Phenotypic Heterogeneity of Potentially Curable Non–Small-Cell Lung Cancer

Cohort Study with Cluster Analysis

Lama Sakr, MD, MSc, FRCPC, David Small, MD, FRCPC,* Goulnar Kasymjanova, MD, PhD,† Samy Suissa, PhD,‡§ and Pierre Ernst, MD, MSc, FRCPC*‡§*

Introduction: Significant differences in outcome are observed among lung cancer patients belonging to the same tumor node metastasis stage, suggesting phenotypic heterogeneity beyond this staging algorithm. We used a cluster analysis approach to classify patients into distinct phenotypes, and we attempted to validate the clinical relevance of these phenotypes by comparing outcome.

Methods: We formed a cohort of all stage I to III non–small-cell lung cancer patients seen between January 2004 and October 2010 in a cancer center and followed until death or last follow-up appointment, with prospectively collected data on clinical and tumor characteristics. Multiple correspondence analysis was followed by hierarchical clustering to form homogenous clusters of patients. Overall survival and disease-free survival estimates were compared among clusters.

Results: The cohort included 367 patients (mean follow-up of 2.5 years), 173 of whom died during that period (191 deaths per 1000 person-years). A four-cluster model was identified, revealing distinct phenotypes with respect to baseline characteristics. Hazard ratios for mortality were 8.1, 5.0, and 3.7 (all statistically significant) for clusters 2, 1, and 3, respectively, when compared with cluster 4—with the most favorable outcome.

Conclusion: Staging of patients with non–small-cell lung cancer for prognostic purposes may be improved by considering phenotypes that exhibit significant differences in clinical course and outcome.

*Division of Pulmonary Diseases, Department of Medicine, Jewish General Hospital, McGill University, Quebec, Canada; †Peter Brojde Lung Cancer Centre, Jewish General Hospital, Montreal, Quebec, Canada; ‡Department of Epidemiology and Biostatistics, McGill University, Quebec, Canada; and §Center for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, McGill University, Montreal, Quebec, Canada

Disclosure: No conflicts of interest exist for any of the authors of this manuscript.Prior abstract presentation/publication: An abstract of the present study has been submitted and accepted for presentation at the American Thoracic Society Meeting in Philadelphia in May 2013.

This work was supported by Fonds de recherche du Québec—Santé (FRQS) Master's Training Award for Applicants with a Professional Degree (Dr. Sakr), and a Clinical Research Award from the Jewish General Hospital (Dr. Sakr).

Address for correspondence: Lama Sakr, MD, MSc, FRCPC, Division of Pulmonary Diseases, Department of Medicine, Jewish General Hospital, McGill University, Pavilion G, Room 203, 3755 Cote-Ste-Catherine Road, Montreal, Quebec H3T 1E2, Canada. E-mail: lama.sakr@mail.mcgill.ca

DOI: 10.1097/JTO.0000000000000505

ISSN: 1556-0864/15/1005-0754

Key Words: Non–small-cell lung cancer, Cluster analysis, Phenotypic heterogeneity.

(*J Thorac Oncol.* 2015;10: 754–761)

Lung cancer is the leading cause of cancer-related deaths in Canada, in both men and women, accounting for over one-quarter of all cancer deaths.[1] Non–small cell lung cancer (NSCLC) is the diagnostic entity in the vast majority of cases. Up to now, the only firmly established prognostic tools are the tumor node metastasis (TNM) staging system and performance status (PS).[2,3] However, the potential of those two prognostic factors to accurately predict overall survival (OS) and disease-free survival (DFS) is limited. In fact, there are significant interindividual differences with respect to disease course and outcome, even within the same TNM stage. Such findings are indicative of the phenotypic heterogeneity of NSCLC, beyond the TNM staging system. Identifying patients with a higher likelihood of poor outcome would allow optimization of patient care by offering individually tailored treatment strategies.

Defining clinical phenotypes of NSCLC with specific patterns of disease presentation, based on a simultaneous analysis of various demographic, clinical, pathologic, molecular, and genetic features, would be an interesting and novel methodology that has never been performed in lung cancer patient cohorts, to the best of our knowledge. Such an approach may be done, using statistical techniques, such as cluster analysis. The latter is a multivariate statistical technique that has been successfully used to classify asthma patients into phenotypes with distinctive clinical, physiologic, and pathologic parameters.[4–7] We hypothesized that NSCLC has distinct phenotypes, each being associated with a different clinical presentation, course, and outcome.

MATERIALS AND METHODS

A review of prospectively collected data from consecutive patients with stages I to III NSCLC, diagnosed between January 2004 and December 2010 and followed at a single tertiary care referral center (Jewish General Hospital, Montreal, Canada), was undertaken. The population analyzed

in this study consisted of patients with a diagnosis of potentially curable, clinically or pathologically proven stages I to III NSCLC. Patients were excluded if they entered the cohort with a second primary, recurrent disease, or distant metastatic involvement (stage IV NSCLC).

We attempted to collect missing data, by reviewing electronic and/or paper medical records, and contacting staff at the Pathology Laboratory and Nuclear Medicine Division.

Cluster and Multiple Correspondence Analysis Methodology

We employed two complementary statistical techniques to define specific phenotypes within our patient cohort: multiple correspondence analysis (MCA) and cluster analysis. MCA is a principal component analysis method that assigns numerical scores to subjects and categories of categorical variables and captures the relative associations among them,[8] by generating a graphical display of categories as data points in a high-dimensional "Euclidian" space. This technique, therefore, allows a reduction in the number of variables needed to summarize the data. The MCA output provides several dimensional solutions; the optimal solution is selected on the basis of the smallest number of dimensions that would account for the largest total explained variance. A set of continuous coordinates is generated, and a continuous similarity measure—the Euclidian distance—is used to determine similarity between observations. A linkage method is then applied to determine distance between clusters or groups of observations, so similar clusters can be merged together. Ward's linkage—using analysis of variance as the distance measure between clusters—was applied. A graphical illustration of the clustering algorithm results in a tree hierarchical diagram or a "dendrogram," displaying the number of likely clusters in a given cohort. Cutting this dendrogram at various levels divides study subjects into a variety of partitions. Once the optimal number of clusters is determined, vertical box plots are then generated to illustrate the disposition of the described clusters, according to specified dimensions. Each dimension related to each corresponding cluster is associated with a particular coordinate score. The two dimensions that best identify each specific cluster, by optimally distinguishing it from other groups, are chosen. Subsequently, MCA plots—consisting of two-dimensional scatter plots displaying each category of each variable as a data-point in space—are generated for each of the clusters, using the best two dimensions previously identified. A definition is, therefore, provided for each of the clusters, according to the position of each category for every corresponding variable on the axis of both dimensions. All analyses were performed using STATA12 statistical software package.

MCA was implemented by choosing 10 variables, consisting of demographic, clinical, pathologic, and metabolic parameters, found—after our review of the literature—to be potentially prognostic in patients with NSCLC. All variables were categorical, either ordinal or nominal, except for age and standardized uptake value (SUV) on positron emission tomography–computed tomography (PET–CT). Such variables were converted to categorical parameters for the purpose of MCA. The variable "age" was subdivided into six categories. As for the SUV, we set the threshold at the median value as previously described.[9] Such threshold divides patients into subjects having a tumor with "high" metabolism or SUV and those with a tumor of "low" SUV. The variable representing "Eastern Cooperative Oncology Group Performance Status" (ECOG PS)—a five-point scale measuring an individual's level of functioning and capability of self-care—was converted to a three-level parameter, with the following categories: PS of "0," PS of "1"—including those whose PS was entered in the database as "0–1" or "1" or "missing"—and a PS of "2 or more," which comprised subjects with a PS of "1–2" or "2 or more." A "missing" category was generated for each variable with missing data. When a given variable contained only few missing observations, such missing data were included in the category with the largest number of observations. Missing categories and observations were taken into account and included in the analysis.

OS and DFS time for each of the identified clusters was estimated from the date of diagnosis to the date of death or disease recurrence, using the Kaplan–Meier method. When neither death nor disease recurrence occurred, data were censored at the date of last follow-up appointment. Cox's proportional hazard modeling was carried out to compare OS and DFS estimates between clusters, before and after adjusting for the variable age. Hazard ratios (HR) for survival estimates were deemed statistically significant when the 95% confidence interval (CI) did not include one, and when p-value was less than 0.05.

We also performed multivariable Cox regression model analyses, separately for OS and DFS with "stage," categorized to three levels (I, II, and III) as the main factor, adjusting for gender, weight loss, PS, differentiation, histology, thyroid transcription factor 1 (TTF-1) staining, SUV, and smoking status. Cross-validation of the Cox regression models was done to assess validation and calibration, by computing Somer's Dxy correlation coefficient and the slope shrinkage, with the bootstrap method.[10] To perform such analysis, we used the R software.[11,12]

RESULTS

Subject Demographics

The initial data set included a cohort of 382 subjects consisting of all consecutive patients diagnosed between January 2004 and December 2010. Fifteen patients were subsequently excluded for the following reasons: recurrent disease at initial entry in the cohort (10 patients), final surgical tissue histology only revealing squamous dysplasia (one patient) or squamous cell carcinoma in situ (one patient), mixed histological entity of sarcomatoid carcinoma and poorly differentiated adenocarcinoma (ADK) (one patient), a second metachronous primary tumor (one patient), and missing and irretrievable important data (one patient). The final study population thus consisted of 367 subjects. Subjects' baseline characteristics are shown in Table 1.

