# Accounting for *EGFR* mutations in epidemiological analyses of non-small cell lung cancers: Examples based on the International Lung Cancer Consortium data

## Running Title (60 characters): Accounting for *EGFR* mutations in epidemiological analyses

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

+ Contributed equally

* Contributed equally

**Institutions**:

1 The Princess Margaret Cancer Centre and University Health Network, University of Toronto, Toronto, ON, Canada
2 Department of Medical Oncology, Cantonal Hospital St.Gallen, St.Gallen, Switzerland
3 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
4 Division of Medical Oncology, Hospital de Base de São José do Rio Preto, SP, Brazil
5 University of Ottawa, Ottawa, ON Canada
6 Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

7 Lunenfeld-Tanenbaum Research Institute, Sinai Health Systems, Toronto, Canada
8 Division of Genome Biology, National Cancer Centre Research Institute, Tokyo, Japan
9 Molecular Oncology Research Center, Barretos Cancer Hospital, Barretos, Brazil
10 Harvard T.H. Chan School of Public Health, Boston, MA, USA
11 Department of Thoracic Surgery, Fudan University Shanghai Cancer Center, Shanghai, China
12 Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
13 Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China
14 Russian N.N. Blokhin Cancer Research Centre, Moscow, Russian Federation
15 Department of Cancer Epidemiology and Prevention, M. Sklodowska-Curie National Research Institute of Oncology
16 Institute of Hygiene and Epidemiology, 1st Faculty of Medicine, Charles University, Prague, Czech Republic
17 The Nofer Institute of Occupational Medicine, Lodz, Poland
18 National Institute of Public Health, Bucharest, Romania.
19 Department of Thoracic Surgery, Clinical Center of Serbia, Belgrade, Serbia
20 Barbara Ann Karmanos Cancer Institute, Wayne State University, Detroit, MI, USA
21 Laboratory of Human Carcinogenesis, Centre for Cancer Research, National Institutes of Health, Bethesda, MD, USA
22 Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
23 Department of Epidemiology, School of Public Health, Nanjing Medical University, Nanjing, China
24 IUOPA, University of Oviedo, and ISPA (Health Research Institute of the Principality of Asturias) and CIBERESP, Asturias, Spain
25 Program in Biostatistics Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
26 Swedish Cancer Institute, Seattle, Washington
27 Roy Castle Lung Cancer Research Programme, The University of Liverpool, Department of Molecular and Clinical Cancer Medicine, Liverpool, UK
28 Department of Preventive Medicine and Public Health, University of Santiago de Compostela, Spain
29 CIBER de Epidemiología y Salud Pública, CIBERESP, Santiago de Compostela, Spain
30 Department of Oncology and Metabolism, University of Sheffield Medical School, Sheffield, UK
31 Sheffield Teaching Hospitals Foundation Trust, Sheffield, UK
32 Catalan Institute of Oncology (ICO), Barcelona, Spain
33 Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain
34 Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany
35 Network of Aging Research, Heidelberg University, Heidelberg, Germany
36 Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany
37 German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany
38 Dartmouth-Hitchcock Medical Center, Lebanaon, NH, USA

39 University of Hawaii Cancer Centre, Hawaii, USA
40 American Cancer Society, Atlanta, GA, USA
41 Program in Epidemiology, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
42 The University of Texas MD Anderson Cancer Center, Houston, Texas, USA
43 Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA
44 The University of California Davis Medical Sciences, Davis, California, USA
45 Kaiser Permanente Northern California, Division of Research, Oakland, California, USA
46 International Agency for Research on Cancer, Lyon, France
47 Life and Health Sciences Research Institute (ICVS), Medical School, University of Minho, Braga, Portuga
48 ICVS/3B's-PT Government Associate Laboratory, Braga, Portugal

**Co-corresponding authors**:

**Geoffrey Liu, (MD MSc FRCPC FISPE)**

Division of Medical Oncology

Princess Margaret Cancer Centre

University Health Network

610 University Ave.

Toronto ON M5G 2M9

Mail: Geoffrey.Liu@uhn.ca

Phone: 416-946-2000 extension 3428

**Wei Xu (PhD)**

Princess Margaret Cancer Centre

University Health Network

610 University Ave.

Toronto ON M5G 2M9

Mail: Wei.Xu@uhnresearch.ca

Phone: 416-946-4501 extension 4497

**COI-Statement:** The authors declare no conflicts of interest for this manuscript

## ABSTRACT

**Introduction**: 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 epidemiological datasets. We developed and tested pragmatic approaches to account for *EGFR*-mutation-status based on variables commonly included in epidemiological datasets and evaluated the clinical utility of these approaches.

**Methods**: Through analysis of the International Lung Cancer Consortium (ILCCO) epidemiological datasets, we developed a regression model for *EGFR*-status; we then applied a clinical-restriction approach using the optimal cutpoint, and a second epidemiological, 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 4231 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 NSCLC patients, 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.

**Conclusion**: Pragmatic analytical methods have the potential to elucidate the impact of *EGFR*-mutation-status in settings where *EGFR*-mutation status is unknown, an occurence common across many epidemiological datasets of NSCLC patients.

## INTRODUCTION

Somatic epidermal growth factor receptor (*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. Clinico-pathologic 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* wild-type) (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 *EGFR*-mutated NSCLC patients 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 tumours represent the largest subgroup whose biology is markedly different than of typical smoking-related NSCLC; proportions of *EGFR*-mutated tumours can range from 10% to upwards of 50% (10-13). Thus, epidemiological studies aiming to gain better understanding of the genetic and environmental etiological 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 *et al.* 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 serological 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 *et al.* 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 versus IIIB-IV) and predominant histological 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 epidemiological studies because they incorporate some variables that are not readily available in existing epidemiological 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 epidemiological studies, using variables generally included in existing datasets. We developed a regression model for *EGFR*-status by analyzing International Lung Cancer Consortium (ILCCO) epidemiological 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 cutpoint 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 epidemiological 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.

## 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 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 epidemiological multiple imputation approach utilizing the same regression model for estimating *EGFR*-status.

***Study population:*** ILCCO harmonizes compatible data from various epidemiological 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 lung cancer patients 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 clinico-pathologic 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. 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 (ORs) and 95% confidence intervals (CIs) 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 area under the curve (AUC) of the receiver operating characteristic curve (ROC). The probability score (PS) was defined based on 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.

***Epidemiological 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 hazard ratios (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 hazard ratios and 95% confidence intervals (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 epidemiological-imputation approaches to identifying an *EGFR*-wildtype subcohort were individually compared to the analysis of the entire ILCCO cohort. In the Asian subgroup, restriction and imputation were also applied to generate a predicted *EGFR*-positive subgroup to be compared to 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 to previous analyses that did not consider *EGFR*-status, for the following two associations: BMI and overall survival (OS) (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 hazard ratios were generated, we summarized the data as mean hazard ratios and 95% confidence intervals. 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, Vienna, Austria). All P values were based on 2-sided tests and considered statistically significant at $P < 0.05$.

## 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, whilst 31,125 patients across 27 studies had unknown *EGFR*-status (**Figure 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 1**). The characteristics of those with known and unknown *EGFR*-status are presented in **Supplementary Table 2**. Of the patients with known *EGFR* status, 1,481 were EGFR-mutated whilst 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%) while American studies had lower prevalence (LCS 21%; Barretos-Brazil 19%); the multicultural Toronto MSH-PMH study had an intermediate prevalence of 42% (**Supplementary Table 3**). 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 4**).

***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 ≥25 $kg/m^2$ and having any smoking history was inversely associated with being *EGFR*-mutated, as was heavy smoking (**Supplementary Table 4**). 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 4**).

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. Concordance indices (C-indices) were very similar across models containing different variables: all between 0.740 and 0.778. 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, 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 (**Figure 2**).

***Choosing a clinically relevant probability score cut-point from the multivariable regression model for being EGFR-wildtype*:** Based on the ROC-curve generated by our model (**Figure 2**) and distribution of PS (**Supplementary Figure 1**) 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), a sensitivity of 69% and specificity of 72.5% could be achieved. Lower PS cut-points would have resulted in decreased specificity.

The *EGFR* status-known dataset of 4231 patients had a 35% *EGFR* mutation prevalence that corresponded to a *EGFR*-mutated positive predictive value (PPV) of 57% and negative predictive value (NPV) of 81%; the NPV was thus reasonably associated with identifying *EGFR*-wildtype NSCLC patients while retaining 2453 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=1879).

When assessing all ILCCO participants (n= 35,356) (**Supplementary Figure 1D**), the PS distribution was very different from the PS 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 2**). This was because there was over-sampling of the *EGFR*-mutated patients amongst all tested patients: until centres 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.

**Overall Survival (OS) of *EGFR*-wildtype patients, as determined by different approaches**: As expected, the OS of *EGFR*-mutated patients was longer, compared to *EGFR*-wildtype patients **(Supplementary Figure 2A-B**). 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 cutpoint of <0.335 selected patients with median OS closer to the known *EGFR*-wildtype patients (**Supplementary Figure 2C**). To avoid confounding by stage, we also performed the same comparisons, but restricted to Stage IV patients only (**Supplementary Figure 2D**). 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 Figures 2E and 2F**) 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 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 epidemiological-imputation approaches.

When re-analyzing our previous studies on the influence of BMI on OS in NSCLC patients by clinical-restriction or epidemiological-imputation approaches, the direction of change remained the same for all BMI levels and interactions. In most cases the magnitude of hazard ratios was similar too; however, in a few subgroups, the overall effect size varied (**Figures 3 and 4; Supplementary Tables 5 and 6**).

***Asian subcohort analyses:*** When using the ILCCO dataset with predominantly European ancestry, there is anticipated low prevalence of *EGFR*-mutation. Thus, there is no cut-point that provides a PPV with sufficiently high accuracy to classify patients confidently as being *EGFR*-mutated based on 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 epidemiological-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 to 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 8**).

## **DISCUSSION**:

Leveraging the variables available in the ILCCO datasets, we built a multivariable regression model to identify *EGFR*-status amongst patients who had missing *EGFR*-status data, based exclusively on clinical parameters readily available in most lung cancer epidemiological 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 epidemiological, multiple imputation approach. Given the underlying population of pooled ILCCO NSCLC patients, 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 to 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.

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). Whereas 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. Firstly, *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 to 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 was 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 NSCLC patients. 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, epidemiological 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 did actually include patients after 2009 for which a finding of *EGFR*-mutation would have resulted in treatment with an EGFR TKI, and only in Stage IV NSCLC. Though *EGFR* status may lead to minimal or small differences in prognosis in most settings, markedly improved survival can be seen in patients treated with EGFR TKI; in the case where patients were treated with the most recent standard of care, first-line

osimertinib, a median overall survival of 38.6 months was observed (24).

We thus suggest that our approaches could be most useful when analyzing contemporary datasets, Stage IV metastatic NSCLC patients, or predominantly Asian NSCLC patients 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 even 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 resected stage IB-IIIA *EGFR*-mutation positive NSCLC will have 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 intrepret results. Further, etiological 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. Firstly, treatment data was 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 to relapsed NSCLC patients without driver mutations. Secondly, as our aim was to build a pragmatic model applicable to most epidemiological 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. Thirdly, in our model we did not consider other lung cancer risk factors such as environmental tobacco exposure(26) or especially radon, for which previoulsy some association with *EGFR* mutations has been shown(27).

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 epidemiological 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 epidemiological studies. With this method, we lay the foundation to refine future epidemiological studies of NSCLC risk and outcome.

Funding:

- 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).
- DC has received funding through an U01 Grant (U01 CA209414).
- Geoffrey Liu was supported by Alan B. Brown Chair and the Lusi Wong Family Fund, Princess Margaret Cancer Foundation. M. Catherine Brown is supported by the Alan B. Brown Chair.
- Sabine Schmid was supported by the Swiss Cancer Research Foundation

**REFERENCES**:


1. 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(36):13306-11.
2. 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(3):231-9.
3. 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.
4. 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(9):1411-22.
5. 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(21):2129-39.
6. 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(5676):1497-500.
7. 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(18):1711-23.
8. 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(10):947-57.
9. 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(2):371-5.
10. 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.
11. 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(48):78985-93.
12. 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(32):22559-70.
13. 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(1):3209.
14. 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(12):1680-6.
15. 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(2):366-72.
16. 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(3):837-45.
17. Akobeng AK. Understanding diagnostic tests 3: Receiver operating characteristic curves. Acta Paediatr. 2007;96(5):644-7.
18. Perkins NJ, Schisterman EF, Vexler A. Receiver operating characteristic curve inference from a sample with a limit of detection. Am J Epidemiol. 2007;165(3):325-33.
19. 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(8):2166-74.
20. 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(6):820-9.

21. 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(9):1594-607.
22. 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.
23. Desai M, Kubo J, Esserman D, Terry MB. The handling of missing data in molecular epidemiology studies. Cancer Epidemiol Biomarkers Prev. 2011;20(8):1571-9.
24. 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(1):41-50.
25. 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(4):252-66.
26. 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.
27. 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(5):1462-70.

**Figures and Tables main manuscript (in order they appear in the text)**

**Figure 1: CONSORT diagram:**

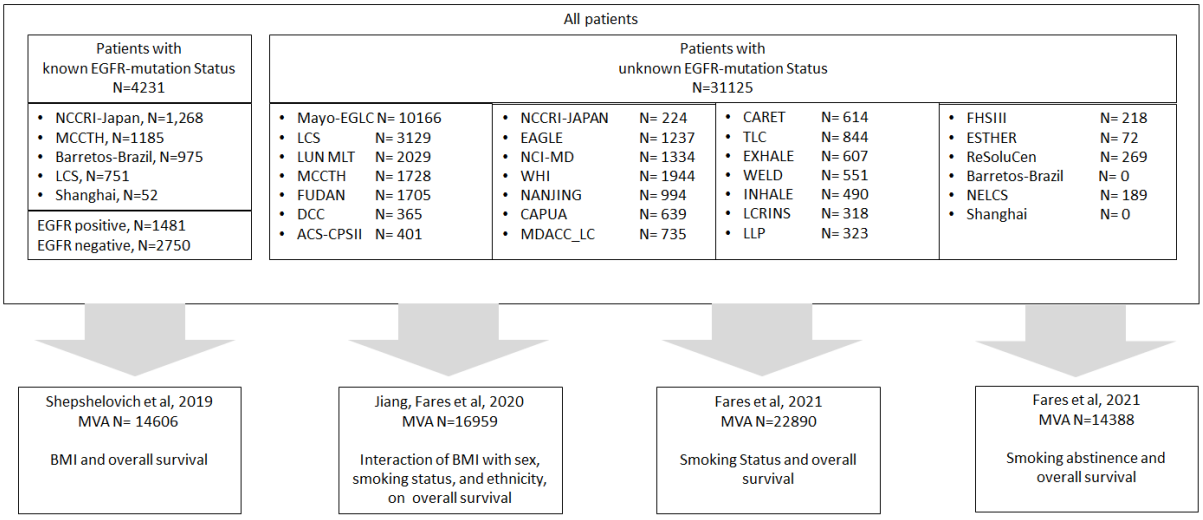

**Table 1: Baseline characteristics of *EGFR* mutation-tested (i.e. mutation-known) cohort, overall and by *EGFR* status**

| Covariate | Category | Patients with *EGFR* mutation-tested tumours: N (%) | | | p-value |
|---|---|---|---|---|---|
| | | Full Sample | *EGFR*-Mutated | *EGFR*-Wildtype | |
| **Total Count (100%)** | | 4231 | 1481 | 2750 | |
| **Age** | Median [Min-Max] | 63 [18-95] | 62 [22-95] | 63 [18-93] | **0.008** |
| **Sex** | Male | 1976 (47) | 510 (34) | 1466 (53) | **<0.001** |
| | Female | 2255 (53) | 971 (66) | 1284 (47) | |
| **Ethnicity** | White | 1513 (43) | 371 (28) | 1142 (53) | **<0.001** |
| | Asian | 1727 (49) | 892 (67) | 835 (39) | |
| | Black/Other | 252 (7) | 65 (5) | 187 (9) | |
| | Unknown | 739 | 153 | 586 | |
| **BMI (kg/m$^2$)** | <18 5 | 1201 (53) | 432 (57) | 769 (50) | **0.0083** |
| | 18 5-< 25 | 159 (7) | 46 (6) | 113 (7) | |
| | >=25 | 925 (40) | 278 (37) | 647 (42) | |
| | Unknown | 1946 | 725 | 1221 | |
| **Smoking status** | Never | 1686 (40) | 964 (66) | 722 (27) | **<0.001** |
| | Former | 1348 (32) | 349 (24) | 999 (37) | |
| | Current | 1133 (27) | 155 (11) | 978 (36) | |
| | Unknown | 64 | 13 | 51 | |
| **Packyears*** | ≤20 | 410 (26) | 179 (48) | 231 (19) | **<0.001** |
| | >20 | 1151 (74) | 195 (52) | 956 (81) | |
| | Unknown | 920 | 130 | 790 | |
| **NSCLC Histology** | Adeno | 3974 (94) | 1455 (98) | 2519 (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 | 1372 (32) | 565 (38) | 807 (29) | **<0.001** |
| | II | 326 (8) | 106 (7) | 220 (8) | |
| | III | 784 (19) | 227 (15) | 557 (20) | |
| | IV | 1749 (41) | 583 (39) | 1166 (42) | |

*** Only among ever-smokers**

**Figure 2: Multivariable Model and Receiver Operator Curve when using optimal cutpoint of 0.335 probability score.** Top: Final multivariable model with included variables. Bottom: ROC-curves of the Training, Validation and Combined datasets of with known *EGFR* status.

| Category | Reference | Odds Ratio (95% Confidence Interval) | 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 | |
| Black/Other | White | 1.13 (0.76-1.66) | 0.55 | <0.001 |
| Unknown | White | 0.90 (0.69-1.18) | 0.45 | |
| Ever Smoker or unknown | Never Smoker | 0.27 (0.23-0.33) | | <0.001 |
| Non-Adeno | Adeno | 0.36 (0.22-0.61) | | <0.001 |

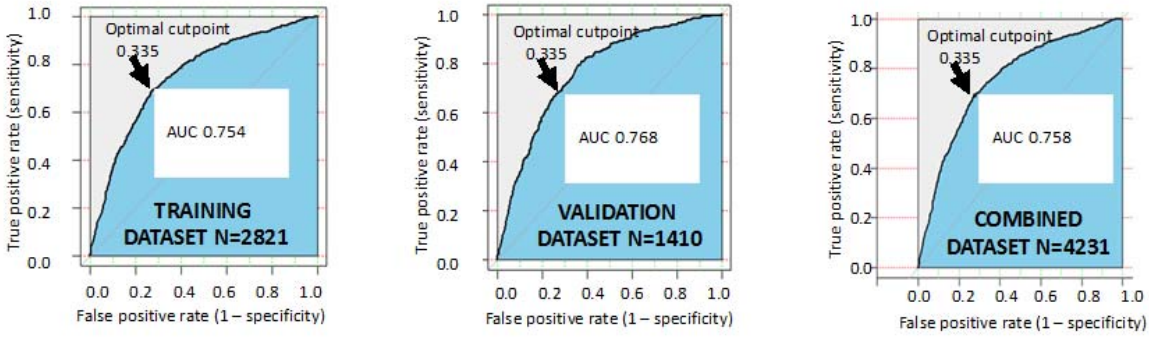

**Figure 3. Forest plots of the association between BMI at diagnosis on survival** for A) results from original publication not accounting for *EGFR*-status. B) results if accounting for *EGFR*-status using the restriction method to identify *EGFR*-wildtype patients; C) results if accounting for *EGFR*-status using multiple imputation to identify *EGFR*-wildtype patients.

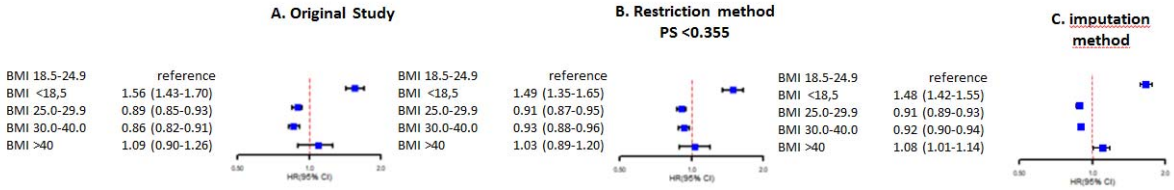

**Figure 4: Forest plots of the association between BMI and survival in depending on sex, smoking status and ethnicity.** Horizontal rows show all results with regard to one patient characteristic of interest (e.g. sex, smoking status and ethnicity. Vertical columns show all results within a certain BMI group comparison (underweight vs normal BMI etc). For each patient characteristic of interest the influence on survival is shown for three different BMI comparison from left to right and for every 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.