Peer Review File

Article information: https://dx.doi.org/10.21037/tp-22-246

Reviewer Comments

Major concerns:

Comment 1: The study population is not fully described. From the discussion I understand that this is a single-center study using data from a hospital in Chongqing. The authors discuss some particularities of the area that they think that could influence the data. This information should be presented much earlier.

Reply 1: Thanks a lot for your suggestion. As you mentioned, this is a retrospective study in a single center. This information should be presented much earlier. In this way, readers can better understand the design and limitation of the study. We have revised the manuscript tittle and added relative description in the introduction section. With the revision, the readers can better understand the particularities of the study. Such as " Why do children in this area receive surgery late? " and " Why do family members in this area want to know the approximate incidence of adverse events before operation?"

Changes in the text: The tittle has been revised as "**Clinical adverse events after repair of tetralogy of Fallot: Prediction Models by Machine Learning of a Retrospective Cohort Study in Western China**". Relative description has been added in the Introduction section such as "**In a developing area, such as western China, many patients receive the surgery beyond the optimal time. The mortality and incidence of postoperative adverse events higher than that in developed areas, just as operative mortality is higher in developing countries than that in developed countries.** " The details of the revision can be find in the first paragraph of the Introduction. We have modified our text as advised (see Page 5, line 103-107)

Comment 2: The study population is not fully described. It is not clear which type of surgical procedure was included/excluded. I suspect that the authors used data from patients that underwent complete repair, but lines 89-91 seem to suggest otherwise. I strongly encourage the authors to provide a table of included/excluded diagnosis and procedures including standard classification codes (e.g. ICD-10 for diagnosis, and SNOMED CT, ICD-10-PCS or OPCS-4 for procedures).

Reply 2: Thanks for your suggestion. We do apologize to let you misunderstanding, all the patients included received a standard primary complete repair. We have added standard classification codes. All patients included in the study were diagnosed as TOF (ICD-10: Q21.3) and underwent standard primary complete repair (ICD-9-CM-3:35.81). Those with TOF variants including pulmonary atresia, major aortopulmonary collateral arteries, atrioventricular septal defect and the patients who received palliative procedure including RV outflow ballooning/stenting or arterial duct stenting were also

excluded. We have added the including standard classification codes and describe the inclusion/exclusion criteria in detail in the methodology.
Changes in the text: We have added the relative description in data and patients section as "**All the patients included were diagnosed as TOF (ICD-10: Q21.3) and underwent standard primary complete repair (ICD-9-CM-3:35.81). Patients were excluded: (1) if they were combined with variants pulmonary atresia, major aortopulmonary collateral arteries, atrioventricular septal defect; (2) the patients who received palliative procedure such as RV outflow ballooning/stenting or arterial duct stenting; (3) Patients who received staged repair with unifocalization of the pulmonary arteries and conduit placement followed by later closure of the ventricular septal defect.** "we have modified our text as advised (see Page 6-7, line 127-134)

Comment 3: The study uses 20 years of data. This causes some concerns that must be addressed. the way that some procedures are carried out may have changed over that time, and hence patients undergoing that same procedure might have been exposed to different risk factors. This is not discussed anywhere. The authors must analyse if the year in which the procedure was carried out influences the outcome (i.e. post-surgery complications). If the annual distribution of complications is not homogenous, the authors may need to control for the year in their analyses; this would make the predictive model less useful.
Reply 3: Thanks for your suggestion. your suggestion is very important, and is very useful to make our study more preciseness. As you suggested, we had added to analyze the influences of the surgery conducted in different year to the outcome, the result showed that the surgery conducted in different era would not influence the outcome (Univariate logistic regression, OR=0.938-1.021, P=0.325). We thought the reasons might be following as: First, all the patients received a standard primary complete repair, and the major procedure did not change over that time radically; Second, the distribution of patient severity is not well balanced. Many patients with severe TOF gave up and did not undergo surgery in early year. Therefore, the basic conditions of early patients are actually better than those of recent years. The improvement of surgical skills and perioperative management may be offset by the severity of the disease in recent year. Third, our hospital was the biggest Children's Hospital in southwest of China, primary complete repair is a mature procedure, all the surgery and nurse were experienced.
We did not analyze this factor in the initial manuscript because this study is mainly to establish a predictable model, and the past years cannot be used as an available predictor.
According to your excellent suggestion, we added relative description to notice readers. We particularly emphasize that the use of the model requires more external validation, especially for those institutions with less surgical experience. The year of surgery is possibly to affect the outcome and should not be completely ignored in these institutions.

Changes in the text: The relative description has been added in the method section as" **The year of the surgery and surgical team were also analyzed to avoid potential bias. However, they were not included in performing models because they cannot be used as available predictors.**" In the result section as" **Significantly difference of Year of surgery (OR=0.938-1.021, P=0.325) and surgical teams (OR=0.379-2.156, P=0.827) were not observed.**" In the discussion section as" **Besides, we had analysed the influences of the surgery conducted in different year and different surgery teams to the outcome, the result showed that the surgery conducted in different year and different surgery teams would not influence the outcome. We thought the reasons might be following as: First, all the patients received a standard primary complete repair, and the major procedure did not change over that time radically; Second, the distribution of patient severity is not well balanced. Many patients with severe TOF gave up and did not undergo surgery in early year. Therefore, the basic conditions of early patients are actually better than those of recent years. The improvement of surgical skills and perioperative management may be offset by the severity of the disease in recent year. Third, our hospital was the biggest Children's Hospital in southwest of China, primary complete repair is a mature procedure, all the surgery and nurse were experienced. We particularly emphasize that the use of the model requires more external validation, especially for those institutions with less surgical experience. The year of surgery and different surgery teams is possibly to affect the outcome and should not be completely ignored in these institutions.** "
we have modified our text as advised (see Page 7-8, line 152-155 in the method section; Page 10, line 206-208 in the result section; Page 15-16, line 318-332 in the discussion section)

Comment 4: The study uses 20 years of data. This causes some concerns that must be addressed. it is possible that those procedures were carried out by different surgical teams. The ability/dexterity of those teams might influence the outcome. The authors must analyse if the year in which the procedure was carried out influences the outcome (i.e. post-surgery complications). If the annual distribution of complications is not homogenous, the authors may need to control for the year/team in their analyses; this would make the predictive model less useful.
Reply 4: Thanks for your suggestion. We had added to analyze the influences of the different surgery teams to the outcome, the result showed that the surgery conducted in different surgery teams would not influence the outcome (Univariate logistic regression, OR=0.379-2.156, P=0.827). We believe that this is due to our doctors' rich experience and strict access authorization. In our institution, the premise for the operation of tetralogy of Fallot is to complete at least 500 operations of ventricular septal defect or atrial septal defect under cardiopulmonary bypass. All the operations in this study were performed by 6

chief surgeons who had at least 10 years of experience before the first radical operation of tetralogy of Fallot.
We did not analyze this factor in the initial manuscript because this study is mainly to establish a predictable model, and the surgical team could not easily be used in external validation. Your excellent opinions suggest that we may take the number of surgeons' operations as an important variable to improve the model in the future. In the relevant part of our discussion, we emphasized the importance of surgeon's technology. We hope that patients would not suffer potential risks of adverse event because of the lack of surgeon's skills and experience. We believe that this is the reason why you put forward your opinions in this regard. This is not only a revision opinion on the writing of the manuscript, but also a requirement for surgeons to improve their own skills. Thank you very much for your help.
Changes in the text: The relative description has been added in the method section as" **The year of the surgery and surgical team were also analyzed to avoid potential bias. However, they were not included in performing models because they cannot be used as available predictors.**" In the result section as" **Significantly difference of Year of surgery (OR=0.938-1.021, P=0.325) and surgical teams (OR=0.379-2.156, P=0.827) were not observed.**" In the discussion section as" **We do hope that all doctors who perform tetralogy of Fallot surgery should receive strict training, otherwise they may be the greatest risk factor for postoperative adverse events**"
we have modified our text as advised (see Page 7-8, line 152-155 in the method section; Page 10, line 206-208 in the result section; Page 16, line 332-334 in the discussion section)

Comment 5: The description of the analyses and results is not clear enough.
Table S1 contains two very different types of info: the parameter settings for the ML algorithms, and the feature importance. These data should be presented in two different tables. It is not clear from the text of the table if the feature importance was determined with the ML approaches of by other means. Also, it is not clear if the feature importance numbers in Table S1 correspond to the coefficients in Figure 2.
Reply 5: Thanks for your suggestion. We do apologize that Table S1 is not clear enough. As you mentioned we have presented the data in two different tables (Table s1 and Table s2). The feature importance was calculated according to LASSO regression for selecting variables. The relative description could be found in Statistical analysis section. The feature importance numbers in Table S1 correspond to the coefficients in Figure 2. We hope that this revision will make the manuscript clearer and more concise.
Changes in the text: We have presented the data in two different tables (Table s1 and Table s2). We renumbered and reordered the supplementary tables. Relevant files have been uploaded to the online system. we have modified our text as advised (see Page 9, line 184 and Page 10,line 210)

Comment 6: The minimum in Figure 3 seems to be around -3.5. This should correspond to a lambda around 0.03. However, 0.058 is given as the lambda minimum in the text. Also, which features were included in the analysis? Both the text and the legend mention 6 features; however, Table 2 and Figure 2 show only 5 features as different between patients with/without adverse events.

Reply 6 Thanks for your suggestion. We checked this part in detail, and there was no error in the previous calculation. The minimum in Figure 3 was -2.85(Gray dotted line on right). In that case the lambda minimum was 0.058 [ln (0.058)=-2.85].

We are sincerely sorry that the previous description was not clear. In table 2, the difference was calculated by Univariate logistic regression. Under this calculation, there are five meaningful variables. We can better find the variables that may affect the occurrence of adverse events through various ways. In figure2, we listed all importance of features. In the results section of figure 2, we only list 5 variables for reasons of omission. We understand that this may cause misunderstanding. In the revised version, we have listed all variables as SPO2, DP, CPB, TP repair, Gender, Age and so on in result section. Moreover, we revised Figure 4 by Venn diagram (Error in previous version).

Generally speaking, we used subgroup analysis (10 variables selected), Univariate logistic regression (5 variables selected) and weight importance by LASSO regression (6 variables selected) to search important variables (Figure 4 by Venn diagram). The variables confirmed by the three methods (SPO2, CPB time, DP, and TP-repair) are considered as important variables affecting the occurrence of adverse events and will be analyzed in more detail (Such as RCS in Figure 5, PSM and Detailed trend analysis in figure 6).

Changes in the text: The relative description has been added in the method section as" **The variables confirmed by the Univariate logistic regression, LASSO regression and subgroup analysis are considered as selected main variables affecting the occurrence of adverse events and will be analysed in more detail by Propensity score matching (PSM), trend analysis and restricted cubic spline (RCS) in detail.**" In the result section as" **Differences in the SPO2, CPB time, Age, PvO, DP, annulus, M-index, Z-index, TP-repair and LVEDI were observed in the two groups." " The order of feature importance was SPO2, DP, CPB, TP repair, Gender, Age and so on "** and**" Based on a Venn diagram, SPO2, CPB time, DP, and TP-repair were selected as main variables according to the three analysis methods mentioned above (Figure 4)."Figure 4** has been revised.we have modified our text as advised (see Page 9, line 176 in the method section; Page 9, line 202-204,line 208-209 and line 212-214 in the result section;)

Comment 7: The authors say both in the Results and Discussion sections that the LR model is the best one. However, this is not true according to Figure 7 and Table S3: the Gaussian NB model is slightly better in terms of AUC, sensitivity, and positive and negative predictive values. LR is only better in specificity. The

performance of the method must be measured in a dataset not used in training. Otherwise, overfitted models would probably be the best ones.
Reply 7: Thanks a lot. Your suggestion is right. As you mentioned, performance of GNB was better than that of LR in testing sets. We can not draw the conclusion that LR was the best model in both training and testing set. We have revised relative description. In fact, it is difficult to comprehensively evaluate which model is the best from the training set and the testing set, especially considering that the results of LR and GNB are similar. External validation is required in future studies. This article only shows that traditional algorithms like LR do not necessarily perform worse than complex AI algorithms in non-large sample studies. We can only say **"The best-performing model for the training set was the LR model and the performance of the LR model in the test set was good as well. " Classical algorithms such as LR still have good application in paediatric surgery research. "**
In the abstract of method, we should not say "to build prediction models and to screen out the **best** model to predict adverse events.**".** The word of **"best"** caused misunderstanding among readers and has been revised as **"good performance"** in introduction.
Changes in the text: The word of **"best"** caused misunderstanding among readers and has been revised as **"good performance"** in abstract. we have modified our text as advised (see Page 4, line 79)

Comment 8: What does the detailed analysis of the LR model entailed?
Reply 8: Thanks a lot. We performed the detailed analysis of LR because we need to provided evidence for the sentence that **"Classical algorithms such as LR still have good application in paediatric surgery research. ".** Hence, in the detailed analysis, the LR model show good Calibration and comprehensive performance. We provide a dynamic nomogram in order to make LR model used easily. Only by detailed analysis of the LR model, could we make readers believe **"Classical algorithms such as LR still have good application in paediatric surgery research. Traditional algorithms like LR do not necessarily perform worse than complex AI algorithms in non-large sample studies. "**
Changes in the text: None.

Minor concerns:
Comment 1: Both in the Introduction and Discussion, the authors claim that communication is difficult because of the parent's lack of knowledge. These statements are not only extremely patronising, but also inaccurate: the burden is on the speaker, not on the audience. Lack of comprehension usually results from the inability of the speakers to tailor their explanations to their audience.
Reply 1: Thanks for your suggestion. your suggestion is very important, and we feel very sorry the misunderstanding. We have noticed that there is a problem with our presentation. We have great respect for patients and their families. We don't want inappropriate expressions to offend anyone. In fact, we do this

research in the hope of better communicating with the patients' families and providing accurate data to help them, rather than just saying that it is possible. Such mistakes are mainly due to our lack of English writing skills. We have revised all relevant descriptions. I apologize to those of us who may be potentially hurt on behalf of co-authors.
Changes in the text: We removed all inappropriate descriptions in Introduction and Discussion. The description in the discussion has been revised as **"Because of the complex haemodynamics of TOF, communicate with parents regarding haemodynamics and cardiac abnormalities would not let parents understand this complex disease, we need a more effective method.** "we have modified our text as advised (see Page 14, line 291-294)

Comment 2: The authors introduce the application of AI into medicine in lines 74-76; however, they not reference any publication supporting their statement.
Reply 2: Thanks for your suggestion. We have noticed that and add some reference to support our statement. Reference 4: AI in health and medicine. Reference 5: Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients
Lassau, Nathalie, etc.
Changes in the text: Reference 4 and 5 have been added. (see Page 6, line 116)

Comment 3: The authors used the Kolmogorov-Smirnov test to test for normality. As far as I know this test is less powerful than the Shapiro-Wilks test. The authors should either justify their choice, or repeat the analyses with a more powerful test.
Reply 3: Amazing question. We really learn a lot from you. I have never noticed the problem in small sample size study. I carefully read the article "**Razali, Nornadiah; Wah, Yap Bee (2011). Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. Journal of Statistical Modeling and Analytics. 2 (1): 21–33.** " It said "**For symmetric distributions with kurtosis less than 3 that is platykurtic distributions, SW outperforms the other three tests. However, for sample size 30 or less the powers at 5% significance level for all four tests are less than 40%. Similarly, SW performs better than AD, KS and LF for symmetric distributions with kurtosis greater than 3 that is leptokurtic distributions. Again the performance of all tests is low for small sample sizes. Overall, generally for symmetric nonnormal distributions, SW is the best test followed by AD, LF and KS tests. Results also show that LF test performs better than the KS test.** "
Hence, your suggestion is completely right. We repeat the analyses with Shapiro-Wilks test.
Fortunately, the results of the normality test did not change significantly. We once again express our sincere thanks to you.
Changes in the text: The description in the method has been revised as **" Shapiro-Wilks test was used to evaluate a normal distribution** "(see Page 8, line 168)

Comment 4: Tables S2 and S3 contain NaN/NAN in the PPV column. This must be wrong if there are values for sensitivity.
Reply 4: Thanks for your notice. This is due to the identification error when we export the data, and we have modified the relevant parts.
Changes in the text: Table S2 and S3 have been revised and re-submitted into the online system.

Language/format concerns:
Comment 1: Mmeans (line 127) must be Means.
Reply 1: Thanks for your suggestion. your suggestion is very important, and we feel very sorry for clerical error.
Changes in the text: we have modified the sentence as advised. (see Page 8, line 167).

Comment 2: "Mann–Whitney U test was used if the was not coincided with normal distribution" sentence (line 130) is not correct.
Reply 2: Thanks for your suggestion. We have carefully analyzed your question and also consulted this question with our statistical analysts, we believe that there is no obvious error in this sentence.
In statistics, the Mann–Whitney U test (also called the Mann–Whitney–Wilcoxon (MWW), Wilcoxon rank-sum test, or Wilcoxon–Mann–Whitney test) is a nonparametric test of the null hypothesis that two samples come from the same population against an alternative hypothesis, especially that a particular population tends to have larger values than the other. Unlike the t-test it does not require the assumption of normal distributions. It is nearly as efficient as the t-test on normal distributions.
Our previous research used similar statistical methods(1,2), we are not clear whether it is because our description is not clear and caused misunderstanding or our statistical knowledge needs to be updated. Thank you very much for your help.
1. Song J, Wang Q, Pan Z, et al. The Safety and Efficacy of the Modified Single Incision Non-thoracoscopic Nuss Procedure for Children With Pectus Excavatum. Front Pediatr 2022;10:831617.
2. Li H, Jin X, Fan S, et al. Behavioural disorders in children with pectus excavatum in China: a retrospective cohort study with propensity score matching and risk prediction model. Eur J Cardiothorac Surg 2019;56:596-603.
Changes in the text: None in the text.

Comment 3: "In order to figure out the different variables" sentence (line 136) does not make much sense. It is possible to investigate/learn the contribution/effect/value of variables, but not the variables themselves as much as I am aware of.
Reply 3: Thanks for your suggestion. Your suggestion is right. As you mentioned, the sentence does not make much sense, and the meaning of this sentence does

not change, so, we decide to delete the sentence.
Changes in the text: We have delete this sentence.

Comment 4: the in line 151 must be The.
Reply 4: Thanks for your suggestion. we feel very sorry for spelling mistake.
Changes in the text: We have modified the sentence as advised. (see Page 9, line 191).

Comment 5: Do not pose questions to the reader (lines 221-222). Make your statements.
Reply 5: Thanks for your suggestion. Your suggestion is right. As you mentioned, we had pose questions to the reader and will leave a bad impression to the readers, so we have modified this sentence.
Changes in the text: We have modified our text as advised. We have modified this sentence into"However, we still have no idea on the issue such as the risk factors of the happening of adverse events and the cut-off points of risk factors" (see Page 12, line 254-255).

Comment 6: The data sharing is available" (line 340) is an incorrect sentence. It should be stated if data will be shared on request, or if the data has been deposited somewhere in order to share.
Reply 6: Thanks for your suggestion, we have realized that our expression will let readers misunderstanding, so we have modified this sentence.
Changes in the text: We have modified our text as advised. We have modified this sentence into"Data will be shared on request" (see Page 18, line 374).

Comment 7: What does "texture" mean in the Figure 3 legend?
Reply 7: Thanks for your suggestion. We use **"**texture**"** to describe the appearance shape of the LASSO regression. We understand the use of the word is inaccurate and lack of standardization. We deleted the word in the revised manuscript.
Changes in the text: Figure 3 legend has been revised.

Comment 8: Labels and axes in Figure 5 are too small.
Reply 8: Thanks for your suggestion. We have increased font size of Labels and axes in Figure 5.
Changes in the text: Figure 5 has been revised, and no change in the text.

Comment 9: The colour legend in Figure 6 should be edited to contain the names of meaningful variables.
Reply 9: Thanks a lot. We have revised Figure 6 as your requirement.
Changes in the text: Figure 6 has been revised and resubmitted.

Comment 10: Text in the Figure 7 colour legend is too small.
Reply 10: Thanks for your suggestion. We have increased size of colour legend in

Figure 7.
Changes in the text: Figure 7 has been revised and resubmitted.

Comment 11: Legend in Figure 9 contains a repeated sentence.
Reply 11: Thanks a lot. The repeated sentence has been removed.
Changes in the text: We have revised the figure legends.

Comment 12: Abbreviations in Tables 1 and 2 should be presented the other way round: first the abbreviation, then the meaning of the abbreviation.
Reply 12: Thanks for your suggestion. your suggestion is very important, we have realized that our expression is not fit the reading habit of the readers, so the sentence has been revised.
Changes in the text: We have modified the sentence as advised.

Comment 13: Figure S1 is never referred to in the text. The figure needs both a full legend and a colour legend so that we know what is plotted.
Reply 13: Thanks a lot for your notice. Figure S1 is another form of lasso regression results. The effect of this form is not good and the useful information has been disclosed in Figure 3. Therefore, we deleted this figure in the revised manuscript.
Changes in the text: Figure S1 has been delated.