepwise selection (i.e., entry, 0.1; remove, 0.2). Subsequently, a model developed via combination of these eight clinicogenomic factors was applied to the independent validation cohorts. The calculated Cox score was transformed into survival function $[S(t)]$ by SAS software to predict recurrence in individual patients.

Results

Gene selection and modeling of recurrence prediction. Initially, 3,112 genes with P values of <0.05 were selected from a log-rank test of microarray data from 54,675 probe sets and further subjected to multiple testing adjustments using the false discovery rate method for the rigorous selection of significant genes. Within these selected genes, we prepared two different gene sets by (a) selecting 20 genes with the smallest q values and (b) selecting 20 assayable genes whose levels of expression were measurable by real-time QPCR. Cox proportional hazard regression was then done on the two gene sets.

When samples in the first set were divided into two groups at the median Cox score (i.e., -9.28), the survival curves for the group with Cox scores higher than the median were significantly separated from curves for the group with Cox scores lower than the median (P < 0.0001; Appendix A Fig. 2A). This result indicated that the model successfully separated the two risk groups. The selected genes and the constructed model were statistically validated using leave-one-out cross-validation.

Within the second gene set (i.e., 20 assayable genes, as defined in the Materials and Materials and Methods; Table 2), the K-M survival curve did not differ significantly from the

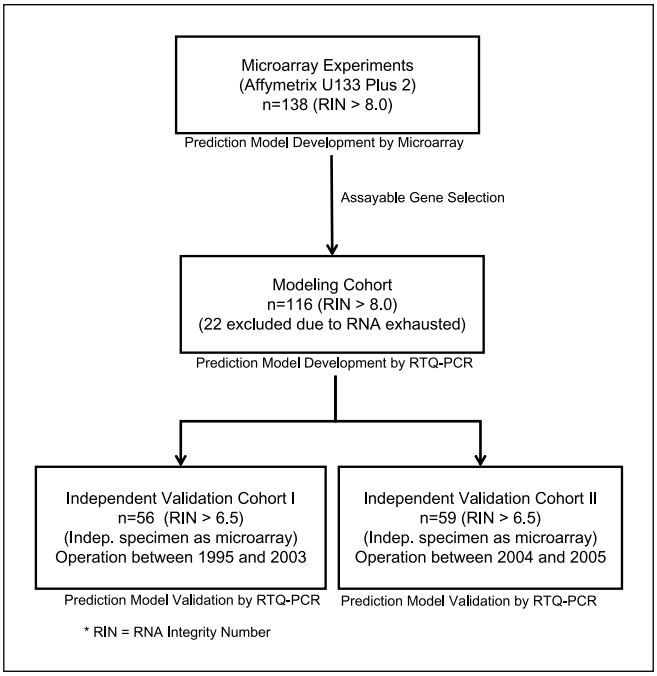

Fig. 1. Development and validation of the recurrence prediction model. The log-rank test of microarray data from 138 patients generated 3,112 genes ($P < 5\%$) or 1,091 genes ($P < 1\%$). Among these, 20 assayable genes with the lowest P values were used in the development of a model that was validated with two independent cohorts.

curve for the 20 genes with the smallest q values (Appendix A Fig. 1B). Time-dependent receiver operating characteristic (ROC) analysis (16) confirmed the performance of both models, with area under the curves values of >0.8 (Appendix A Fig. 1C). Thus, we defined the "gene-based model" as one that consists of assayable genes.

Correlation between microarray and real-time QPCR data. *LOC402110* and *C14orf78* were not included in the model development process because of insufficient RNA levels. Results for the remaining 18 genes were verified by correlating microarray and real-time QPCR data (i.e., correlation coefficients, 0.55-0.90; Table 2).

Comparison of clinical, genomic, and clinicogenomic models for prognosis prediction. We developed a prediction model from the modeling cohort ($n = 116$) using seven selected clinical variables. This clinical model gave a P value of 0.0518 in the modeling cohort (Appendix A Fig. 2A), and P values of 0.2644 and 0.7204 in validation cohorts I and II, respectively (Fig. 2A and B). Next, we developed a gene-based prediction model from the same modeling cohort. This genomic model provided more significant results ($P < 0.0001$; Appendix A Fig. 2B), and P values were 0.0710 and 0.1333 in the validation cohorts I and II, respectively (Fig. 2C and D).

To construct a recurrence prediction model that is applicable to clinical practice, we sought to include both clinical variables and genetic information. Through the stepwise selection of variables in both classes, two clinical variables (i.e., pStage and "degree of differentiation" of the tumor cells) and six genes (i.e., *CALB1*, *MMP7*, *SLC1A7*, *GSTA1*, *CCL19*, and *IFI44*) were incorporated. This clinicogenomic model was more effective at separating the recurrence risk group from others ($P < 0.0001$; Appendix A Fig. 2C) and provided more statistically significant results in validation cohorts I and II ($P = 0.0314$ and $P = 0.0305$, respectively; Fig. 2E and F). Time-dependent ROC analysis further verified the superior performance of the clinicogenomic model (i.e., area under the curve, >0.75) within the independent validation cohorts compared with the clinical

Table 2. List of assayable 20 genes and microarray to PCR correlation

No.	Probe set ID	Gene title	Gene symbol	q value	Correlation
1	219271_at	UDP-*N*-acetyl-α-D-galactosamine: polypeptide *N*-acetyl galactosaminyl transferase 14 (GalNAc-T14)	*GALNT14*	0.0003	0.7065
2	227952_at	Hypothetical LOC402110	*LOC402110*	0.0006	
3	201641_at	Bone marrow stromal cell antigen 2	*BST2*	0.0012	0.6452
4	205626_s_at	Calbindin 1, 28 kDa	*CALB1*	0.0024	0.7344
5	204259_at	Matrix metallopeptidase 7 (matrilysin, uterine)	*MMP7*	0.0042	0.8127
6	243623_at	Solute carrier family 1 (glutamate transporter), member 7	*SLC1A7*	0.0042	0.8272
7	219956_at	UDP-*N*-acetyl-α-D-galactosamine: polypeptide *N*-acetyl galactosaminyl transferase 6 (GalNAc-T6)	*GALNT6*	0.0043	0.5472
8	205267_at	POU domain, class 2, associating factor 1	*POU2AF1*	0.0043	0.7667
9	225342_at	Adenylate kinase 3-like 1	*AK3L1*	0.0065	0.6931
10	211748_x_at	Prostaglandin D2 synthase 21 kDa (brain)///prostaglandin D2 synthase 21 kDa (brain)	*PTGDS*	0.0071	0.7926
11	212992_at	Chromosome 14 open reading frame 78	*C14orf78*	0.0084	
12	210809_s_at	Periostin, osteoblast specific factor	*POSTN*	0.0123	0.6765
13	203924_at	Glutathione *S*-transferase A1	*GSTA1*	0.0125	0.8963
14	210072_at	Chemokine (C-C motif) ligand 19	*CCL19*	0.0135	0.7193
15	211668_s_at	Plasminogen activator, urokinase///plasminogen activator, urokinase	*PLAU*	0.0137	0.8087
16	205513_at	Transcobalamin I (vitamin B12 binding protein, R binder family)	*TCN1*	0.0156	0.6906
17	214453_s_at	IFN-induced protein 44	*IFI44*	0.0157	0.7171
18	213664_at	Solute carrier family 1 (neuronal/epithelial high affinity glutamate transporter, system Xag), member 1	*SLC1A1*	0.0159	0.7604
19	219869_s_at	Solute carrier family 39 (zinc transporter), member 8	*SLC39A8*	0.0162	0.6961
20	223551_at	Protein kinase (cyclic AMP–dependent, catalytic) inhibitor β	*PKIB*	0.0163	0.627

NOTE: q values were calculated by false discovery rate method using log-rank test P value.

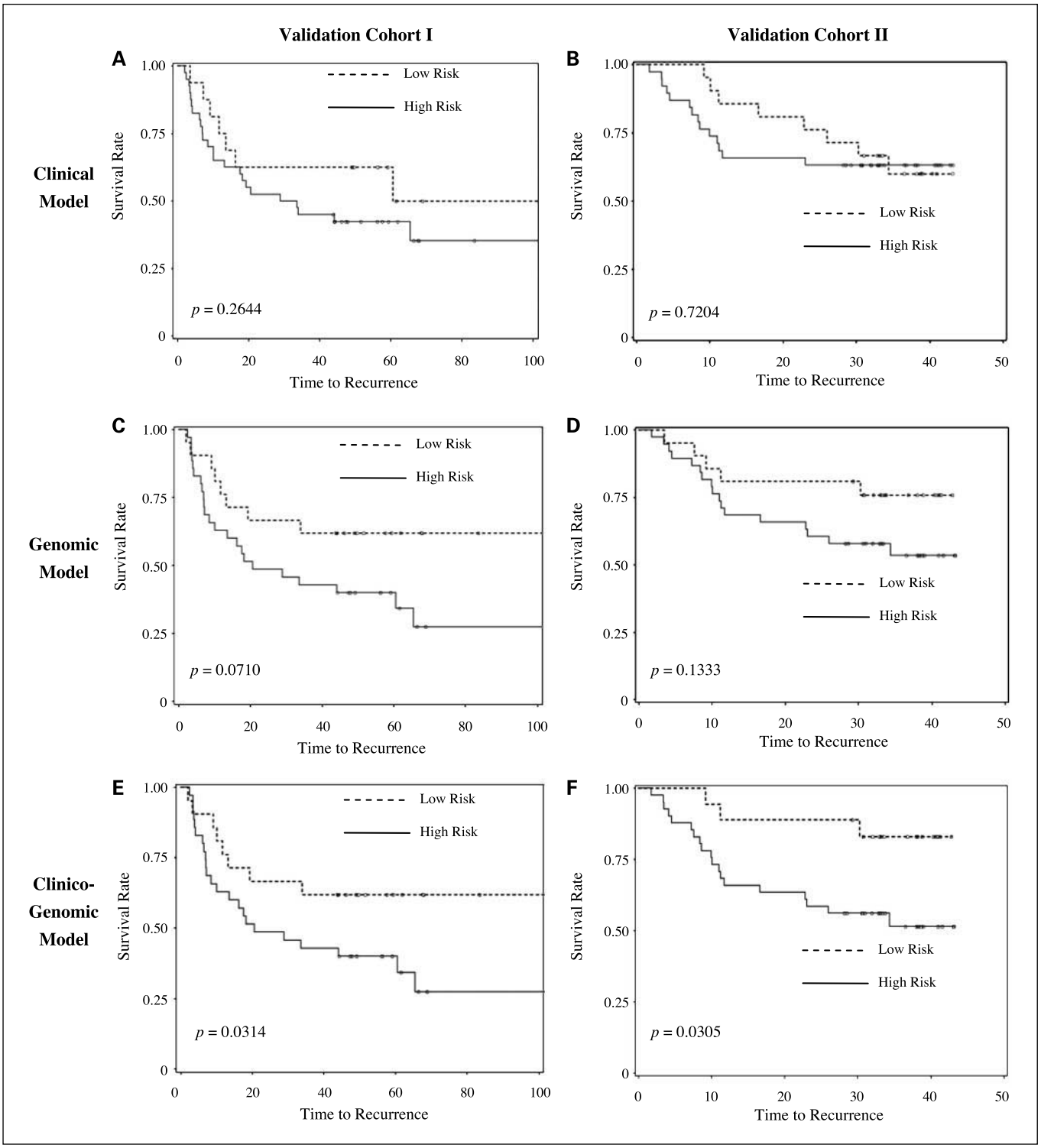

Fig. 2. Comparison of clinical, genomic, and clinicogenomic model. For the clinical model with the modeling cohort (*A* and *B*) and validation cohorts I and II. For the genomic model (*C* and *D*) and for the clinicogenomic model (*E* and *F*) with the two validation cohorts, respectively.

or genomic models alone (Appendix A Fig. 3). Cox score equations for the three models are listed in Model Equations in Appendix A.

Recurrence-free survival predictions for individual patients. Using the Cox score calculated from the model, individualized Kaplan-Meier recurrence-free survival graphs (K-M graphs) were plotted to predict individual survival and calculate median recurrence-free time (Fig. 3; Appendix A Table 2).

Our clinicogenomic model gave different recurrence predictions for individual patients even when patients had similar

clinical findings. Many patients with short predicted median recurrence-free survival times (PMRFS) developed recurrences soon after surgery, whereas many patients with long PMRFS survived >5 years of follow-up without recurrence. The PMRFS for each patient correlated well with the actual survival time (Fig. 3; Appendix A Table 2).

The overall accuracy of this predictive model was calculated from the relationship between the PMRFS and the actual TTR for independent validation cohorts I and II (n = 115). The sensitivity of recurrence within 2 years and the specificity of nonrecurrence for 5 years were 72.1% and 90.0%, respectively (Appendix A Fig. 4).

Discussion

Fundamental knowledge of the molecular bases for and the consequent molecular classifications of NSCLC could aid in the prediction of patient outcomes. Gene expression studies that focus on cancer require a high-throughput technology, such as the microarray, to enable simultaneous analysis of tens of thousands of genes. The introduction of microarrays has inspired several approaches to the analysis of gene expression data related to patient survival (17–25). Conventional analytic methods have incorporated two to three classes of division in models for risk prediction. However, recurrence is not a fixed variable (e.g., man/woman), but rather a continuous, time-dependent variable. Such limitations prompted us to consider other ways of integrating TTR information into a gene-based risk prediction model. Several approaches have been proposed for obtaining TTR information from microarray data, such as the Median Cut Approach (26), Supervised Principal Component Analysis (26), Lasso-Cox Approach, or CASPAR algorithm (27). Particularly, when the Median Cut Approach is used, patients with short follow-up periods are classified into high-risk groups. Such drawbacks might be overcome by using log-rank tests that select significant genes based on information regarding the continuous time-dependent property of recurrence. We used this strategy to generate 3,112 genes with P values of <0.05.

Microarray analysis is a highly promising technique, but its use in daily clinical practice is limited because of the necessary benchwork, equipment, analytic methods, and technical precision needed to generate reproducible results (3, 4, 6, 9, 28–39). Nevertheless, we used microarray analysis to (*a*) obtain the overall gene expression profile from which genes of interest (particularly target genes) were selected. Once we identified the target genes, we were able to (*b*) confirm the microarray data via real-time QPCR analysis of a finite number of genes and (*c*) construct a prediction model based on the results of the real-time QPCR analysis, which can be obtained within hours of mRNA preparation. Real-time QPCR analysis provides a quantitative measurement of mRNA in tumor tissues and a sensitive measurement of fold changes in the expression of individual genes. Furthermore, real-time QPCR is a simple method that can generate highly reproducible results in routine clinical settings (13, 14, 29, 40–43).

Real-time QPCR-based models have been described in the literature; however, high degrees of correlation between microarray data and real-time QPCR data have not been shown (22, 25). Gene selection is the most important step in the

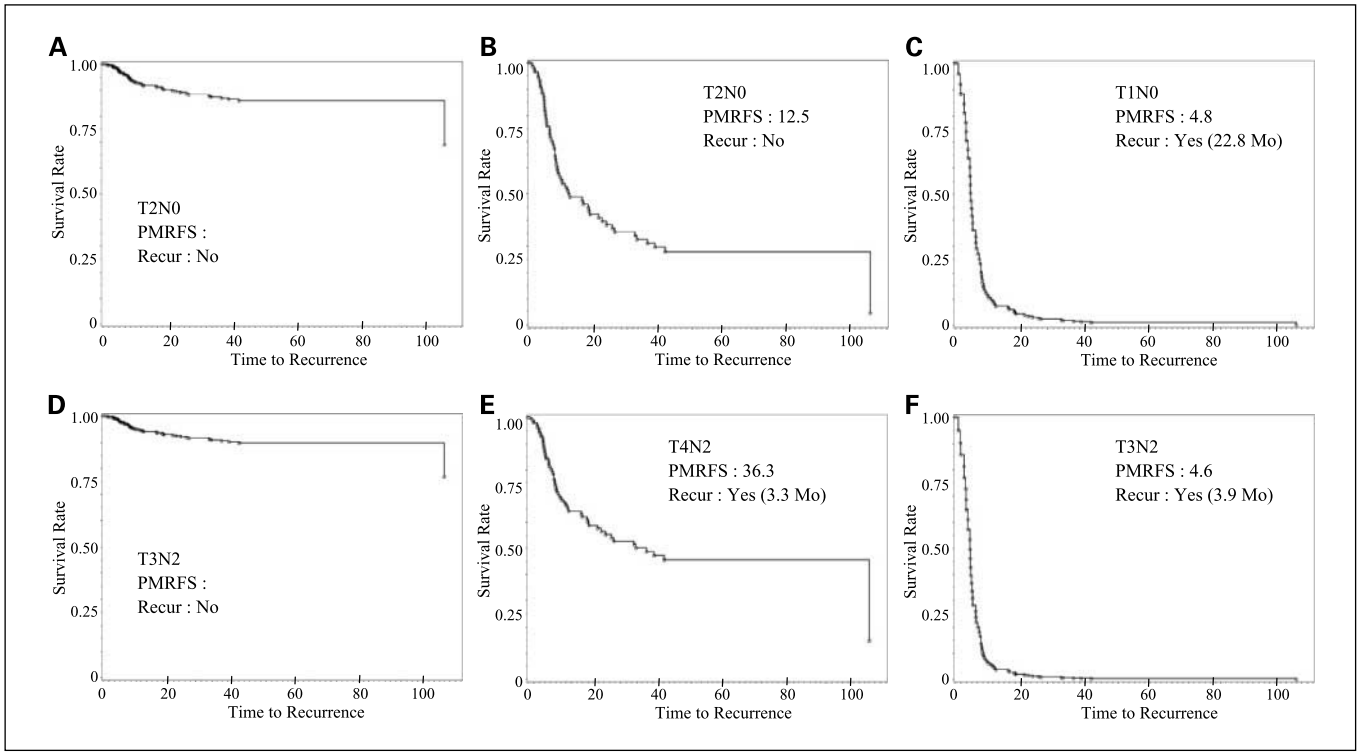

Fig. 3. Kaplan-Meier recurrence-free survival curves of individualized prediction. Exemplary six cases at two different pathologic stage: pStage I (patients a, b, and c) and pStage III (patients d, e, and f). Patients classified into the same pStage gave different recurrence predictions according to our recurrence prediction model. MRFS, median recurrence-free survival time. In cases when Kaplan-Meier curve does not reach below 50%, "." was used for MRFS.

development of a real-time QPCR-based model. Based on our experience and that of others (44), mRNA levels of a common housekeeping gene can vary between patients by as much as 10-fold. Construction of a model using real-time QPCR-based data requires that differences in the expression of component genes are large enough to overcome background noise from differences in diagnostic technology and variation in the expression of housekeeping genes. We therefore chose assayable genes from the pool of microarray data that showed a minimum of a 100-fold difference in gene expression between recurrence and nonrecurrence patients. This step enhanced subsequent detection of gene expression changes via real-time QPCR.

In this study, we measured gene expression levels in 138 tumor specimens using the Affymetrix U133 Plus 2.0 microarray with 54,675 probes. We then constructed a model and validated it using two independent cohorts (n = 56 and n = 59) to overcome issues concerning sample size and number of probes (3, 4, 10–12, 38, 45, 36).

Recently, gene expression profiles have been compared with conventional clinical predictors, such as age, gender, tumor stage, and histologic features, to determine the accuracy of predicting prognoses of patients with NSCLC (17–25). It was not clear whether a gene expression profile might be an independent specific prognostic factor in itself or whether such a profile might be influenced by other prognostic factors (4). In 2006, Potti et al. showed that the genomic model adds prognostic value to the widely accepted clinical predictors of outcome (e.g., age, gender, tumor size, stage of disease, histologic subtype, and smoking history; ref. 23). In the current study, we constructed a model using both clinical and genomic prognostic factors that fundamentally differed in metagene construction and binary prediction tree analysis (23). Future studies should explore the biological significance of each gene used in our predictive model.

Recent advances in genetics, molecular medicine, and biotechnology foretell that increasingly complex biomedical and molecular information will soon underlie a coherent system of personalized medicine. Cancer diagnosis, treatment planning, and customized adjuvant therapy will focus on stratified or individual patients rather than on broader population cohorts. Many risk prediction studies divide a test group into two subgroups of poor or good prognosis and evaluate the outcomes of each subgroup. However, many patients might prefer individualized clinical decisions, because the random assignment of a patient into one or the other subgroup may not reflect a true prognosis.

To address this issue, we constructed recurrence-free survival curves for individual patients, which were validated by the predicted PMRFS and by the actual TTR of independent validation cohorts (Fig. 3; Appendix A Table 2). Of 115 validation cohorts, three patients with each of two types of pathologic stage were chosen: pStage I (i.e., patients a, b, c) or pStage III (i.e., patients d, e, f). Although patients in the same pStage had similar clinical findings, our clinicogenomic model gave different PMRFS (Fig. 3), so that the PMRFSs of the patients with pStage I disease were over 120, 12.5, and 4.8 months, respectively, and those of the patients at pStage III were over 120, 36.3, and 4.6 months, respectively. Patients c and f, who had short PMRFS, suffered recurrences soon after surgery. However, patients a and d, whose survival did not fall under 50% on the K-M graph, survived >5 years of follow-up without recurrence. Relationships between the PMRFS and the actual TTR of cohorts I and II (n = 115) are shown in Appendix A Fig. 4. The sensitivity of recurrence within 2 years and the specificity of nonrecurrence for 5 years were found to be 72.1% and 90.0%, respectively. Thus, the PMRFS for each patient correlated well with the actual data.

Our results suggest that the proposed model enhances the current system of personalized recurrence risk prediction. Although further improvement is needed, our model allows a customized lung cancer management strategy that can be used in day-to-day clinical practice.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank Matthew Meyerson, Roel Verhaak, Sanghyuk Lee, and Won Ki Kang for useful comments suggestions, and discussions.

References

1. Parkin DM, Bray FI, Devesa SS. Cancer burden in the year 2000. The global picture. Eur J Cancer 2001;37 Suppl 8:S4–66.
2. Hoffman PC, Mauer AM, Vokes EE. Lung cancer. Lancet 2000;355:479–85.
3. Pollack JR. A perspective on DNA microarrays in pathology research and practice. Am J Pathol 2007;171:375–85.
4. Sun Z, Yang P. Gene expression profiling on lung cancer outcome prediction: present clinical value and future premise. Cancer Epidemiol Biomarkers Prev 2006;15:2063–8.
5. Gandara DR, Lara PN, Lau DH, Mack P, Gumerlock PH. Molecular-clinical correlative studies in non-small cell lung cancer: application of a three-tiered approach. Lung Cancer 2001;34 Suppl 3:S75–80.
6. Jaluria P, Konstantopoulos K, Betenbaugh M, Shiloach J. A perspective on microarrays: current applications, pitfalls, and potential uses. Microb Cell Fact 2007;6:4.
7. Schmidt U, Begley CG. Cancer diagnosis and microarrays. Int J Biochem Cell Biol 2003;35:119–24.
8. Kaprio J. Science, medicine, and the future. Genetic epidemiology. BMJ 2000;320:1257–9.
9. Yang P, Sun Z, Aubry MC, et al. Study design considerations in clinical outcome research of lung cancer using microarray analysis. Lung Cancer 2004;46:215–26.
10. Sarkar IN, Planet PJ, Bael TE, et al. Characteristic attributes in cancer microarrays. J Biomed Inform 2002;35:111–22.
11. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. Nat Genet 1999;21:10–4.
12. Lander ES. Array of hope. Nat Genet 1999;21:3–4.
13. Shi L, Tong W, Fang H, et al. Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. BMC Bioinformatics 2005;6 Suppl 2:S12.
14. Mackay IM, Arden KE, Nitsche A. Real-time PCR in virology. Nucleic Acids Res 2002;30:1292–305.
15. Choi N, Son DS, Lee J, et al. The signature from messenger RNA expression profiling can predict lymph node metastasis with high accuracy for non-small cell lung cancer. J Thorac Oncol 2006;1:622–8.
16. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 2000;56:337–44.
17. Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A 2001;98:13790–5.
18. Garber ME, Troyanskaya OG, Schluens K, et al. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci U S A 2001;98:13784–9.
19. Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816–24.
20. Tomida S, Koshikawa K, Yatabe Y, et al. Gene expression-based, individualized outcome prediction for surgically treated lung cancer patients. Oncogene 2004;23:5360–70.
21. Sun Z, Yang P, Aubry MC, et al. Can gene expression profiling predict survival for patients with squamous cell carcinoma of the lung? Mol Cancer 2004;3:35.

22. Blackhall FH, Wigle DA, Jurisica I, et al. Validating the prognostic value of marker genes derived from a non-small cell lung cancer microarray study. Lung Cancer 2004;46:197–204.

23. Potti A, Mukherjee S, Petersen R, et al. A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. N Engl J Med 2006;355:570–80.

24. Raponi M, Zhang Y, Yu J, et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer Res 2006;66: 7466–72.

25. Chen HY, Yu SL, Chen CH, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med 2007;356:11–20.

26. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol 2004;2:E108.

27. Kaderali L, Zander T, Faigle U, Wolf J, Schultze JL, Schrader R. CASPAR: a hierarchical bayesian approach to predict survival times in cancer from gene expression data. Bioinformatics 2006;22:1495–502.

28. Ach RA, Floore A, Curry B, et al. Robust interlaboratory reproducibility of a gene expression signature measurement consistent with the needs of a new generation of diagnostic tools. BMC Genomics 2007;8:148.

29. Barbacioru CC, Wang Y, Canales RD, et al. Effect of various normalization methods on Applied Biosystems expression array system data. BMC Bioinformatics 2006;7:533.

30. Bosotti R, Locatelli G, Healy S, et al. Cross platform microarray analysis for robust identification of differentially expressed genes. BMC Bioinformatics 2007; 8 Suppl 1:S5.

31. Carter SL, Eklund AC, Mecham BH, Kohane IS, Szallasi Z. Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. BMC Bioinformatics 2005;6:107.

32. Centeno BA, Enkemann SA, Coppola D, Huntsman S, Bloom G, Yeatman TJ. Classification of human tumors using gene expression profiles obtained after microarray analysis of fine-needle aspiration biopsy samples. Cancer 2005;105:101–9.

33. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A 2006; 103:5923–8.

34. Gordon GJ, Jensen RV, Hsiao LL, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res 2002;62:4963–7.

35. Hoffmann R, Seidl T, Dugas M. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. Genome Biol 2002;3:RESEARCH0033.

36. Nimgaonkar A, Sanoudou D, Butte AJ, et al. Reproducibility of gene expression across generations of Affymetrix microarrays. BMC Bioinformatics 2003;4:27.

37. Shi L, Reid LH, Jones WD, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 2006;24:1151–61.

38. Tamayo P, Scanfeld D, Ebert BL, Gillette MA, Roberts CW, Mesirov JP. Metagene projection for cross-platform, cross-species characterization of global transcriptional states. Proc Natl Acad Sci U S A 2007;104:5959–64.

39. Warnat P, Eils R, Brors B. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. BMC Bioinformatics 2005;6:265.

40. Arya M, Shergill IS, Williamson M, Gommersall L, Arya N, Patel HR. Basic principles of real-time quantitative PCR. Expert Rev Mol Diagn 2005;5: 209–19.

41. Canales RD, Luo Y, Willey JC, et al. Evaluation of DNA microarray results with quantitative gene expression platforms. Nat Biotechnol 2006;24:1115–22.

42. Wilhelm J, Pingoud A. Real-time polymerase chain reaction. Chembiochem 2003;4:1120–8.

43. Wong ML, Medrano JF. Real-time PCR for mRNA quantitation. Biotechniques 2005;39:75–85.

44. Saviozzi S, Cordero F, Lo Iacono M, Novello S, Scagliotti GV, Calogero RA. Selection of suitable reference genes for accurate normalization of gene expression profile studies in non-small cell lung cancer. BMC Cancer 2006;6:200.

45. Ioannidis JP, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG. Genetic associations in large versus small studies: an empirical assessment. Lancet 2003; 361:567–71.

46. Ntzani EE, Ioannidis JP. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Lancet 2003;362:1439–44.