Use of Multiple Imputation to Correct for Bias in Lung Cancer Incidence Trends by Histologic Subtype

Mandi Yu¹, Eric J. Feuer¹, Kathleen A. Cronin¹, and Neil E. Caporaso²

Abstract

Background: Over the past several decades, advances in lung cancer research and practice have led to refinements of histologic diagnosis of lung cancer. The differential use and subsequent alterations of nonspecific morphology codes, however, may have caused artifactual fluctuations in the incidence rates for histologic subtypes, thus biasing temporal trends.

Methods: We developed a multiple imputation (MI) method to correct lung cancer incidence for nonspecific histology using data from the Surveillance, Epidemiology, and End Results Program during 1975 to 2010.

Results: For adenocarcinoma in men and squamous in both genders, the change to an increasing trend around 2005, after more than 10 years of decreasing incidence, is apparently an artifact of the changes in histopathology practice and coding system. After imputation, the rates remained decreasing for adenocarcinoma and squamous in men, and became constant for squamous in women.

Conclusions: As molecular features of distinct histologies are increasingly identified by new technologies, accurate histologic distinctions are becoming increasingly relevant to more effective “targeted” therapies, and therefore, are important to track in patients. However, without incorporating the coding changes, the incidence trends estimated for histologic subtypes could be misleading.

Impact: The MI approach provides a valuable tool for bridging the different histology definitions, thus permitting meaningful inferences about the long-term trends of lung cancer by histologic subtype. Cancer Epidemiol Biomarkers Prev; 23(8); 1–13. ©2014 AACR.
Classification of Diseases for Oncology (ICD-O). In the 1990s, pathologists tended not to report NSC carcinomas with specificity because their treatments and prognoses were considered similar, thus an increasing number of cases are coded with 8010 (carcinoma, NOS) since 1980. In recognition of this trend, 8046 (NSC carcinoma) was added into ICD-O-3 in 2001 to group cases that could not be classified beyond the exclusion of small cell. Collectively, the percentage of cases coded with 8010 or 8046 increased dramatically, from 5% in 1982 to more than 22% in 2005 (11). Some of these cases could have been derived from one of the specific histologic subtypes, which would have subsequently reduced their incidence rates. However, this increasing use of nonspecific codes did not continue. In light of the advances in cancer research and therapy, increasingly NSC cases have been diagnosed with more histologic specificity (12) over the last few years, which may have driven up the rates for squamous or adenocarcinoma. Such differential use of nonspecific morphology codes could bias the estimated temporal trends of histologic subtypes and complicate interpretation. Appropriate statistical adjustments are necessary to improve the quality of inferences using the authoritative cancer registry data, which otherwise has been compromised by the unavoidable limitations imposed by the imperfect earlier classification system.

Multiple imputation (MI) has been shown to be a useful approach for handling measurement or coding changes for settings both in the presence (13–16) and absence (17) of calibration data (observations that are measured in all measurement scales or coding systems). When calibration data (usually on a random subsample) are available, one can generate plausible values in all measurement scales from an imputation model and analyze the imputed data using the preferred scale. For the issue associated with the change in the use of nonspecific morphology codes, that is 8010 and 8046, 2 types of calibration data could be useful for correcting coding inconsistency. The first type comprises cancer cases that are originally assigned to a nonspecific code, but are updated with a specific code through reexamination. Such data provide information about the association between nonspecific and specific histologies that one can use to recover the missing histology for all cases with a nonspecific code. Because nonspecific codes no longer exist in the imputed data, the trend analysis of incidence by histology is valid (provided that the imputation model is correct). The second type consists of cancer cases coded in multiple classification systems. Using these data as a bridge, one can convert data from one system to another. Although nonspecific codes still exist, temporal comparisons of imputed histology in any classification system is valid because coding consistency is maintained. However, neither type of calibration data could be easily obtained because of practical reasons, such as budget constraints and the lack of diagnostic data sources. Thus, this problem becomes a missing data issue where the specific histologies for cases with a nonspecific morphology code are missing and an assumption about the association between the missing specific histology and observed data (18–20) is required. We make a reasonable assumption that for cancer cases with similar tumor, treatment, survival, patients’ demographic characteristics, the distribution of nonspecific and specific histology is similar. Based on this assumption, we developed an MI approach using the sequential regression imputation method (SRMI; ref. 21) to redistribute cases without specific histology to one of specific subtypes, thus correcting the biased estimates of incidence rates.

Materials and Methods

Data sources


Data analysis

We treated the cases with 8010 or 8046 as missing data that we dealt with by MI (22). This MI approach took each case with missing histology and imputed it with a specific histologic subtype. Cases coded with 8010 were imputed with one of the 5 carcinoma subtypes, that is small cell, squamous, adenocarcinoma, large cell, and other NSC. For cases coded with 8046, the imputation was limited to one of the NSC subtypes, that is excluding small cell. This process was repeated independently 10 times to create 10 completed datasets to account for imputation uncertainty. Age-adjusted incidence rates (using the 2000 U.S. standard population in 19 age groups) were estimated from each completed dataset in the same way as using the original dataset, thus producing 10 sets of estimates. We then combined these estimates to produce MI estimates.
For a single incidence rate, the MI point estimate was the average of 10 imputed data estimates. The associated standard error was calculated by combining the average of the squared standard errors of the 10 estimates and the variance of the 10 rate estimates (22). Joinpoint linear regression models (23) were used to fit connected linear trends on a log scale with up to 4 joinpoints using the Joinpoint regression program version 3.5.0 developed by the NCI. Annual percentage change (APC) with a corresponding 95% confidence interval (CI) was calculated to describe each joined trend.

**Imputation method**

The nonspecific histologic diagnoses are highly likely to have nonrandom characteristics. For example, patients may not merit further histologic diagnostic procedures because they have diseases too advanced to permit curative surgery (i.e., stage IIIIB or greater) or because their medical status preclude surgery or other modalities with curative intent. When surgery is not a clinical option, obtaining adequate tissue to establish a histologic subtype may be impossible and, in this circumstance, clinicians may elect to forgo further histologic classification. Therefore, we considered using the information that is predictive of histology and the missingness of specific histology to recover the incomplete specific histology. We assumed the missingness is random conditional on this information, and this assumption has been shown to be reasonable in most practical situations (24, 25).

Specifically, we selected the covariates to be included in the imputation model following the principle of reducing missing data bias in a statistical analysis (26). Sociodemographic covariates include age, gender, race, Hispanic origin, nativity, and marital status. Covariates describing tumor characteristics and treatment include tumor size (27), grade, stage, survival time, and receipt of cancer-directed surgery. Certain therapies have shown to be more responsive in some histologic subtypes, thus making them important predictors. However, such information can only be made available for patients 65 years and older through the linked SEER-Medicare database (28) for 1991 and later. Considering the lack of analytics tools to handle the dynamics of the availability and access to particular regimen over time and patients’ age, we did not include more detailed treatment variables in the model. We also did not include lymph node involvement in the final model because it is highly collinear with stage. We included a nominal variable of 9 SEER registries to reflect the variability among registries in the use of nonspecific morphology codes. Cancer diagnosis year was entered into the models as a nominal variable (instead of a continuous variable) to relax the temporal assumption about the intervariable relationships. Smoking and socioeconomic deprivation are also strongly predictive of histology (29), but they are not routinely collected in SEER. To substitute, we used county-level smoking prevalence estimates obtained from the Model-based Small Area Estimates Projects of NCI (http://sae.cancer.gov/; ref. 30), and poverty prevalence estimates from the 2000 U.S. Census Bureau (31).

Because missing histology cannot be imputed for cases that are associated with missing covariates using simple regression-based imputation approaches, we developed an algorithm using SRMI technique to deal with multivariate missing data with arbitrary missing patterns. Specifically, SRMI fits a conditional model for each variable at a time on the remaining variables sequentially for multiple rounds to achieve convergence. The form of conditional model depends on the type of variable imputed. Our algorithm offers 2 new capacities beyond what is available in existing SRMI-based imputation packages, such as IVEware (http://www.isr.umich.edu/src/smp/ive/) and MICE (http://cran.r-project.org/web/packages/mice/index.html). First, for imputing binary data (categorical variables with more than 2 levels can be expressed as a series of nested dummy variables), we used ridge-penalized logistic regressions (32, 33) to improve imputation precision in the presence of binary outcome with skewed distribution and highly correlated covariates (34). The standard approach for imputing missing binary data is usually based on a logistic regression model (21, 35). However, the adequacy of logistic models could highly depend upon the extent to which the binary outcome is balanced and there is an absence of collinearity. In the presence of either condition or both at the same time, logistic regression coefficients may still be unbiased, but the precision could be very low, which could lead to poorly imputed data. The proposed approach improves the imputation by estimating a penalized log likelihood to obtain coefficients estimates with minimum prediction errors. Optimizing the penalty parameters is critical and usually requires intensive cross-validation studies (36). We follow the simplified approach proposed by Yu (34) and obtain the optimized parameters directly from the data by estimating the unrestricted log likelihood. The remaining steps are similar to those when standard logistic models are used (21). Second, we added a module to impute discrete right-censored survival data. For the data we chose for this study, more than 25% of survival time was censored because the patient was still alive at the end of study or died from other causes. Because both survival and censoring are highly correlated with histology as well as other covariates such as age, stage, tumor size, and grade, it is problematic to use relatively simple approaches, such as the indicator method where censoring is taken care of by including a censoring indicator (37, 38). The proposed method applies the MI principle to impute the censored time with a plausible future survival time. Specifically, to generate the imputed values, we first aggregate continuous survival time (in month) into several meaningful categories and sort them in an increasing order of survival. We then define an imputing risk set for each censored case as the cases with observed survivals no shorter than the censoring time. Using data from this imputing risk set, we finally estimate the predictive conditional distributions of survival...
categories, from which we randomly draw a value to be the imputed survival. Note that the possible value of an imputed survival is always equal to or longer than the censoring time category. This is reasonable because a censored case could only die at a later time in its own survival category or be still alive and die at a future category, but not die at a past category. This imputation process starts with censored cases in the first survival category and cycles through all categories to complete one imputed survival data. Because the survival is now a discrete variable, we estimate its predictive conditional distribution using nested ridge-penalized logistic models similar to what we have outlined for categorical data. Furthermore, to deal with the inconsistency in stage definitions over time, we conducted the imputation separately for 1975 to 1982, 1983 to 1987, and 1988 to 2010, so that staging is comparable within each period.

**Simulation study**

To explore information recovery from the MI in estimating the distribution of histology, we generated a simulated dataset from the analysis data with only complete observations included ($n = 10,659$). We considered a situation similar to the main analysis where histology is missing at random and the probability of the induced missingness is determined by a logistic regression model with the coefficients estimated using the analysis data. The rate of induced missing data was 8.4% (the observed missing rate was 10.0% for the portion of data with all covariates observed). Twenty imputed datasets were generated using the proposed approach and the standard logistic regression method, respectively.

The ridge-penalized logistic regression model outperformed the standard logistic regression model in recovering the missing information based on the Akaike information criterion (AIC; ridge-penalized method: AIC = 31,008 and standard method: AIC = 31,065). The imputed distributions of histology obtained using the proposed method were similar to the complete data distribution (with absolute difference less than 2% in estimating the percentage of cases in each histology and gender group). We also calculated the overlap probability (39) to evaluate how much the associated 95% CI estimated from the imputed and complete data overlap. Suppose ($L_{imp}$, $U_{imp}$) and ($L_{com}$, $U_{com}$) are the 95% CIs for estimating $P$, the percentage of adenocarcinoma among men, using the imputed and complete data, respectively. The probability overlap in the CIs for $P$ is $I = \frac{1}{2} \int_{L_{imp}}^{U_{imp}} f_{imp}(t)dt + \frac{1}{2} \int_{L_{com}}^{U_{com}} f_{com}(t)dt$, where $f_{imp}$ and $f_{com}$ are the distributions of $P$ computed under the imputed and complete data, respectively. Note that $f_{com}$ could take a different form of distribution depending on the type of statistics for which one wish to obtain estimates, but $f_{imp}$ is always $t$-distributed according to Rubin’s rules (22). $I$ takes value 0.95 if 2 CIs overlap perfectly and 0 if they do not overlap at all. A large value in $I$ suggests that the imputed data highly maintains the analytical properties of the complete data. This measure provides more information than a simple comparison of 2 point estimates by also considering the standard errors. Estimates with large standard errors might still have a high CI overlap even if their point estimates differ considerably from each other because the CI will increase with the standard error of the estimate. In this simulation study, most overlap probabilities (for estimating the distributions of cases by histology and gender) were more than 0.8, which suggested a very strong agreement, with a few exceptions in which the probabilities were around 0.75, which still suggested a strong agreement. These evaluation results provided

### Table 1. The numbers and percentages of lung cancer cases by histologic type and histologic confirmation status, SEER 9<sup>a</sup>, 1975 to 2010

<table>
<thead>
<tr>
<th>Histologic confirmation status (column%)</th>
<th>Overall $[n = 522,416 (100.0%)]$</th>
<th>Confirmed $[n = 470,326 (90.0%)]$</th>
<th>Not confirmed $[n = 38,657 (7.4%)]$</th>
<th>Unknown $[n = 13,433 (2.6%)]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small cell carcinoma</td>
<td>14.4 (15.7)</td>
<td>15.7 (7.0)</td>
<td>1.7 (0.3)</td>
<td>3.4 (0.6)</td>
</tr>
<tr>
<td>NSC carcinoma</td>
<td>68.5 (75.2)</td>
<td>75.2 (7.0)</td>
<td>7.0 (2.3)</td>
<td>8.9 (2.3)</td>
</tr>
<tr>
<td>Squamous</td>
<td>22.6 (24.8)</td>
<td>24.8 (7.0)</td>
<td>1.9 (2.3)</td>
<td>2.3 (2.3)</td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>32.1 (35.3)</td>
<td>35.3 (7.0)</td>
<td>3.1 (2.3)</td>
<td>4.0 (2.3)</td>
</tr>
<tr>
<td>Large-cell</td>
<td>5.6 (6.2)</td>
<td>6.2 (7.0)</td>
<td>0.3 (2.3)</td>
<td>0.6 (2.3)</td>
</tr>
<tr>
<td>Other specified NSC</td>
<td>3.1 (3.4)</td>
<td>3.4 (7.0)</td>
<td>0.2 (2.3)</td>
<td>0.3 (2.3)</td>
</tr>
<tr>
<td>8046 (NSC carcinoma)</td>
<td>5.1 (5.5)</td>
<td>5.5 (7.0)</td>
<td>1.5 (2.3)</td>
<td>1.7 (2.3)</td>
</tr>
<tr>
<td>8010 (carcinoma, NOS)</td>
<td>12.5 (7.6)</td>
<td>7.6 (7.0)</td>
<td>61.5 (43.0)</td>
<td>43.0 (44.7)</td>
</tr>
<tr>
<td>Other specified and unspecified types</td>
<td>4.6 (1.4)</td>
<td>1.4 (7.0)</td>
<td>29.8 (44.7)</td>
<td>44.7 (44.7)</td>
</tr>
</tbody>
</table>

Abbreviation: NOS, not otherwise specified.

<sup>a</sup>The SEER 9 registries include Atlanta, Connecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco–Oakland, Seattle–Puget Sound, and Utah.
Table 2. Distribution of histologically confirmed lung cancer cases by histology and selected covariates, SEER 9th, 1975 to 2010

<table>
<thead>
<tr>
<th>Overall</th>
<th>Small cell</th>
<th>Squamous</th>
<th>Adeno-carcinoma</th>
<th>Large cell</th>
<th>Other specified NSC</th>
<th>8010 (carcinoma, NOS)</th>
<th>8046 (NSC carcinoma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>463,609 (100.0%)</td>
<td>73,994 (100.0%)</td>
<td>116,775 (100.0%)</td>
<td>166,006 (100.0%)</td>
<td>29,123 (100.0%)</td>
<td>15,914 (100.0%)</td>
<td>35,954 (100.0%)</td>
<td>25,843 (100.0%)</td>
</tr>
</tbody>
</table>

**Age**
- <50 y: 6.5, 5.4, 7.8, 8.5, 14.6, 6.3, 5.7
- 50–60 y: 17.9, 19.5, 19.1, 20.6, 19.7, 16.4, 16.6
- 60–70 y: 32.4, 35.3, 31.5, 33.0, 30.5, 30.4, 26.8
- 70–80 y: 31.2, 30.3, 29.6, 28.4, 25.9, 32.3, 32.7
- ≥80 y: 12.0, 9.4, 12.3, 9.5, 9.4, 14.6, 18.2

**Sex**
- Male: 59.8, 56.1, 71.4, 53.4, 63.0, 53.2, 62.1, 55.8
- Female: 40.2, 43.9, 28.6, 46.6, 37.0, 36.8, 37.9, 44.2

**Race**
- White: 84.0, 88.2, 83.3, 83.1, 84.6, 86.1, 79.5, 83.0
- Black: 10.4, 7.8, 12.1, 9.9, 11.2, 9.5, 12.6, 11.2
- Other: 5.5, 4.0, 4.5, 6.9, 4.1, 4.1, 7.7, 5.8
- Missing: 0.1, 0.1, 0.1, 0.1, 0.1, 0.3, 0.2, 0.1

**Ethnicity**
- Non-Hispanic: 2.7, 2.4, 2.4, 3.0, 2.6, 3.3, 2.6, 3.9
- Hispanic: 9.0, 8.2, 8.0, 9.0, 8.2, 9.7, 9.0, 9.9
- Missing: 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1

**Marital status**
- Single: 9.0, 8.2, 8.0, 9.0, 8.2, 9.7, 9.0, 9.9
- Married: 58.2, 57.2, 59.0, 59.0, 60.4, 59.2, 9.2, 62.2
- Sep/Div/Wid: 29.7, 31.6, 29.1, 29.0, 28.4, 27.1, 56.1, 51.1
- Missing: 3.1, 3.0, 3.1, 3.1, 3.0, 3.0, 3.0, 3.0

**Nativity**
- Native-born: 81.0, 86.1, 83.1, 77.9, 84.9, 72.8, 83.7, 74.2
- Foreign-born: 8.2, 7.0, 7.9, 9.0, 8.1, 7.2, 9.0, 8.0
- Missing: 10.8, 7.0, 9.1, 13.1, 7.0, 20.1, 7.3, 17.8

**Data source**
- Non-hospital: 2.7, 2.4, 2.4, 3.0, 2.6, 3.3, 2.6, 3.9
- Hospital: 4.0, 4.1, 4.1, 7.8, 0.2, 4.9, 0.2, 0.2
- Missing: 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1

**Grade**
- Grade 1: 4.0, 0.1, 4.1, 7.8, 0.2, 4.9, 0.2, 0.2
- Grade 2: 13.1, 0.7, 24.8, 18.2, 0.5, 2.6, 0.6, 1.8
- Grade 3: 27.4, 5.9, 36.0, 30.6, 21.8, 8.0, 30.5, 30.5
- Grade 4: 27.4, 44.7, 2.2, 1.8, 48.3, 41.9, 2.8, 2.4
- Missing: 42.3, 48.7, 32.9, 41.7, 29.3, 42.7, 57.0, 65.0

**Tumor size**
- <2 cm: 8.3, 4.6, 5.9, 12.2, 5.3, 17.1, 4.2, 8.1
- 2–<3 cm: 10.9, 6.3, 9.1, 15.0, 8.8, 12.4, 7.4, 11.6
- 3–<4 cm: 10.2, 6.4, 10.2, 12.3, 9.6, 8.4, 8.4, 11.9
- 4–<5 cm: 8.1, 5.7, 9.1, 8.4, 8.3, 6.0, 7.1, 10.4
- ≥5 cm: 19.6, 17.9, 24.4, 15.7, 23.0, 14.7, 18.8, 28.2
- Missing: 43.0, 59.0, 41.3, 36.4, 45.0, 41.4, 45.1, 45.1

**Stage**
- Localized: 19.7, 7.9, 24.7, 23.8, 16.5, 33.0, 11.4, 11.9
- Regional: 27.8, 24.7, 36.2, 25.6, 30.0, 21.7, 21.4, 23.1
- Distant: 46.5, 61.6, 31.7, 46.1, 47.1, 40.3, 56.1, 62.0

(Continued on the following page)
Table 2. Distribution of histologically confirmed lung cancer cases by histology and selected covariates, SEER 9	extsuperscript{a}, 1975 to 2010 (Cont’d)

<table>
<thead>
<tr>
<th>Overall</th>
<th>Small cell</th>
<th>Squamous</th>
<th>Adeno-careinoma</th>
<th>Large cell</th>
<th>Other specified NSC</th>
<th>8010 (carcinoma, NOS)</th>
<th>8046 (NSC carcinoma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>463,609</td>
<td>(100.0%)</td>
<td>73,994</td>
<td>116,775</td>
<td>166,006</td>
<td>29,123</td>
<td>15,914</td>
<td>35,954</td>
</tr>
<tr>
<td>Overall</td>
<td>Missing</td>
<td>Surgery</td>
<td>Performed</td>
<td>Not performed</td>
<td>Missing</td>
<td>Survival</td>
<td>&lt;1 y</td>
</tr>
<tr>
<td></td>
<td>6.0</td>
<td>27.4</td>
<td>69.1</td>
<td>3.5</td>
<td>43.8</td>
<td>11.2</td>
<td>3.7</td>
</tr>
<tr>
<td></td>
<td>5.9</td>
<td>5.9</td>
<td>89.2</td>
<td>4.9</td>
<td>53.7</td>
<td>15.5</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>7.4</td>
<td>31.9</td>
<td>64.3</td>
<td>3.9</td>
<td>39.9</td>
<td>11.4</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>4.4</td>
<td>38.6</td>
<td>58.8</td>
<td>2.6</td>
<td>38.7</td>
<td>9.9</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>6.4</td>
<td>25.6</td>
<td>69.0</td>
<td>5.4</td>
<td>52.1</td>
<td>10.7</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td>5.0</td>
<td>43.6</td>
<td>52.5</td>
<td>3.8</td>
<td>54.6</td>
<td>7.3</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>11.1</td>
<td>11.7</td>
<td>83.4</td>
<td>4.9</td>
<td>45.5</td>
<td>10.1</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>3.1</td>
<td>10.3</td>
<td>89.4</td>
<td>0.3</td>
<td>4.5</td>
<td>10.3</td>
<td>3.2</td>
</tr>
</tbody>
</table>

NOTE: All two-way associations are significant at the 0.001 level.
Abbreviations: Atlanta, Atlanta metropolitan; Seattle, Seattle–Puget Sound; SMS, San Francisco–Oakland.

	extsuperscript{a}The SEER 9 registries include Atlanta, Connecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco–Oakland, Seattle–Puget Sound, and Utah.
strong evidences for model adequacy in the proposed method.

Results

Table 1 shows the distribution of histologic categories by histology confirmation status. Ninety percent of cases are histologically confirmed. Among the cases that are not confirmed and the cases for which the confirmation status is unknown, 8010 accounts for about 50% of the total whereas 8046 only accounts for less than 2%. Possible explanation for the differential use of 8010 and 8046 could be that the latter is mainly used when histologic diagnosis, although not quite specific, exists, and the former is also used when the diagnosis is not available.

Table 2 shows the distributions of lung cancer cases by histology and selected covariates. All covariates are closely associated with histology. Men and older patients were more likely to be diagnosed with squamous type. Squamous and adenocarcinoma tumors tended to be more well-differentiated than large cell and other specific NSCLC tumors. Squamous and large cell tumors tended to be larger at diagnosis. Small cell tumors were likely detected at a later stage (61.6%) as compared with other types. In contrast, tumors of squamous and adenocarcinoma types tended to be detected at early stage. There are also a few notable differences in the use of nonspecific codes across registries. For example, a lower use of 8046 (15.2% in 8046 compared with the overall percentage of 20.8%) is observed in Detroit, and a higher use of both 8010 (16.9% compared with the overall percentage of 15.0%) and 8046 (19.9%) is observed in Seattle. The use of nonspecific code is also slightly higher for cases not reported by a hospital (2.8% in 8010 and 2.9% in 8046 compared with the overall percentage of 1.8%). These variables are also predictive to the use of nonspecific morphology codes. As we expected, tumors without specific histologic diagnosis tended to be less well differentiated, diagnosed at a late stage, had shorter survivals, and were less likely to be candidates for surgery.

Figure 1 shows the percentages of cases coded with 8046 and 8010 by year of diagnosis for men and women separately. The temporal distributions are similar for both genders. The percentage of cases coded with 8010 had increased from 1982 until the introduction of 8046 into ICD-O-3 in 2001, when it dropped to around 3%. There seems to be a smooth compensation between 8010 and 8046 in 2001, which suggests that 8010 and 8046 are probably used interchangeably in practice.

Figure 2 shows the rates of incidence by imputed histology among cases coded with 8010 or 8046. Overall, the amount of imputed histology differs by histologic subtype and year. For both 8010 and 8046, the rates of incidence raised by imputation were greatest for adenocarcinoma and squamous. For both histologic subtypes, the rates followed an n-shaped pattern over the most recent 15 years. Small cell was the third most raised category, although only contributed from imputing 8010 cases, and the amount of increases was relatively stable over time.

Figure 3 compares the before and after imputation temporal trends in age-adjusted incidence rate of lung cancer by histology for men and women separately (see Table 3, for detailed results of the joinpoint trends analysis) The numbers listed over (imputed) or under (original) each segment represents the APC for that portion of the trend and an asterisk indicates a statistically significant trend at 0.05 level. The rates for 8010 (small cell type) and 8010 and 8046 combined (NSC subtypes) are also included in these plots to help examine how cases are distributed by the imputation procedure.

The imputation adjustment affected the incidence trends differently for each histologic subtype. For small cell in both genders, the original and imputed trends are similar. For squamous cell cancer in both genders and adenocarcinoma in men, the trends showed a similar pattern overall from 1970 to early 1990s before and after imputation. From early 1990s to 2005, the decreasing trends also remained unchanged after imputation, but the pace of decline slowed. After 2005, the increasing trends based on the original data had been replaced by

Figure 1. Percentages of histologically confirmed lung cancer cases coded as 8010 and 8046, SEER 9, 1975 to 2010.
the steady continuations of earlier decreasing trends for squamous and adenocarcinoma in men, a constant trend for squamous in women, after imputation. For adenocarcinoma in women, the trends, before and after imputation, exhibited similar patterns overall before early 1990s. From 1992 to 2007, the plateau followed by an increasing trend started in 2004 changed to a continuously increasing trend after imputation. It is also worth noting that the imputed rates showed a nonsignificant decreasing tendency during the most recent 3 years starting in 2007. For large cell cancer and cancer in other specified NSC type, the imputed rates were similar to the original rates and the imputation did not change the overall trends.

To rule out the possibility that changes in trends may be because of the absence of cases that are not histologically confirmed or have missing confirmation status, we conducted a sensitivity analysis on all cases. The imputation affected the trends similarly (see Supplementary Fig. S1 and Table S1, for detailed results on the rates and jointpoint analysis), which suggests that excluding these cases does not affect the overall findings and conclusions.

Discussion

In cancer surveillance data collections, it is common for the morphological classification systems to change to reflect the contemporary pathology practice. Hence, the data often comprise cancer cases coded one way at one time and others a different way at another time. When classification systems differ in coding histology, temporal inferences by histologic subtype can be misleading and difficult to interpret. Without access to calibration data to inform the underlying distribution of histology among cases coded without specificity or the association in histology between editions of classification systems, we carefully developed an MI approach to correct for biases in statistical inferences about temporal trends of lung cancer incidence based on the MAR assumption.

Although this assumption is not empirically testable, we argue that MAR is reasonable in our setting because we have identified and included into the imputation models an extensive set of auxiliary variables that can explain the missingness of specific histology, for example receipt of cancer-directed surgery, and that are correlates of histology, for example the stage, grade, and size of a tumor, as well as patient survival. Other important
variables that could enhance the MAR assumption plausibility are patients' smoking status and socioeconomic status (40, 41), for which we substituted county level estimates at 2000 (pooled estimates from 2000 to 2003 for smoking) from the decennial census because they are not routinely collected in SEER. Although such estimates are not available for every diagnosis year, we believe the ranking of a county in smoking prevalence or poverty level relative to the rest of the country remains relatively unchanged over time. The potential confounding between smoking status and poverty (40) is not likely a cause for concern in our analysis because both are aggregate measures and neither is a strong predictor to histology after conditional on other patient-level information.

Ensuring the plausibility of MAR assumption imposed 2 modeling challenges of handling a large number of variables with missing data and a general missing data pattern, which often cannot be adequately addressed by simple imputation methods (21). The proposed MI approach based on SRMI is particularly suitable to this complex situation because of its flexibility in specifying and fitting conditional distributions. The search for refined ridge-penalized logistic regression imputation models is necessary because the standard SRMI approach (based on logistic regressions) might be inadequate in handling a categorical outcome with a skewed distribution (e.g., certain histology categories only contain 3%-6% of cases) and correlated covariates (e.g., stage and survival). The simulation study demonstrated the adequacy and prediction benefits of the proposed semiparametric models.

The amount of lung cancer cases lacking specific histologic subtypes was predominantly associated with the year of diagnosis, which reflected the evolution of SEER coding algorithms and recent changes in diagnostic practice. The imputation raised the incidence rates across the entire study period for both genders and histology subgroups. However, the magnitudes of the elevations varied. Of the various histologic subtypes, the most impacted were squamous and adenocarcinoma, on which the
<table>
<thead>
<tr>
<th></th>
<th>Trend 1</th>
<th>Trend 2</th>
<th>Trend 3</th>
<th>Trend 4</th>
<th>Trend 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>APC (95% CI)</td>
<td>APC (95% CI)</td>
<td>APC (95% CI)</td>
<td>APC (95% CI)</td>
<td>APC (95% CI)</td>
</tr>
<tr>
<td><strong>Years</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Men</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small cell</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original</td>
<td>1975–1981</td>
<td>5.6</td>
<td>(3.8 to 7.4)</td>
<td>1988–2010</td>
<td>−3.2</td>
</tr>
<tr>
<td>Imputed</td>
<td>1975–1978</td>
<td>7.0</td>
<td>(0.3 to 14.2)</td>
<td>1981–1986</td>
<td>1.7</td>
</tr>
<tr>
<td>Squamous</td>
<td>1975–1982</td>
<td>1.7</td>
<td>(0.6 to 2.7)</td>
<td>1982–1990</td>
<td>−2.1</td>
</tr>
<tr>
<td>Imputed</td>
<td>1975–1982</td>
<td>0.6</td>
<td>(−0.1 to 1.3)</td>
<td>1982–1992</td>
<td>−1.8</td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>1975–1978</td>
<td>10.8</td>
<td>(4.3 to 17.6)</td>
<td>1978–1992</td>
<td>2.0</td>
</tr>
<tr>
<td>Imputed</td>
<td>1975–1978</td>
<td>8.0</td>
<td>(2.6 to 13.7)</td>
<td>1978–1992</td>
<td>2.1</td>
</tr>
<tr>
<td>Large cell</td>
<td>1975–1980</td>
<td>17.5</td>
<td>(12.1 to 23.2)</td>
<td>1980–1988</td>
<td>2.3</td>
</tr>
<tr>
<td>Other specific NSC</td>
<td>1975–1977</td>
<td>−17.9</td>
<td>(−28.7 to −5.3)</td>
<td>1977–1990</td>
<td>−6.4</td>
</tr>
<tr>
<td><strong>Women</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imputed</td>
<td>1975–1987</td>
<td>6.3</td>
<td>(5.3 to 7.4)</td>
<td>1987–1997</td>
<td>0.4</td>
</tr>
<tr>
<td>Squamous</td>
<td>1975–1984</td>
<td>5.8</td>
<td>(4.8 to 6.8)</td>
<td>1984–1995</td>
<td>1.0</td>
</tr>
<tr>
<td>Imputed</td>
<td>1975–1988</td>
<td>4.3</td>
<td>(3.7 to 4.9)</td>
<td>1988–2010</td>
<td>0.1</td>
</tr>
<tr>
<td>Imputed</td>
<td>1975–1990</td>
<td>4.7</td>
<td>(4.3 to 5.0)</td>
<td>1990–2007</td>
<td>1.9</td>
</tr>
</tbody>
</table>

(Continued on the following page)
Table 3. Joinpoint analysis for histologically confirmed malignant lung cancers by imputation status, gender, and histology, SEER 9a, 1975 to 2010 (Cont’d)

<table>
<thead>
<tr>
<th>Trend 1</th>
<th>Trend 2</th>
<th>Trend 3</th>
<th>Trend 4</th>
<th>Trend 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>APC (95% CI)</td>
<td>APC (95% CI)</td>
<td>APC (95% CI)</td>
<td>APC (95% CI)</td>
<td>APC (95% CI)</td>
</tr>
<tr>
<td>Years</td>
<td>Years</td>
<td>Years</td>
<td>Years</td>
<td>Years</td>
</tr>
<tr>
<td>Large cell</td>
<td>Other specific NSC</td>
<td>Other specific NSC</td>
<td>Other specific NSC</td>
<td>Other specific NSC</td>
</tr>
<tr>
<td>40.4 (24.1 to 59.0)</td>
<td>37.3 (20.8 to 56.2)</td>
<td>3.5 (1.9 to 5.1)</td>
<td>3.9 (1.6 to 6.1)</td>
<td>2.9 (1.7 to 4.1)</td>
</tr>
<tr>
<td>6.6 (2.9 to 10.3)</td>
<td>6.6 (3.9 to 9.3)</td>
<td>6.6 (3.9 to 9.3)</td>
<td>6.6 (3.9 to 9.3)</td>
<td>6.6 (3.9 to 9.3)</td>
</tr>
<tr>
<td>2.1 (0.3 to 3.9)</td>
<td>2.1 (0.3 to 3.9)</td>
<td>2.1 (0.3 to 3.9)</td>
<td>2.1 (0.3 to 3.9)</td>
<td>2.1 (0.3 to 3.9)</td>
</tr>
<tr>
<td>9.9 (5.8 to 14.0)</td>
<td>9.9 (5.8 to 14.0)</td>
<td>9.9 (5.8 to 14.0)</td>
<td>9.9 (5.8 to 14.0)</td>
<td>9.9 (5.8 to 14.0)</td>
</tr>
<tr>
<td>12.4 (7.0 to 17.8)</td>
<td>12.4 (7.0 to 17.8)</td>
<td>12.4 (7.0 to 17.8)</td>
<td>12.4 (7.0 to 17.8)</td>
<td>12.4 (7.0 to 17.8)</td>
</tr>
</tbody>
</table>

The SEER 9 registries include Atlanta, Connecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco–Oakland, Seattle–Puget Sound, and Utah.

In summary, molecular, genetic, and etiologic features are increasingly associated with histologic distinctions (3, 4, 44). Progress in linking molecular features to morphology will facilitate mechanistic understanding and further characterization of the molecular and genetic features specific to histologic subtypes in lung cancer. These considerations, along with the emergence of targeted therapies within specific histologic subtypes especially adenocarcinoma, clearly indicates that accurate population tracking of trends by lung cancer histology will be increasingly important in the future, and that the MI technique applied in this study can help refine these trends. Planned data collections for bridge data in the future will further enhance the quality of data augmented by MI.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Authors’ Contributions
Conception and design: M. Yu, E.J. Feuer, K.A. Cronin, N.E. Caporaso
Development of methodology: M. Yu, E.J. Feuer, K.A. Cronin
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): M. Yu
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): M. Yu, E.J. Feuer, K.A. Cronin, N.E. Caporaso
Writing, review, and/or revision of the manuscript: M. Yu, E.J. Feuer, K.A. Cronin, N.E. Caporaso
Yu et al.

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): M. Yu

Study supervision: M. Yu, K.A. Cronin

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.

Received February 5, 2014; revised April 22, 2014; accepted May 5, 2014; published OnlineFirst May 22, 2014.

References


# Use of Multiple Imputation to Correct for Bias in Lung Cancer Incidence Trends by Histologic Subtype

Mandi Yu, Eric J. Feuer, Kathleen A. Cronin, et al.

*Cancer Epidemiol Biomarkers Prev* Published OnlineFirst May 22, 2014.

<table>
<thead>
<tr>
<th>Updated version</th>
<th>Access the most recent version of this article at: doi:10.1158/1055-9965.EPI-14-0130</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplementary Material</td>
<td>Access the most recent supplemental material at: <a href="http://cebp.aacrjournals.org/content/suppl/2014/05/28/1055-9965.EPI-14-0130.DC1">http://cebp.aacrjournals.org/content/suppl/2014/05/28/1055-9965.EPI-14-0130.DC1</a></td>
</tr>
</tbody>
</table>

E-mail alerts | Sign up to receive free email-alerts related to this article or journal. |
Reprints and Subscriptions | To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org. |
Permissions | To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org. |