A 14-gene gemcitabine resistance gene signature is significantly associated with the prognosis of pancreatic cancer patients

Xing Wei1, Xiaochong Zhou1, Yun Zhao3,4,6, Yang He2,4,5, Zhen Weng2,3,4,6* & Chunfang Xu1,

To identify a gemcitabine resistance-associated gene signature for risk stratification and prognosis prediction in pancreatic cancer. Pearson correlation analysis was performed with gemcitabine half maximal inhibitory concentration (IC50) data of 17 primary pancreatic cancer lines from Genomics of Drug Sensitivity in Cancer (GDSC) and the transcriptomic data from GDSC and Broad Institute Cancer Cell Line Encyclopedia, followed by risk stratification, expression evaluation, overall survival (OS) prediction, clinical data validation and nomogram establishment. Our biomarker discovery effort identified a 14-gene signature, most of which featured differential expression. The 14-gene signature was associated with poor OS in E-MTAB-6134 (HR 2.37; 95% CI 1.75–3.2; \( p < 0.0001 \)), pancreatic cancer-Canada (PACA-CA) (HR 1.76; 95% CI 1.31–2.37; \( p = 0.00015 \)), and 4 other independent validation cohorts: pancreatic cancer-Australia (PACA-AU) (HR 1.9; 95% CI 1.38–2.61; \( p < 0.0001 \)), The Cancer Genome Atlas (TCGA) (HR 1.73; 95% CI 1.11–2.69; \( p = 0.014 \)), GSE85916 (HR 1.97; 95% CI 1.14–3.42; \( p = 0.014 \)) and GSE62452 (HR 1.82; 95% CI 1.02–3.24; \( p = 0.039 \)). Multivariate analysis revealed that the 14-gene risk score was an independent pancreatic cancer outcome predictor in E-MTAB-6134 \( ( p < 0.001) \) and TCGA \( ( p = 0.006) \). A nomogram including the 14-gene was established for eventual clinical translation. We identified a novel gemcitabine resistance gene signature for risk stratification and robust categorization of pancreatic cancer patients with poor prognosis.

Pancreatic cancer is one of the most dangerous malignancies worldwide, with a 5-year overall survival of less than 5% and an approximately 6–8-month median survival after diagnosis1,2. Surgical resection followed by adjuvant treatment with the nucleoside analog drug gemcitabine is the standard management of pancreatic cancer patients in clinical practice3. Moreover, nab-paclitaxel plus gemcitabine (Nab-Gem) and FOLFIRINOX (folinic acid, 5-fluorouracil, irinotecan and oxaliplatin) represent the standard regimens for the first-line treatment of locally advanced and metastatic pancreatic cancer4,5. However, the majority of patients treated with gemcitabine chemotherapy eventually develop gemcitabine resistance6. Obviously, the effectiveness of gemcitabine chemotherapy is related to the prognosis of pancreatic cancer and therefore could be employed as a baseline for treatment and prognosis improvement of this fatal disease.

Many studies have been carried out to elucidate the possible mechanisms involved in gemcitabine resistance7–9, which are considered to be related to transport and metabolism behavior and are thought to involve multiple enzymes and signaling pathways, including human concentrative nucleoside transporters (hCNTs), human equilibrative nucleoside transporters (hENTs), deoxycytidine kinase (dCK), and ribonucleotide reductase (RR), or integrated activity of the detoxifying enzyme cytidine deaminase (CDA), and PI3K/Akt, MAPK and NF-κB pathways. Based on these findings, changes in multiple genes at the functional and expression levels could be a reasonable phenomenon during the development of gemcitabine resistance. Recent advances in microarray gene chip technology and high-throughput sequencing have resulted in protocols for evaluating the expression level changes of multiple genes at the same time. Moreover, the generation of large public databases with
Results
Identification of the gemcitabine resistance- and survival-related gene signature, establishment of the prognostic signature and evaluation of the expression of the identified genes. A total of 1208 (550 positively correlated and 658 negatively correlated genes) and 1983 (1294 positively correlated and 689 negatively correlated genes) gemcitabine resistance-related genes were identified from the E-MTAB-3610 and CCLE datasets, respectively (Fig. 1A). The intersection of these two datasets revealed 509 genes (292 positively correlated and 217 negatively correlated). After combination analysis of the clinical data from E-MTAB-6134 and PACA-CA, 92 and 70 survival-related genes were obtained, respectively, and 22 intersecting genes were found (Fig. 1B). Finally, the efficacy of a 14-gene signature was calculated by Lasso Cox regression (Fig. 1C,D), and the detailed gene screening data of each step are shown in the Supplementary file.

These genes included CCDC148, SH3RF2, CACNA1D, POLD3, PARP1, AP1M2, C4orf19, ANO1, VGLL1, SCEL, INPP4B, NET1, INSIG2 and BVES, and 6 of these genes were reported by previous studies (detailed information is provided in the Supplementary file; Table 1). Next, we evaluated the expression profiles of these 14 genes in 4 different Gene Expression Omnibus (GEO) datasets and out of the 14, the numbers of differentially expressed genes in the GSE140077 (gemcitabine-resistant cells; FPKM: Fragments Per Kilobase of transcript per Million mapped reads; (E), GSE62165 (PDAC and normal tissues; (F), GSE91035 (PDAC, benign and normal tissues; (G) and GSE19650 (multiple different pre-pancreatic carcinoma stage tissues; IPMA: intraductal papillary-mucinous adenoma; IPMC: intraductal papillary-mucinous carcinoma; IPMN: intraductal papillary-mucinous neoplasm; (H)).

Identification of a 14-gene risk stratification signature for the prognosis of patients with pancreatic cancer and exploration of the expression pattern. Subsequently, we used multiple datasets to validate the 14-gene signature. The results of Kaplan–Meier curve uncovered significantly favorable overall survival in patients with a low risk score from the E-MTAB-6134 (HR 2.37; 95% CI 1.75–3.24; p = 0.00015), PACA-CA (HR 1.76; 95% CI 1.31–2.37; p = 0.00015), PACA-AU (HR 1.9; 95% CI 1.38–2.61; p < 0.0001), TCGA (HR 1.73; 95% CI 1.11–2.69; p = 0.014), GSE85916 (HR 1.97; 95% CI 1.14–3.42; p = 0.014) and GSE62452 (HR 1.82; 95% CI 1.02–3.24; p = 0.039) (Fig. 2A,D,G,J,M,P) datasets but not in the GSE71729 dataset (HR 0.98; 95% CI 0.63–1.51; Supplementary Fig. 1). Moreover, the prediction efficacy of the 14-gene signature was also evaluated by ROC curves; in most of the datasets, the AUCs for 1-, 3- and 5-year OS was larger than 0.6 and close to 0.7. The highest AUC was found in the E-MTAB-6134 set, and the AUCs for 1-, 3- and 5-year OS were 0.757, 0.716 and 0.717, respectively, while the AUCs for 3- and 5-year OS were relatively low in the TCGA set (0.560 and 0.458, respectively) (Fig. 2B,E,H,K,N,Q and Supplementary Fig. 1). Detailed information on the risk score and survival status can be found in Supplementary Fig. 2. In addition, we evaluated the expression pattern of these 14 genes in the above datasets, and the heatmap results showed a differential expression of these genes in patients with high and low risk scores (Fig. 2C,F,I,L,O,R and Supplementary Fig. 1). These results suggested...
Identification of the 14-gene signature as an independent risk and survival outcome predictor in pancreatic cancer patients. We then performed univariate and multivariate Cox regression analyses on this gene panel by including various clinicopathological features of pancreatic cancer patients in 3 datasets with available clinical information. In the univariate analysis, the 14-gene risk score was identified as a significant predictor of overall survival in all 5 datasets (E-MTAB-6134: \( p < 0.001 \); TCGA: \( p = 0.015 \); GSE62452: \( p = 0.042 \); PACA-AU: \( p < 0.001 \); PACA-CA: \( p < 0.001 \)) (Fig. 3A–C and Supplementary Fig. 3). Further multivariate analysis revealed that the 14-gene risk score was an independent predictor in pancreatic cancer patients from the E-MTAB-6134 (\( p < 0.001 \)) and TCGA (\( p = 0.006 \)) datasets but not the GSE62452 dataset (\( p = 0.241 \)) (Fig. 3A–C). Since there were relatively large numbers of patients in the E-MTAB-6134 and TCGA datasets, our results verified that our 14-gene signature is quite robust and can be used as an independent factor for survival prediction in pancreatic cancer patients.

Identification of the 14-gene signature as an independent risk and survival outcome predictor in pancreatic cancer patients. We then performed univariate and multivariate Cox regression analyses on this gene panel by including various clinicopathological features of pancreatic cancer patients in 3 datasets with available clinical information. In the univariate analysis, the 14-gene risk score was identified as a significant predictor of overall survival in all 5 datasets (E-MTAB-6134: \( p < 0.001 \); TCGA: \( p = 0.015 \); GSE62452: \( p = 0.042 \); PACA-AU: \( p < 0.001 \); PACA-CA: \( p < 0.001 \)) (Fig. 3A–C and Supplementary Fig. 3). Further multivariate analysis revealed that the 14-gene risk score was an independent predictor in pancreatic cancer patients from the E-MTAB-6134 (\( p < 0.001 \)) and TCGA (\( p = 0.006 \)) datasets but not the GSE62452 dataset (\( p = 0.241 \)) (Fig. 3A–C). Since there were relatively large numbers of patients in the E-MTAB-6134 and TCGA datasets, our results verified that our 14-gene signature is quite robust and can be used as an independent factor for survival prediction in pancreatic cancer patients.

The 14-gene signature robustly identifies poor molecular subtypes in pancreatic cancer. Moreover, recently discovered transcriptomic data-based molecular subtypes, including the squamous subtype defined by Bailey\(^8\), the quasimesenchymal (QM) subtype by Collisson\(^9\), the basal-like subtype by Moffitt\(^10\), and the pure basal-like/stroma activated subtype by Puleo\(^11\), are predicted to have poor survival outcomes among PDAC patients. Due to the incomplete clinical information in some of the datasets, we only verified the effectiveness of our 14-gene signature for molecular subtype prediction in the E-MTAB-6134 and GSE71729 sets, and an AUC value between 0.8 and 0.96 could be achieved by using our signature (Fig. 4).

Establishment of a risk nomogram for survival prediction in pancreatic cancer patients. To further strengthen the accuracy of the predictive power of our 14-gene signature, we established a ready-to-use and clinically applicable risk nomogram for survival prediction in pancreatic cancer patients from E-MTAB-6134 to assess the performance of this signature via a combination of other univariate significant patient characteristic parameters (e.g., grade and N-stage). As shown in Fig. 5A, a higher total score calculated by adding the assigned numbers of each factor included in the nomogram was associated with worse 3-year and 5-year overall survival rates. For example, a patient with a higher grade and a higher 14-gene score could generate a total of 168 points
Figure 2. Identification of 14-gene risk stratification signature for prognosis of patients with pancreatic cancer and exploration of the expression pattern. Kaplan–Meier curve, receiver operator characteristic curves for 1, 3 and 5-year overall survival prediction and 14 genes expression heatmaps of different datasets were shown. (A–C) E-MTAB-6134; (D–F) PACA-CA; (G–I) PACA-AU; (J–L) TCGA; (M–O) GSE85916; (P–R) GSE62452.
Figure 3. Performance of 14-gene signature as a predictor of risk and survival outcomes in pancreatic cancer patients by univariate and multivariate Cox analysis including clinicopathological features. (A) E-MTAB-3164; (B) TCGA; (C) GSE62452.
(68 points for high grade and 100 points 14-gene high), with predicted 1-year, 3-year and 5-year overall survival rates of 73.0%, 25.0% and 13.0%, respectively. To validate the prediction model, the performance of the nomogram was evaluated by the discrimination index calculation and calibration plot adjusting for the 1-year, 3-year and 5-year survival. The accuracy of the predictive nomogram was higher than that of the 14-gene signature alone, with an accompanying C-statistic discriminatory index value of 0.67. This higher C-statistic revealed that the 14-gene signature in combination with sex, tumor grade, resection margin and N-stage was considerably robust in discriminating subjects with variable outcomes. The 284-patient bootstrapped calibration plots for the prediction of 1-year, 3-year and 5-year overall survival rates are shown in Fig. 5B–D. The calibration plots of the 14-gene signature with sex, grade, resection margin and N stage exhibited excellent consistency between the observed outcomes and predicted survival. These results further support the clinical significance of our 14-gene signature in combination with sex, grade, resection margin and N stage, which has superior overall predictive power for distinguishing survival outcomes in pancreatic cancer patients.

Discussion
In addition to the application of conventional clinicopathological parameters or AJCC staging for prognosticating disease progression, molecular prognostic markers with the features of easy detection and simple quantification by standard methods could be employed not only for early diagnosis and prognosis prediction but also for novel therapeutic target discovery. Moreover, the combination of a panel of molecular biomarkers could overcome the hurdles resulting from tumor heterogeneity. In order to strengthen the survival and prognosis of pancreatic cancer patients, elucidate the mechanisms involved in gemcitabine resistance and discover novel diagnostic, prognostic and therapeutic-related biomarkers, we performed a preset study focusing on the relationship of gemcitabine resistance and overall survival. A novel 14-gene signature comprising CCDC148, SH3RF2, CACNA1D, POLD3, PARP1, AP1M2, C4orf19, ANO1, VGLL1, SCEL, INPP4B, NET1, INSIIG2 and BVES and a corresponding risk score formula were established for pancreatic cancer risk stratification and prognosis prediction. Patients in the low-risk group had significantly better prognosis than those in the high-risk group. In addition, we confirmed the effectiveness of the 14-gene signature in predicting 1-year, 3-year and 5-year overall survival via AUC calculation.

As a widely employed prognosis evaluation modality in clinical oncology, nomograms can integrate different prognostic determinants, including molecular and clinicopathological characteristics. The different probabilities of clinical events can be visualized and calculated into a total score for risk illustration. Compared to conventional staging methods, nomograms may more effectively strengthen clinical decision-making regarding prognosis prediction and thereby be beneficial for implementing timely treatment. The C-index and calibration curves identified that the 14-gene signature was effective for predicting 1-year, 3-year and 5-year overall survival. An integrated nomogram using the 14-gene signature and clinicopathological characteristics was set up for accurate overall survival prediction. As a supplement to AJCC staging and conventional clinicopathological characteristics, the 14-gene signature and nomogram could be employed as useful high-risk indicators and overall survival predictors.

With the advances in gene sequencing technologies, several groups have reported their own mRNA-level prognostic gene signatures that have the ability to predict overall survival in pancreatic cancer. Birnbaum et al. suggested a 25-gene signature based on clinicopathological parameters and gene expression data that predicts postoperative OS independent of classical factors and molecular subtypes. Raman et al. reported a 5-gene prognostic model comprising ADM, ASPM, DCBLD2, E2F7, and KRT6A for accurate prediction of overall survival using the PAAD datasets from the TCGA. Yan et al. confirmed that a 4-gene signature (CDC6, IGF2BP2,
KNTC1, and LYRM) was significantly related to the progression and prognosis of pancreatic cancer. Most recently, Wu et al. identified a nine-gene signature comprising ANKRD22, ARNTL2, CEP55, COL17A1, ITGB6, MCOLN3, MET, KLK10, and SLC25A45 for predicting overall survival of pancreatic cancer. In addition, Kandimalla et al. identified a 15-gene immune, stromal and proliferation (ISP) gene signature comprising TNFRSF4, TNFSF10, RFC4, PVRL3, PLD4, KDM6B, INHBA, IL32, IL4, IFIT3, FOXP3, CDC20, CD160, and AHR with favorable prediction of poor OS and identification of the molecular subtype of PDAC. Considering that recent progress in combination therapies includes various chemotherapeutic regimens, especially the frequent use of gemcitabine in neoadjuvant and palliative therapies of pancreatic cancer patients, genes related to gemcitabine resistance could be a group of strong candidate genes in the identification of robust prognostic and predictive risk-stratification biomarkers for pancreatic cancer patients. Moreover, plenty of the available clinical and gene expression information from previous studies (including International Cancer Genome Consortium [ICGC], TCGA and GEO) could be used to validate the predictive power of these genes.

Among our panel of 14 genes included in the signature, 6 genes were previously studied in pancreatic cancers. PARP-1 plays critical roles in DNA damage repair during intrinsic cell death, and cytoplasmic PARP-1 was recently confirmed to promote pancreatic cancer tumorigenesis and resistance. ANO1 located on amplicon 11q13 is usually amplified in human cancers with poor prognosis and is pivotal in PDAC cell migration. SCEL is a potential biomarker for the prognosis of pancreatic cancer confirmed by Cheng et al. INPP4B, which was first described by Zhai et al., is an oncogenic gene in pancreatic cancer and could serve as a potential diagnostic marker and an independent prognostic marker, suggesting that it is a novel therapeutic target for pancreatic cancer. INSIG2 overexpression is related to the malignant phenotype of pancreatic cancer under hypoxic conditions. BVES, as a novel regulator of the Rac1 and Cdc42 signaling cascades, controls cell shape and movement in multiple cancers, including pancreatic cancer. In addition, CACNA1D was found to be
expressed on pancreatic islets and play a vital role in insulin secretion, which is involved in β-cell physiology and pathophysiology\(^3\), and APIM2 expression was found to be upregulated only in pancreatic cancer\(^5\). Other genes, such as C3CDG148\(^8\), SH3RF2\(^9\), VGLL1\(^10\) and NET1\(^11\), were found to regulate different cell behaviors (such as cell proliferation and migration, cell death and cell cycle) involved in the development and progression of gastric cancer, whereas POLD3 was reported to be required for cell cycle progression and DNA synthesis and to be responsible for the high frequency of genomic duplications in human cancers\(^12\). No study was found on role of C4orf19 in cancers. Since most of the genes were previously reported to be involved in cancer biology, the findings of our panel are believable and could be useful. Moreover, since the 14 genes were selected using the Lasso regression method, the inconsistency in the expression patterns of the 14 genes at different levels (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected. In addition, we did not observe statistically significant correlations between the gemcitabine IC\(_{50}\) data and the mRNA expression (cell lines, normal vs. cancer, and specimens of different disease stages) could be neglected.
and an optimal 14-gene combination was obtained using Ln (λ.min). Risk scores were divided into high- and low-risk groups based on the median risk score. The flowchart of the data processing procedure is shown in Fig. 6.

**Statistical analysis.** Statistical analysis was carried out using R version 3.6.0. Student’s t-test was used for between-group comparisons, while one-way ANOVA was used for multiple-group comparisons. Kaplan–Meier (KM) analysis was carried out to determine survival outcomes. The median values were used as cutoff thresholds to plot the KM curves, and the statistical significance was evaluated by the log rank test (survival and survminer R packages). The survival probability prediction for 1-, 3- and 5-year survival was calculated by receiver operator characteristic (ROC) curves and the area under the curve (AUC) (survivalROC and ROCR R packages). Univariate and multivariate Cox regression analyses with all the available clinical information were performed for survival predictive nomogram construction and validation (forestplot and rms R packages). The hazard ratio (HR) and 95% confidence interval (CI) were employed to confirm genes associated with overall survival. Unless specifically mentioned, \( P < 0.05 \) was considered statistically significant.

### Table 3. Basic information of patients from the included datasets.
*The downloaded data from PACA-AU on T stage, N stage and AJCC stage was not complete and the number represents the available data. NA Not Available.*

|                  | E-MTAB-6134 (n = 288) | PACA-CA (n = 214) | PACA-AU* (n = 266) | TCGA (n = 148) | GSE85916 (n = 79) | GSE62452 (n = 65) |
|------------------|-----------------------|-------------------|--------------------|---------------|------------------|-------------------|
| **Sex**          |                       |                   |                    |               |                  |                   |
| Male             | 166                   | 128               | 142                | 79            | NA               | NA                |
| Female           | 122                   | 86                | 124                | 69            | NA               | NA                |
| Age (mean ± SD, years) | NA   | 65.2 ± 11.0      | 66.6 ± 10.6        | 65.0 ± 10.8   | NA               | NA                |
| **Status**       |                       |                   |                    |               |                  |                   |
| Alive            | 107                   | 34                | 106                | 67            | 22               | 16                |
| Dead             | 181                   | 180               | 160                | 81            | 57               | 49                |
| Follow-up (mean ± SD, Months) | 29.8 ± 26.9 | 22.6 ± 20.8  | 18.3 ± 13.7        | 17.5 ± 13.9   | 28.0 ± 30.5      | 20.2 ± 16.7       |
| **T-stage**      |                       |                   |                    |               |                  |                   |
| T1               | 12                    | NA                | NA                 | 5             | NA               | NA                |
| T2               | 39                    | NA                | 3                  | 16            | NA               | NA                |
| T3               | 237                   | NA                | 4                  | 123           | NA               | NA                |
| T4               | 0                     | NA                | NA                 | 3             | NA               | NA                |
| **N-stage**      |                       |                   |                    |               |                  |                   |
| N0               | 72                    | NA                | 5                  | 39            | NA               | NA                |
| N1               | 216                   | NA                | 1                  | 108           | NA               | NA                |
| Nx               | 0                     | NA                | 6                  | 1             | NA               | NA                |
| **M-stage**      |                       |                   |                    |               |                  |                   |
| M0               | NA                    | NA                | NA                 | 67            | NA               | NA                |
| M1               | NA                    | NA                | NA                 | 4             | NA               | NA                |
| Mx               | NA                    | NA                | NA                 | 77            | NA               | NA                |
| **AJCC stage**   |                       |                   |                    |               |                  |                   |
| I                | NA                    | 55                | 1                  | 12            | NA               | 4                 |
| II               | NA                    | 89                | 2                  | 127           | NA               | 44                |
| III              | NA                    | 8                 | NA                 | 3             | NA               | 10                |
| IV               | NA                    | 0                 | NA                 | 4             | NA               | 6                 |
| Uncertain        | NA                    | 62                | NA                 | 2             | NA               | 1                 |
| **Grade**        |                       |                   |                    |               |                  |                   |
| G1               | 110                   | NA                | NA                 | 20            | NA               | 2                 |
| G2               | 130                   | NA                | NA                 | 84            | NA               | 32                |
| G3               | 48                    | NA                | NA                 | 43            | NA               | 29                |
| G4               | 0                     | NA                | NA                 | 1             | NA               | 1                 |
| Gx               | 0                     | NA                | NA                 | 0             | NA               | 1                 |
Figure 6. Flowchart of the data processing procedure.

Received: 9 October 2020; Accepted: 3 March 2021
Published online: 17 March 2021

References
1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. CA Cancer J. Clin. 70, 7–30. https://doi.org/10.3322/caac.21590 (2020).
2. Jia, X. et al. Pancreatic cancer mortality in china: characteristics and prediction. Pancreas 47, 233–237. https://doi.org/10.1097/MPA.0000000000000976 (2018).
3. Jones, R. P. et al. Patterns of Recurrence after resection of pancreatic ductal adenocarcinoma: a secondary analysis of the ESPAC-4 randomized adjuvant chemotherapy trial. JAMA Surg. 154, 1038–1048. https://doi.org/10.1001/jamasurg.2019.3337 (2019).
4. Von Hoff, D. D. et al. Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine. N. Engl. J. Med. 369, 1691–1703. https://doi.org/10.1056/NEJMoa1304369 (2013).
5. Kim, S. et al. Comparative effectiveness of nab-paclitaxel plus gemcitabine vs FOLFIRINOX in metastatic pancreatic cancer: a retrospective nationwide chart review in the United States. Adv Ther. 35, 1564–1577. https://doi.org/10.1007/s12325-018-0784-z (2018).
6. Karasic, T. B. et al. Effect of gemcitabine and nab-paclitaxel with or without hydroxychloroquine on patients with advanced pancreatic cancer: a phase 2 randomized clinical trial. JAMA Oncol. 5, 993–998. https://doi.org/10.1001/jamaoncol.2019.0684 (2019).
7. Amrutkar, M. & Gladhaug, I. P. Pancreatic cancer chemoresistance to gemcitabine. Cancers https://doi.org/10.3390/cancers9110157 (2017).
8. de Sousa Cavalcante, L. & Monteiro, G. Gemcitabine: metabolism and molecular mechanisms of action, sensitivity and chemoresistance in pancreatic cancer. Eur. J. Pharmacol. 741, 8–16. https://doi.org/10.1016/j.ejphar.2014.07.041 (2014).
Vol.: (1234567890)

The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-85680-x.

The authors declare no competing interests.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-85680-x.
