Lung cancer is the deadliest form of cancer in the United States, and it causes more deaths each year than colon, breast, and prostate cancer combined. The deaths of television news anchor Peter Jennings and Christopher Reeve’s widow, Dana Reeve, have recently put lung cancer in the national spotlight. Using lung cancer as an example, we illustrate important statistical issues in discovering and validating high throughput cancer biomarkers for predicting a cancer patient’s risk of cancer recurrence.
Lung Cancer Is a Heterogeneous Disease

Jennings was an older male with a long history of smoking, while Reeve was a younger female who never smoked. Yet, the difference among lung cancer patients can be even more striking than the apparent differences of its famous patients; lung cancer is a common term for several heterogeneous diseases.

The heterogeneity of lung cancer patients lies at the levels of histology, pathological stage, molecular characteristics, and genetics. Accurate classification of lung cancer patients into subtypes is the key for effective clinical management and disease prognosis. In terms of histology, the vast majority of lung cancers are carcinomas—malignancies that arise from epithelial cells. There are two main types of lung carcinoma: non-small cell (80%) and small-cell (17%) lung carcinoma. The non-small cell lung carcinoma can be further classified into three main sub-types: squamous cell lung carcinoma, adenocarcinoma, and large-cell lung carcinoma.

Besides histology, cancer staging is the most important factor in deciding a patient's prognosis and treatment. The current staging system uses information about tumor size, lymph nodes, and metastasis to assess whether cancer has spread from a primary site and, if so, how extensive the spread is. For non-small cell lung cancer (NSCLC), for example, stage I cancer is confined to lung tissue, stage II cancer is confined to lung tissue and lymph nodes in the lung, stage III cancer is found in lung tissue and lymph nodes outside the lung, and stage IV cancer has distant spread to other sites, such as the liver, glands, bone, and brain. Patients with advanced NSCLC (stages III and IV) have a five-year survival rate of less than 5%, as compared to 50% for patients with stage I NSCLC.

In current clinical practice, lung cancer patients who have the same histology and pathological stage may receive the same chemo- and radiation therapy. However, a substantial proportion of treated patients do not respond to the recommended treatment. It is now believed that lung cancer is indeed a combination of hundreds of distinct diseases. The heterogeneous response to treatment is due to the differences in patients' genetic makeup, and the most effective treatment is tailored to each patient.

In recent years, targeted agents have been developed to turn a specific genetic pathway on or off. The development of targeted lung cancer therapy has generated much interest in discovering markers to identify the subgroup of patients who benefit most from these therapies. The key questions are what biomarker reveals the distinct patient's genetic profile, how can such a biomarker be developed, and how effective is the biomarker in classifying patients into different subgroups of genetic profiles?

Lung Cancer Biomarkers

What exactly is a cancer biomarker? According to the 2001 NIH Biomarker Definitions Working Group, a biomarker is "a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention." A robust and reproducible biomarker that identifies patients with unique genetic or molecular features lays the foundation for personalizing treatment for lung cancer patients. One may categorize biomarkers into the following three groups, depending on their intended use in cancer treatment:

1. **Prognostic biomarkers**, which predict the outcome of patients in terms of a clinical endpoint. A validated prognostic biomarker provides opportunity to identify patients at high risk and a possibility for early intervention. For example, several genomic-based classifiers have been developed recently to predict the risk of cancer recurrence for stage I NSCLC patients following surgical resection, and a randomized clinical trial is planned to evaluate whether chemotherapy would benefit high-risk patients.

2. **Predictive biomarkers**, which predict the effect of a specific treatment on a clinical endpoint for patients. As an example, over-expression of Cyclo-oxygenase-2 (COX-2) is associated with poor prognosis of overall survival in advanced NSCLC, and patients with over-expressed COX-2 benefited significantly more from receiving celecoxib (a COX-2 inhibitor) and standard chemotherapy relative to those receiving standard chemotherapy only. In other words, COX-2 serves as a biomarker that is both prognostic and predictive.

3. **Surrogate biomarkers**, which replace a clinical endpoint in clinical trials carried out to evaluate the effect of a specific treatment on patients. Surrogate biomarkers can be used as intermediate indicators of treatment efficacy in cancer treatment studies. For example, standardized uptake values (SUV) calculated from fluorodeoxyglucose–positron emission tomography (FDG-PET) are believed to measure tumor metabolism, and a decrease in SUV is being evaluated as a surrogate endpoint for progression-free survival in a clinical trial for advanced non-small cell lung cancer.

Depending on the nature of a specific biomarker (prognostic, predictive, or surrogate), a lung cancer biomarker can be useful for risk stratification, early cancer detection, treatment selection, prognostication, or monitoring. In recent years, there have been significant advances in our understanding of the molecular and genetic changes involved in lung carcinogenesis, including circulating DNA, genetic mutations, gene hypermethylation, gene expression, and proteins. Meanwhile, there have also been considerable technological and scientific advances in fields such as gene expression profiling, proteomics, and molecular imaging. Progress as such has allowed efficient characterization of the changes underpinning disease progress and drug response. We are in an unprecedented time to discover and validate biomarkers and the corresponding targeted agents for lung cancer.

Development of Lung Cancer Biomarkers

To illustrate some important statistical issues in discovery and validation of lung cancer biomarkers, we use as an example the development of a genomics-based prognostic biomarker for predicting cancer recurrence risk in stage I NSCLC patients.