Peer Review File

Article Information: https://dx.doi.org/10.21037/tlcr-22-714

Review Comments

Comment 1: First, the title needs to indicate the clinical research design of this study, i.e., a retrospective cohort study.

Reply 1: We would like to thank the reviewer very much for this valuable input.We have changed the title of our research manuscript as follows:

> *"High Expression of Transcription Factor POU2F1 Confers Improved Survival on Smokers with Lung Adenocarcinoma –Retrospective Study of Two Cohorts".*

Comment 2: Second, the abstract needs some revisions. It did not indicate the knowledge gaps on the prognostic biomarkers of smokers with lung adenocarcinoma and what the limitations of prior studies are. It did not describe the inclusion of subjects, assessment of baseline clinical factors including the test of POU2F1, follow up procedures and measurements of prognosis outcomes. The results need to quantify the findings on the prognostic role of POU2F1 by using HR and P values and clearly indicate the clinical covariates adjusted. The conclusion needs comments for the clinical implications of the findings.

Reply 2: Likewise, we see this input of the reviewer of high relevance. We have adjusted the abstract accordingly including knowledge and gaps of know prognostic factors in NSCLC. We also added the missing information on evaluation methods with regards to POU2F1 expression. Moreover, Hazard Ratios and p-values are included in the abstract, now. In these contexts, our short conclusion should now be able to explain why POU2F1 is a biomarker of high interest in NSCLC. The abstract now reads as follows:

> "Non-Small Cell Lung Cancer (NSCLC) is the leading cause of cancer-related mortality worldwide and its most important risk factor is tobacco smoking. While smoking is associated with inferior outcome in NSCLC patients, smoking also correlates with a higher tumor mutational burden. In contrast to adenocarcinomas (ADC) of non-smokers, that frequently harbor targetable gain-of-function mutations, NSCLC smokers largely present with non-targetable loss-of-function mutations of genes associated with DNA-damage repair.
>
> The transcription factor Unc-86 (POU) domain class 2 transcription factor 1 (POU2F1) is a widely expressed bipotential stabilizer of repressed and inducible transcriptional states and frequently deregulated in cancer. By immunohistochemical staining and analysis of a tissue micro array of 217 operable stage I-III NSCLC patients, we revealed that high protein levels of POU2F1 in primary NSCLC tumor tissue improved outcome of smokers with ADC (Hazard ratio [HR] 0.30 [0.09 – 0.99], $p = 0.035$). We validated this finding, analyzing survival data of 1144 NSCLC patients from a publicly available gene expression dataset, separating POU2F1 expression levels at median. In this cohort, we confirmed favorable outcome of high POU2F1 expression in smokers with ADC (HR 0.41 [0.24 – 0.69], $p < 0.001$). Moreover, retrovirally induced overexpression of POU2F1 in A549 cells significantly reduced both, clonogenic growth as well as proliferation of NSCLC cells, whereas CRISPR-Cas 9 mediated knockdown of the protein did not have any impact.

Our data suggest that high expression of POU2F1 mediates a less aggressive cancer phenotype in smokers with ADC NSCLC. Pharmacological induction of genes and signaling pathways controlled by POU2F1 may provide novel avenues for future targeted NSCLC therapies in smokers."

Comment 3: Third, the introduction of the main text needs to review known prognostic factors and biomarkers in NSCLC, have comments on the limitations of previous studies and their knowledge gaps, have comments on the potential strengths of POU2F1 in comparison to other known biomarkers, and explain why it deserves to be studied.

Reply 3: We would also like to thank the reviewer very much for this highly important suggestion. The role of frequent genetic aberrations including changes of EGFR, ALK and KRAS in smokers as well as non-smokers with NSCLC is depicted in the revised version of our manuscript, now. Moreover, we elaborated the common loss-of-function mutations in genes involved in DNA-damage repair in smokers with NSCLC. These aberrations are currently not targetable. Thus, novel biomarkers and future therapeutic targets need to be identified, and we have further emphasized the potential future role of POU2F1 in this context, now.

"Lung cancer remains the leading cause of cancer-related mortality, and non-small cell lung cancer (NSCLC) accounts for approximately 85 % of all lung cancer patients1,2. Several studies have reported a negative impact of cigarette-smoking on prognosis of lung cancer patients3,4. Overall tumor mutational burden is manifold higher in NSCLC tumors of smokers compared to never-smokers3,5–8, and recently, the genomic landscape of NSCLC in smokers has been unveiled as so called "tobacco signature"5,9,10. In NSCLC of non-smokers, epidermal growth factor receptor (EGFR) mutations and tumor protein 53 (TP53) mutations are most common10 and in adenocarcinoma (ADC) of the lung targetable genetic changes are frequently found including EGFR mutations (i.e., 71.6 % in non-smokers vs. 40.4 % in smokers) and anaplastic lymphoma kinase (ALK) rearrangements (i.e., 7.0 % in non-smokers vs. 3.4 % in smokers)3. Moreover, non-smoking squamous cell carcinoma (SCC) patients may harbor an EGFR mutation in up to 8 % of the cases, but in smokers its prevalence is at only 2.5 %. While tobacco-induced genetic damages as predictor for targetable driver mutations has already been applied to treatment guidelines for NSCLC11, the identification of novel biomarkers for smokers with NSCLC remains challenging. Solely, activating mutations of Kirsten rat sarcoma virus (KRAS) are more prominent in smokers with ADC than in non-smokers (i.e., 14.0 % vs. 3.4 %, respectively)3 and can be targeted by sotorasib that has been approved for p.G12C-mutated KRAS-driven pre-treated NSCLC12. However, the clinical efficacy of KRAS-inhibitors falls short of expectations13. Other than KRAS, most of the smoking-associated mutations comprise loss-of-function of tumor-suppressive deoxyribonucleic acid (DNA)-damage repair genes, such as excision repair complementary complex 5 and 8 (ERCC5, ERCC8), DNA ligase 4 (LIG4), mutY DNA glycosylase (MUTYH), DNA polymerase beta (POLB), DNA repair protein REV1 as well as G/T mismatch-specific thymine DNA glycosylase (TDG)9. While some of these genetic aberrations might predict responsiveness to platinum-based chemotherapies14–17, none of them have been translated into targeted NSCLC therapies yet. Thus, discovery of novel targets for the treatment of smoking-related lung cancer is of primary importance.

The Pit-1, Oct1/2, Unc-86 (POU) domain class 2 transcription factor 1 (POU2F1) is a bipotential transcriptional regulator of induced and repressed states9. POU2F1 is widely expressed in healthy as well as in malignant tissues18 and a versatile acting transcription factor in cancer19,20. Whereas in other malignant entities21–23, including breast24,25, gastric26, cervical27, and prostate cancer28, an oncogenic role of POU2F1 has been described, its role in patients with NSCLC remains to be determined.

Here, we evaluated the prognostic role of POU2F1 in NSCLC, retrospectively analyzing protein and gene expression data of two large patient cohorts as well as its functional impact on human A549 lung cancer cell line."

Moreover, we added a small section at the end of the discussion (*c.f.*, p. 20 ll. 410), emphasizing a potential future therapeutic approach towards targeting of POU2F1:

"While direct pharmacological targeting of POU2F1 with the intention to increase its expression and/or activity is out of reach, our data suggest POU2F1-regulated signaling pathways and genes as promising novel targets for future therapeutic approaches in NSCLC."

Comment 4: Fourth, the methodology of the main text needs to clearly describe the clinical research design, subjects' inclusion, assessment of clinical factors, follow up procedures, and measurements of prognosis outcomes, including the sample used in the databases. In statistics, it is important to exclude the confounding effects of other clinical covariates to ascertain the independent prognostic role of POU2F1, so in the multiple statistical analysis, please describe the adjustment of clinical covariates. Please ensure $P<0.05$ is two-sided. Please further clarify why "Due to the explorative character of the analysis, an adjustment to multiplicity was not determined".

Reply 4: We thank the reviewer for the highly relevant input and addressed the issues raised. We have added the exclusion criteria and the gathered clinical information, that in detail are also depicted in table 1. As the analysis is of a retrospective character, information on follow-up can likewise be taken from table 1. Due to its explorative character, the protein-expression analysis was used to generate a hypothesis on POU2F1's role in smokers with adenocarcinoma of the lung. Hence, p-value adjustments to multiple testing were not performed and stated by the sentence on p. 10, ll. 238. We used a second gene-expression validation cohort, to confirm the previously generated hypothesis on POU2F1. Moreover, we performed in vitro experiments based on retroviral/ lentiviral induction of POU2F1 overexpression and knockdown in human A549 NSCLC cells to further address the influence of protein expression levels of this transcription factor in human lung adenocarcinoma. We also revised the first section 'study collectives' of the material and methods section.

"Study collectives

To generate a hypothesis on POU2F1's prognostic role in NSCLC, we retrospectively analyzed 311 patients, diagnosed, and surgically treated for non-small cell lung cancer (NSCLC) between December 1998 and November 2004 at the Thoracic Surgery department of the St. Georg's Clinic Ostercappeln. Study approval by the ethical committees of Osnabrück, Münster and Mainz were obtained for the collection of paraffin-embedded tissue samples for biomarker testing. Patients under 18 years of age, with an unknown smoking status, with stage IV disease, with an incomplete resection status (R1- or R2-resection) or with an indeterminate histological growth pattern (i.e., NSCLC not otherwise specified) were excluded from the primary analysis. Due to possible interaction with POU2F1 expression, exclusion of 94 neoadjuvant treated cases resulted in a final cohort size of 217 patients, see **figure 1A**. Gathered clinical data included age at NSCLC diagnosis, gender, Eastern Cooperative Oncology Group (ECOG) performance status at diagnosis, smoking status and – if attainable – packyears, neoadjuvant treatment regimen, adjuvant treatment regimen, date of surgery, histopathological diagnosis, overall tumor stage, pT-, pN- and pM-status, tumor grading, resection status (R0, R1, or R2), date and status of relapse, as well as date and status of death or last contact.

With regard to histopathology, 75 of 217 patients presented adenocarcinoma (ADC) growth pattern and 105 of 217 had squamous cell carcinoma (SCC) histology. The remaining 37 patients were diagnosed with large cell carcinoma (LCC). Staging was performed according to the sixth edition of Tumor Nodule Metastasis (TNM) system proposed by the "Union Internationale Contre le Cancer" (UICC)[29]. An update towards later TNM staging system editions was not possible due to ethical concerns regarding patient privacy and data protection. Moreover, complete tissue specimen were paraffinized and no fresh frozen tissues were

available. Hence, mutations in the EGFR-[30], and BRAF-[31], as well as translocations and fusions in ALK-[32], and ROS1-[33] gene loci could retrospectively not be assessed.

We used the online messenger ribonucleic acid meta-analysis database 'kmplot.com' for NSCLC as a validation cohort to evaluate the prognostic effect of POU2F1 gene expression (affymetrix Probe ID 227254_at) on the mRNA level in 672 ADC patients and 271 SCC patients, see **figure 1B**. Regarding prior history of smoking, subcohort analyses of the ADC cohort were performed for 140 non-smokers and 231 smokers[34]."

Additionally, we have clarified the section 'statistical analysis' in the material and methods section, now.

"Statistical analysis

To describe the cohort and cell culture analyses, we used mean, standard deviation (SD), median, interquartile range (IQR; Q1-Q3), 95 percent confidence interval (CI) as well as raw count and frequencies. Two-fold associations between categorical variables were analyzed via Fisher's exact test or Chi-square test, if applicable. Continuous and ordinal variables were tested using either unpaired t-test or Mann-Whitney-U test, depending on the normality of the data. If applicable, two-tailed p-values were recorded.

Overall survival (OS) included the time (days [protein expression cohort] or months [gene expression cohort]) between histopathological diagnosis and death or censoring. Progression free survival (PFS) was defined as time (days [protein expression cohort] or months [gene expression cohort]) between histopathological diagnosis and first relapse, progress after initial treatment, death, or censoring (last visit alive, lost to follow up), depending on the first chronological appearance. Univariate survival analyses compared OS and PFS between groups by using Log rank tests. Kaplan-Meier plots helped to visualize survival differences. We performed univariate Cox Regression analyses using the inclusion variable selection. Hazard Ratios (HR) are presented with 95 % confidence intervals (95 % CI). Due to a small sample size, a multivariate survival analysis was not performed in the protein expression cohort. Due to patient privacy, 'kmplot.com' extracted data did likewise not allow for multivariate survival testing. Data collection as well as calculations were performed using IBM® SPSS® Statistics Version 27 (released 2020, IBM Corp., Armonk, NY, USA). The Kaplan-Meier plotter online tool ('kmplot.com') automatically analyzes data by using the statistical software R with its underlying 'survplot' command of the survival Bioconductor package[34,43]. Data of the Kaplan-Meier plotter online tool ('kmplot.com') were transferred to IBM® SPSS® Statistics via text output of the underlying raw data. Survival analyses in IBM® SPSS® Statistics were thus performed, as mentioned before. The local significance level was set to 0.05. Due to the explorative character of the protein-expression analysis, an adjustment to multiplicity was not determined."