

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/357818491>

# Accounting for EGFR Mutations in Epidemiologic Analyses of Non-Small Cell Lung Cancers: Examples Based on the International Lung Cancer Consortium Data

Article in *Cancer Epidemiology Biomarkers & Prevention* · January 2022

DOI: 10.1158/1055-9965.EPI-21-0747

CITATION

1

READS

119

57 authors, including:



**Sabine Schmid**

Inselspital, Universitätsspital Bern

90 PUBLICATIONS 2,078 CITATIONS

[SEE PROFILE](#)



**Mei Jiang**

153 PUBLICATIONS 2,808 CITATIONS

[SEE PROFILE](#)



**M. Catherine Brown**

The Princess Margaret Hospital

191 PUBLICATIONS 898 CITATIONS

[SEE PROFILE](#)



**Aline Fares**

University Health Network

43 PUBLICATIONS 267 CITATIONS

[SEE PROFILE](#)

# Accounting for *EGFR* Mutations in Epidemiologic Analyses of Non–Small Cell Lung Cancers: Examples Based on the International Lung Cancer Consortium Data



Sabine Schmid<sup>1,2</sup>, Mei Jiang<sup>3</sup>, M. Catherine Brown<sup>1</sup>, Aline Fares<sup>4</sup>, Miguel Garcia<sup>1</sup>, Joelle Soriano<sup>1,5</sup>, Mei Dong<sup>1,6</sup>, Sera Thomas<sup>7</sup>, Takashi Kohno<sup>8</sup>, Leticia Ferro Leal<sup>9</sup>, Nancy Diao<sup>10</sup>, Juntao Xie<sup>11</sup>, Zhichao Wang<sup>12,13</sup>, David Zaridze<sup>14</sup>, Ivana Holcatova<sup>15</sup>, Jolanta Lissowska<sup>16</sup>, Beata Świątkowska<sup>17</sup>, Dana Mates<sup>18</sup>, Milan Savic<sup>19</sup>, Angela S. Wenzlaff<sup>20</sup>, Curtis C. Harris<sup>21</sup>, Neil E. Caporaso<sup>22</sup>, Hongxia Ma<sup>23</sup>, Guillermo Fernandez-Tardon<sup>24</sup>, Matthew J. Barnett<sup>25</sup>, Gary Goodman<sup>26</sup>, Michael P.A. Davies<sup>27</sup>, Mónica Pérez-Ríos<sup>28,29</sup>, Fiona Taylor<sup>30,31</sup>, Eric J. Duell<sup>32,33</sup>, Ben Schoettker<sup>34,35</sup>, Hermann Brenner<sup>34,35,36,37</sup>, Angeline Andrew<sup>38</sup>, Angela Cox<sup>30</sup>, Alberto Ruano-Rayana<sup>28,29</sup>, John K. Field<sup>27</sup>, Loic Le Marchand<sup>39</sup>, Ying Wang<sup>40</sup>, Chu Chen<sup>41</sup>, Adonina Tardon<sup>24</sup>, Sanjay Shete<sup>42</sup>, Matthew B. Schabath<sup>43</sup>, Hongbing Shen<sup>23</sup>, Maria Teresa Landi<sup>22</sup>, Brid M. Ryan<sup>21</sup>, Ann G. Schwartz<sup>20</sup>, Lihong Qi<sup>44</sup>, Lori C. Sakoda<sup>45</sup>, Paul Brennan<sup>46</sup>, Ping Yang<sup>12</sup>, Jie Zhang<sup>11</sup>, David C. Christiani<sup>10</sup>, Rui Manuel Reis<sup>9,47,48</sup>, Kouya Shiraishi<sup>8</sup>, Rayjean J. Hung<sup>6,7</sup>, Wei Xu<sup>1,6</sup>, and Geoffrey Liu<sup>1,6</sup>

## ABSTRACT

**Background:** Somatic *EGFR* mutations define a subset of non–small cell lung cancers (NSCLC) that have clinical impact on NSCLC risk and outcome. However, *EGFR*-mutation-status is often missing in epidemiologic datasets. We developed and tested pragmatic approaches to account for *EGFR*-mutation-status based on variables commonly included in epidemiologic datasets and evaluated the clinical utility of these approaches.

**Methods:** Through analysis of the International Lung Cancer Consortium (ILCCO) epidemiologic datasets, we developed a regression model for *EGFR*-status; we then applied a clinical-restriction approach using the optimal cut-point, and a second epidemiologic, multiple imputation approach to ILCCO survival analyses that did and did not account for *EGFR*-status.

**Results:** Of 35,356 ILCCO patients with NSCLC, *EGFR*-mutation-status was available in 4,231 patients. A model regressing known *EGFR*-mutation-status on clinical and demographic

variables achieved a concordance index of 0.75 (95% CI, 0.74–0.77) in the training and 0.77 (95% CI, 0.74–0.79) in the testing dataset. At an optimal cut-point of probability-score = 0.335, sensitivity = 69% and specificity = 72.5% for determining *EGFR*-wildtype status. In both restriction-based and imputation-based regression analyses of the individual roles of BMI on overall survival of patients with NSCLC, similar results were observed between overall and *EGFR*-mutation-negative cohort analyses of patients of all ancestries. However, our approach identified some differences: *EGFR*-mutated Asian patients did not incur a survival benefit from being obese, as observed in *EGFR*-wildtype Asian patients.

**Conclusions:** We introduce a pragmatic method to evaluate the potential impact of *EGFR*-status on epidemiological analyses of NSCLC.

**Impact:** The proposed method is generalizable in the common occurrence in which *EGFR*-status data are missing.

<sup>1</sup>The Princess Margaret Cancer Centre and University Health Network, University of Toronto, Toronto, Ontario, Canada. <sup>2</sup>Department of Medical Oncology, Cantonal Hospital St. Gallen, St. Gallen, Switzerland. <sup>3</sup>State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Health, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China. <sup>4</sup>Division of Medical Oncology, Hospital de Base de São José do Rio Preto, SP, Brazil. <sup>5</sup>University of Ottawa, Ottawa, Ontario, Canada. <sup>6</sup>Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada. <sup>7</sup>Lunenfeld-Tanenbaum Research Institute, Sinai Health Systems, Toronto, Ontario, Canada. <sup>8</sup>Division of Genome Biology, National Cancer Centre Research Institute, Tokyo, Japan. <sup>9</sup>Molecular Oncology Research Center, Barretos Cancer Hospital, Barretos, Brazil. <sup>10</sup>Harvard T.H. Chan School of Public Health, Boston, Massachusetts. <sup>11</sup>Department of Thoracic Surgery, Fudan University Shanghai Cancer Center, Shanghai, China. <sup>12</sup>Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota. <sup>13</sup>Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China. <sup>14</sup>Russian N.N. Blokhin Cancer Research Centre, Moscow, Russian Federation. <sup>15</sup>Institute of Hygiene and Epidemiology, 1st Faculty of Medicine, Charles University, Prague, Czech Republic. <sup>16</sup>Department of Cancer Epidemiology and Prevention, M. Skłodowska-Curie National Research Institute of Oncol-

ogy, Warsaw, Poland. <sup>17</sup>The Nofer Institute of Occupational Medicine, Lodz, Poland. <sup>18</sup>National Institute of Public Health, Bucharest, Romania. <sup>19</sup>Department of Thoracic Surgery, Clinical Center of Serbia, Belgrade, Serbia. <sup>20</sup>Barbara Ann Karmanos Cancer Institute, Wayne State University, Detroit, Michigan. <sup>21</sup>Laboratory of Human Carcinogenesis, Centre for Cancer Research, National Institutes of Health, Bethesda, Maryland. <sup>22</sup>Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland. <sup>23</sup>Department of Epidemiology, School of Public Health, Nanjing Medical University, Nanjing, China. <sup>24</sup>IUOPA, University of Oviedo, and ISPA (Health Research Institute of the Principality of Asturias) and CIBERESP, Asturias, Spain. <sup>25</sup>Program in Biostatistics Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington. <sup>26</sup>Swedish Cancer Institute, Seattle, Washington. <sup>27</sup>Roy Castle Lung Cancer Research Programme, The University of Liverpool, Department of Molecular and Clinical Cancer Medicine, Liverpool, United Kingdom. <sup>28</sup>Department of Preventive Medicine and Public Health, University of Santiago de Compostela, Galicia, Spain. <sup>29</sup>CIBER de Epidemiología y Salud Pública, CIBERESP, Santiago de Compostela, Galicia, Spain. <sup>30</sup>Department of Oncology and Metabolism, University of Sheffield Medical School, Sheffield, United Kingdom. <sup>31</sup>Sheffield Teaching Hospitals Foundation Trust,

## Introduction

Somatic *EGFR* mutations define a unique subset of non-small cell lung cancers (NSCLC) and have clinical impact on NSCLC outcomes; further, genetic and environmental risk factors may be different in patients with *EGFR*-mutated and *EGFR*-wildtype NSCLCs. Clinicopathologic factors such as being a lifetime never-smoker, female, of Asian ancestry, and having a histology of adenocarcinoma have each been independently associated with a greater likelihood of having *EGFR*-mutated NSCLC (1, 2). In contrast, heavy smoking, male sex, and squamous carcinoma histology are associated with NSCLC without *EGFR* mutations (i.e., *EGFR* wildtype; refs. 3, 4). Up to 90% of *EGFR* mutations are sensitizing mutations, therefore being strongly predictive of response to tyrosine kinase inhibitors (TKI) targeting the mutated *EGFR* protein; *EGFR* TKIs are used commonly in advanced or metastatic incurable patients with *EGFR*-mutated NSCLC to improve overall survival (5, 6), and recently to improve disease-free survival in early stage, resected patients (7).

Molecular detection of *EGFR* mutations itself only became widely available in routine clinical practice after publication of the seminal IPASS study in 2009 (8), which established *EGFR*-TKIs as the preferred treatment for patients with incurable stage IIIB/IV *EGFR*-mutated NSCLC; further, the availability of *EGFR* testing depended on the speed of clinical uptake, which varied across the world (9). Therefore, many epidemiological research databases have not historically collected *EGFR* mutation data or detailed treatment data. Consequently, interpretation of both risk and survival outcomes could be impacted by this lack of available information, especially for lung adenocarcinoma.

Among NSCLC subgroups, individuals carrying *EGFR*-mutated tumors represent the largest subgroup whose biology is markedly different than of typical smoking-related NSCLC; proportions of *EGFR*-mutated tumors can range from 10% to upwards of 50% (10–13). Thus, epidemiologic studies aiming to gain better understanding of the genetic and environmental etiologic factors will likely need to study *EGFR*-mutated and *EGFR*-wildtype NSCLCs separately.

To account for missing data, there have been prior efforts to predict *EGFR*-status based on clinical and demographic variables. Chang and colleagues developed a predictive model for being *EGFR*-mutated exclusively in an Asian population based on seven variables, namely sex, adenocarcinoma histology, smoking history, N-stage, M-stage, presence of brain metastases, and elevated CYFRA 21–1 serologic levels (14). With a sensitivity of 95% and specificity of 32.3%, their

model achieved a positive predictive value (PPV) of 85.1% and a negative predictive value (NPV) of 65.6%. Another nomogram, proposed by Girard and colleagues for adenocarcinomas based on a non-Asian population, incorporated age, sex, smoking pack-years, time interval between smoking cessation and NSCLC diagnosis, disease stage (I–IIA vs. IIIB–IV) and predominant histologic subtype (solid, papillary, or bronchioalveolar); this study achieved a concordance index of 0.84 (15). However, despite acceptable accuracy, these two published predictive models cannot be easily applied in most epidemiologic studies because they incorporate some variables that are not readily available in existing epidemiologic or clinical studies, such as predominant histologic subtypes and CYFRA 21–1 levels.

The overarching aim of this study was to develop and evaluate a pragmatic approach to account for *EGFR*-status in the analysis of epidemiologic studies, using variables generally included in existing datasets. We developed a regression model for *EGFR*-status by analyzing International Lung Cancer Consortium (ILCCO) epidemiologic datasets. With this regression model, we applied two approaches, a clinical approach and an epidemiological approach. In the clinical approach, we identified a regression value cut-point from which we dichotomized patients into those who were most likely or least likely to have an *EGFR*-mutated NSCLC; we have termed this the restriction method because it “restricts” the entire population into a smaller dataset most likely to have or have not an *EGFR* mutation. The alternative epidemiologic approach utilized a multiple imputation approach to differentiate between the likely *EGFR*-mutated from patients who were less likely to have *EGFR*-mutated NSCLCs. We used these two approaches to represent approaches widely familiar with either clinicians or epidemiologists, respectively, and to demonstrate that these two approaches could yield in consistent results. We then applied these two different approaches to previous survival analyses to compare how much change in results would occur had we used these two approaches to separate our datasets into those most and least likely to carry *EGFR* mutations.

## Materials and Methods

### Study design

We first developed a pragmatic multivariable regression model with the outcome of *EGFR*-status, in an ILCCO subcohort dataset that included only patients with known *EGFR* mutation-status (*EGFR*-wildtype vs. *EGFR*-mutated). We then applied this regression model to predict *EGFR*-status in patients with NSCLC in the larger ILCCO

Sheffield, United Kingdom. <sup>32</sup>Catalan Institute of Oncology (ICO), Barcelona, Spain. <sup>33</sup>Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. <sup>34</sup>Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>35</sup>Network of Aging Research, Heidelberg University, Heidelberg, Germany. <sup>36</sup>Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany. <sup>37</sup>German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>38</sup>Dartmouth-Hitchcock Medical Center, Lebanon, New Hampshire. <sup>39</sup>University of Hawaii Cancer Center, Honolulu, Hawaii. <sup>40</sup>American Cancer Society, Atlanta, Georgia. <sup>41</sup>Program in Epidemiology, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington. <sup>42</sup>The University of Texas MD Anderson Cancer Center, Houston, Texas. <sup>43</sup>Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, Florida. <sup>44</sup>The University of California Davis Medical Sciences, Davis, California. <sup>45</sup>Kaiser Permanente Northern California, Division of Research, Oakland, California. <sup>46</sup>International Agency for Research on Cancer, Lyon, France. <sup>47</sup>Life and Health Sciences Research

Institute (ICVS), Medical School, University of Minho, Braga, Portugal. <sup>48</sup>ICVS/3B's-PT Government Associate Laboratory, Braga, Portugal.

**Note:** Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (<http://cebp.aacrjournals.org/>).

S. Schmid and M. Jiang contributed equally to this article.

W. Xu and G. Liu contributed equally to this article.

**Corresponding Authors:** Geoffrey Liu, Division of Medical Oncology, Princess Margaret Cancer Centre, University Health Network, 610 University Avenue, Toronto, Ontario M5G 2M9, Canada. E-mail: Geoffrey.Liu@uhn.ca; and Wei Xu, E-mail: Wei.Xu@uhnresearch.ca

Cancer Epidemiol Biomarkers Prev 2022;31:679–87

doi: 10.1158/1055-9965.EPI-21-0747

©2022 American Association for Cancer Research

dataset, using two different approaches: a clinical restriction approach where the probability of having either *EGFR*-wildtype or *EGFR*-mutated NSCLC was estimated through an optimal cut-off determined by the multivariable regression model, and an epidemiologic multiple imputation approach utilizing the same regression model for estimating *EGFR*-status.

### Study population

ILCCO harmonizes compatible data from various epidemiologic studies worldwide to facilitate collaborative lung cancer epidemiology research in large combined datasets (details are available on <http://ilcco.iarc.fr>). Twenty-seven ILCCO studies participated in prior survival analyses, and among the participating studies the majority of patients with lung cancer were male, ever-smokers, and of European ancestry, suggesting that the majority of cases would not carry a somatic *EGFR* mutation. Thus, our primary goal was to identify a subset of patients who are not likely to carry the mutation (i.e., *EGFR*-wildtype), so that we can perform sensitivity analyses to compare any main results in the entire ILCCO cohort (regardless of *EGFR*-status) to results generated in a predicted *EGFR*-wildtype subcohort to better understand possible influence of *EGFR*-status on survival outcomes. To explore possible utility in an Asian population with higher prevalence of *EGFR*-mutation, we performed additional analyses in our Asian subgroup accounting for *EGFR*-status. Ethics approval was obtained by each participating study from local review boards.

### Analysis

Summary statistics were provided with continuous and categorical variables presented as median with range and as frequency with percentage (%), respectively. Comparisons of baseline clinicopathologic profiles among different groups were performed using Kruskal–Wallis and Chi-square tests, as appropriate.

### Multivariable regression model development

We first developed a multivariable regression model that incorporated basic clinico-epidemiological variables that are typically captured in most observational studies. We developed this regression model using only patients with known *EGFR*-status (*EGFR*-wildtype or *EGFR*-mutated). To develop the best regression models of clinico-demographic-pathologic variables and *EGFR*-status, we randomly divided data from patients with known *EGFR* status into a training set (comprised of two-thirds of the patients), which was used for prediction model development, and a testing set (including the remaining one-third) for model validation. In addition, the selected model was also validated using bootstrap resampling methods. The candidate variables in the regression model for *EGFR* status included age, gender, ethnicity, stage, smoking history, and histology. We used the backward selection algorithm with the Akaike information criterion to select the variables in the regression model. Odds ratios (OR) and 95% confidence intervals (CI) of each variable in the model were calculated.

### Clinical or restriction approach to identify an *EGFR*-wildtype subcohort (as well as an *EGFR*-positive subcohort in Asian population-specific subanalyses)

As this regression model served to predict *EGFR*-status, the discriminatory ability of the model was quantified using the AUC of the ROC. The probability score (PS) was defined on the basis of the weighted summary of the variables in the model weighted by the corresponding regression coefficients. The optimal cut-point

value of the PS for distinguishing high probability *EGFR*-wildtype lung cancers from others was determined using the ROC curve. The ROC of a perfect test passes through the left-upper corner of the ROC plot, the point where both sensitivity and specificity are equal to 1; the optimal cut-off point is the point on the ROC curve that has the smallest distance to this left-upper corner (16–18).

Those with a PS for having a specific *EGFR*-status that was greater than the optimal cut-off point was given that *EGFR*-status.

### Epidemiologic or multiple imputation approach to identify an *EGFR*-wildtype subcohort (as well as an *EGFR*-positive subcohort in Asian population-specific subanalyses)

As a second approach, we used a multiple imputation algorithm to generate HRs (by applying the multivariable regression model). For each patient with unknown *EGFR* status, we compared the probability of *EGFR* status based on the predicted model and the generated random number with uniform distribution; if greater, then the patient of predicted *EGFR* status was assigned as positive, otherwise negative. The association between predicted *EGFR* status and overall survival was examined by using Cox regression. The above procedure was repeated 100 times and we summarized the data as mean HRs and 95% CIs (19, 20).

### Application of both restriction and imputation approaches to prior ILCCO outcome analyses

Data on the relationship between BMI and survival outcomes from the ILCCO dataset were utilized for these assessments. For each sensitivity analysis, the clinical-restriction and epidemiologic-imputation approaches to identifying an *EGFR*-wildtype subcohort were individually compared with the analysis of the entire ILCCO cohort as previously published based on patient data from 16 centers; as some centers had since provided additional patient data and additional centers (now up to 27) had provided data, we analyzed this updated version of the dataset, as we found no logical reason to exclude these additional patients. As most *EGFR*-mutated tumors are adenocarcinomas, we conducted an additional sensitivity analyses exclusively in the adenocarcinoma subset of our cohort. In the Asian subgroup, restriction and imputation were also applied to generate a predicted *EGFR*-positive subgroup to be compared with the analyses of the entire Asian population of the ILCCO cohort. Application to Kaplan–Meier curves and Cox proportional hazards regression models were used in illustrative examples to demonstrate the potential impact of taking into account *EGFR*-status (restriction approach, imputation approach) when compared with previous analyses that did not consider *EGFR*-status, for the following two associations: BMI and overall survival (OS; ref. 21) and interaction of BMI with smoking, gender, and ethnicity on OS as measured through subset analyses (22).

In the restriction approach, we estimated hazard ratios on a restricted dataset that analyzed only predicted *EGFR*-wildtype patients based on the optimal PS cut-point as determined from the generated ROC curves. In the multiple imputation approach, after 100 HRs were generated, we summarized the data as mean HRs and 95% CIs. For the Asian subgroup, analyses were also performed using both approaches to identify both *EGFR*-wildtype and *EGFR*-mutated patient subgroups.

All statistical analyses were performed using R 4.0.1 (<http://CRAN.R-project.org>, The R Foundation for Statistical Computing). All *P* values were based on two-sided tests and considered statistically significant at *P* < 0.05.

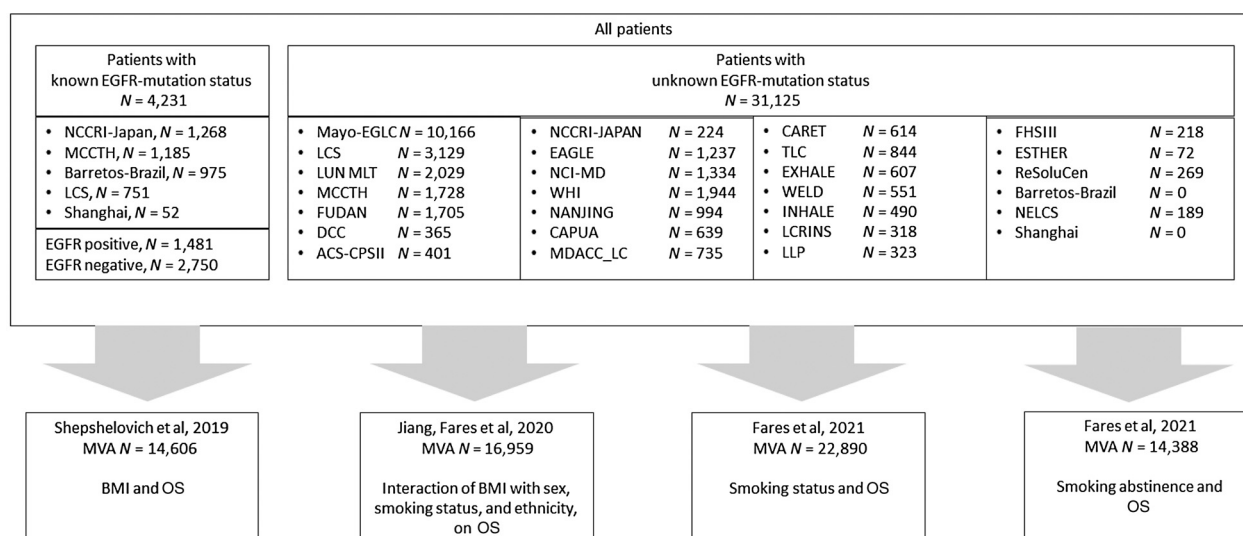


Figure 1.

CONSORT diagram. Top two boxes illustrate patient data flow into the two datasets used (left: dataset including only patients with known EGFR status for model development, right: dataset including patients with unknown EGFR status for main analysis). Boxes below illustrate publications derived from these datasets.

## Results

### Baseline characteristics

Overall, there were 35,356 patients with lung cancer in the ILCCO database, of which *EGFR*-status was available in a subset of 4,231 patients across five studies, whereas 31,125 patients across 27 studies had unknown *EGFR*-status (Fig. 1). The majority of studies included in this analysis had completed the major part of their recruitment before 2009; however, *EGFR* testing became more available as standard of care only after 2009 (Supplementary Table S1). The characteristics of those with known and unknown *EGFR*-status are presented in Supplementary Table S2. Of the patients with known *EGFR* status, 1,481 were *EGFR*-mutated whereas 2,750 were *EGFR*-wildtype (*EGFR*-mutation prevalence of 35%). Studies from Asia had higher prevalence of *EGFR*-mutated patients (NCCRI-Japan 48%; Shanghai 56%) whereas American studies had lower prevalence (LCS 21%; Barretos-Brazil 19%); the multicultural Toronto MSH-PMH study had an intermediate prevalence of 42% (Supplementary Table S3). As expected, baseline characteristics differed significantly between *EGFR*-mutated and *EGFR*-wildtype patients with respect to age, sex, ethnicity, and smoking status (Table 1; Supplementary Table S4).

### Multivariable regression model development

In univariable analysis, being female and Asian were associated with higher chance of being *EGFR*-mutated, whereas non-adenocarcinoma histology, BMI  $\geq 25$  kg/m<sup>2</sup> and having any smoking history was inversely associated with being *EGFR*-mutated, as was heavy smoking (Supplementary Table S4). In this dataset, earlier stage was more likely to be associated with being *EGFR*-mutated, which was due to ascertainment bias, as the Asian studies were mostly from thoracic surgeon practices of early stage, resected lung cancers (Supplementary Table S4).

Multivariable regression models were primarily assessed for their ability to create accurate *EGFR*-wildtype cohorts, using different combinations of variables that have been shown to be significant in univariable analyses; we also evaluated several models that contained

interaction terms (Ethnicity  $\times$  smoking status; ethnicity  $\times$  stage; ethnicity  $\times$  sex; Supplementary Table S5) based on known associations between several key clinico-demographic factors and presence/absence of *EGFR* mutation. Concordance indices (C-indices) were very similar across models containing different variables: all between 0.740 and 0.778 (Supplementary Table S5). Therefore, we selected a pragmatic model that included only variables available for most ILCCO patients to maximize statistical power. Our final model included age, sex, ethnicity, histology, and smoking status (see parameters and estimates of final model in Supplementary Table S6), which achieved a C-index of 0.75 (95% CI, 0.74–0.77) in the training dataset and 0.77 (95% CI, 0.74–0.79) in the testing dataset (Fig. 2). Model performance was also validated using bootstrap resampling methods confirming model performance (Supplementary Table S7; Supplementary Fig. S1).

### Choosing a clinically relevant probability score cut-point from the multivariable regression model for being *EGFR*-wildtype

On the basis of the ROC-curve generated by our model (Fig. 2) and distribution of PS (Supplementary Fig. S2), we evaluated various possible cut-points to determine which patients should be classified as *EGFR*-mutated versus *EGFR*-wildtype. With a PS cut-point of 0.335 (optimal cut-point from a statistical standpoint determined from the ROC curves generated by the regression model), there was a sensitivity of 69% and specificity of 72.5%. Lower PS cut-points would have resulted in decreased specificity.

The *EGFR* status-known dataset of 4,231 patients had a 35% *EGFR* mutation prevalence that corresponded to a *EGFR*-mutated PPV of 57% and NPV of 81%; the NPV was thus reasonably associated with identifying patients with *EGFR*-wildtype NSCLC whereas retaining 2,453 patients that would be considered *EGFR*-wildtype in the analysis. With a more conservative probability-score cut-point of 0.25, NPV increased to 85%, but at the expense of a substantially smaller sample size of patients that would be considered *EGFR*-wildtype ( $N = 1,879$ ).

When assessing all ILCCO participants ( $n = 35,356$ ; Supplementary Fig. S2D), the PS distribution was very different from the PS

**Table 1.** Baseline characteristics of *EGFR* mutation-tested (i.e., mutation-known) cohort, overall and by *EGFR*-status.

Covariate	Category	Patients with <i>EGFR</i> mutation-tested tumors: N (%)			P value
		Full sample	<i>EGFR</i> -mutated	<i>EGFR</i> -wildtype	
Total count (100%)		4,231	1,481	2,750	
Age	Median [Min–Max]	63 [18–95]	62 [22–95]	63 [18–93]	0.008
Sex	Male	1,976 (47)	510 (34)	1,466 (53)	<0.001
	Female	2,255 (53)	971 (66)	1,284 (47)	
Ethnicity	White	1,513 (43)	371 (28)	1,142 (53)	<0.001
	Asian	1,727 (49)	892 (67)	835 (39)	
	Black/other	252 (7)	65 (5)	187 (9)	
	Unknown	739	153	586	
BMI (kg/m <sup>2</sup> )	<18.5	1,201 (53)	432 (57)	769 (50)	0.0083
	18.5–<25	159 (7)	46 (6)	113 (7)	
	≥25	925 (40)	278 (37)	647 (42)	
	Unknown	1,946	725	1,221	
Smoking status	Never	1,686 (40)	964 (66)	722 (27)	<0.001
	Former	1,348 (32)	349 (24)	999 (37)	
	Current	1,133 (27)	155 (11)	978 (36)	
	Unknown	64	13	51	
Packyears <sup>a</sup>	≤20	410 (26)	179 (48)	231 (19)	<0.001
	>20	1,151 (74)	195 (52)	956 (81)	
	Unknown	920	130	790	
NSCLC histology	Adeno	3,974 (94)	1,455 (98)	2,519 (92)	<0.001
	Squamous	149 (4)	11 (1)	138 (5)	
	Large cell	33 (1)	3 (0)	30 (1)	
	Not specified	75 (2)	12 (1)	63 (2)	
Stage	I	1,372 (32)	565 (38)	807 (29)	<0.001
	II	326 (8)	106 (7)	220 (8)	
	III	784 (19)	227 (15)	557 (20)	
	IV	1,749 (41)	583 (39)	1,166 (42)	

<sup>a</sup>Only among ever-smokers.

distribution observed in the *EGFR*-status-known cohort, which was also reflected in different distributions in characteristics associated with *EGFR* status (Table 1; Supplementary Table S2). This was because there was oversampling of the *EGFR*-mutated patients among all tested patients: until centers started to perform routine testing for *EGFR*-status in all patients, patients would often be selected for testing on the basis being a never-smoker, or being of Asian ethnicity. Thus, in our overall ILCCO dataset, we anticipated an *EGFR* mutation prevalence lower than 35%. As a sensitivity analysis, we artificially reduced the *EGFR*-mutation prevalence to 15% while keeping the same test sensitivity and specificity and recalculated the following: the NPV increased to 92% at a PS cut-point of 0.335 ( $n = 23,434$ ), and to 94% ( $n = 18,484$ ) at a PS cut-point of 0.25.

#### OS of *EGFR*-wildtype patients, as determined by different approaches

As expected, the OS of *EGFR*-mutated patients was longer, compared with *EGFR*-wildtype patients (Supplementary Figs. S3A and S3B). We then compared Kaplan–Meier curves of known *EGFR*-wildtype patients (median OS: 2.67 years) with those defined on the basis of PS <0.335 (median OS: 2.49 years) and PS <0.25 (median OS: 1.91 years), and found that the optimal cut-point of <0.335 selected patients with median OS closer to the known *EGFR*-wildtype patients (Supplementary Fig. S3C). To avoid confounding by stage, we also performed the same comparisons, but restricted to stage IV patients only (Supplementary Fig. S3D). We then compared Kaplan–Meier curves and median OS of true *EGFR*-wildtype patients with the predicted *EGFR*-wildtype patients in all ILCCO patients (Supplementary Figs. S3E and S3F) and demonstrated high concordance. The

patterns and relationships of OS were similar across all the different approaches and sensitivity analyses.

#### Assessing the clinical utility of our clinical-restriction and epidemiological-imputation approaches

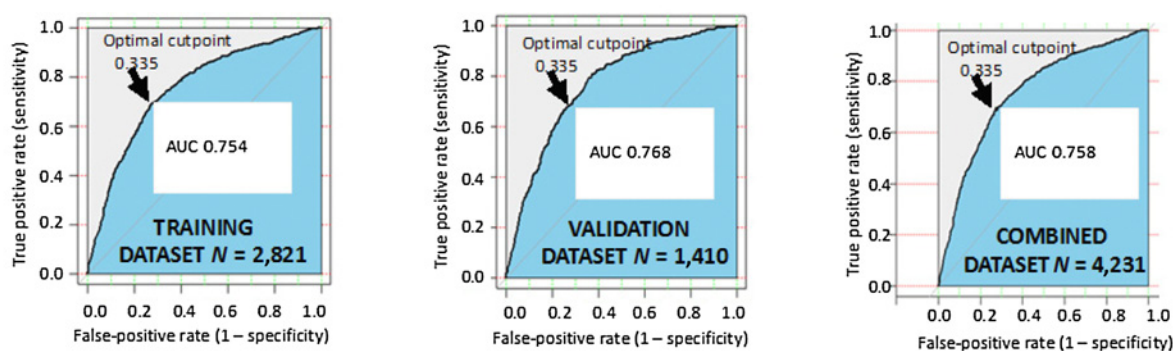
We re-analyzed previously published ILCCO-analyses on BMI-OS hypotheses described in the Materials and Methods section. Although test characteristics (sensitivity, specificity) of our model do not change with changes in *EGFR* prevalence, PPV and NPV, and therefore accuracy (true positives and true negatives, all divided by total evaluated) will change with changes in *EGFR* prevalence. As our overall model only had sufficient accuracy to predict patients with *EGFR*-wildtype status (being a largely Caucasian, smoking dataset) but lacked adequate PPV to identify *EGFR*-mutated patients in the overall population, we focused our re-analysis only on the *EGFR*-wildtype cohort using both clinical-restriction and epidemiologic-imputation approaches.

When re-analyzing our previous studies on the influence of BMI on OS in patients with NSCLC by clinical-restriction or epidemiologic-imputation approaches, the direction of change remained the same for all BMI levels and interactions. In most cases, the magnitude of HRs was similar too; however, in a few subgroups, the overall effect size varied (Figs. 3 and 4; Supplementary Tables S8 and S9). Results remained comparable in a sensitivity analysis exclusively in patients with known adenocarcinoma histology (Supplementary Tables S10 and S11).

#### Asian subcohort analyses

When using the ILCCO dataset with predominantly European ancestry, there is anticipated low prevalence of *EGFR*-mutation.

Category	Reference	OR (95% CI)	P value	Global P value
Age in years	Per 10 year increase	1.07 (0.99–1.16)		0.10
Female	Male	1.43 (1.19–1.71)		<0.001
Asian	White	2.38 (1.95–2.91)	<0.001	<0.001
Black/other	White	1.13 (0.76–1.66)	0.55	
Unknown	White	0.90 (0.69–1.18)	0.45	
Eversmoker or unknown	Never smoker	0.27 (0.23–0.33)		<0.001
Non-Adeno	Adeno	0.36 (0.22–0.61)		<0.001



**Figure 2.**

Multivariable model and ROC when using optimal cut-point of 0.335 probability score. Top: Final multivariable model with included variables. Bottom: ROC-curves of the training, validation, and combined datasets of patients with known *EGFR* status.

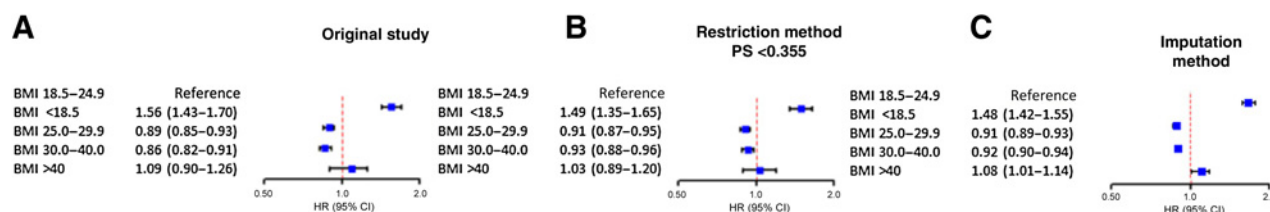
Thus, there is no cut-point that provides a PPV with sufficiently high accuracy to classify patients confidently as being *EGFR*-mutated on the basis of our multivariable regression model. However, we did explore both *EGFR*-mutated and *EGFR*-wildtype patients in the Asian subcohort because of the higher prevalence of *EGFR*-mutations in this population, which therefore leads to a higher PPV and accuracy.

When exploring these sensitivity analyses in an exclusively Asian subpopulation, we applied clinical-restriction and epidemiologic-imputation methods to generate predicted *EGFR*-wildtype and *EGFR*-mutated cohorts. The relationship between BMI and OS remained similar, when stratified by *EGFR* status, with one exception. In the subset of Asian patients with BMI >30, the BMI–OS relationship remained comparable with the original study (HR, 0.70) for predicted *EGFR*-negative patients by both restriction and imputation methods (0.65 and 0.72, respectively); however, the direction and magnitude of

the BMI–OS relationship in predicted *EGFR*-positive patients was quite different (Supplementary Table S12).

## Discussion

Leveraging the variables available in the ILCCO datasets, we built a multivariable regression model to identify *EGFR*-status among patients who had missing *EGFR*-status data, based exclusively on clinical parameters readily available in most lung cancer epidemiologic studies. We utilized two approaches to predict for *EGFR*-status in individual patients based on the regression model: the first utilized a clinically-focused, restriction approach based on identifying an optimal cut-off point to distinguish between *EGFR*-mutated and *EGFR*-wildtype subgroups; a second approach was based on an alternative epidemiologic, multiple imputation approach. We find these two approaches complementary. Although the multiple



**Figure 3.**

Forest plots of the association between BMI at diagnosis on survival for (A) results from original publication not accounting for *EGFR*-status, (B) results if accounting for *EGFR*-status using the restriction method to identify *EGFR*-wildtype patients, and (C) results if accounting for *EGFR*-status using multiple imputation to identify *EGFR*-wildtype patients.



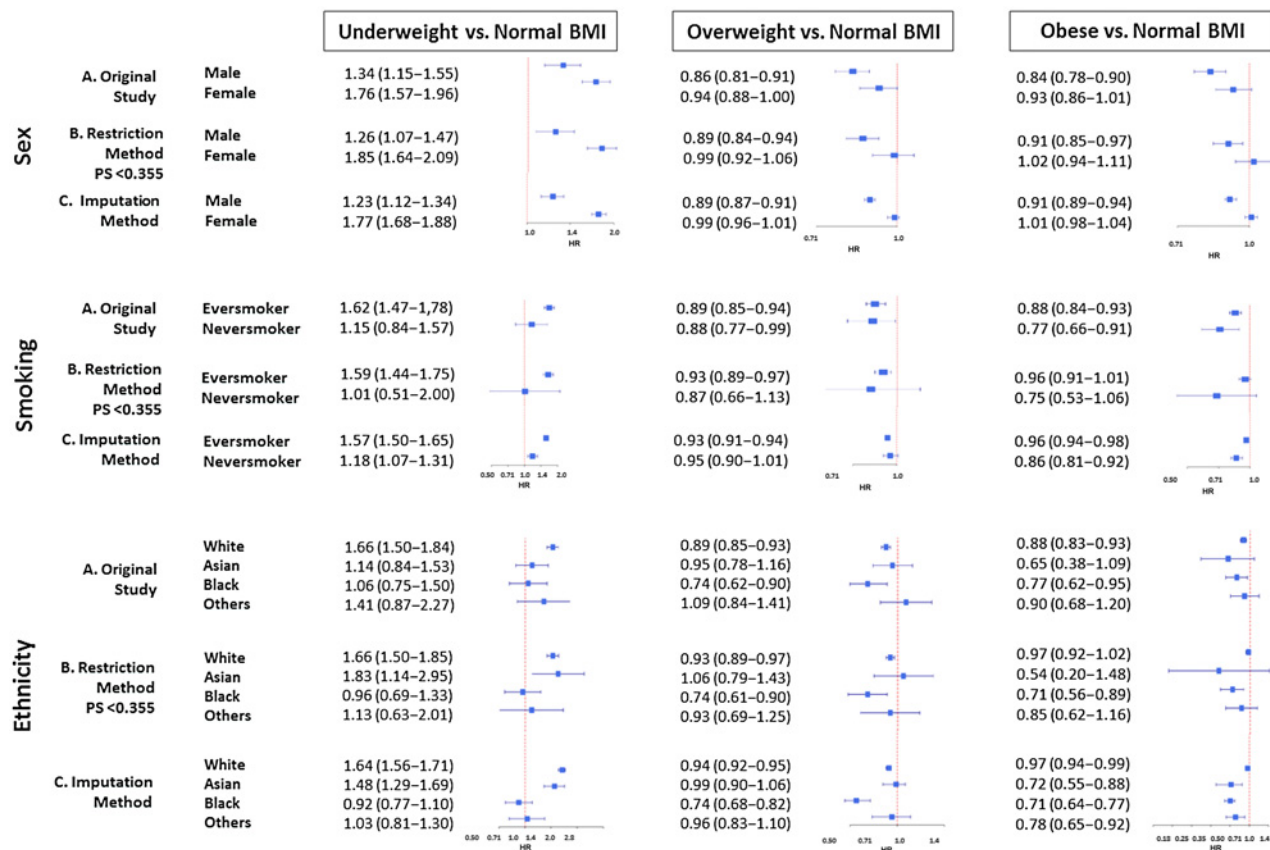


Figure 4.

Forest plots of the association between BMI and survival depending on sex, smoking status, and ethnicity. Horizontal rows show all results with regard to one patient characteristic of interest (e.g., sex, smoking status, and ethnicity). Vertical columns show all results within a certain BMI group comparison (underweight vs. normal BMI, etc.). For each patient characteristic of interest, the influence on survival is shown for three different BMI comparisons from left to right, and for each of these three different comparisons results are given for results from original publication not accounting for *EGFR*-status, results if accounting for *EGFR*-status using the restriction method to identify *EGFR*-wildtype patients, and results if accounting for *EGFR*-status using multiple imputation to identify *EGFR*-wildtype patients.

imputation approach is preferred in the epidemiologic world, the restriction approach may be more acceptable to clinicians who are uncomfortable with the concept of assigning values to missing data, no matter how scientifically rigorous this process may be.

Given the underlying population of pooled ILCCO patients with NSCLC, we focused on evaluating the utility of defining an *EGFR*-wildtype subcohort through these two approaches. We then tested the potential clinical utility of our two approaches to compare *EGFR*-wildtype subcohorts with our original full-cohort analyses on two separate hypotheses on the influence of BMI on survival; here, we confirmed that our prior full-cohort analyses had similar direction and magnitude of associations when compared with the same analyses in our *EGFR*-wildtype subcohorts. This remained largely true in an exploratory analysis of exclusively Asian subcohort where we included predicted *EGFR*-mutated and *EGFR* wildtype patients; however, some differences especially in patients with BMI >30 were observed. We cannot readily explain this difference seen in the Asian compared with the nonselected population; however, there may be residual confounding specifically relevant to the Asian population due to confounder variables we have not collected in our Caucasian-predominant dataset and therefore not adjusted for.

Missing variables are a common problem in epidemiology studies and they are commonly categorized into three different categories

depending on their relation to observed and unobserved data: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). Although for MCAR variables the probability of being missing is the same for all cases, MAR variables are missing in specific subgroups captured by the available data and NMAR variables are missing because of certain other variables not captured by the available dataset. However, methods to deal with missing data as multiple imputation are rarely utilized to account for these variables introducing bias (23).

In our dataset, *EGFR*-status was widely missing for several reasons. First, *EGFR*-testing was not standard clinical practice when most of the studies were designed or had started recruiting. However, when testing became available, oversampling bias occurred early during *EGFR*-test implementation, whereby patients selected for testing by clinicians tended to be those who had clinico-epidemiologic characteristics that enhanced the patient's probability of having an *EGFR*-mutated NSCLC, therefore enriching the population for *EGFR*-positive patients and in consequence leading to higher prevalence of *EGFR*-positive NSCLC in the tested group compared with what would be expected in the overall population. Further, availability of testing was very heterogeneous worldwide for some time. Therefore, missing *EGFR*-status data in our study population were likely a mixture of MCAR (these mutations were only identified in 2004, and broad clinical testing took



a number of years and technological advances) and MAR (testing only in selected groups); and these are the two type of missing data patterns that can be addressed by multivariable and multiple imputation techniques. When we re-analyzed our previous ILCCO analyses on influence of BMI on OS, by restricting to an *EGFR*-wildtype subcohort, the overall direction, magnitude, and significance did not change much; this result was expected, given that majority of our ILCCO patients did not fit the clinico-demographic profile of *EGFR*-mutated patients with NSCLC. Results in our Asian subcohort including predicted *EGFR*-positive patients do suggest possible differences between the predicted *EGFR*-mutated and *EGFR*-wildtype patients, substantiating our hypothesis that in *EGFR*-mutated enriched populations, epidemiologic associations may truly vary by *EGFR* status. However, these exploratory findings will need to be validated in larger datasets of Asian patients.

Several factors should be taken into account. Many of the studies that comprised the ILCCO dataset involved patients diagnosed before 2009 when the seminal IPASS trial was published and therefore during a time when testing was not standard of care in most places worldwide. Therefore, only a small proportion of these studies did actually involve patients with stage IV disease after 2009 for which a finding of *EGFR*-mutation would have resulted in treatment with an *EGFR* TKI, which consequently may lead to markedly improvement survival (24). The vast majority of these ILCCO dataset patients would not have been affected.

We thus suggest that our approaches could be most useful when analyzing contemporary datasets, patients with stage IV metastatic NSCLC, or predominantly Asian patients with NSCLC or NSCLCs in other ethnicities with known higher *EGFR*-mutation prevalence, or in any dataset where a large fraction of patients are expected to be *EGFR*-mutated and/or treated with TKI. Note that when the proportion of patients with *EGFR*-mutations is high, even the early stage resected *EGFR*-mutated patients can influence results, as some of these patients invariably will relapse over time and be treated with *EGFR* TKIs; already, patients with early stage resected (stage IB–IIIA) *EGFR*-mutation positive NSCLC will have a new standard of care TKI therapy soon, based on a recent trial (7). In such instances, our approaches to deal with missing *EGFR*-status may become critical to interpret results properly. Further, etiologic studies of NSCLC also need to determine the potential impact of *EGFR*-status on results, given that most scientists and clinicians consider *EGFR*-wildtype and *EGFR*-mutated NSCLCs to be two separate carcinogenesis pathways (25). Having established approaches to dealing with missing *EGFR*-status and the use of these approaches in sensitivity analyses provides potential pragmatic solutions to these issues.

Our analysis has several limitations. First, treatment data were only available in a small fraction of study participants, too small to incorporate into our analyses. However, this underlines the importance of accounting for *EGFR*-status, as *EGFR*-mutated patients who initially or later relapse into late stage will then likely receive TKI therapy, thereby potentially increasing survival outcomes when compared with relapsed patients with NSCLC without driver mutations. Second, as our aim was to build a pragmatic model applicable to most epidemiologic studies, we could only include a small number of very basic clinical variables that have been collected in most of the studies; however, we are satisfied that the resultant concordance indices are quite reasonable. Third, in our model we did not consider other lung cancer risk factors such as environmental tobacco exposure (26) or especially radon, for which previously some association with *EGFR* mutations has been shown (27). Finally, the sample size of our Asian subpopulation analyses was small and potential residual confounding cannot be excluded.

In conclusion, we introduce a pragmatic, step-wise method that uses both restriction and multiple imputation approaches in sensitivity analyses to evaluate the potential impact of *EGFR*-status on epidemiologic analyses of NSCLC. Our model only incorporates readily available variables and therefore trades off some accuracy for the ability to be applied across a broad set of clinical circumstances in many other populations. This method is generalizable in the common occurrence in which *EGFR*-status data are missing from epidemiologic studies. With this method, we lay the foundation to refine future epidemiologic studies of NSCLC risk and outcome.

## Disclaimer

Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer/World Health Organization.

## Authors' Disclosures

S. Schmid reports other support from Swiss Cancer Research Foundation, AstraZeneca, MSD, BMS, Boehringer Ingelheim; grants from AstraZeneca and BMS; and personal fees from Boehringer Ingelheim, Takeda, and MSD outside the submitted work. L. Ferro Leal reports grants from AstraZeneca - Brazil outside the submitted work. A.S. Wenzlaff reports grants from NIH during the conduct of the study. L. Le Marchand reports grants from NCI during the conduct of the study. A.G. Schwartz reports grants from NIH during the conduct of the study. L.C. Sakoda reports grants from NCI and California Tobacco-Related Disease Research Program; personal fees from NIH; and other support from National Lung Cancer Roundtable outside the submitted work. G. Liu reports other support from Princess Margaret Cancer Foundation during the conduct of the study as well as grants and personal fees from AstraZeneca, Takeda, and Roche; personal fees from Pfizer, Novartis, Bristol Myers Squibb, EMD Serono, Merck, Amgen, AbbVie, and Jazz Pharmaceuticals; and grants from Boehringer outside the submitted work. No disclosures were reported by the other authors.

## Authors' Contributions

**S. Schmid:** Conceptualization, formal analysis, investigation, methodology, writing—original draft. **M. Jiang:** Conceptualization, data curation, formal analysis, investigation, methodology, writing—original draft. **M.C. Brown:** Conceptualization, data curation, formal analysis, investigation, methodology, writing—original draft. **A. Fares:** Conceptualization, data curation, methodology, writing—review and editing. **M. Garcia:** Conceptualization, data curation, methodology, writing—review and editing. **J. Soriano:** Data curation, investigation, writing—review and editing. **M. Dong:** Data curation, formal analysis, methodology, writing—review and editing. **S. Thomas:** Data curation, investigation, writing—review and editing. **T. Kohno:** Data curation, investigation, writing—review and editing. **L.F. Leal:** Data curation, investigation, writing—review and editing. **N. Diao:** Data curation, investigation, writing—review and editing. **J. Xie:** Data curation, investigation, writing—review and editing. **Z. Wang:** Data curation, investigation, writing—review and editing. **D. Zaridze:** Data curation, investigation, writing—review and editing. **I. Holcatova:** Data curation, investigation, writing—review and editing. **J. Lissowska:** Data curation, investigation, writing—review and editing. **B. Świątkowska:** Data curation, investigation, writing—review and editing. **D. Mates:** Data curation, investigation, writing—review and editing. **M. Savic:** Data curation, investigation, writing—review and editing. **A.S. Wenzlaff:** Data curation, investigation, writing—review and editing. **C.C. Harris:** Data curation, investigation, writing—review and editing. **N.E. Caporaso:** Data curation, investigation, writing—review and editing. **H. Ma:** Data curation, investigation, writing—review and editing. **G. Fernandez-Tardon:** Data curation, investigation, writing—review and editing. **M.J. Barnett:** Data curation, investigation, writing—review and editing. **G. Goodman:** Data curation, investigation, writing—review and editing. **M.P.A. Davies:** Data curation, investigation, writing—review and editing. **M. Pérez-Ríos:** Data curation, investigation, writing—review and editing. **F. Taylor:** Data curation, investigation, writing—review and editing. **E.J. Duell:** Data curation, investigation, writing—review and editing. **B. Schoettker:** Data curation, investigation, writing—review and editing. **H. Brenner:** Data curation, investigation, writing—review and editing. **A. Andrew:** Data curation, investigation, writing—review and editing. **A. Cox:** Data curation, investigation, writing—review

and editing. **A. Ruano-Ravina**: Data curation, investigation, writing-review and editing. **J.K. Field**: Data curation, investigation, writing-review and editing. **L. Le Marchand**: Data curation, investigation, writing-review and editing. **Y. Wang**: Data curation, investigation, writing-review and editing. **C. Chen**: Data curation, investigation, writing-review and editing. **A. Tardon**: Data curation, investigation, writing-review and editing. **S. Shete**: Data curation, investigation, writing-review and editing. **M.B. Schabath**: Data curation, investigation, writing-review and editing. **H. Shen**: Data curation, investigation, writing-review and editing. **M.T. Landi**: Data curation, investigation, writing-review and editing. **B.M. Ryan**: Data curation, investigation, writing-review and editing. **A.G. Schwartz**: Data curation, investigation, writing-review and editing. **L. Qi**: Data curation, investigation, writing-review and editing. **L.C. Sakoda**: Data curation, investigation, writing-review and editing. **P. Brennan**: Data curation, investigation, writing-review and editing. **P. Yang**: Data curation, investigation, writing-review and editing. **J. Zhang**: Data curation, investigation, writing-review and editing. **D.C. Christiani**: Data curation, investigation, writing-review and editing. **R.M. Reis**: Data curation, investigation, writing-review and editing. **K. Shiraishi**: Data curation, investigation, writing-review and editing. **R.J. Hung**: Conceptualization, data curation, investigation, writing-review and editing. **W. Xu**: Conceptualization, formal

analysis, methodology, writing-original draft. **G. Liu**: Conceptualization, data curation, formal analysis, supervision, investigation, methodology, writing-original draft.

## Acknowledgments

This study was partially supported by the Public Ministry of Labor Campinas (Research, Prevention, and Education of Occupational Cancer), FINEP - CT-INFRA (02/2010). We thank all members of the GTOP group (Translational Group of Pulmonary Oncology - Barretos Cancer Hospital, Brazil). Caret-Study was funded by the NCI, NIH, through grants U01-CA063673, UM1-CA167462, and U01-CA167462. D.C. Christiani has received funding through an U01 Grant (U01 CA209414). G. Liu was supported by Alan B. Brown Chair and the Lusi Wong Family Fund, Princess Margaret Cancer Foundation. M.C. Brown is supported by the Alan B. Brown Chair. S. Schmid was supported by the Swiss Cancer Research Foundation.

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 August 20, 2021; revised October 28, 2021; accepted January 5, 2022; published first January 13, 2022.

## References

- Pao W, Miller V, Zakowski M, Doherty J, Politi K, Sarkaria I, et al. EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc Natl Acad Sci U S A* 2004;101:13306–11.
- Tsao AS, Tang XM, Sabloff B, Xiao L, Shigematsu H, Roth J, et al. Clinicopathologic characteristics of the EGFR gene mutation in non-small cell lung cancer. *J Thorac Oncol* 2006;1:231–9.
- Chapman AM, Sun KY, Ruestow P, Cowan DM, Madl AK. Lung cancer mutation profile of EGFR, ALK, and KRAS: meta-analysis and comparison of never and ever smokers. *Lung Cancer* 2016;102:122–34.
- Socinski MA, Obasaju C, Gandara D, Hirsch FR, Bonomi P, Bunn P, et al. Clinicopathologic features of advanced squamous NSCLC. *J Thorac Oncol* 2016;11:1411–22.
- Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med* 2004;350:2129–39.
- Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 2004;304:1497–500.
- Wu YL, Tsuboi M, He J, John T, Grohe C, Majem M, et al. Osimertinib in resected EGFR-mutated non-small-cell lung cancer. *N Engl J Med* 2020;383:1711–23.
- Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, Saijo N, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med* 2009;361:947–57.
- Xue C, Hu Z, Jiang W, Zhao Y, Xu F, Huang Y, et al. National survey of the medical treatment status for non-small cell lung cancer (NSCLC) in China. *Lung Cancer* 2012;77:371–5.
- Cai L, Chen Y, Tong X, Wu X, Bao H, Shao Y, et al. The genomic landscape of young and old lung cancer patients highlights age-dependent mutation frequencies and clinical actionability in young patients. *Int J Cancer* 2021.
- Zhang YL, Yuan JQ, Wang KF, Fu XH, Han XR, Threapleton D, et al. The prevalence of EGFR mutation in patients with non-small cell lung cancer: a systematic review and meta-analysis. *Oncotarget* 2016;7:78985–93.
- Korpanty GJ, Kamel-Reid S, Pintilie M, Hwang DM, Zer A, Liu G, et al. Lung cancer in never smokers from the Princess Margaret Cancer Centre. *Oncotarget* 2018;9:22559–70.
- Leal LF, de Paula FE, De Marchi P, de Souza Viana L, Pinto GDJ, Carlos CD, et al. Mutational profile of Brazilian lung adenocarcinoma unveils association of EGFR mutations with high Asian ancestry and independent prognostic role of KRAS mutations. *Sci Rep* 2019;9:3209.
- Chang H, Liu YB, Yi W, Lu JB, Zhang JX. Development and validation of a model to predict tyrosine kinase inhibitor-sensitive EGFR mutations of non-small cell lung cancer based on multi-institutional data. *Thorac Cancer* 2018;9:1680–6.
- Girard N, Sima CS, Jackman DM, Sequist LV, Chen H, Yang JC, et al. Nomogram to predict the presence of EGFR activating mutation in lung adenocarcinoma. *Eur Respir J* 2012;39:366–72.
- DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–45.
- Akobeng AK. Understanding diagnostic tests 3: receiver operating characteristic curves. *Acta Paediatr* 2007;96:644–7.
- Perkins NJ, Schisterman EF, Vexler A. Receiver operating characteristic curve inference from a sample with a limit of detection. *Am J Epidemiol* 2007;165:325–33.
- Ren J, Xu W, Su J, Ren X, Cheng D, Chen Z, et al. Multiple imputation and clinico-serological models to predict human papillomavirus status in oropharyngeal carcinoma: an alternative when tissue is unavailable. *Int J Cancer* 2020;146:2166–74.
- Habbous S, Chu KP, Qiu X, La Delfa A, Harland LT, Fadhel E, et al. The changing incidence of human papillomavirus-associated oropharyngeal cancer using multiple imputation from 2000 to 2010 at a Comprehensive Cancer Centre. *Cancer Epidemiol* 2013;37:820–9.
- Shepshelovich D, Xu W, Lu L, Fares A, Yang P, Christiani D, et al. Body mass index (BMI), BMI change, and overall survival in patients with SCLC and NSCLC: a pooled analysis of the International Lung Cancer Consortium. *J Thorac Oncol* 2019;14:1594–607.
- Jiang M, Fares AF, Shepshelovich D, Yang P, Christiani D, Zhang J, et al. The relationship between body-mass index and overall survival in non-small cell lung cancer by sex, smoking status, and race: a pooled analysis of 20,937 International lung Cancer consortium (ILCCO) patients. *Lung Cancer* 2020;152:58–65.
- Desai M, Kubo J, Esserman D, Terry MB. The handling of missing data in molecular epidemiology studies. *Cancer Epidemiol Biomarkers Prev* 2011;20:1571–9.
- Ramalingam SS, Vansteenkiste J, Planchard D, Cho BC, Gray JE, Ohe Y, et al. Overall survival with osimertinib in untreated, EGFR-mutated advanced NSCLC. *N Engl J Med* 2020;382:41–50.
- Reungwetwattana T, Weroha SJ, Molina JR. Oncogenic pathways, molecularly targeted therapies, and highlighted clinical trials in non-small-cell lung cancer (NSCLC). *Clin Lung Cancer* 2012;13:252–66.
- Torres-Duran M, Ruano-Ravina A, Kelsey KT, Parente-Lamelas I, Leiro-Fernandez V, Abdulkader I, et al. Environmental tobacco smoke exposure and EGFR and ALK alterations in never smokers' lung cancer. Results from the LCRINS study. *Cancer Lett* 2017;411:130–5.
- Ruano-Ravina A, Torres-Duran M, Kelsey KT, Parente-Lamelas I, Leiro-Fernandez V, Abdulkader I, et al. Residential radon, EGFR mutations and ALK alterations in never-smoking lung cancer cases. *Eur Respir J* 2016;48:1462–70.