EXTENDED MATERIAL AND METHODS

Clinical trial details

Patient Characteristics. This open-label randomized phase II trial (NCT03367741), conducted through the National Cancer Institute Experimental Therapeutics Clinical Trials Network, assessed the activity of nivo combined with cabo (Arm A) versus nivo alone (Arm B). Patients who experience disease progression in Arm B were allowed to cross-over to combination therapy as part of exploratory Arm C. This exploratory arm also included carcinosarcoma patients as well as patients previously treated with IO therapies. Dosing, route, and drug schedule are indicated in **Fig. 1A**. This manuscript reports on the biomarker cohort which includes all with available biomarker data (n = 77). A total of 35 patients in Arm A, 17 in Arm B, and 22 in Arm C were included in this analysis. From previously published cohort, we exclude one patient in Arm B who withdrew consent and one patient in Arm C who did not have biomarker data available. One patient in Arm A as lost to follow-up, not evaluated for response, and therefore was not included in the survival analysis presented in this manuscript.

Main eligibility criteria included measurable disease according to Response Evaluation Criteria in Solid Tumors (RECIST; version 1.1), regardless of the histologic subtype, and radiologic progression after at least one line of previous platinum-based chemotherapy. There was no restriction on the number of prior treatment lines, but prior immunotherapy was not allowed for arms A and B. Eligible patients were randomized (2:1) to either combination or monotherapy and stratified according to microsatellite status assessed by genomic analysis, or MMR status defined from archival tissue according to local guidelines. An exploratory cohort (Arm C) included patients with carcinosarcoma histology and patients previously treated with IO agents, including, but not limited to, anti-PD-1, anti-programmed cell death ligand-1 (*PD-L1*), or anti-*TIM3* therapy). Patients on Arm B who progressed, were allowed to cross over to exploratory Arm C. For translational research, each patient had a baseline biopsy before starting treatment and blood samples (specifically peripheral blood mononuclear cells and plasma) were collected at baseline (prior to cycle 1 day 1 or C1), prior to cycle 2 day 1 (C2), cycle 4 day 1 (C4) and at the time of progression (P). A biopsy at the time of progression was mandated for patients who progressed on Arm B and chose to cross-over to Arm C. These patients were analyzed in Arm C from the time of crossover (**Fig. 1B**). The study protocol was compliant with Good Clinical Practice guidelines and the Declaration of Helsinki. Ethics approval was obtained in the USA and Canada, and for all participating centers. All patients provided written informed consent for research specimen collection and data sharing.

Biomarker statistical analysis plan

Statistical and data analyses

Quality controls. The analysis for all datasets (Olink, Serology, CyTOF, WES) was performed in python and python implementations of R statistical software using mixed-effects linear models to adjust for relevant clinical variables and demographics if not randomized. The data distributions for markers and cell populations for all assays were investigated as part of routine quality control to identify biases, and corrected as follows: (1) Olink analytes with zero variance were excluded; (2) Olink analytes with abundances below the limit of detection in more than >50% of samples were excluded; (3) CyTOF cell populations unassigned by Astrolabe were ignored. QC was performed using Principal Component Analysis and Principal Variation Component analysis using scanpy package in python.

Longitudinal data analysis. Longitudinal protein expression analysis was performed in python using packages Pymer (lme4 R implementation in python), The mixed-effect linear models included treatment arm, time, and its interaction as fixed effects as well as other relevant covariates needed for adjustment. The model also included a random intercept for each patient and a compound symmetry correlation structure was assumed. Models were fitted using REML and the Emmeans package was used to extract marginal mean estimates and test the hypothesis of interest, namely differences over time within treatment arms, tumor EC histology, prior IO status, tumor genomics or adverse event status, as well as different across the groups defined for those factors. Marginal mean and 95% confidence intervals were estimated using emmeans and visualized using the forestplot package. For Olink the response variables were individual protein levels (NPX, log2). For CyTOF the response variables were cell frequencies, the surface markers mean MFI values or subset frequencies defined automatically by Astrolabe. The results were visualized using seaborn and matplotlib packages.

Whole exome sequencing analysis. We used WES somatic mutation calling pipeline nf-core/sarek implemented in the nextflow pipeline management system (*25*). Identified somatic mutations by Mutect2 and Strelka2 were annotated using VEP and filtered for known tumor drivers in TP53 and POLE genes. For patients with matched tumor/normal samples we performed formal copy-number analysis using combined outcomes from ASCAT, ControlFreq, CNVkit, mutation signature quantification by projecting SBS/INDEL signatures on COSMIC database. MSI calling was done with MSIsensor2 implemented within the nf-core/sarek pipeline.

Survival Analysis:

Survival Analysis for the Secondary RCT Endpoint: Since OS was not presented in the parent trial manuscript, we are including a complete analysis for this important endpoint before assessing the biomarkers defining the treatment response. We had included the Kaplan Meier modelling censored at crossover, as presented in the clinical trial protocol, but included sensitivity analysis using both phases of the trial study. The analysis population includes all patients with available biomarker data (n=35 in Arm A, n=17 in Arm B), which differs from the parent trial only on 1 patient per arm.

Censored Survival Analysis. To obtain hazard ratios for OS, we analyzed patients as randomized. We performed "on-treatment" analysis in which survival data was censored at the time of crossover for patients in Arm B as per the clinical trial protocol. Both Kaplan-Meier (KM) and Cox proportional hazard models were performed with a single factor (treatment).

Inverse probability of censorship weighting (IPCW). To account for departure from randomization upon crossover, we used IPCW approach to balance fixed clinical covariates in the OS analysis, using R package "WeightIt" (*26-28*). Briefly, in IPCW approach, subjects who crossed over from Arm B to C were censored at the time of crossover, the crossover event (a binary response variable) was modeled as a logistic regression with patient clinical covariates (*cross ~ age + race + bsa + ecog*). Only subjects on Arm B were modeled, as participants randomized to Arm A were allowed to cross over. *Covariate balance for the* Average effect of Treatment on the Treated (ATT) *with method "weighting" was evaluated using R function "bal.tab".* Propensity scores were estimated via logistic regression balancing on *age*, *bsa* and *ecog* (*cross ~ age + bsa + ecog*) using the non-parametric covariate balancing Propensity scores (npcbps) with estimand ATT available in the weightit function. The resultant weights are

added to subjects from Arm B, while for Arm A weights are set to 1 and weights were subsequently applied to fit KM/COX models of OS.

Rank preserving structural failure time RPSFT model: was use to account for cross-over was also applied as a sensitivity analysis. The standard RPSFTM classifies the observed event time (Ti) for each patient into two components: TS, the time spent on the single agent nivolumab, and TC, the time spent on combination nivolumab with cabozantinib. For patients randomized to Arm A, TS is zero. For those in Arm B who did not cross-over, either because they did not progress or chose not to cross over, TC is zero. However, for patients who crossed over from Arm B to Arm C, both TSi and TCi are greater than zero. This model uses the relationship between Ti and the hypothetical event time (Ui) that would have been observed in the absence of cross-over t to explore causal effects within this framework (*29*). We estimate the parameter Ψ, the parameter which balances Ui across treatment arm, with e considered an acceleration factor. The outcome of interest was overall survival, as was calculated as time from first dose until death or censoring. Survival data was not recensored. This method relies on the assumption that the effect of treatment is equal across patients, i.e, the "common treatment effect assumption".

Association of biomarkers with OS survival and time-to-AE: Univariable and multivariable regression models were used to estimate the hazard ratios (HRs) and corresponding 95% confidence intervals (CIs) for time-to-event outcomes. The log-rank test was used to assess the significance of the difference between endpoints for overall survival (OS), Progression-free survival (PFS), and time to first adverse event in KM analysis. The univariable models were used to determine which covariates should be kept in the multivariable models. Significance was defined as adjusted p-values or False Discovery Rate (FDR) <0.05. Regression models were fit, survival analyses performed, and visualizations were created in a python environment using lifelines, matplotlib, seaborn and pandas packages. The constructed pipeline is available as a jupyter notebook upon request (vladimir.roudko@gmail.com)

Adjusting for multiple comparisons. In multi-omic assays (Olink and CyTOF) we applied moderate t-test statistics. We adjusted p-values using the Benjamini & Hochberg method (1995), controlling the false discovery rate, the expected proportion of false discoveries amongst the rejected hypotheses. Nevertheless, throughout the manuscript we show nominally significant results as $p<0.05$ and results significant after adjustment for multiple testing as FDR<0.05.

Correlation analyses. We used corr function prebuilt in the pandas package to compute Pearson (normal distribution assumptions) and Spearman (rank-based) correlations between analytes and endpoints were discussed and needed.

Data Sharing Statement. In accordance with NIH's Genomic Data Sharing Policy, the DNA sequencing data used to support the findings of this study will be deposited under controlled access in the database of Genotypes and Phenotypes (dbGaP) under accession number phs003414.v1. Genomic, clinical, mass cytometry, and protein analyte data from this study used to support this publication will be made available upon reasonable request from a qualified medical or scientific professional for the specific purpose laid out in that request and may include de-identified individual participant data. Requests for secondary use of this data will require completing a data use agreement (https://osp.od.nih.gov/wp-content/uploads/Model_DUC.pdf) and submitting a data access request to the NIH.

Code Sharing Statement. All code used for analysis is available upon request at vladimir.roudko@gmail.com