

Supplemental Online Content

Popat S, Liu SV, Scheuer N, et al. Association between smoking history and overall survival in patients receiving pembrolizumab for first-line treatment of advanced non-small cell lung cancer. *JAMA Netw Open*. 2022;5(5):e2214046. doi:10.1001/jamanetworkopen.2022.14046

eAppendix. Variable Definitions and Comments

eMethods.

eFigure 1. Schoenfeld Residuals and Log-Negative-Log (LNL) Survival Curves for Tests of Proportional Hazards Before and After IPTW-Adjustment

eFigure 2. Distribution of IPTW Weights

eFigure 3. Delta-based Tipping Point Analysis for ECOG PS

This supplemental material has been provided by the authors to give readers additional information about their work.

eAppendix. Variable Definitions and Comments

Variable definitions for baseline covariates included in this study.

- Age at index date (years) – Categorized as <65 or ≥65 years of age
- Sex – Male or female
- Race – White, Other (Asian, African America/Black, or Latino/Hispanic) or Unknown
- Cancer stage at diagnosis – Categorized as stage I-III or IV
- Tumor histology – non-squamous, squamous or unknown
- ECOG PS – Categorized as 0, 1 or 2+
 - Baseline ECOG PS was measured between 30 days prior to and/or 7 days after index date
- Time from diagnosis to index date (months)
- Comorbidities – identified using ICD-9/10 codes from the Charlson comorbidity index for comorbidities related to liver disease, pulmonary disease, cardiovascular disease, diabetes or rheumatic conditions measured within 90 days prior to index date. Categorized as 0 (none) or 1+ (any).
- Metastases – identified using ICD-9/10 codes. Categorized as present or absent.
- Recorded PD-L1 positivity based on a test at or prior to index date
- Insurance status – insured or uninsured at index date
- Practice type – community or academic practice

In the real-world dataset, dates of death were not recorded and only month and year were available. Therefore, the 15th day of the month was used to derive overall survival. Smoking status information is collected at baseline, and never smokers are patients that have never smoked in their entire lifetime up to baseline. In contrast, ever-smokers are those who reported having smoked at least once in their entire lifetime, and this may have occurred in the past, and/or they may still be smoking at baseline. We do not have information regarding smoking duration, pack years, intensity, cessation time in the Flatiron Health dataset used. We did not consider this to be a limitation as our research question focussed on the most simple comparison between a group that had any smoking history at all, versus one that had none in their entire lifetime up to baseline.

eMethods.

Inverse probability treatment weighting (IPTW)

IPTW was used to adjust for differences in baseline covariates between ever-smoker and never-smoker groups, and the chosen estimand was the average treatment effect (ATE). This allowed comparison of overall survival between these groups after minimizing imbalances in prognostic variables. A logistic regression was used to model the probability $P(S|L)$ of having a positive smoking status $S = 1$ at diagnosis conditional on baseline covariates L . Patients were weighted by the inverse of $P(S|L)$ stabilized by the marginal probability of being a former and/or current smoker $P(S = 1)$, i.e., $\frac{P(S=1)}{P(S=1|L)}$ to create stabilized IPTW weights. For the primary analysis, variables included in the logistic regression model were – age, sex, ECOG PS, cancer stage, histology and PD-L1 positivity. IPTW based on this model yielded sufficient balance across all covariates listed in Table 1 used in the primary analysis except ECOG PS and race, which were therefore also included as covariates in the Cox model. Though there was residual imbalance in the absence of CNS metastases variable, it is known to be under-recorded and therefore was not adjusted for in the primary analysis. Sensitivity analysis adjusting for this variable in addition to ECOG PS and race; and separately including patients with comorbidities was performed. IPTW modelling was blinded to the study outcomes. Robust standard errors were estimated to account for weighting.

Multiple imputation and pooling

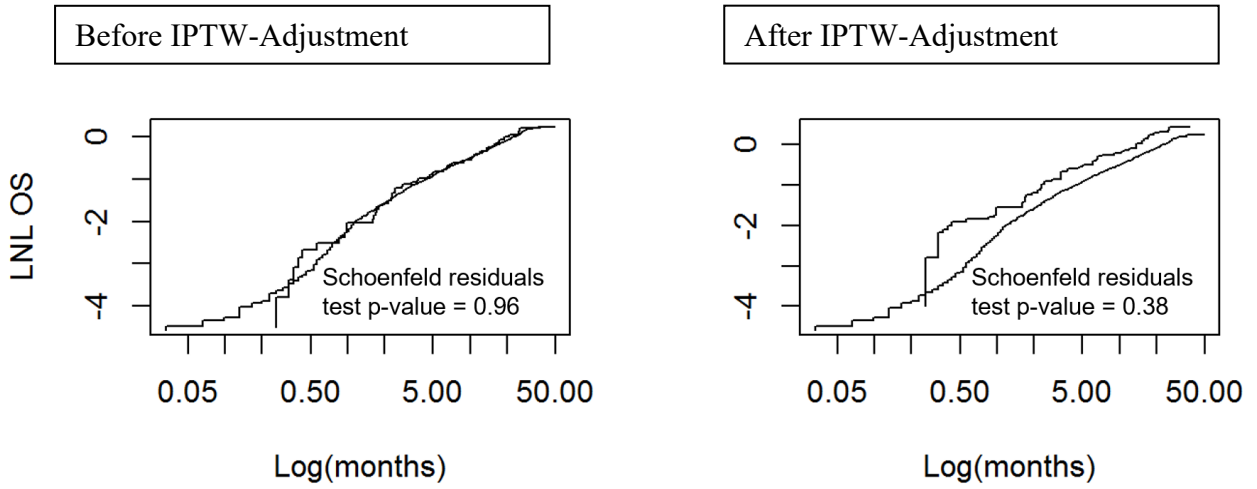
Assuming missingness at random (MAR), missing data for measured baseline covariates were imputed using multiple imputation with chained equations. 25 imputed data sets were generated to account for random error in the imputation of missing values. All variables used in propensity score estimation and Cox regression were included in multiple imputation, including outcome variables. Hazard ratios and standard errors were computed for each imputed data set independently and then pooled using Rubin rules to account for intra- and inter-data set variance.

Delta (δ)-based tipping point analysis

Given the non-negligible proportion of missing data (>10%) in baseline ECOG PS, we performed a tipping point-based bias analysis using δ -based shifts under the assumption that ECOG PS was missing not at random (MNAR) for the ever-smokers group, which was larger. For δ adjustments δ was an additive term applied to the ordered logistic regression model for ECOG PS (Y) and $j = \{0, 1, 2, 3, 4\}$ representing $\log \frac{p(Y \leq j)}{p(Y > j)}$.

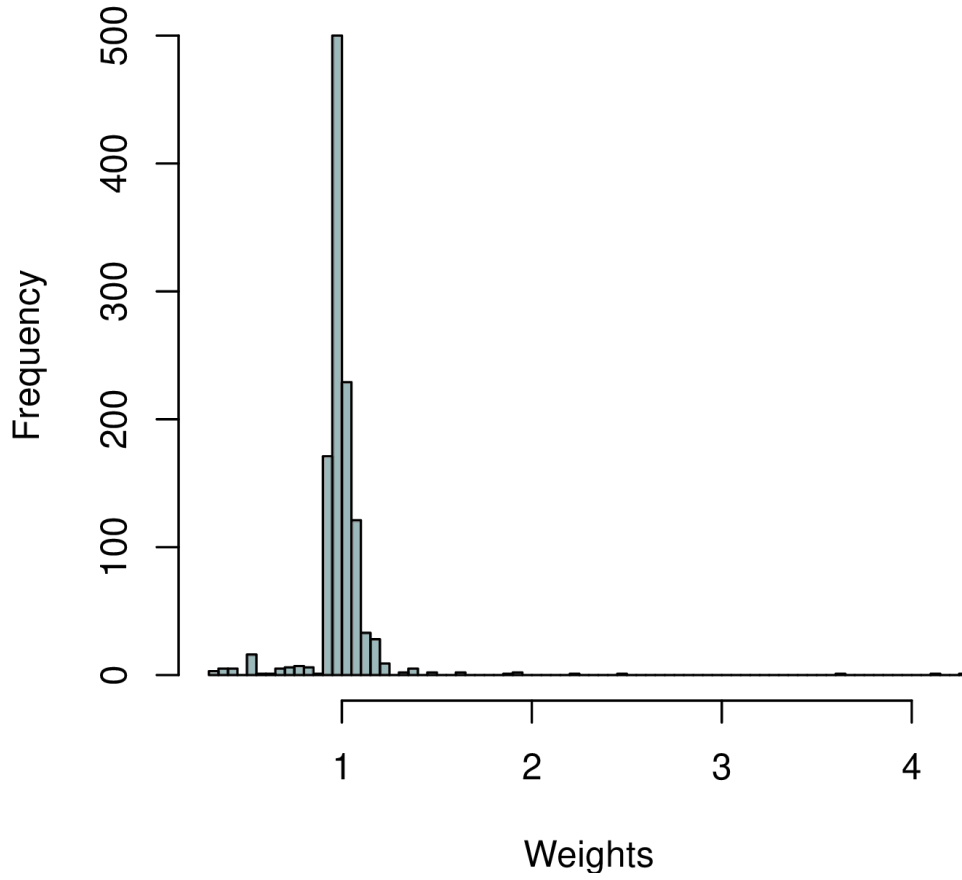
For the adjustments, fixed constant values of δ of -1, 0, 1, 2, 3, 4 or 5 were added to the ordered logistic regression imputation model for ECOG PS. Positive values for δ probabilistically shifted imputed ECOG PS to be more favorable than expected under MAR (i.e., assigning a lower ECOG PS than imputed given other measured covariates) for those missing ECOG PS. Conversely, a negative δ randomly shifted imputed ECOG PS to be poorer than expected under MAR. 20 datasets were multiply imputed for each setting of the δ parameter. For interpretability of results, instead of the log-odds defined by δ , we report the resulting mean shift in imputed ECOG PS for each setting of δ .

eFigure 1. Schoenfeld Residuals and Log-Negative-Log (LNL) Survival Curves for Tests of Proportional Hazards Before and After IPTW-Adjustment



eFigure 2. Distribution of IPTW Weights

Weights ranged from 0.3 to 6 and no extreme weights were observed.



eFigure 3. Delta-based Tipping Point Analysis for ECOG PS

Mean ECOG shift represents the observed shift under the ordinal regression model amongst the ever-smoker group. Negative shift simulates scenarios where ECOG PS was increasingly more favorable (i.e., lower) amongst ever-smokers missing baseline ECOG PS.

