Peer Review File

Article Information: https://dx.doi.org/10.21037/tlcr-23-737

Reviewer A

Lou et al have developed a machine learning model for prognosing outcomes of patients with EGFR-mutated NSCLC.

Reply: We appreciate the reviewer for recognizing the value of our paper on machine learning model developed for predicting outcomes of patients with *EGFR*-mutated NSCLC treated with third-generation EGFR-TKI.

Some of the author's meaning in the work may have been lost in the quality of the writing. The premise for developing the algorithm is that EGFR testing alone does not predict outcomes to EGFR-TKI with 100% accuracy, and therefore, either the algorithm may help by adding predictive value on top of EGFR testing alone; OR the algorithm may help to stratify EGFR-mutated NSCLC into risk groups (have prognosticating value), and thereby select patients who may not benefit from EGFR TKI.

Reply: We apologize for not making this point clear. We have added this point in Abstract section in **Lines 36-40, Page 2**.

Changes in the text:

We revised "Epidermal growth factor receptor (*EGFR*) -sensitive mutations are crucial selection criteria for EGFR- tyrosine kinase inhibitors (TKIs) treatment, and *EGFR* T790M mutation is a common genotype variation that results in resistance to first- and second-generation EGFR-TKIs. While not all T790M-positive patients respond to third-generation EGFR-TKIs and tools for efficacy evaluation of third-generation EGFR-TKIs are lacking." into "Epidermal growth factor receptor (*EGFR*) T790M mutation is the standard predictive biomarker for third-generation EGFR-TKI treatment. While not all T790M-positive patients respond to third-generation EGFR-TKIs and have a good prognosis, it necessitates novel tools to supplement *EGFR* genotype detection for predicting efficacy and stratifying *EGFR*-mutant patients with various prognoses."

Firstly, the aim of developing the algorithm, CoxMoE, has not been clearly stated in the abstract. The abstract background states the need for tools to optimize ability to better predict patients with EGFR-mutated NSCLC who may or may not benefit from EGFR TKI, it mentions nothing of the goal of the study to evaluate the said algorithm.

Reply: Thanks for your professional suggestions. We added the description of CoxMoE in Abstract section.

Changes in the text: We added in **Lines 41-45, Page 2** as follows: "Mixture-of-experts (MoE) is designed to disassemble a large model into many small models. Meanwhile, it is also a model ensembling method that can better capture multiple patterns of intrinsic subgroups of enrolled patients. Therefore, the combination of MoE and Cox algorithm has the potential to predict efficacy and stratify survival in NSCLC patients with *EGFR* mutations."

Line 86-89, sentence "Whereas around 30% of patients….. in our experience." The sentence does not make sense grammatically.

Reply: We apologize for our mistake. We revised this sentence in **Lines 85-88, Page 4**.

Changes in the text: We revised the sentence from **"**Whereas around 30% of patients with T790M mutation may fail to respond to third-generation EGFR-TKI (Barnet et al, 2017; Soria et al, 2015), which might be even lower in clinical trials of EGFR-TKI according to our experience." to "Whereas around 30% of patients with T790M mutation may fail to respond to third-generation EGFR-TKI (Barnet et al, 2017; Soria et al, 2015), based on our experience, it might be even higher in clinical trials of EGFR-TKI."

Line 89-91, following the above sentence "Meanwhile, almost half of….EGFR sequencing." This sentence is not entirely true. Whilst EGFR sequencing alone may not be highly sensitive in detecting early resistance to EGFR-TKI, but there is evidence that the use of circulating-tumor DNA may reveal early drug resistance, depending on the test. In this principle, the data on ctDNA monitoring may actually be more mature than what is being studied in this trial.

Reply: Thanks for your kind suggestion. We agree with the reviewer that circulating-tumor DNA may reveal early drug resistance and we also realize that it's not approtiate to express it this way. In order to eliminate possible misunderstandings, we have revised the previous sentence. Meanwhile, we have emphasized the importance of EMR data in the third paragraph of the introduction, according to the reviewer's comments.

Changes in the text: We revised the sentence from **"**Meanwhile, almost half of the secondary T790M-mutated patients will progress within 18 months (Ramalingam et al, 2020; Soria et al, 2018), and thus, they cannot be assessed through *EGFR* sequencing." to "Meanwhile, *EGFR*-sensitive patients inevitably develop drug resistance, suggesting *EGFR* testing alone is insufficient (Soria et al., 2018 , Ramalingam et al., 2020). Due to tumor heterogeneity and difficulties obtaining tissue from advanced-stage patients, non-invasive biomarkers that could stratify NSCLC patients with a specific *EGFR* mutation are needed to aid in targeted therapy administration." (**Lines 88-93, Page 4**)

The introduction section would be more informative by a more succinct summary of the inability of EGFR testing alone to predict response to EGFR-TKI with high accuracy, and probability room for AI to contribute. Artificial intelligence should not have to be defined as such definitions are well known in popular media and also easily searchable. Comparing an algorithm based on laboratory testing vs radiomics may be one way to highlight the shortfalls of other existing ML models, but whilst "the potential variability of diagnoses across hospitals and operators due to manual checkout (line 104-105)" may be lower for laboratory-based results vs imaging-based results, the process is not entirely automated and practices still vary across hospitals. E.g. are all laboratories accredited the same way? Are the tests results generated utilizing the same kits, etc.

Reply: Thank you for your professional suggestions. The hospital laboratory is required to pass the CNASISO15189 criteria to ensure that the results are comparable between laboratories. However, the results vary from hospital to hospital due to differences in procedures and kits. Thus we rewrote this paragraph and deleted this sentence which may cause misleadings.

Changes in the text: We revised "Artificial intelligence (AI) attempts to simulate human brain behaviors based on a neural network to learn from large amounts of data automatically. Currently, the primary tool monitoring EGFR-TKI future risk is computed tomography (CT), which exhibits tumor features in CT imaging non-invasively. AI combined with CT has shown potential for predicting EGFR-TKI responses and optimizing treatment decisions. For example, previous studies have proposed a fully automated artificial intelligence system (FAIS) that mines lung information from CT images focusing on *EGFR* mutation status prediction to identify patients sensitive to EGFR-TKI (Deng et al, 2022; Mu et al, 2020; Song et al, 2020; Wang et al, 2022). However, the high cost and complex operational procedures required for CT, combined with the potential variability of diagnoses across hospitals and operators due to manual checkout, may limit the widespread implementation and adoption of CT-based AI models (Yip & Aerts, 2016). " to "Presently, *EGFR* genotype detection of tumor tissues is considered as the gold standard for EGFR-TKIs treatment in NSCLC. While not all *EGFR*-mutant NSCLC patients respond to EGFR-TKI therapy and complete responses are rare. Moreover, *EGFR*-sensitive patients inevitably develop drug resistance, suggesting *EGFR* testing alone is not enough. Efforts have been made to develop new approaches for predicting efficacy and prognosis stratification. Currently, the primary tool monitoring EGFR-TKI future risk is computed tomography (CT), which exhibits tumor features in CT imaging non-invasively. AI combined with CT has shown potential for predicting EGFR-TKI responses and optimizing treatment decisions. For example, previous studies have proposed a fully automated artificial intelligence system (FAIS) that mines lung information from CT images focusing on *EGFR* mutation status prediction to identify patients sensitive to EGFR-TKI (Deng et al, 2022; Mu et al, 2020; Song et al, 2020; Wang et al, 2022). These results illustrate that the construction of a machine learning model based on clinical data has great potential in predicting the efficacy of EGFR-TKI." (**Lines 94-108, Page 4-5**)

Similarly, it should not be necessary to define what is an EMR in the third paragraph of the introduction. The authors' meaning here is vaguely grasped, but will require rewriting for more specificity and clarity about what work others have done.

Reply: Thank you for your professional suggestions. We rewrote this paragraph in our revised manuscript.

Changes in the text: We revised in **Lines 109-134, Page 5-6** as follows: "More recently, there has been a groundswell of interest in using artificial intelligence and laboratory values obtained from Electronic Medical Record (EMR) data to develop risk models for disease diagnosis and prognosis prediction. For instance, a previous study developed a modified version of the well-validated 2012 Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial risk model (mPLCOm2012) using Extreme Gradient Boosting (XGBoost) algorithm based mainly on routine laboratory test data. MPLCOm2012 was designed to diagnose NSCLC, and the performance of mPLCOm2012 was evaluated in 6,505 NSCLC patients and 189,597 control subjects with an AUC of 0.79 and a sensitivity of 27.9% at a specificity of 95% (Gould et al., 2021). Furthermore, a gradient-boosted decision tree (GBDT) model incorporating patient demographic features (age, sex, race) with 27 routine laboratory tests to predict an individual's SARS-CoV-2 infection status with AUCs of 0.838-0.854 (Yang et al., 2020). Moreover, EMR has been commonly and economically used as inclusion and exclusion criteria in clinical trials.

In contrast, the majority of clinical laboratory tests utilize established reference values for defining thresholds, which may not always be suitable for a particular study for being either too strict or too permissive. EMR itself usually cannot fully identify patients' responses to the drug, while sophisticated analytics methods could assist in making full use of the EMR data to identify high-risk patients' subsets, probably with poor prognoses. For example, Trial Pathfinder, has been developed to associate EMR with survival hazard ratios (Liu et al., 2021). In addition, machine learning (ML) algorithms combined with EMR and genotype data are a potentially helpful tool for providing clinicians with early toxicity prediction in Phase I clinical trials (Bedon et al., 2022). Thus, leveraging AI algorithms to analyze EMR data in early clinical trials might hold great potential for diagnosing disease and predicting efficacy and toxicity, aiding in patient selection and improving the success rates of clinical trials."

Line 129-131: "This non-invasive prognostic system performs better than ….in clinical practice." What traditional methods are the authors referring to? How is this current model better? If it is indeed better, this should be discussed in the final Discussion section, as the Introduction should mainly lay the background highlighting unmet need, and the goal of the study, which is to test the performance of CoxMoE against a certain standard for either EGFR-mutated NSCLC prognostication, or ability to predict response to EGFR-TKI.
Reply: We apologize for not making this point clear. The traditional methods refer to previously released CoxNet [PMID: 34458659, PMID: 27065756], CoxSVM [Fast Training of Support Vector Machines for Survival Analysis], and DeepSurv [PMID: 29482517]. In comparison with these methods, CoxMoE performed the best. We have added the comparison of these models in Results section, which is also discussed in the Discussion section.
Changes in the text: We added these results in **Lines 318-330, Page 12-13** as follows: "Furthermore, CoxMoE performed better than typical machine-learning models (CoxNet and CoxSVM) for survival analysis and another deep-learning model, DeepSurv (Table S5 and S6). We evaluated the performance of two machine-learning models (CoxNet and CoxSVM) and two deep-learning models (CoxMoE and DeepSurv). CoxMoE achieved the highest C-index in the training cohort with an averaged C-index score of 0.6761 for cross-validation (Table S5). CoxNet performed worst with an averaged C-index of 0.6443 in the training cohort and 0.5817 in the validating cohort (Table S5).

Table S5. Performance comparison of different models in survival analysis using the selected 4 features

Method	**Cross-validation (Averaged)**
CoxNet	0.6443
CoxSVM	0.6663
DeepSurv	0.6681
CoxMoE	0.6761

As shown in Table S6, CoxMoE performed better than DeepSurv in predicting PFS (C-index for CoxMoE and DeepSurv reached 0.6732 and 0.6527, respectively) and efficacy (ACC: 0.7714 and 0.7564, respectively; AUC: 0.8181 and 0.7814, respectively) for cross-validation."

Table S6. Performance comparison of two deep-learning models in multi-task modeling

Method	**Cross-validation (Averaged)**	
	Risk Score	Treatment Response

	Prediction	Prediction	
	C-Index	ACC	AUC
DeepSurv	0.6527	0.7564	0.7814
CoxMoE	0.6732	0.7714	0.8181

In Discussion section, we revised in **Lines 407-418, Page 15-16** as follows: “In this study, CoxMoE model performed better than DeepSurv because the design of CoxMoE was more conducive to capturing different intrinsic subgroup patterns of enrolled patients. Machine-learning methods generally performed worse on every score than deep-learning methods. While deep-learning methods are much more prone to be overfitted on given data than machine-learning methods, there are also plenty of ways to prevent it (e.g., add a dropout layer or add a regularization loss item). More importantly, deep-learning methods can quickly implement different tasks in a single model, but machine-learning methods cannot. After the new task was added, the predictive performance decrement of models for the original task was almost negligible, mainly due to the correlation between the two tasks. Also, our proposed model, CoxMoE, has shown its advantages in this multi-task modeling experiment.”

The first paragraph of the Methods section also requires some specificity. Whilst the entire section requires grammatical review to ensure meaning has not been altered as a result of grammatical misuse.

Reply: We apologize for not making this point clear. We revised the primary efficacy endpoint to ensure specificity. Meanwhile, we checked and corrected the grammatical misuse of this section.

Changes in the text: We revised “The workflow of this study is graphically summarized in Figure 1. Initially, pre-processed single feature data were fed into two machine learning models, namely CoxNet and CoxSVM, to reduce the number of features and identify features with good performance. Subsequently, we computed the probability of patients being poor/good responders and the risk score of survival. ” into “The workflow of this study is graphically summarized in Figure 1. Initially, we assembled 177 patients in Abivertinib phase I clinical trial into a training cohort (n=177) and patients from phase II clinical trial of Abivertinib were used as validation cohort 1 (n=106). Forty-three patients were randomly selected from the BPI-7711 phase I clinical trial as validation cohort 2. The preprocessed single feature data of training cohort were fed into CoxNet, and the top 15 features were selected based on C-index. Then, these 15 features were shrunk into 4 features based on C-index calculated by CoxSVM. Subsequently, we trained CoxMoE in the training cohort and computed the probability of patients being responder (R)/ non-responder (NR) and the risk score of survival in two validation cohorts.” in **Lines 150-159, Page 6**.

We added “We classified patients with CR or PR as R and patients with SD or PD were defined as NR. ” in **Lines 169-170, Page 7**.

The study endpoints are not just ORR/PR/CR/ survival, but also as are compared between poor and good responders.

Reply: Thanks for your professional suggestion.

Changes in the text: We added the definition of poor and good responders in **Lines 169-170, Page 7** as follows: "We classified patients with CR or PR as R and patients with SD or PD were defined as NR. "

The authors also explained their reason for the calculation methods in the Methods section, whereas this is usually more appropriate in the Background or Discussion section. Again it is not necessary to define the Shapley value, or the value of an MoE design, just as we would not usually define e.g. a t-test in a study methods section. What are the CoxNET and CoxSVM models, as well as DeepSurv algorithm? Are they validated? References to prior validation reports should be given.

Reply: We apologize for not making this point clear. We removed the calculation methods of CoxMoE and added these descriptions in Abstract and Introduction sections. We have removed the description of Shapley value. Because CoxMoE is an algorithm we developed ourselves and first proposed, we described the calculating formula of this algorithm. The CoxNet [PMID: 34458659, PMID: 27065756], CoxSVM [Fast Training of Support Vector Machines for Survival Analysis], as well as DeepSurv [PMID: 29482517] algorithms were proposed and validated in previous studies. We have added the references of each algorithm in our revised manuscript.

Changes in the text: We added the references of each algorithm in **Lines 248-253, Page 10** as follows: "CoxNet is a model based on ElasticNet, an improved version of CoxPH, a linear regression model that uses L1 and L2 priors as regularization matrices (Bellal et al., 2021 , Simon et al., 2011), while CoxSVM is a nonlinear Cox model (Pölsterl et al., 2015). DeepSurv is a deep-learning-based model utilizing a multilayer perceptron (MLP) to fit features and NLL function for loss calculation (Katzman et al., 2018)."

We added "Since mixture-of-experts (MoE) contains a neural network, similar feature input will yield a similar output. Therefore, more similar samples were assigned to the same expert model to realize the data's automatic grouping and clustering. To some extent, this property aligns with our perception of the real world. For example, men and women have different prognostic patterns in certain diseases." In **Lines 135-139, Page 6**.

And "Mixture-of-experts (MoE) is designed to disassemble a large model into many small models. Meanwhile, it is also a model ensembling method that can better capture multiple patterns of intrinsic subgroups of enrolled patients. Therefore, the combination of MoE and Cox algorithm has the potential to predict efficacy and stratify survival in NSCLC patients with *EGFR* mutations." in **Lines 41-45, Page 2**.

"Our modeling can perform well even when the sample size is small.", because of what?

Reply: We realized that this sentence is not appropriate, so we removed this sentence to avoid any unnecessary misunderstandings.

The term NLL first appears line 213, but is only elaborated in line 236 - this should be reversed.

Reply: We apologize for not making this point clear. We have revised this point accordingly.

Changes in the text: "The negative log-likelihood (NLL) and cross-entropy function were used for loss calculation. " in **Lines 227-228, Page 9** and "DeepSurv is a deep-learning-based model utilizing multilayer perceptron (MLP) to fit features and NLL function for loss calculation." in

Lines 251-252, Page 10.

There are again sections of the Results section, which could be omitted (e.g. defining APTT, lines 281-287 is redundant for the purpose of this study where the biological relationship of the features selected is not the focus, as the selection is based on calculation of probabilities). I would recommend rephrasing the sentence spanning lines 297-301, as it is not immediately obvious that the risk groups are being stratified by PFS (versus e.g. OS). It is also not very obvious what the first sentences under the sections "Decision Curve Analysis" and "Interpretation of CoxMoE by WES and Shapley values", is trying to convey, although it is possible to understand the ensuing explanation following those sentences.

Reply: Thanks for your kind suggestions. We have made modifications according to the reviewer's comments

Changes in the text: We revised "Based on the risk score calculated for the training cohort, we divided the Abivertinib trial cohort into high-risk (median [range], 4.2 [1.0-35] months) and low-risk (median [range], 6.0 [1.0-23.3] months) groups and the two groups exhibited significant distinct PFS (HR, 0.56; 95% CI, 0.40–0.78; P = .0013) (Figure 3A)." into "Based on the risk score calculated for the training cohort, we divided the Abivertinib trial cohort into high- and low-risk groups and the two groups exhibited significant distinct PFS (HR, 0.56; 95% CI, 0.40–0.78; P = .0013) (Figure 3A)." (**Lines 330-332, Page 13**)

We revised "According to the DCA, the prediction of therapeutic response could achieve better clinical benefits than PFS prediction across the risk probabilities of 12%-60%, revealing the necessity for early efficacy prediction (Figure 4A). " into "The decision curve analysis (DCA) indicated that the prediction of therapeutic response could achieve better clinical benefits than PFS prediction across the risk probabilities of 12%-60% (Figure 4A), revealing the necessity for early efficacy prediction." in **Lines 342-345, Page 13**.

We revised "CoxMoE has the advantage of identifying the heterogeneity of the population, thus, we speculated that this ability may be accomplished by the selected features. In the 18 patients who underwent whole exome sequencing (WES), we found that four features were associated with distinct altered pathways contributing to tumor aggressiveness and metabolism." into "In the 18 patients in validation cohort 2 who underwent WES detection, we found that the four deep-learning features were associated with distinct altered pathways contributing to tumor aggressiveness and metabolism. " in **Lines 349-351, Page 13**.

It does appear that the model has some additive prognostic value on top of EGFR testing alone. Discussion section could also benefit from rewriting. There is no need to highlight the unmet need anymore as this has already been done so in the Introductions/Background section. The non-invasive nature of this method is not a practical highlight, as it will not obviate the need for blood testing, molecular profiling, nor imaging. NR patients may have a prolonged APTT as it likely reflects pre-existing organ dysfunction and is a prognosticating feature. It could be useful, instead, to focus the discussion more on what existing models or algorithms have to offer or compare to the one being studied. The focus should highlight on limitations and strengths of the algorithm, rather than the technology of ML versus other technologies.

Reply: Thanks for your professional suggestion. We agree with the reviewer's suggestion and we rewrite this paragraph. The revised paragraph focused on the limitations and strengths of the algorithm in Discussion section.

Changes in the text:

We revised in **Lines 397-418, Page 15-16** as follows: "In contrast with previous artificial intelligence-based models, CoxMoE simultaneously predicts efficacy and personalized prognosis. A previous study demonstrated that *EGFR* genotype and prognostic information cannot be obtained only from tumor tissues. Macro-level changes were also correlated with therapeutic efficacy and prognosis (Wang et al., 2022). The good performance of CoxMoE further proved this point. Unlike previous studies that extract tumor information from pre-therapy CT images as the input, this study is the first to explore EMR data by artificial intelligence for efficacy and prognosis prediction of EGFR-TKI. To ensure the robustness of CoxMoE, we built and validated CoxMoE in two prospective multicenter cohorts collected from 16 hospitals and 12 hospitals, respectively.

In this study, CoxMoE model performed better than DeepSurv because the design of CoxMoE was more conducive to capturing different intrinsic subgroup patterns of enrolled patients. Machine-learning methods generally performed worse on every score than deep-learning methods. While deep-learning methods are much more prone to be overfitted on given data than machine-learning methods, there are also plenty of ways to prevent it (e.g., add a dropout layer or add a regularization loss item). More importantly, deep-learning methods can quickly implement different tasks in a single model, but machine-learning methods cannot. After the new task was added, the predictive performance decrement of models for the original task was almost negligible, mainly due to the correlation between the two tasks. Also, our proposed model, CoxMoE, has shown its advantages in this multi-task modeling experiment."

It is also important to describe how the data was extracted. Is the source data reliable and accurate. The authors said that their model is good for small sample size. How did they arrive at an appropriate sample size estimation? The discussion section should also talk in more detail about how the model should fit into the current clinical or trial workflow. The authors should also highlight whether data is confidential.

Reply: Thank you for your suggestion. Physical examinations and clinical laboratory tests were performed at screening. All subjects had a morning fasting blood sample for laboratory tests and subsequently, these data were extracted. Pharmacokinetics parameters, including maximum concentration (Css_max), and minimum concentration (Css_min), were directly extracted from pharmacokinetics analysis as previously described [PMID: 29626621, PMID: 35181498]. These data were strictly detected and analyzed according to the procedure of clinical trials. We agree that the patient size of validation cohort 2 in this study is not big enough. The sample size of validation cohort 2 was estimated using the R package 'pwr' to achieve 80% power with $\alpha = 0.0005$. It was observed that the optimal sample size was 43 (Fig. S2). Moreover, this is par for the course in the field, in which studies with about 340 patients routinely are published [PMID: 33331920]. Taken together, we enrolled a total of 326 patients regarding previously published literature and the results of sample size calculation. Though the validation cohort 2 is relatively small, it is completely independent of training and validation cohort 1. What's more, the patients of these cohorts were collected from multiple centers across China. We could provided survival data of these patients.

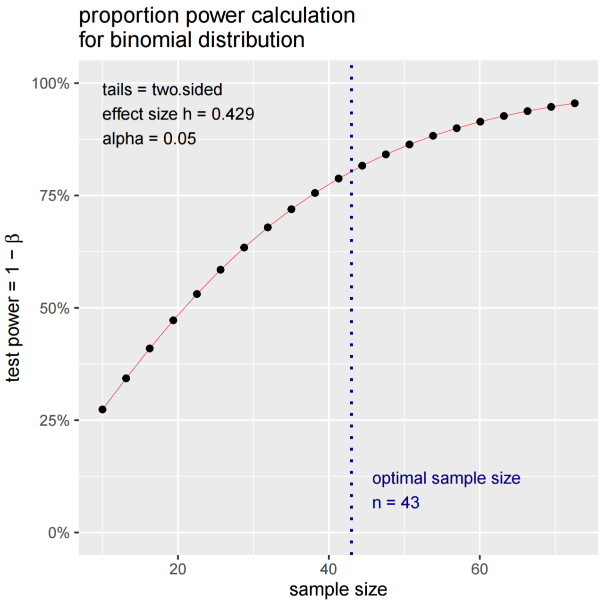

Fig. S2 The plot of sample size and power

Changes in the text: We added the description of the sample size calculation in **Lines 275-278, Page 11** as follows: "It was observed that the optimal sample size was 43 (Figure S2) to achieve 80% power with α = 0.0005. Thus, we randomly selected 43 patients from BPI-7711 trial as validation cohort 2 (mean [SD] age, 59 [10] years; 31 [68.8%] female)."

We revised in Discussion section "The CoxMoE model effectively supplements *EGFR* genotype detection, which could aid in selecting appropriate patients for EGFR-TKI treatment. Patients confirmed to have an *EGFR* mutation by gene sequencing and predicted to be R to EGFR-targeted therapy by CoxMoE showed good prognosis. However, those with a confirmed *EGFR* mutation by gene sequencing but predicted to be NR showed a poor prognosis. Importantly, the CoxMoE model provides personalized PFS predictions for patients undergoing EGFR-TKIs, offering a means to stratify *EGFR*-mutant genotypes based on individual therapeutic responses. Consequently, the CoxMoE system represents a considerable expansion to gene sequencing." (**Lines 453-462, Page 17**)

Reviewer B

The authors use one cohort to develop a prediction tool and two other cohorts to validate/test this tool. The aim is to use this tool to predict a poor response to a 3rd generation tyrosine kinase inhibitor used in the presence of the T790M mutation. The development of such a tool would indeed be of great clinical interest.

We appreciate that the reviewer recognizes the clinical value of our paper, which develops this prediction tool to predict a poor response to a 3rd generation tyrosine kinase inhibitor used in the presence of the T790M mutation.

Major comments:

The tool was developed with a cohort having relatively small numbers, and the two validation cohorts also have small numbers. Did the authors take this into account? If so, how? Were the numbers calculated or estimated?

Reply: We agree that the patient size of validation cohort in this study is not big enough. For validation cohort 1, we directly collected patients (n=106) with available EMR data. The sample size of validation cohort 2 was estimated using the R package 'pwr' to achieve 80% power with α = 0.0005. It was observed that the optimal sample size was 43 (Fig. S2). Moreover, this is par for the course in the field, in which studies with about 340 patients routinely are published [PMID: 33331920]. Taken together, we enrolled a total of 326 patients regarding previously published literature and the results of sample size calculation. Though the validation cohort 2 is relatively small, it is completely independent of training and validation cohort 1. What's more, the patients of these cohorts were collected from multiple centers across China.

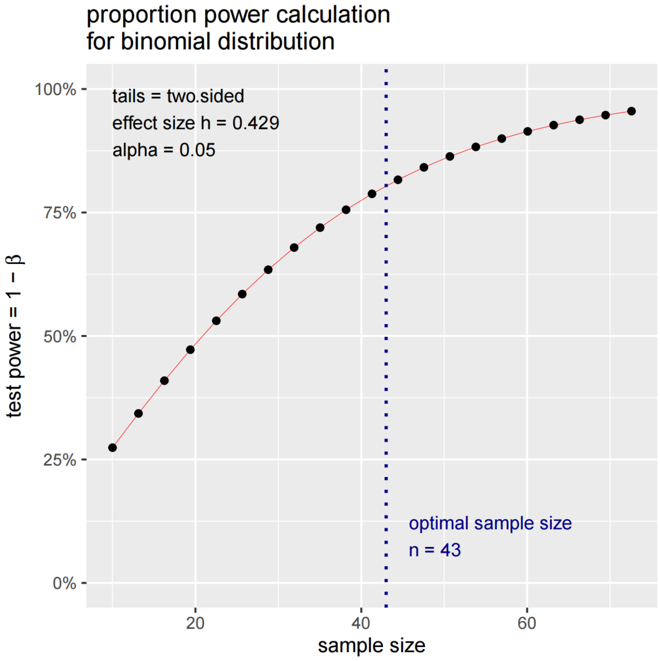

Fig. S2 The plot of sample size and power

Changes in the text:

We added the description of the sample size calculation in **Lines 275-278, Page 11** as follows:

"It was observed that the optimal sample size was 43 (Figure S2) to achieve 80% power with α = 0.0005. Thus, we randomly selected 43 patients from BPI-7711 trial as validation cohort 2 (mean [SD] age, 59 [10] years; 31 [68.8%] female)."

The cohorts come from two different studies. The authors do not mention how the T790M mutation was detected in patients in each study. However, there may be heterogeneity in the presence of this mutation and differences in detection sensitivity; this parameter may explain the poor response to TKI in some patients.

Reply: Thanks for your professional suggestion. For Abivertinib study, T790M status was conducted by a central laboratory from a tissue biopsy specimen or plasma samples using an amplification refractory mutation system (ARMS) [PMID: 34740925, PMID: 29626621]. For BPI-7711 research, patients were centrally confirmed *EGFR* T790M mutation according to either tumor tissues or plasma samples using the cobas *EGFR* mutation test [PMID: 35181498]. Both of these studies are derived from clinical trial projects, which underwent stringent *EGFR* genotype testing and screening when enrolling patients, and the relative papers have been published [PMID: 34740925, PMID: 35181498, PMID: 29626621].

Changes in the text:

We added the description of T790M detection in **Lines 174-177, Page 7**: “T790M status was conducted by a central laboratory from a tissue biopsy specimen or plasma samples using an amplification refractory mutation system (ARMS)(Ma et al., 2018 , Zhou et al., 2022) or the cobas *EGFR* mutation test (Shi et al., 2022b).”

Part of the study involved WES of samples, analysis of the mutations and signalling pathways involved. Although there is a brief description of the sequencing and analysis (supplemental), much important information is missing. For example, which tissues were sequenced? If on the primary tumor, what percentage (to account for heterogeneity)? Was the T790M mutation found? How were the signaling pathways identified? Was there a statistical study? If so, which one? How was figure 5 obtained? This part must be detailed and better explained.

Reply: We apologize for not making this point clear. We sequenced the tumor tissue with a percentage of tumors of more than 80%. We confirmed that T790M mutation could be founded by WES in these patients. We applied Spearman correlation analysis to explore the association between four features and genetic mutations. KEGG signaling pathway enrichment analyses were conducted by using DAVID using the genetic mutations which were significantly associated ($p \leqslant 0.05$) with each feature. Then unsupervised hierarchical clustering (Pearson correlation, average-linkage method) was performed on the correlation coefficients of Spearman correlation analysis.

Changes in the text:

We added these description in **Lines 259-265, Page 10**: “We applied Spearman correlation analysis to explore the association between four features and genetic mutations. KEGG signaling pathway enrichment analyses were conducted using DAVID using the genetic mutations that were significantly associated ($p \leq 0.05$) with each feature. Then, unsupervised hierarchical clustering (Pearson correlation, average-linkage method) was performed on the correlation coefficients of Spearman correlation analysis. ”

Minor comments:

English needs revision, some sentences are incomplete or the meaning is difficult to understand.

Reply: Thank you for your suggestion. We have checked and corrected the typos throughout the manuscript according to the editor’s suggestions. The language in our paper has been smoothed by a native English speaker.