Online Supplement:

Predicting Postoperative Lung Function Following Lobectomy:

A New Method to Adjust for Inherent Selection Bias

Narda Ontiveros,[4] David Eapen-John, [1] Natasha Osorio,[4] Juhee Song PhD,[2] Liang Li PhD,[2] Ajay Sheshadri MD MPH,[1] Xin Tian, BS,[1] Natasha Ghosh, MPH,[1] Ara Vaporciyan,[3] MD, Arlene Correa, PhD,[3] Garrett Walsh, MD,[3] Horiana B. Grosu MD,[2] David E. Ost MD MPH[1]

1. Department of Pulmonary Medicine, MD Anderson Cancer Center, Houston Texas.
2. Department of Biostatistics, MD Anderson Cancer Center, Houston Texas
3. Department of Thoracic Surgery, MD Anderson Cancer Center, Houston Texas
4. School of Medicine and Health Sciences, Tecnológico de Monterrey, Monterrey, Nuevo León, México; work done while at MD Anderson Cancer Center

Corresponding Author Information: David E. Ost, MD, The University of Texas MD Anderson Cancer Center, Pulmonary Department, 1515 Holcombe Blvd, Unit 1462, Houston Tx, 77030; Email: dost@mdanderson.org; phone: 713-745-8775; fax: 713-749-4922.

Conflicts of Interest: none to declare
Funding: Statistical analysis work supported in part by the Cancer Center Support Grant (NCI Grant P30 CA016672).

Methods

Prediction Models

Our prediction models use % of predicted rather than raw numbers, such as liters for FEV1 or ml/min/mm Hg for DLCO. The reasons for this are several. First, existing guidelines use % of predicted.(1) Using liters for the postoperative value is not as informative, since the same number may constitute a sufficient value of postoperative function for a smaller person while being insufficient for a larger person. Using raw numbers makes graphical interpretation of the data less informative and makes the LOA not as useful clinically. Second, when assessing selection bias, it is easy to miss the impact of selection bias when you evaluate only the raw numbers rather than the % of predicted. For example, consider the scatter plots in figure 5 and e-figures 1-3. If we had used liters for FEV1 in the scatter plot, the truncation of the normal distribution would not be as obvious. Using % of predicted, it is clear that patients with ppo values under 40% are not likely to get surgery. But if you use liters for FEV1, because there is wide variation in patient height and weight, the truncation becomes less obvious unless you plot the distribution in 3 or 4 dimensions. There is still selection bias present, but the bias is not based on a metric that uses liters, but rather on a metric based on % of predicted. The same applies for DLCO.

We can verify this by looking at the scatter plots from other studies. Scatter plots from studies that use liters for FEV1 rather than % of predicted do not demonstrate truncation as clearly,(2) since no single number in liters constitutes the prediction threshold below which surgery would be ill advised. Conversely, those studies that use

% predicted demonstrate truncation in their scatter plots (3), with the thresholds corresponding to local clinical practice guidelines (e.g. 30% of predicted).

Assessing and Correcting for Selection Bias

Because MDACC is a tertiary referral center, many patients have all their follow-up care provided locally. Because of this only a fraction of patients return for longitudinal care. This is not problematic in terms of analyzing the data if we look only at patients that had surgery (Cohort A) since the data is missing completely at random (MCAR). However, it leads to a problem when including surgical candidates that did not have surgery because of limited pulmonary reserve, since we do not know which patients would have returned and which would have had their care locally. We cannot include all of them since this would overweight the natural distribution towards the low end of the spectrum. But we do know the percentage of patients that returned for longitudinal care among patients that had surgery. We therefore randomly selected the same percentage of non-surgery patients to combine with the surgery patients to generate cohort B.

Multiple Imputation

In cohort B, patients that did not have surgery due to limited pulmonary reserve had preoperative data and predictions but no postoperative FEV1 or DLCO. It was therefore necessary to impute what the postoperative values would have been had surgery been performed. We did this by using the multiple imputation method (proc MI procedure in SAS).

The SAS multiple imputation procedure assumes that the missing data is missing at random (MAR), that is, the probability that an observation is missing may depend on the observed part of the data but not on a missing part of the data.(4) So it is acceptable if the missing postoperative FEV1 depends on ppoFEV1 using the quantitative perfusion (Q) or the segment counting 18 (SC18) or segments counting 19 (SC19) methods. There were 4 variables associated with FEV1: the actual postoperative FEV1 and three different predictions of it (Q, SC18, and SC19 models). Patients that had surgery had all 4 of these variables while patients that did not have surgery had 3 of these variables. We used a parametric regression method for monotone missing data patterns for FEV1 imputation. Using the data from the patients that had surgery, we fit a model where actual postoperative FEV1 was the dependent variable and the other three variables (ppoFEV1 using Q, SC18, and SC19 models) were the independent variables. For each variable with missing values, the previous variables were used as covariates. The process was then repeated sequentially for variables with missing values.

We generated 30 sets of imputed data. The statistical efficiency of multiple imputation methods is maximized when the number of repetitions (M) is infinite. If the fraction of missing information is modest (<20%), multiple imputation methods based on as few as M=5 or M=10 repetitions will achieve >96% of the maximum statistical efficiency. However, for inferential goals, such as confidence intervals and p-values, a small number of imputations might not be adequate.(4) The work of Graham, Olchowski, and Gilreath, suggests that multiple imputation with 20 imputations for 30% missing information is sufficient to achieve a 1% power fall off tolerance as compared to

an infinite number of imputations.(5) We therefore chose to generate 30 imputed sets of data.

Once we had the 30 sets of imputed data we compared agreement between “actual” postoperative values (which consisted of real or imputed values) and predicted values using Bland-Altman plots as described above. Note that in each set the surgical patients (corresponding to Cohort A) would remain the same but what varied between the 30 sets was the imputed values for the non-surgical patients. With 30 imputations, we compute 30 different sets of the point and variance estimates for bias. The point estimate of bias from multiple imputations is the average of the 30 bias estimates. The variance estimate associated with the mean bias is the sum of within-imputation variance, which is the average of the 30 variance estimates, and (1+1/m) multiplied by the between-imputation variance where m is the number of imputations.(6)

Determining Threshold values

Among patients that had surgery, the predicted and actual postoperative values were truncated because of selection bias as noted above. What we need to do is adjust for the truncation of the normal distribution that was created by the selection bias in order to derive the parameters of the non-truncated/non-biased distributions. We used the data from patients that had surgery (cohort A) and the tmvnorm function in R to estimate the parameters (mean=(m1, m2), sigma (v1, v2, covariance between actual and predicted) of the true bivariate normal distribution.

We then use m1, m2, s1, s2, and r (s1, s2, and r can be estimated from sigma) (s1, s2: standard deviation) to calculate the conditional distribution desired. The conditional

distribution of actual post FEV1 given predicted FEV1=x will be distributed as a normal distribution with mean=($m_1+s_1/s_2*r*(x-m_2)$) and variance=$(1-r^2)s_1^2$.

Clinically, we want to identify x which satisfies Pr(Actual FEV1 > minimum sufficient postoperative FEV1 | ppoFEV1 = x) = desired level of clinical certainty. This can be read as the probability that the actual postoperative FEV1 will be greater than our minimum sufficient posteroperative FEV1 is equal to our desired level of clinical certainty, given that the ppoFEV is x. Note that the minimum sufficient postoperative value must be determined by the physician, although based on current guidelines this would usually be either 30% or 40%.(1) The desired level of clinical certainty is chosen by the surgeon and multidisciplinary team and must consider other factors such as alternative treatment modalities, their relative efficacy, and patient preferences. For this analysis we chose to analyze three levels of clinical certainty, 95%, 97.5%, and 99%, but the same method can be used to derive thresholds for any chosen level of certainty. We used the same method for DLCO.

To visualize the adjusted bivariate normal distribution, we used scatter plots and contour plots. In each instance we began with the original data, then generated 1,000 pairs of the underlying bivariate normal distributions that were obtained using the tmvnorm function as described above.

Bland-Altman Limits of Agreement and Confidence Intervals

For table 3 using imputed data, the 95% limits of agreement are provided by the formula: mean(diff)+/-1.96*$SD_{imputed}$(diff), assuming normality. The lower limit ± $t_{(n-1, 0.025)}$*sqrt (3*$Var_{imputed}$(diff)/n); describes a possible error in the estimate due to a

sampling error. The upper limit ± $t_{(n-1, 0.025)}$*sqrt (3*$Var_{imputed}$(diff)/n); describes a possible error in the estimate due to a sampling error. The $SD_{imputed}$(diff) is estimated utilizing a method of combining m imputed data sets which is described by Yang in http://www.ats.ucla.edu/stat/sas/library/multipleimputation.pdf.(6)

Results

Cohort B Missing Data Patterns

We first examined the data to describe the pattern of missingness. The missing data followed a monotone missing pattern. Monotone patterns are those in which missingness occurs when data is available for all assessments until a time at which the patient drops out and provides no further assessments. An intermittent pattern is when there are missing observations in between assessments that are observed. A mixed patterns is one which starts off as mixed and then becomes monotone. Since all the patients with missing data were missing their last measurement (i.e. postoperative value) the pattern was a pure monotone pattern.

Next we compared patients that had missing data to those that had complete data to determine if the data was missing completely at random (MCAR). Patients that had surgery had complete information on all 4 variables (Q, SC18, SC19, and actual postoperative value) while those with missing data (i.e. non-surgical) had information on 3 of the variables (Q, SC18, and SC19). The missing data had a very different distribution of values (e-tables 1 and 2), indicating that the data may not be MCAR.

We then identified auxiliary variables correlated with actual postoperative values. All three models (Q, SC18, and SC19) were strongly correlated with each other and with actual postoperative values (e-tables 3 and 4) and were associated with missingness.

Using the proc MI procedure in SAS to perform parametric regression we generated M=30 sets of missing FEV1 and DLCO data. Variance information for FEV1 and DLCO imputation is shown in e-table 5. The relative efficiency with M=30 was 99% (as compared to an infinite number of repetitions) for both FEV1 and DLCO.

Selection Bias

The most recent ACCP guidelines use % of predicted rather than raw values (e.g. liters/second for FEV1).(1) The care pathway for preoperative assessment for lobectomy at MD Anderson includes a quantitative perfusion scan as part of the assessment if patients have borderline pulmonary function or performance status. The segment counting methods are not really used. Our institutional guidelines use a threshold ppoFEV1 and ppoDLCO of 40% of predicted to guide decision making. When both ppoFEV1 and ppoDLCO are greater than 40% of predicted patients are deemed as acceptable surgical risks from a pulmonary perspective. When either or both ppoFEV1 or ppoDLCO are 30-39% of predicted, pulmonary exercise testing is used to further inform the decision. If either ppoFEV1 or ppoDLCO are less than 30% this is considered high risk for surgery.

Looking at the scatter plots of predicted vs. actual postoperative data in figure 5a for FEV1 and e-figure 2a for DLCO, we see that there are very few points below the horizontal line where the ppo values are equal to 40% of predicted. This is as per

institutional and ACCP guidelines since those patients are less likely to get surgery. But while ppoFEV1 and ppoDLCO are normally distributed, once we eliminate patients with a ppo value of <40%, we have essentially cut off the left side of the normal distribution, truncating the distribution.

How will this impact the Bland Altman (BA) plot and interpretation? This is easier to understand by visualizing where the eliminated patients would be in the BA plot. This is shown in e-figures 4 for DLCO and e-figure 5 for FEV1. The BA plots in panel A include only patients that actually had surgery. To analyze the BA plot we need to see if the bias changes as the observed postoperative value changes, so we regress the difference between predicted and observed on the actual observed value. This is the solid blue line in panel A of e-figures 4 and 5. Note that the slope is significantly different than 0, so we know that the bias is not constant. Hence we have to report regression-based 95% limits of agreement (LOA; shown by the blue dashed lines) rather than considering the LOA as fixed (as shown by the horizontal dashed lines).(7) The regression-based 95% LOA are shown by the dashed blue lines.

Where would patients with a ppo value of 40% or less be located? The heavy red dashed line (e-figure 4) represents patients that have a ppo value of exactly 40%. Since difference on the vertical axis is ppo value minus actual postoperative value, patients with a ppo value of less than 40% will lie below and to the left of this line, while patients with ppo values greater than 40% will be above and to the right of it. There are almost no patients in this area since as we observed above the distribution is truncated based on guidelines and institutional thresholds.

How does missing data impact the slope of the regression line? At the right hand side of the BA plot, when the actual postoperative value is above approximately 70%, this has no impact. But on the left hand side, this results in an asymmetry since only patients above the red line are likely to be observed. The result is that the regression line becomes steeper and more negative than it should be because of the missing data. The underlying cause is that for any given patients whose actual postoperative values are in the low to mid-range, a prediction which overestimates the true value is more likely to be included in the analysis (i.e. received surgery and be observed) than predictions that underestimate the true value (i.e. no surgery and therefore unobservable). It is this asymmetry of inclusion based on the direction of the prediction error (i.e. over vs. underestimates) that leads to inaccurate estimates of the true prediction bias.

What happens when we include patients with ppo values that are less than 40%? Panel B in e-figures 4 and 5 shows this. The red dots are the imputed values for a proportional sample of patients that did not receive surgery because assessed pulmonary reserve was deemed inadequate. Note that many, although not all, are to the left of the red dashed line. This is because in our institution a value of less than 40% on either FEV1 or DLCO is enough to warrant further evaluation and many patients have only one value (either FEV1 or DLCO) that is less than 40%. Note the solid blue line represents the regression line which includes the original patients (blue dots) and the patients with ppo values less than 40% (red dots). Because the red dots are more often down and to the left in the BA plot, the regression line is less steep.

Clinical Implications: Threshold Values

Scatter and contour plots of the bivariate normal distribution for FEV1 are shown in figure 5 of the main manuscript and e-figures 1. DLCO plots are show in e-figures 2 and 3. Note that the original observed data (e-figure 1a) clearly shows the impact of selection bias, manifest as left sided truncation, because there are very few patients with a predicted FEV1 less than 40% by Q scan. E-figure 1b shows the impact of correction for truncation (i.e. selection bias) using the tmvnorm function. Figure c and d show the contour plots of the corrected underlying bivariate normal distribution for desired postoperative values of 30% and 40% respectively and their corresponding predicted value thresholds.

e-Table 1. Missing Data patterns for FEV1

Group		Group Means			
	Frequency (%)	ppoFEV1 Q Model	ppoFEV1 SC18 Model	ppoFEV1 SC19 Model	Actual Postoperative FEV1
Complete Data (Surgery)	79 (69)	65.9	65.9	65.8	68.4
Missing Data (No surgery)	35 (31)	41.1	43.2	42.5	

Prediction of postoperative FEV1 (ppoFEV1) models are based on quantitative perfusion scans (Q), segment counting with 18 segments (SC18) or segment counting with 19 segments (SC19). Group means are for FEV1 % of predicted.

e-Table 2. Missing Data Patterns for DLCO

Group		Group Means			
	Frequency (%)	ppoDLCO Q Model	ppoDLCO SC18 Model	ppoDLCO SC19 Model	Actual Postoperative DLCO
Complete Data (Surgery)	78 (69)	60.9	61.0	60.9	61.3
Missing Data (No surgery)*	36 (31)*	35.1	37.0	36.5	

Prediction of postoperative DLCO (ppoDLCO) models are based on quantitative perfusion scans (Q), segment counting with 18 segments (SC18) or segment counting with 19 segments (SC19). Group means are for DLCO % of predicted.

* One patient did have surgery and had spirometry without a DLCO. So one patient with missing DLCO data did have surgery.

e-Table 3. Correlation Between Prediction Models and Actual Postoperative FEV1 % of Predicted

	Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
Prediction Models	**SC18 Model**	**SC19 Model**	**Actual postoperative FEV1**
Q Model	0.95830 <.0001 114	0.95058 <.0001 114	0.74584 <.0001 79
SC18 Model		0.99420 <.0001 114	0.78218 <.0001 79
SC19 Model			0.79499 <.0001 79

Prediction of postoperative lung function models are based on quantitative perfusion scans (Q), segment counting with 18 segments (SC18) or segment counting with 19 segments (SC19).

e-table 4. Correlation Between Prediction Models and Actual Postoperative DLCO % of Predicted

	Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
	SC18 Model	**SC19 Model**	**Actual postoperative DLCO**
Prediction Models	0.96111 <.0001 114	0.95422 <.0001 114	0.66339 <.0001 78
Q Model		0.99566 <.0001 114	0.70806 <.0001 78
SC18 Model			0.71062 <.0001 78

Prediction of postoperative lung function models are based on quantitative perfusion scans (Q), segment counting with 18 segments (SC18) or segment counting with 19 segments (SC19).

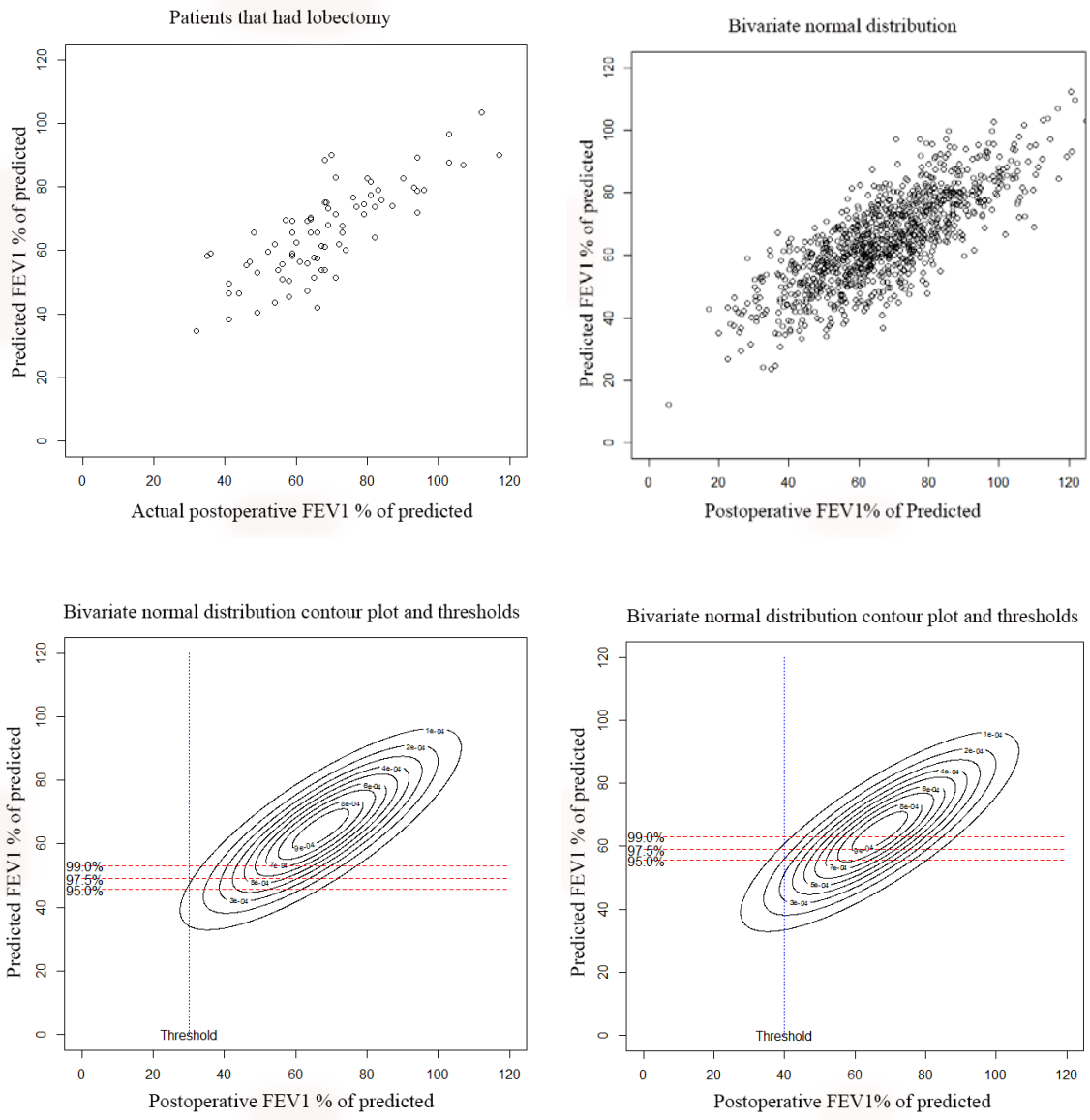
e-table 5. Variance Information on Imputation (M=30)

Variance Information (30 Imputations)							
Variable	**Variance**			**DF**	**Relative Increase in Variance**	**Fraction Missing Information**	**Relative Efficiency**
	Between	**Within**	**Total**				
Postoperative FEV1	1.101	3.638	4.776	72.572	0.3128	0.2412	0.9920
Postoperative DLCO	1.331	4.083	5.458	70.293	0.3368	0.2559	0.9916

References

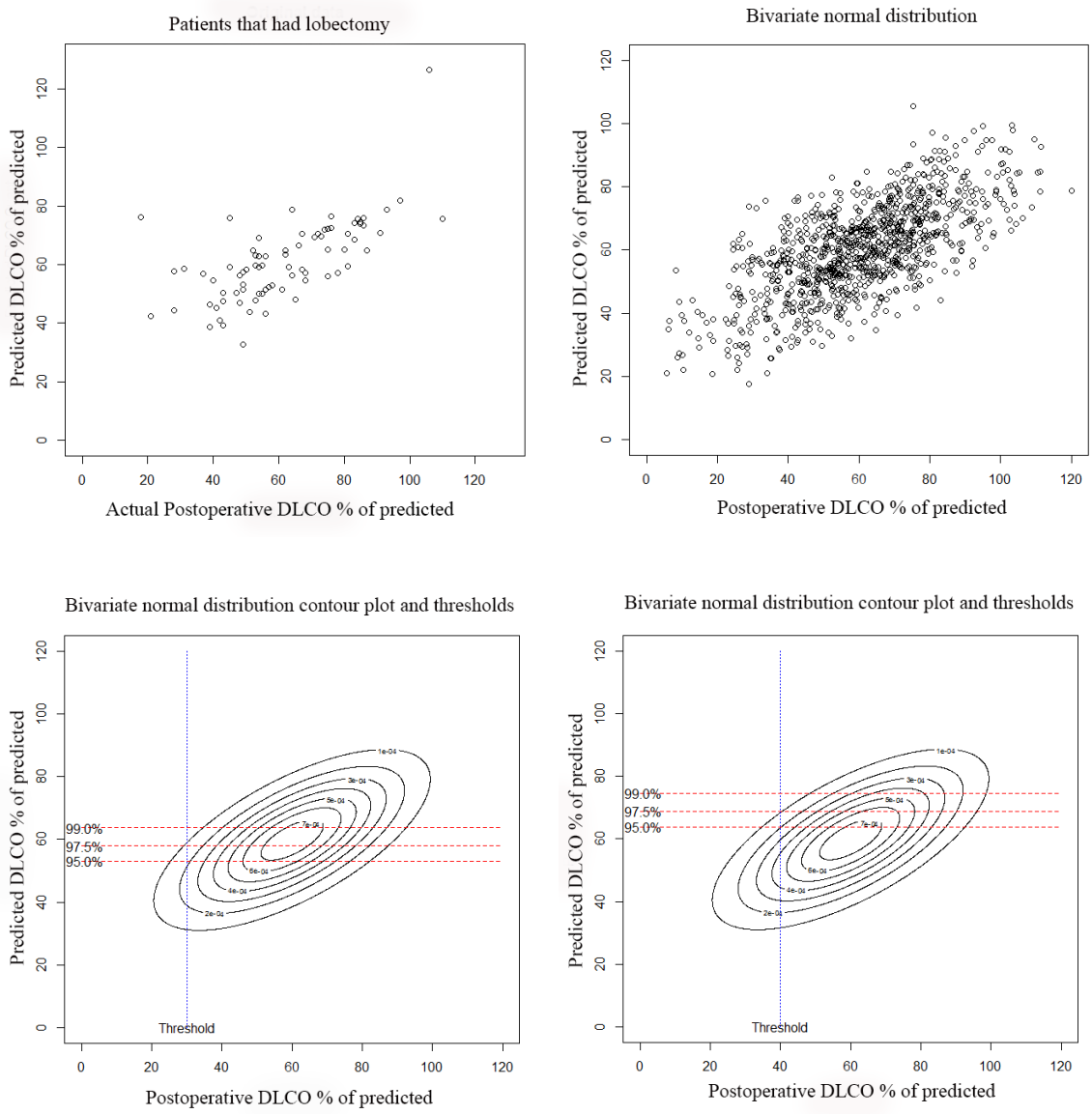
1. Brunelli A, Kim AW, Berger KI, Addrizzo-Harris DJ. Physiologic evaluation of the patient with lung cancer being considered for resectional surgery: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* 2013; 143: e166S-190S.
2. Bolliger CT, Guckel C, Engel H, Stohr S, Wyser CP, Schoetzau A, Habicht J, Soler M, Tamm M, Perruchoud AP. Prediction of functional reserves after lung resection: comparison between quantitative computed tomography, scintigraphy, and anatomy. *Respiration* 2002; 69: 482-489.
3. Ohno Y, Seki S, Koyama H, Yoshikawa T, Matsumoto S, Takenaka D, Kassai Y, Yui M, Sugimura K. 3D ECG- and respiratory-gated non-contrast-enhanced (CE) perfusion MRI for postoperative lung function prediction in non-small-cell lung cancer patients: A comparison with thin-section quantitative computed tomography, dynamic CE-perfusion MRI, and perfusion scan. *J Magn Reson Imaging* 2015; 42: 340-353.
4. SAS Institute Inc. The MI Procedure. SAS/STAT 141 User's Guide. Cary, NC: SAS Institute, Inc.; 2015. p. 5918-5919.
5. Graham JW, Olchowski AE, Gilreath TD. How Many Imputations Are Really Needed? *Prevention Science* 2007; 8: 206-213.
6. Yuan YC. Multiple Imputation for Missing Data: Concept and New Development. [cited 2017 3/8/2017]. Available from: http://www.ats.ucla.edu/stat/sas/library/multipleimputation.pdf.
7. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999; 8: 135-160.

e-Figure 1. Scatter plot of actual FEV1 vs. predicted FEV1 using SC19 method



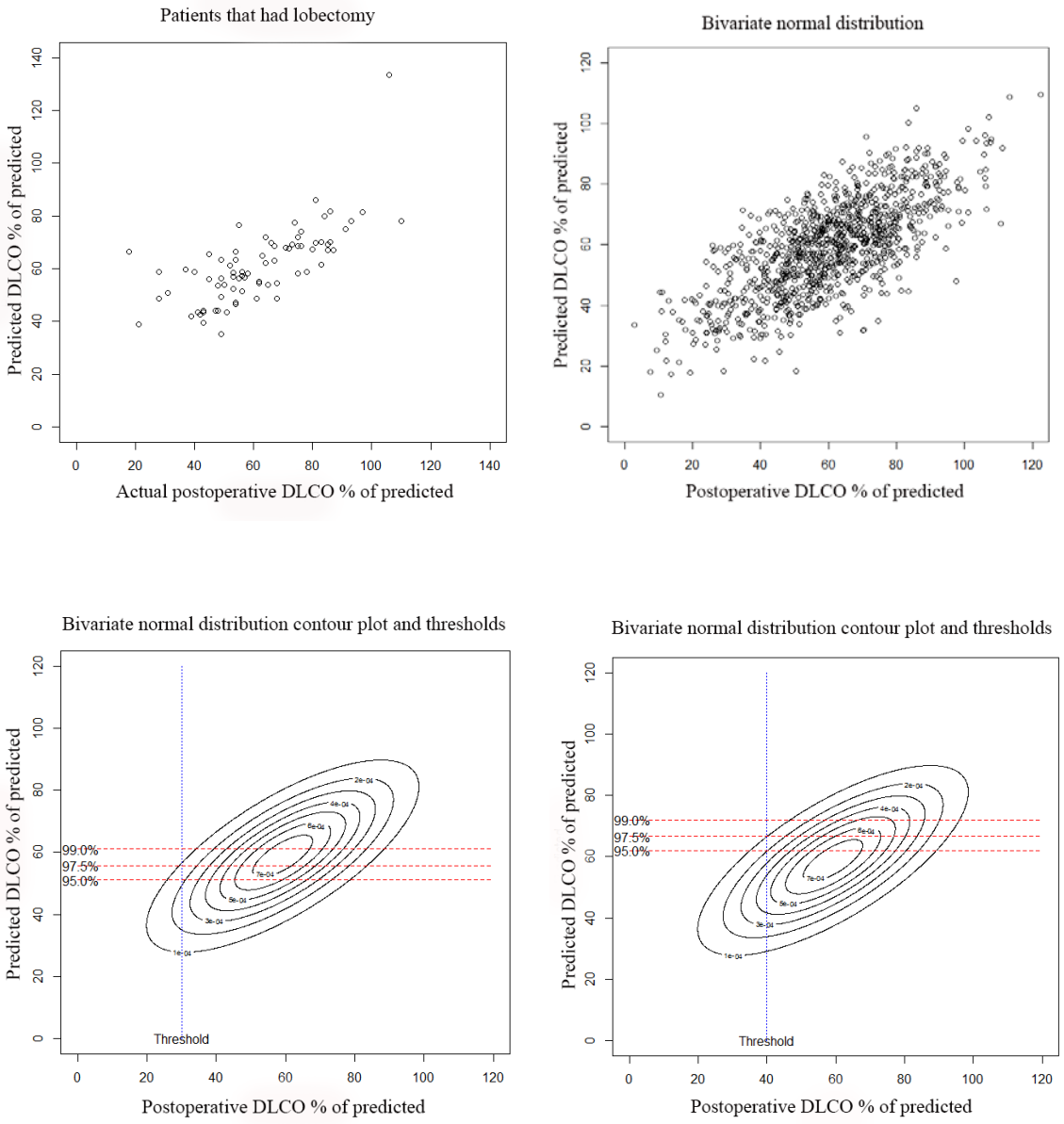
a) Top left: Scatter plot of actual FEV1 vs. predicted FEV1 using SC19 method.
b) Top right: Random samples of 1000 pairs were generated from underlying bivariate normal distribution
c) Bottom left: Contour plot of underlying bivariate normal distribution (actual FEV1 vs. predicted FEV1 SC19 method) for actual threshold of 30%
d) Bottom right: Contour plot of underlying bivariate normal distribution (actual FEV1 vs. predicted FEV1 SC19 method) for actual threshold of 40%

e-Figure 2. Scatter plot of actual DLCO vs. predicted DLCO using quantitative perfusion scans



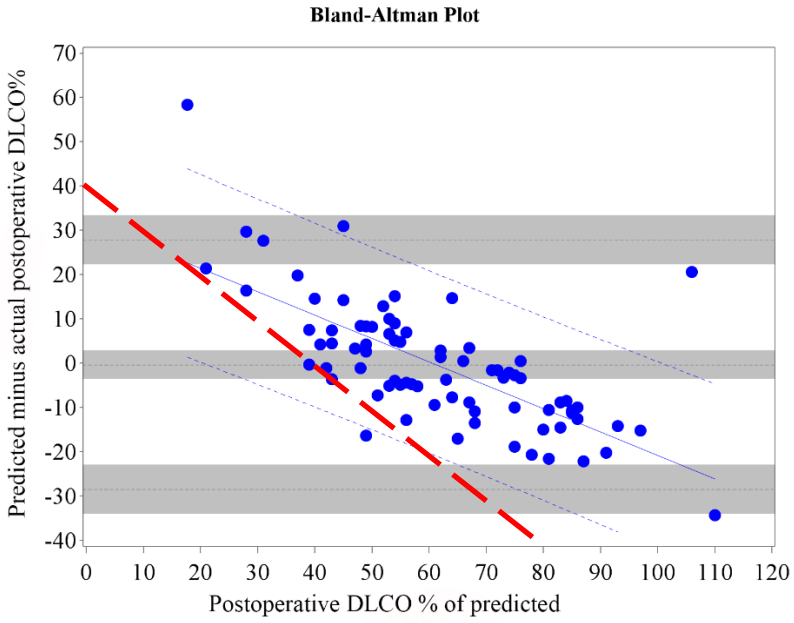
Top left: Scatter plot of actual DLCO vs. predicted DLCO using Q method.
Top right: Random samples of 1000 pairs were generated from underlying bivariate normal distribution
Bottom left: Contour plot of underlying bivariate normal distribution (actual DLCO vs. predicted DLCO Q method) for actual threshold of 30%
Bottom right: Contour plot of underlying bivariate normal distribution (actual DLCO vs. predicted DLCO Q method) for actual threshold of 40%

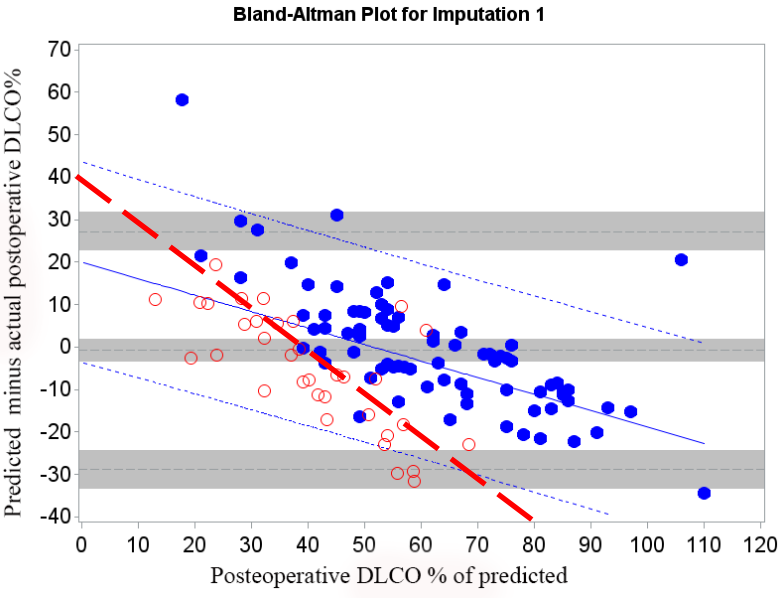
e-Figure 3. Scatter plot of actual DLCO vs. predicted DLCO using SC19 method



Top left: Scatter plot of actual DLCO vs. predicted DLCO using SC19 method.
Top right: Random samples of 1000 pairs were generated from underlying bivariate normal distribution
Bottom left: Contour plot of underlying bivariate normal distribution (actual DLCO vs. predicted DLCO SC19 method) for actual threshold of 30%
Bottom right: Contour plot of underlying bivariate normal distribution (actual DLCO vs. predicted DLCO SC19 method) for actual threshold of 40%

e-Figure 4. Bland Altman Plot of ppoDLCO minus actual postoperative DLCO vs. actual postoperative DLCO





e-Figure 5. Bland Altman Plot of ppoFEV1 minus actual postoperative FEV1 vs. actual postoperative FEV1