TCR TRANSLATIONAL CANCER RESEARCH
ADVANCES CLINICAL MEDICINE TOWARD THE GOAL OF IMPROVING PATIENTS' QUALITY OF LIFE

JOURNAL CITATION REPORTS® IMPACT FACTOR 1.241 © 2021 Clarivate Analytics

Peer Review File

Article information: https://dx.doi.org/10.21037/tcr-22-883

Reviewer A

Kang et al. present an interesting paper which reliably predicts BRAF V600E/K variants from gene (mRNA) expression data in thyroid carcinoma, colon adenocarcinoma, and cutaneous melanoma, all cancers known to have relatively high incidence of BRAF V600E/K alterations. The authors have done similar work to predict PIK3CA and homologous recombination deficiency using the TCGA mRNA expression data. The following points should be addressed before consideration for publication.

1. The AUC for ROC and precision-recall curve differ between the abstract and results. If the abstract is correct, then AUC for ROC of 0.85 in cutaneous melanoma, not 0.86. Also ROC for precision-recall should be 0.98, 0.71, and 0.65 for thyroid carcinoma, colon adenocarcinoma and cutaneous melanoma, respectively, in the abstract.

Response: Thank you for pointing this out. The main text results are correct, and we have revised the abstract. The revised text reads as follows on page 3 line 15-17.

2. Cancer types are excluded from the final model. Are genes such as ETV1, AKT2 etc, overexpressed in all cancers with BRAF V600E/K variant, or differently expressed between thyroid carcinoma, colon adenocarcinoma, and cutaneous melanoma? Can these coefficient values be applied to other cancers types with BRAF V600E/K variant? Is there difference of mRNA expression of these genes between BRAF V600E and BRAF V600K cancers?

Response: Thank you for pointing this out. It is interesting questions about many aspects of the predictor genes of our model including whether they are over-expressed or differently expressed and whether they can be generalized to all cancer types. Although we agree that this is an important consideration, it is beyond the scope of this manuscript because we were primarily interested in the prediction but not in the cancer biology of BRAF V600E and BRAF V600K cancers. Generally speaking, the coefficient of the logistic regression model associated with a predictor X is the expected change in log odds of having the outcome (presence of BRAF V600E/K

variant). In prediction models such as machine learning, the interpretation of coefficient is not rigorous because the prediction modeling process does not consider assumptions such as independence of errors, linearity in the logit for continuous variables, and absence of multicollinearity in contrast with a statistical model.
Our prediction model can not be applied to other cancer types because the performance for other cancer types was poor. Therefore the coefficient value can not be generalized to other cancer types.

3. Minor grammatical errors are observed. Page 8, line 9 should be "which pathways are important in predicting BRAF V600E/K variants." Page 8 lines 20-21 should be "The coefficient values of genes that were included in the final model are summarized in S2 table".

Response: Thank you for pointing this out. The reviewer is correct. The revised text reads as follows on page 8 line 10 and page 8 line 21- page 9 line 1.

4. Cutaneous means skin, so skin cutaneous melanoma should be changed to cutaneous melanoma.

Response: Thank you for pointing this out. We used skin cutaneous melanoma because TCGA uses that term. As pointed out by the reviewer, skin cutaneous melanoma is redundant. We change skin cutaneous melanoma to cutaneous melanoma. Page 6 line 15, etc.

5. In S1 table, Number of case should be Number of cases.

Response: Thank you for pointing this out. The reviewer is correct, and we have changed Number of case to Number of cases.

6. Page 10, line 2. S2 table should be S3 table.

Response: Thank you for pointing this out. The reviewer is correct, and we have changed S2 table to S3 table. page 10 line 1

Reviewer B

In the present study, authors attempt to develop a prediction model to evaluate BRAF V600 variants by a penalized logistic regression analysis of mRNA data in various cancer types. Data processing approaches used included kNN Imputation for missing

values, Yeo-Johnson Transformation for skewness correction in the raw variables, synthetic minority over-sampling technique for imbalanced data, and hyperparameter optimization with a grid search. The model showed a good performance in predicting BRAF V600E/K variants in thyroid cancer, skin cutaneous melanoma, and colon adenocarcinoma as it achieved a high area under the curve of the receiver operating characteristic curve and a high the area under the precision-recall of the test set.
The study was interesting and the writing was clear. The authors presented a potentially useful prediction model to evaluate the occurrence of BRAF variants. Since the model was built mainly based on the TGCA dataset, the reviewer is curious whether it could be applied to a different dataset.
My specific comments are below.

1. At lines 12-13 page 5, it states the methods for BRAF variant detection. IHC and digital PCR may be included. The newly developed digital PCR is an alternative approach that can be easily used for quantitative detection of the presence of BRAF V600E variant (doi:10.1001/jamanetworkopen.2021.27243).

Response: Thank you for this suggestion. We have changed the manuscript as your suggestion. Page 5 line 14-15.

2. At lines 1-2 page 7, “The training set included 1136 cases, 376 cases in the first scheme was unseen test set 9377 cases.” The sentence sounds unusual and revision is required.

Response: Thank you for pointing this out. The reviewer is correct. The revised text reads as follows on page 7 line 2-3.

As there are differential effects between BRAF V600E and V600K, please specify the numbers of BRAF V600E, V600K cases, and total cases along with the prevalence at 0.57 for BRAF V600E/K variants in Thyroid carcinoma. Also list the detailed numbers for skin cutaneous melanoma and colon adenocarcinoma.

Response: While reviewing the data according to the reviewer's opinion, we found that V600K is annotated as “V600E, V600M”. Therefore, our model predicts V600E only. We revised V600E/K to V600E according to the context in the manuscript including the title. We specified the case number of BRAF V600E on page 7 line 5-6.

3. At lines 20-21 page 8, the statement “The cancer types were excluded from the final model.” was duplicated. And it’s unclear what cancer types were meant.

Response: Thank you for pointing this out. We removed the duplicated sentence. Pange We initially included the cancer type (colon cancer, cutaneous melanoma, thyroid cancer) as predictors. The cancer types were excluded during the model training process. Our model training process selects predictors for final models.

4. The coefficient values of predictors ranged from +0.300852407 to -0.280122156 in S2 table. Please define the coefficient value specifically. How were the values generated? And what were their differential biological significances of positive and negative values relevant to BRAF V600E/K?

Response: Generally speaking, the coefficient of the logistic regression model associated with a predictor X is the expected change in log odds of having the outcome (presence of BRAF V600E/K variant). In prediction models such as machine learning, the interpretation of coefficient is not rigorous because the prediction modeling process does not consider assumptions such as independence of errors, linearity in the logit for continuous variables, and absence of multicollinearity in contrast with a statistical model.

5. Can the authors comment on the clinical utility of BRAF V600E mutation prediction model based on mRNA expression in a real world?

Response: We agree that this is a potential limitation of the study. Currently, gene expression tests are limitedly used in clinical fields and are expensive, so it is difficult to apply our predictive model right away. This study shows that a specific gene mutation can be estimated with a gene expression test, it will have value when the gene expression test is widely used.

Reviewer C

It is an interesting study to predict the BRAF V600E/K mutation using gene expression data. Clinically, the BRAF mutation will be decided by using PCR. Some comments are as follows.

1) It is better to list the number of BRAF nutation # of patients and percentiles in each type of cancer.

Response: Thank you for pointing this out. The reviewer is correct, and we add a number of cases with BRAF V600E variant and prevalence on page 7 line 5-6.

2) As the authors mentioned, it is more interesting to understand the molecular

mechanisms or consequent signaling related to the BRAF V600E/K mutation vs other genotypes. Therefore, it is important to investigate the top-ranked genes, and annotate these genes using pathway or GO term enrichment analysis.

Response: Thank you for your suggestion. We found the following significantly overrepresented pathways using gene ontology test; Insulin/IGF pathway-protein kinase B signaling cascade, PI3 kinase pathway, Endothelin signaling pathway, Integrin signaling pathway, Apoptosis signaling pathway, T cell activation, CCKR signaling map, Inflammation mediated by chemokine and cytokine signaling pathway, Gonadotropin-releasing hormone receptor pathway.

3) The survival analysis using these top-ranked genes might be also informative.

Response: Thank you for your suggestion. But we believe that survival analysis with individual gene expression has limitations because it is difficult to validate the survival difference according to the gene expression in this dataset.