Peer Review File
Article information: https://dx.doi.org/10.21037/jtd-21-1484.

Reviewer A:

Comment 1: Regarding the risk predicting model for PAL, IPAL score reported by the French Society of Thoracic and Cardiovascular Surgery in 2011 is well-known (Reference 6). Compared to the analysis of more than 30,000 patients in the French study, the 6,000 cases in this study is a smaller number. However, unlike the French study, patients who received wedge resection and bullectomy were excluded, which seems to be a more clinically relevant setting because they have low risk for PAL. It is a new finding that middle lobe resection and GGO lesion were identified as protective factors for PAL. Unfortunately, the C statistic of 0.63 is negative data for the study; however, the results were based on large data from multiple facilities and are considered valuable.

Reply 1: thank you for your comment. We completely agree. There are numerous predictive models for the postoperative risk of PAL in the literature and many have numbers of patients and events that are certainly higher than those we currently have. However, models are often subject to overfitting and this makes them not very robust and poorly performing when tested in a different cohort of patients.

In our previous article (13), we analyzed relevant predictive models for PAL (5-8) developed through classic internal validation procedures. These models showed low discrimination values both during the internal validation, performed by the authors, and during our external validation. These performances do not allow a clinical application of these predictive models. The reason for this lies both in the high level of randomness to the occurrence of postoperative air leak, and in the necessary further research of pre- or intraoperative variables that can predict this event.

In our article we have shown how computationally burdensome internal validation techniques, such as machine learning, result in outcomes that are not different from those coming from standard validations. The novelty in this study is the automatic selection of the variables by means of the RFE algorithm in place of the manual choice of the biostatistician performed by means of statistical correlation and biological plausibility.

Minor comments

Comment 2: In Table 1, not only N and % but also median and IQR are listed.

Reply 2: thank you for your suggestion. To distinguish more clearly the continuous variables expressed as median and IRQ, compared to the discrete variables expressed as percentages, we decided to change the type of parenthesis used.
Results are expressed as counts and percentages (%) of patients for categorical variables, and as medians and interquartile ranges [IQR] for numeric variables. We have corrected Table 1 (see Table 1).

Comment 3: Table 2 has a lot of unnecessary information and needs improvement.
Reply 3: thank you for your suggestion.
Table 2 shows all the variables provided to the RFE algorithm to select the most influential variables in the PAL predictive model. The top ranked predictors of the final model were selected among the variables contained in this table. The variables missing more than 5% have already been removed.
We agree that this table contains data that is not necessary for the reader, however it contains important data from a methodological point of view, therefore, we have decided to move Table 2 in the supplementary material.

Comment 4: Line 213-215 and Figure 2 are difficult to understand. What do the numbers indicate?
Reply 4: thank you for your suggestion. When using machine learning models, it is important to understand which predictors are more influential on the outcome variable. ROC curve analysis is conducted on each predictor. The trapezoidal rule is used to compute the area under the ROC curve. This area is used as the measure of variable importance. All measures of importance are scaled to have a maximum value of 1. For more information: Kuhn M. Building predictive models in R using the caret package. Journal of statistical software. 2008 Nov 10;28(1):1-26.
We simplified as requested Figure 2 (see Figure 2)

Comment 5: P10 line 195 & 200. Please correct IQR 63 75 and IQR (35) to 63-75 and 3-5.
Reply 5: thank you for your suggestion. We have corrected as requested (see line 221& 227).

Reviewer B:

Major comment:
Comment 1: Various statistical methods were used, and 6 significant relating factors were found in the study. However, as a result, the generated score model was not satisfactory and cannot be applied to clinical practice. We have little to gain from their study.
Reply 1: thank you for your comment. We understand the reviewer's perplexity. We recently

published a study (13) where we evaluated both the clinical and statistical performances of the best 4 current PAL risk models (5-8). We demonstrated that even these larger studies have a C statistic <0.65 and that score models can be applied to clinical practice but with a high rate of false positives and a low positive predictive value. With our current study, we wanted to assess risk factors for PAL by analyzing data from the Italian VATS group registry. Also in these patients we have shown that current risk factors are not reliable. We state in our conclusions that a combination of preoperative risk identification and intraoperative objective assessment of alveolar air leakage is needed to improve the performance of the current PAL risk models.

Comment 2: The authors should seek the opinion of a statistical expert to see if there are any problems with the methodology, for example, how to split the patients into a derivation and validation cohort.
Reply 2: thank you for your comment. In our team of authors, Mr. Luca Bertolaccini is a certified statistical expert. As is known, when developing a prediction model, several factors may lead to models yielding optimistic apparent performance. It is therefore important that a more honest estimate of the model's performance from the development data set is obtained. This optimism diminishes when the sample size become large or applying the so-called "internal validation. Internal validation is achieved preferably using resampling techniques, such as bootstrapping, or cross-validation methods.
The RFE algorithm performs during the development of a predictive model a 5-fold cross-validation encapsulated inside an outer layer of resampling (10-fold cross-validation). The predictive model selected by the algorithm considers both the variation due to feature selection and the internal validation.
Given its ubiquity we decided to additionally adopt the classical split-sample internal validation approach. The limited empirical evidence to support investigators in guiding their sample size choice for validation studies suggests a minimum of 100 events and 100 nonevents. Our validation cohort includes 1106 nonevents and 140 events.
For more information:

Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Annals of internal medicine. 2015 Jan 6;162(1):W1-73.

Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis. New York: Springer; 2001.

Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external valida- tion studies of predictive logistic regression models. J Clin Epide- miol. 2005;58:475-83. [PMID: 15845334]

Kuhn M. Building predictive models in R using the caret package. Journal of statistical software. 2008 Nov 10;28(1):1-26.

Minor comment:

Comment 3: As a rule, the authors should describe their content in the past tense in RESULTS.

Reply 3: thank you for your suggestion. We agree and we have properly corrected the section of RESULTS (see line 245-259).

Comment 4: In Table 1, the category and number of surgical approaches may be wrong.

Reply 4: thank you for your suggestion. The row "surgical approaches" is wrong and was completely deleted (see Table1).

Comment 5: The authors should provide the data of the number of surgeon and their experience. I want to know whether the incidence of PAL is different by the surgeon's experience.

Reply 5: thank you for your comment. From our database it is not possible to directly trace all the surgeons involved in the different procedures. However, the experience of the first surgeon is recorded for each procedure and expressed as the number of VATS lobectomies performed in his/her career. There are three possible categories: 0 - 20 procedures performed, 21 - 50 and> 50.

As shown by the table below, there is no significant correlation between the experience of the first surgeon and the occurrence of postoperative PAL.

name	*All_patients*	*PAL_yes*	*PAL_no*	*p.value*
Experience of first surgeon	0 – 20 (1042, 16.7%) 21 – 50 (1188, 19.1%) > 50 (4006, 64.2%)	0 – 20 (117, 16.7%) 21 – 50 (123, 17.5%) > 50 (462, 65.8%)	0 – 20 (925, 16.7%) 21 - 50(1065, 19.2%) > 50(3544, 64%)	0.528

Comment 6: The authors should provide the patient algorism as Figure.

Reply 6: thank you for your suggestion. We have added the patient algorithm as Figure as required. See Figure 1.

Comment 7: I think that adhesiolysis is associated with DLCO/VA ratio, male-sex, and COPD. Therefore, the authors should provide the data in each patient with or without adhesiolysis.

Reply 7: thank you for your suggestion. Below you will find the table with the correlations between adhesiolysis and other variables. There are numerous correlations.

However, we do not consider these correlations relevant to the purpose of our analysis. We have therefore decided not to include this data in the text of our article.

name	*adhesions_yes*	*adhesions_no*	*p.value*
5-day PAL	FALSE(1414,83.8%) TRUE(274,16.2%)	FALSE(4120,90.6%) TRUE(428,9.4%)	<0.001
Age	70 (64-75)	69 (62-75)	<0.001
Gender	F(593,35.1%) M(1095,64.9%)	F(1918,42.2%) M(2630,57.8%)	<0.001
BMI (Kg/m2)	25.505 (23.15-28.6)	25.86 (23.32-28.45)	0.081
Smoking history	Ex(706,41.8%) No(405,24%) Yes(577,34.2%)	Ex(1738,38.2%) No(1468,32.3%) Yes(1342,29.5%)	<0.001
Zubrod Score >=2	FALSE(1125,66.6%) TRUE(563,33.4%)	FALSE(3241,71.3%) TRUE(1307,28.7%)	<0.001
Congestive heart failure	No(1619,95.9%) Yes(69,4.1%)	No(4390,96.5%) Yes(158,3.5%)	0.254
COPD	No(1246,73.8%) Yes(442,26.2%)	No(3579,78.7%) Yes(969,21.3%)	<0.001
Diabetes	End-organ damage(6,0.4%) No(1428,84.6%) Uncomplicated(254,15%)	End-organ damage(20,0.4%) No(3968,87.2%) Uncomplicated(560,12.3%)	0.013
FEV1, % predicted	92 (79-104)	94 (80-107)	<0.001
Pulmonary resection	Lower_lobe(595,35.2%) Upper_lobe(979,58%) Middle_lobe(114,6.8%)	Lower_lobe(1687,37.1%) Upper_lobe(2505,55.1%) Middle_lobe(356,7.8%)	0.093
Side	Left(667,39.5%) Right(1021,60.5%)	Left(1867,41.1%) Right(2681,58.9%)	0.283
Pulmonary pathology	Benign(87,5.2%) Malignant(1601,94.8%)	Benign(148,3.3%) Malignant(4400,96.7%)	<0.001
Open Conversion	No(1450,85.9%) Yes(238,14.1%)	No(4249,93.4%) Yes(299,6.6%)	<0.001

Reviewer C:

Comment 1: The authors are to be congratulated on a well performed study, using the large multicenter Italian thoracic surgery registry, with rigorous methodology, and valuable results. The fact that the predictive model is not particularly discriminatory, is certainly useful information. Predicting air leak is particularly challenging.
Reply 1: thank you for your comment.

Comment 2: The following comments are intended to improve the overall quality of the paper. The title of the paper refers to patients following lobectomy, however included are all patients undergoing anatomic pulmonary resection. Thus, I would consider a simpler and more accurate

title, such as “predicting prolonged air leak after video-assisted thoracoscopic pulmonary resection”.

Reply 2: thank you for your suggestion. We have properly corrected the Title (See the Title, line 2).

Comment 3: The fact that a multicenter study of over 6000 patients fail to be able to produce a reliable predictive model with a receiver operating curve area under the curve of not even 0.65 is indeed telling, and merits greater comment regarding why. One possibility is that there is simply a high level of randomness to the occurrence of postoperative air leak that is irreducible. It might make sense to discuss that briefly with a couple of sentences in the conclusion. For example, tears in pulmonary staple lines or parenchyma may be largely random. As clinicians and scientists, we don’t like to admit when we cannot predict, however it is useful to know when that is indeed the case.

Reply 3:thank you for your brilliant suggestion. We added the suggested comment in the discussion section (see line 372-376).

Comment 4: The other possibility is that the authors did not include all of the variables that could predict postop of our leak. Relevant to this question, the authors elected to take a firm approach to missing data, and eliminate any variables that had more than 5% missing data. This led to not considering 200 of the 320 variables. Could there be predictive variables within that 200 that could have significantly improved the predictive model? I can’t think of any variables that might have been eliminated that could have predicted postop air leak, but nonetheless there is the potential that critical variables were eliminated because of missingness. This simple should be mentioned in the limitations paragraph.

Reply 4: thank you for your brilliant suggestion. We added the suggested comment in the limitation section (see line 290-293).

Comment 5: When the authors comment in the second sentence of the abstract that “A useful risk predictor model can help recognize those patients who might benefit from additional preventive procedures”, they are effectively and succinctly communicating the rationale to create predictive models in general. The value of prediction is dependent on the ability to prevent. With postoperative air leak, there are no proven preventative measures that are beneficial for all; thus, a targeted, strategic approach makes eminent sense. While the focus of this investigation is on prediction, the actionability of prediction is another matter. This might be addressed by the authors. For example, I fully agree with the authors when they state in the conclusions that the future

includes the combination of preoperative risk identification and intraoperative assessment alveolar air leakage. The actionability of that assessment, occurring at the end of the pulmonary resection is ideal with regards to the immediate use of sealants.
Reply 5: thank you for your brilliant suggestion. We added the suggested comment in the discussion section (see line 379-380).

Comment 6: The author states in the second paragraph of the Background that one way to prevent air leaks is through "adopting meticulous surgical technique as appropriate tissue manipulation and retraction, mobilization of all intrapleural adhesions, division of the inferior pulmonary ligament, routine pre compression of staple lines, fissureless/fissurelast technique, and select use of surgical sealant", referencing one review paper. However, the evidence to support these interventions is not at all established, and I suggest less firm language. The sentence is also not grammatically correct. Sealants perhaps have the best evidence and to support their cause, the authors should mention a recent meta-analysis of trials relating to sealants, published in this journal. McGuire AL, Yee J. Clinical outcomes of polymeric sealant use in pulmonary resection: a systematic review and meta-analysis of randomized controlled trials. J Thorac Dis. 2018;10(Suppl 32):S3728-S3739. doi:10.21037/jtd.2018.10.48
Reply 6: thank you for your brilliant suggestion. We changed the statement in the Background section and we added the suggested recent meta-analysis in the References section (see line 109-113, 417-419).

Comment 7: I like the occasional artful and humorous additions to the otherwise scientific writing, for example, "to "air" is to leak - to prevent is Devine"; we need more such witticisms in our publications.
Reply 7: thank you for your comment. We completely agree.

Comment 8: On page 13, line 273, the authors comment on gender and sex influences. Perhaps the authors are misusing the words? Gender is generally considered to be self-reported, where is sex is biologic. Presumably prior studies have looked at sex, not gender.

Reply 8: thank you for your suggestion. We corrected as suggested (see line 337).

Comment 9: The authors have evaluated DLCO/VA rather than percent predicted DLCO, which is

more commonly used for risk stratification by surgeons. Does DLCO/VA linearly correlate with percent predicted DLCO, or not? Perhaps it would be useful to comment on that briefly on page 13 when discussing DLCO/VA.

Reply 9: thank you for your suggestion. DLCO/VA doesn't correlate always linearly with percent predicted DLCO. They provide complementary information but may differ based on comorbidities (20). We added the statement as suggested (see line 344-346).

Comment 10: On page 14, line 289, the author stated that "smoking history" was not found to be predictive. Did this include current smoking (i.e. right up to the date of the operation), or any history of past smoking, or both lumped together? I suggest you overtly state whether it was the presence of existing smoking, or past smoking that was not predictive.

Reply 10: thank you for your suggestion. Smoking history includes both current smoking and any past smoking history. We overtly stated it (see line 356).

Comment 11: It is important to point out that some variables do not appear among selected features not because they are not predictive, but for two reasons. On the one hand, their contribution is negligible compared to the variables selected. On the other hand, they may have been removed as they significantly correlated with another variable. When a correlation between predictors exceeds 0.5 (Pearson), the most statistically relevant variable is retained, the other is discarded.

Reviewer D:

In this manuscript, Divisi et al. report their findings on developing a model that predicts prolonged air leak after VATS lobectomy using the Italian VATS group registry. They utilize machine learning on a very large database (n=6236) assessing many variables. I agree with the authors that prolonged air leak is a source of nuisance to thoracic surgeons, and having the ability to predict this would have value. The authors demonstrate more or less what we expected/knew, but in a scientifically sound fashion using a large database. I congratulate the authors on their efforts. Please see below my comments.

Major

Comment 1: Main critique is what the authors state themselves – the model was not strong enough in the validation set to be used in clinical practice. So is there an important take-home message here?

Reply 1: thank you for your comment. We understand the reviewer's perplexity. We recently published a study (13) where we evaluated both the clinical and statistical performances of the best 4 current PAL risk models (5-8). We demonstrated that even these larger studies have a C statistic <0.65 and that score models can be applied to clinical practice but with a high rate of false positives and a low positive predictive value. The reason for this lies both in the high level of randomness to the occurrence of postoperative air leak, and in the necessary further research of pre- or intraoperative variables that can predict this event.

With our current study, we wanted to assess risk factors for PAL by analyzing data from the Italian VATS group registry using the automatic selection of the variables by means of the RFE algorithm in place of the manual choice of a biostatistician based on statistical correlation and biological plausibility.

Also in these patients we have shown that current risk factors are not reliable. We state in our conclusions that a combination of preoperative risk identification and intraoperative objective assessment of alveolar air leakage is needed to improve the performance of the current PAL risk models.

Comment 2: Why was the database split into 80% derivation and 20% validation? Was there a statistical method in splitting it as such? Do you think in the end the model did not perform in the validation cohort given the much smaller size?

Reply 2: thank you for your comment. Here is a brief excerpt from the guidelines TRIPOD. "The limited empirical evidence to support investigators in guiding their sample size choice for validation studies suggests a minimum of 100 events and 100 nonevents."

Our validation cohort includes 1106 nonevents and 140 events.

For more information:

Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Annals of internal medicine. 2015 Jan 6;162(1):W1-73.

Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis. New York: Springer; 2001.

Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external valida- tion studies of predictive logistic regression models. J Clin Epide- miol. 2005;58:475-83. [PMID: 15845334]

Comment 3: I am not sure why the authors excluded open approaches but included VATS cases converted to open. I feel they should be both included or excluded.
Reply 3: thank you for your comment. We excluded open approaches because the Italian VATS group database includes only VATS anatomical resections. We decided to include VATS cases converted to open, because analyzing the database of the best 4 current PAL risk models (5-8), all included these cases.

Comment 4: For the factors that are continuous variables, it is unclear if this was analyzed as continuous vs binary based on some cutoff. Can the authors clarify?
Reply 4: thank you for your question. Continuous variables were not categorized. They have been scaled and centered to have a standard deviation of one and a mean of zero. We added the statement as suggested (see line 176-177).

Comment 5: It makes sense that adhesiolysis is a significant factor, but this is also very subjective. We know there are adhesions that are very soft and easy to take down, and we know there are adhesions that are dense and results in possible lung parenchymal damage during adhesiolysis. Does the Italian VATS registry have a specific definition for adhesiolysis (for example, time spent to take down the adhesions)?
Reply 5: thank you for your comment. We completely agree with your statement. But, as reported in the limitation section, our Italian VATS group database is not designed for a specific research purpose, and we cannot retrieve these data from our database. Unfortunately, no changes have been made.

Comment 6: COPD is another factor that makes sense, but we also know some physicians put this diagnosis based just on clinical picture rather than objective evidence based on PFT's and such. Do the authors have a sense of what is meant by COPD in their database?
Reply 6: thanks for your comment, it will be useful for the next databases. Unfortunately, in our current registry we can only report the presence or absence of COPD. No definition is provided.

Comment 7: Authors mention IAL. Was this something that is recorded consistently in the Italian practice? If so, do the surgeons have a consistent way of measuring this?

Reply 7: thanks for your comment. This is one of the most important crucial points in the prevention and treatment of PAL. In Italy it is currently not practice to objectively register IAL. The aim of our current studies is to promote the objective intraoperative measurement of IAL through mechanical ventilation testing (12).

Comment 8: Does "ground glass opacity" indicate pure ground glass lesions, or any lesions that has ground glass components?
Reply 8: thanks for your comment. Unfortunately, in our current registry we can only report the presence or absence of GGO. No definition is provided. We added this definition in the discussion session (see line 353-354).

Minor
Comment 9: Does this cohort include robotic approach or the standard VATS? This should be mentioned as some believe prolonged air leak is less with the robotic approach.
Reply 9: thank you for your suggestion. The Italian VATS group database includes only standard VATS procedures. In the Methods section we added that no robotic approaches were included (see line 149).

Comment 10: There are a couple spelling and basic errors. Examples include:
a. Last sentence of Methods in Abstract – "loess".
Reply: we are sorry, but we disagree with the reviewer. Loess calibration is a regression method.

b. Line 116-117 sounds like one of the authors' own comments. Was this meant to be included?
Reply: thank you for your suggestion. We intended to include it, but we have decided to delete it (see line 130).

c. Lines 245-246 "Our goal ~ ~~resections~~" doesn't make sense.
Reply: thank you for your comment. I agree with the reviewer that it is not grammatically correct, but it is a joke and is a quote from Mr. Cerfolio RJ (15). It is the favorite statement of Reviewer C. We have decided not to correct it.

Comment 11: Authors state "exclusion criteria" in Abstract when this is not cleared defined.
Reply 11: we are sorry, but we disagree with the reviewer. We have reported the exclusion criteria in the Methods section under Participants (see line 148-160).