Meta-Analysis of 1,200 Transcriptomic Profiles Identifies a Prognostic Model for Pancreatic Ductal Adenocarcinoma

Vandana Sandhu, PhD1,2; Knut Jorgen Labori, MD, PhD1; Ayelet Borgida1; Ilinca Lungu2; John Bartlett, PhD3; Sara Hafezi-Bakhtiar, MD4; Robert E. Denroche2; Gun Ho Jang, PhD5; Danielle Pasternack6; Faridah Mbaabali2; Matthew Watson5; Julie Wilson, PhD1; Elin H. Kure, PhD2,6; Steven Gallinger, MD1,5; and Benjamin Haibe-Kains, PhD1,5,7

POURPOSE With a dismal 8% median 5-year overall survival, pancreatic ductal adenocarcinoma (PDAC) is a highly lethal malignancy. Only 10% to 20% of patients are eligible for surgery, and more than 50% of these patients will die within 1 year of surgery. Building a molecular predictor of early death would enable the selection of patients with PDAC who are at high risk.

MATERIALS AND METHODS We developed the Pancreatic Cancer Overall Survival Predictor (PCOSP), a prognostic model built from a unique set of 89 PDAC tumors in which gene expression was profiled using both microarray and sequencing platforms. We used a meta-analysis framework that was based on the binary gene pair method to create gene expression barcodes that were robust to biases arising from heterogeneous profiling platforms and batch effects. Leveraging the largest compendium of PDAC transcriptomic data sets to date, we show that PCOSP is a robust single-sample predictor of early death—1 year or less—after surgery in a subset of 823 samples with available transcriptomics and survival data.

RESULTS The PCOSP model was strongly and significantly prognostic, with a meta-estimate of the area under the receiver operating curve of 0.70 (P = 2.6E–22) and D-index (robust hazard ratio) of 1.9 (range, 1.6 to 2.3; ( = 1.4E–04) for binary and survival predictions, respectively. The prognostic value of PCOSP was independent of clinicopathologic parameters and molecular subtypes. Over-representation analysis of the PCOSP 2,619 gene pairs—1,070 unique genes—unveiled pathways associated with Hedgehog signaling, epithelial–mesenchymal transition, and extracellular matrix signaling.

CONCLUSION PCOSP could improve treatment decisions by identifying patients who will not benefit from standard surgery/chemotherapy but who may benefit from a more aggressive treatment approach or enrollment in a clinical trial.

Clin Cancer Inform. © 2019 by American Society of Clinical Oncology

INTRODUCTION

Pancreatic ductal adenocarcinoma (PDAC) is a highly lethal malignancy with a 5-year overall survival rate of less than 8%.1 Disease in a majority of patients—more than 80%—is inoperable as a result of locally advanced or metastatic disease at the time of diagnosis. Completion of multimodality treatment—surgery combined with adjuvant or neoadjuvant chemotherapy—is the standard of care for treatment of PDAC. However, even after surgical resection with curative intent, median survival does not exceed 28 months and one half of those patients who undergo surgery develop recurrent disease and die within 1 year after surgery.2–4 Therefore, there is a need for a robust prognostic model to identify patients with a high risk of early death on the basis of molecular profiles of their tumors. Such a prognostic model could assist clinicians in identifying patients who may not benefit from surgery and standard adjuvant chemotherapy but who may benefit from a more aggressive approach or enrollment in a clinical trial.

Various clinical factors are prognostic after PDAC surgery, such as lymph node metastasis status,3 tumor grade,6 margins,7 degree of differentiation,8 and protein biomarker CA-19-9.9 However, the prognostic value of these clinical variables are insufficient to accurately stratify patients on the basis of risk of disease recurrence.10,11 With the advent of high-throughput next-generation molecular profiling technologies, multiple studies have released transcriptomic profiles of PDAC to the public domain. These gene expression profiles have been leveraged to identify molecular subtypes of PDACs.12–16 Whereas overlap between these subtypes15 supports the biologic relevance of these published classification schemes,15 they have not been designed to optimize prognostic value.
We generated a compendium of 17 PDAC data sets, including 1,236 gene expression profiles and 823 patients with survival data, as a resource for future PDAC analyses. We built the Pancreatic Cancer Overall Survival Predictor (PCOSP), a single-sample prognostic model robust to heterogeneous gene expression profiling platform and normalization methods for identifying patients with PDAC who are at high risk of early death.

Relevance
Endoscopic ultrasound biopsies could be used before curative surgery to estimate the prognosis of patients with PDAC using PCOSP to assist clinicians in predicting high-risk patients and making treatment decisions for this population.

Previously published prognostic models were developed from a small number of samples that lacked proper validation in multiple data sets. Attempts have been made recently to build a prognostic gene signature using pooled samples from multiple cohorts to identify patients who are at high risk of short-term survival postsurgery. However, they used samples that were profiled using either an array- or sequencing-based method as the learning cohort; therefore, the classifiers may perform better for patients whose samples were profiled using only one of the two platforms.

To address these issues, we took advantage of a unique set of 89 PDACs that were profiled using both microarray and sequencing technologies to develop the Pancreatic Cancer Overall Survival Predictor (PCOSP) model. Using an independent set of PDAC transcriptomic profiles from 823 primary resected patients, we show that PCOSP is a robust single-sample predictor of early death—1 year or less—after surgery that could be used as a potential tool to assist clinicians in decision making.

MATERIALS AND METHODS
The meta-analysis pipeline used to develop the PCOSP model and evaluate its prognostic value is provided in Figure 1.

Data Sets
We surveyed the literature and curated 17 data sets, which included 1,236 patients with PDAC, from the public domain for which transcriptome data of PDAC were available (Data Supplement). We further filtered samples on the basis of the availability of overall survival (OS) and sample size (n = 10 or greater) data after dichotomization into high- and low-survival groups on the basis of an OS cutoff of 1 year (Fig 2). The different cohorts had similar clinical presentation and were treated with curative surgery followed by adjuvant chemotherapy (Data Supplement).

Prognostic Model
To develop a robust predictor for early death, we used gene expression profiles of 89 samples from patients with PDAC whose tumors had been profiled using both microarray and sequencing platforms within the International Cancer Genome Consortium (ICGC) cohort. Human research ethical approval was granted as previously published. Approxi-mately one half of patients of the training cohort who were eligible for surgery experienced relapse within 1 year; we used this threshold to predict which patients with PDAC had high risk of early death—1 year or less—postsurgery. We excluded seven samples from the training cohort as these patients were censored before 1 year of follow-up.

To make gene expression profiles comparable between the training and validation sets, we transformed the original gene expression profiles into binary gene pair barcodes. We implemented k-Top scoring disjoint pairs classifier predictor using the Wilcoxon rank sum method as a filtering function in the SwitchBox package (version 1.12.0) (Data Supplement). To assess whether the prognostic value of the PCOSP model could be achieved by random chance alone, we tested two permutation tests (Data Supplement).

Early Death Prediction
Meta-analysis was performed for the PDAC sequencing cohorts, PDAC array-based cohorts, and overall combined cohorts to assess and statistically compare the performance of the PCOSP. Patient samples were dichotomized into two groups on the basis of the outcome variable—time from surgery to death of 1 year or less. Samples censored before 1 year of follow-up were excluded from the analysis of the meta-estimate of the area under the receiver operating characteristics curve (AUROC). AUROC plots the sensitivity versus 1-Specificity and is used as a criterion with which to measure the discriminatory ability of the model. AUROC was computed using pROC package (version 1.10.0), and we estimated the P value using the Mann-Whitney test statistics that estimated whether the AUROC curve...

Figure 1. The meta-analysis pipeline used to develop the PCOSP model and evaluate its prognostic value is provided in (Data Supplement).
estimate is significantly different from 0.5 (random classifier). The meta-estimate of AUROC was estimated using the random effect model implemented in survcomp package (version 1.26.0).29,30

Survival Prediction

Prognostic value and statistical significance of survival difference between the predicted classes were assessed using the $\alpha$-Index, which is a robust estimate of the traditional Cox hazard ratio (HR). The main advantage of $\alpha$-index compared with HR is a result of the fact that it is a robust and interpretable scale-free measure of separation between two independent survival distributions under the proportional hazards assumption.31 This makes $\alpha$-index a suitable estimate of prognostic value in a meta-analysis setting in which the heterogeneity of different cohorts must be accounted for. In addition, we used the concordance index (C-index), which estimates the probability that, for a random pair of patients, the PCOSP score for the patient with shorter survival is higher than that of the patient with longer survival.32 Both the $\alpha$-index and C-index were calculated using the survcomp package. We calculated the meta-estimate of the $\alpha$-index and C-index for the PDAC sequencing cohorts, the PDAC array-based cohorts, and the combined PDAC sequencing and array-based cohorts using the random effect model implemented in the survcomp package. Patients were stratified into low- and high-risk groups using median PCOSP score as a threshold. Kaplan-Meier curves

FIG 1. Pipeline showing the approach used for building the Pancreatic Cancer Overall Survival Predictor. ICGC, International Cancer Genome Consortium; k-TSP, k-Top scoring disjoint pair; OS, overall survival; OUH, Oslo University Hospital; PCSI, Pancreatic Cancer Sequencing Initiative; TCGA, The Cancer Genome Atlas; UNC, University of North Carolina.
Comparison With Existing Classifiers

We calculated the Birnbaum signature scores$^{22}$ and Chen signature scores$^{23}$ using the published coefficients of the 25 and 15 classifier genes, respectively, as weight parameters in the sig.score function implemented in the genefu R package (version 2.10.0).$^{35}$ We used Haider signature scores courtesy of the author.$^{24}$ We computed the C-index and d-index for the three classifiers using eight validation cohorts, excluding the cohorts used for training by PCOSP and other classifiers in comparison. Furthermore, we compared meta-estimates of the C-index of each classifier with PCOSP at $P < .05$ (one-sided $t$ test) as implemented in the survcomp package.

RESULTS

OS Predictive Model

To predict patients with early death (1 year or less after surgery), we trained the PCOSP model on the 89 ICGC cohort samples that were profiled using both microarray and sequencing transcriptomic profiles. We tested the prognostic value of the PCOSP score in three independent sequencing cohorts, including the PCSI,$^{36}$ TCGA-Pancreatic Adenocarcinoma (PAAD),$^{15}$ and Kirby,$^{37}$ cohorts, and seven independent array-based cohorts composed of ICGC-array (excluding the 89 samples used for training),$^{38}$ University of North Carolina (UNC),$^{13}$ OUH,$^{39}$ Chen,$^{23}$ Zhang,$^{40}$ Winter,$^{41}$ and Collisson$^{12}$ cohorts. We first tested the predictive value of early death by calculating the AUROC for each data set separately. PCOSP was significant overall (AUROC, 0.70; $P < 2.6E-22$; Fig 3A) but was higher in the data sets that were generated using sequencing platforms compared with microarrays (AUROC, 0.72 v 0.68 for sequencing and array data sets, respectively) at $P = .09$, which suggests that RNA sequencing might be a better assay for PCOSP than microarray platforms. PCOSP was significantly predictive of early death in all cohorts (AUROC $\in [0.67, 0.76]$; $P < .05$), with the exception of the Winter and OUH cohorts ($P > .48$), and was almost significant for the Collisson cohort (AUROC, 0.69; $P = .051$). To determine whether the early death predictive value of the PCOSP model can be achieved by random chance alone, we first computed meta-estimates of AUROC by randomly shuffling the class labels—early deaths—1,000 times and applying the same training procedure used for the PCOSP model. We observed that the expression profiles were significantly associated with survival as none of the random models could yield a predictive value greater or equal to PCOSP ($P < .001$; Appendix Fig A1A). We further assessed whether the gene pairs selected in the PCOSP model were robustly associated with early death events by randomly assigning genes to the PCOSP model. We again observed that the genes selected in PCOSP yielded significantly more predictive information than the models comprised of random genes ($P < .001$; Appendix Fig A1B), which supported the biologic relevance of the PCOSP gene set.
**FIG 3.** Predictive value of the Pancreatic Cancer Overall Survival Predictor for early death and overall survival. (A) Area under the operating characteristics curve (AUC) for all cohorts and meta-estimates for sequencing cohorts, array-based cohorts, and for both platforms combined. (B and C) Forest plot reporting (B) the concordance indices (C-index) and (C) the D-index (robust hazard ratio) for all cohorts and the meta-estimates for sequencing cohorts (orange), array-based cohorts (blue), and for both the platforms combined (gray). Squares in the forest plot represent the point estimates, horizontal bars represent CIs, and the diamond is the meta-estimate. ICGC, International Cancer Genome Consortium; OUH, Oslo University Hospital; PCSI, Pancreatic Cancer Sequencing Initiative; TCGA, The Cancer Genome Atlas; UNC, University of North Carolina.
Prognostic Relevance of the PCOSP Model

To assess the prognostic value of the PCOSP model, we calculated the C-index and D-index using OS data for all cohorts. The C-index is significant overall (C-index, 0.63; \( P = 1.8 \times 10^{-12} \); Fig 3B). In agreement with results of early death prediction, the PCOSP prognostic value was higher for the sequencing data sets compared with arrays (C-index, 0.65; \( P < 3.8 \times 10^{-14} \) vs C-index, 0.61; \( P < 1.6 \times 10^{-12} \), respectively; Fig 3B). Similar to the C-index, PCOSP D-index was strong and significant overall (D-index, 1.95; \( P = 1.4 \times 10^{-04} \); Fig 3C) and stronger for the sequencing data sets (D-index, 2.24 vs 1.83; Fig 3C). To assess whether the prognostic value of PCOSP depends on PDAC molecular subtypes, we stratified PDAC samples into basal and classic subtypes using the Moffitt classifier \(^1\) and calculated meta-estimates of C-index and D-index. We found that PCOSP was prognostic in validation cohorts independent of molecular subtypes (Appendix Figs A2A and A2B). We further assessed whether PCOSP prognostic value was complementary to clinicopathologic parameters and molecular subtypes by fitting both a multivariable Cox proportional hazards model to predict survival and a logistic regression model to predict binary outcome—death at less than 1 year or not (Data Supplement).

To further illustrate the prognostic value of PCOSP, we stratified patients into low- and high-risk groups and plotted Kaplan-Meier curves for each cohort (Figs 4A-4J). OS was significantly different between risk groups for all sequencing cohorts and two microarray cohorts \( P < .05 \) and borderline significant for three microarray cohorts \( .05 \leq P < .10 \); Figs 4A-4J), with a 10-month difference in median OS between risk groups.

Clinicopathologic Model to Predict OS

The logistic regression model fitted using these clinicopathologic features was used to predict early death of patients with PDAC. The clinicopathologic model was not significant overall (C-index, 0.55; \( P = .17 \); Fig 5A). In contrast to PCOSP, the clinicopathologic model was not predictive in the sequencing cohort (C-index, 0.53 and 0.58 with \( P = .75 \) and .05 for the sequencing and array data sets, respectively; Fig 5A). Only nodal status, tumor grade, and molecular classes were significant in the univariable analysis (Data Supplement). We compared the prognostic value of the clinicopathologic model with PCOSP (Figs 5B and 5C). PCOSP was significantly more prognostic than the clinicopathologic model (one-sided \( t \) test, \( P < .01 \); Fig 5D).

Comparison With Published Prognostic Models

We compared the prognostic value of PCOSP with three published PDAC prognostic models, referred to as Birnbaum, \(^22\) Chen, \(^23\) and Haider. \(^24\) Overall prognostic value of the three published models was significant (Figs 6A and 6C). PCOSP significantly outperformed published prognostic models in all cases \( P < .05 \); Figs 6C and 6D), with the exception of the D-index of the Chen classifier where the superiority of the PCOSP prognostic value demonstrated a trend to significance (one-sided \( t \) test, \( P = .10 \)).

Pathway Analysis of Prognostic Genes

Gene enrichment analysis for PCOSP signature genes \( n = 1,070 \) found that the extracellular matrix (ECM), epithelial–mesenchymal transition (EMT), and hedgehog signaling pathway genes were enriched in the PCOSP model at false-discovery rate of less than 5% (Data Supplement).

DISCUSSION

We performed a meta-analysis of the transcriptomic profiles of 1,236 patients with PDAC and developed the PCOSP, a new prognostic model with which to identify patients who are at high risk of early death after surgery. The model is built from a unique set of 89 patients profiled using both array-based and sequencing platforms and validated on a compendium of 10 independent data sets that included 823 patients. The prognostic value of the PCOSP model was highly significant for both early death—1 year or less—and OS \( P < .001 \); Fig 3).

Contrary to published prognostic signatures that were fitted on a small number of samples and that lack validation in large independent data sets, \(^17 -21\) PCOSP has been trained and validated on a large compendium of data sets. Comparison of PCOSP with existing classifiers \(^22 -24\) demonstrated that the Birnbaum, Chen, and Haider models yielded significant but significantly weaker prognostic value than PCOSP (Figs 6C and 6D). Of importance, PCOSP performs significantly better than existing classifiers for both microarray and sequencing platforms, likely because of simplifying the continuous expression space into binary pair barcodes. This enables PCOSP to be used as a single sample predictor robust to profiling platforms, potential batch effects, and normalization methods compared with other classifiers.

Comparison of PCOSP against known prognostic clinicopathologic variables demonstrated that PCOSP outperformed the clinicopathologic model in predicting early death (Fig 5). PCOSP prognostic value was significant, even after adjusting for molecular subtyping (classic vs basal) and clinicopathologic parameters (age, sex, TNM status, differentiation grade of tumor, and molecular classes; Appendix Figs A2A and A2B and Data Supplement).

The PCOSP model incorporates 2,619 unique gene pairs, totaling 1,070 unique genes. Functional analysis of 1,070 genes demonstrated enrichment of Hedgehog signaling and ECM and EMT pathways. Numerous studies have suggested the involvement of EMT in the invasion and metastasis of PDAC. \(^42\) EMT enhances cell motility via loss of cell–cell adhesion, escaping from the ECM and overcoming the apoptosis process. \(^42\) The ECM and EMT pathways are not only associated with the metastatic spread of tumor but also with chemoresistance, which leads to worse survival. \(^43\)
PDAC is a heterogeneous and genetically highly complex disease, which supports the molecular and morphologic characterization of a given tumor as an important cornerstone for the development of future therapies. We provide the largest compendium of PDAC data sets as a gold standard for future PDAC analyses. The new meta-analysis framework implemented in PCOSP maximizes robustness and performance across cohorts. To implement PCOSP as a clinical assay, we tested different feature set sizes for the k-Top scoring disjoint pairs models and compared the performance of the reduced models. We achieved accuracy that was comparable to the 1,070-gene PCOSP model by including only 256 unique genes, which supports the potential for the implementation of a smaller PCOSP-like model for the clinic setting (Appendix Fig A3).

Endoscopic ultrasound biopsies could be used before curative surgery to estimate the prognosis of patients with PDAC using PCOSP. This may assist clinicians in the selection of patients for surgery and help to identify those patients with high-risk progressive disease for whom an operation has little oncologic benefit.

The current study has potential limitations. First, there are inherent tumor sample collection biases as the different data sets were collected and sampled at different centers and have heterogeneous standard-of-care across different hospitals. Levels of tumor cellularity varied highly across cohorts as PCSI and Collisson data sets were generated using laser microdissection before to sequencing; Kirby and Chen data sets were macrodissected; and TCGA, Winter, and J Collisson. Curves show the P values from log-rank test. Overall survival difference between low- and high-risk groups is 13 and 23 months, respectively, at P < .05.

FIG 4. Kaplan Meier overall survival curves for (A) Pancreatic Cancer Sequencing Initiative (PCSI), (B) The Cancer Genome Atlas (TCGA), (C) Kirby, (D) International Cancer Genome Consortium (ICGC) array, (E) University of North Carolina (UNC), (F) Chen, (G) Oslo University Hospital (OUH), (H) Zhang, (I) Winter, and (J) Collisson. Curves show the P values from log-rank test. Overall survival difference between low- and high-risk groups is 13 and 23 months, respectively, at P < .05.
FIG 5. Comparison of the prognostic value of the clinicopathologic model and Pancreatic Cancer Overall Survival Predictor (PCOSP). (A) Bar plot reporting the Area under the operating characteristics curve (AUC) for the clinical model and the PCOSP model. (B and C) Forest plot reporting the (B) concordance index (C-index) and (C) D-index (robust hazard ratio) of validation cohorts computed using PCOSP and clinicopathologic model. Squares
ICGC, OUH, Zhang, and Winter data sets used bulk tumors for profiling. Second, transcriptomic profiles in our data compendium were generated using different gene expression profiling technologies for sequencing (Illumina HiSEquation 2000/2500; Illumina, San Diego, CA) and microarray platforms (Agilent Technologies, Santa Clara, CA; Affymetrix, Santa Clara, CA; and Illumina). Third, all samples were normalized using the published processing methods, which depend on the profiling platforms (Data Supplement). Fourth, there may be a loss of information with regard to the coexpression and magnitude of differential expression between genes by converting expression data into binary barcodes information. However, there are statistical benefits to the binary barcodes approach over predictions that are based on continuous gene expression data. The binary barcode approach produces single-sample predictions that are insensitive to monotonic transformation of the gene expression data, which is particularly relevant in the meta-analysis of heterogeneous cohorts in which continuous gene expression–based prediction approaches need scaling of data for comparison across cohorts.

Despite these limitations, PCOSP yielded robust prognostic value across the heterogeneous data sets, indicating that the gene expression barcode transformation is robust to the inevitable biases that are present in large meta-analyses.

ICGC, OUH, Zhang, and Winter data sets used bulk tumors for profiling. Second, transcriptomic profiles in our data compendium were generated using different gene expression profiling technologies for sequencing (Illumina HiSEquation 2000/2500; Illumina, San Diego, CA) and microarray platforms (Agilent Technologies, Santa Clara, CA; Affymetrix, Santa Clara, CA; and Illumina). Third, all samples were normalized using the published processing methods, which depend on the profiling platforms (Data Supplement). Fourth, there may be a loss of information with regard to the coexpression and magnitude of differential expression between genes by converting expression data into binary barcodes information. However, there are statistical benefits to the binary barcodes approach over predictions that are based on continuous gene expression data. The binary barcode approach produces single-sample predictions that are insensitive to monotonic transformation of the gene expression data, which is particularly relevant in the meta-analysis of heterogeneous cohorts in which continuous gene expression–based prediction approaches need scaling of data for comparison across cohorts.

Despite these limitations, PCOSP yielded robust prognostic value across the heterogeneous data sets, indicating that the gene expression barcode transformation is robust to the inevitable biases that are present in large meta-analyses.
However, exploring other factors, such as germline variants, epigenetics, copy number alterations, noncoding RNAs, protein abundance and epidemiologic and environmental factors, will be necessary to further improve the prediction accuracy of predictive models.

Lack of available clinical and treatment information across cohorts is also a limiting factor in our meta-analysis, which prevents us from investigating this source of heterogeneity further. However, comparison of cohort-specific clinical information for the cohort was not significantly different across cohorts (Data Supplement). During the time period of sample collection, standard-of-care treatment of PDAC was curative-intent surgery followed by adjuvant chemotherapy with gemcitabine or fluorouracil. New approaches using doublet and triplet chemotherapy regimens are now becoming standard of care in the adjuvant setting. The survival benefit observed with FOLFIRINOX (folinic acid, fluorouracil, irinotecan, oxaliplatin) in the adjuvant setting highlights the importance of systemic therapy in curing patients with resectable PDAC. The role of neoadjuvant chemotherapy is also being evaluated in many centers; thus, heterogeneity in treatment is expected within and between different cohorts. We will need to test our PCOSP model using new clinical data sets or preferably within the context of randomized trials.

In conclusion, we leveraged the largest compendium of PDAC transcriptomes to develop PCOSP, a prognostic model that identifies patients with PDAC at high risk of early death independent of, and superior to, clinicopathologic features and molecular subtypes. PCOSP may be useful in the clinical setting as a single sample classifier to identify patients who could be at higher risk of early death after surgery and adjuvant chemotherapy, potentially facilitating treatment decisions.

AFFILIATIONS
1University Health Network, Toronto, Ontario, Canada
2Oslo University Hospital, Institute for Cancer Research, Oslo, Norway
3Oslo University Hospital, Oslo, Norway
4Mount Sinai Hospital, Toronto, Ontario, Canada
5Ontario Institute for Cancer Research, Toronto, Ontario, Canada
6University of South-Eastern Norway, Ba in Telemark, Norway
7University of Toronto, Toronto, Ontario, Canada

Preprint version available on bioRxiv.

CORRESPONDING AUTHOR
Benjamin Haibe-Kains, PhD, University of Toronto, 101 College St, PMCR T11-310, M5G1L7, Toronto, ON M5G1L7, Canada; Twitter: @bhaibeka; @OICR_news; @UHN; e-mail: bhaibeka@uhnresearch.ca.

SUPPORT
Supported by the Ontario Institute for Cancer Research (OICR; PanCure Translational Research Initiative) through funding provided by the Government of Ontario, and a charitable donation from the Canadian Friends of the Hebrew University (Alex U. Soyka); by grants from The Radium Hospital Foundation, Oslo University Hospital, and the PanCure Translational Research Initiative at the OICR (V.S.); and by the Gattuso Slaight Personalized Cancer Medicine Fund at Princess Margaret Cancer Centre, the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada, and the Ministry of Economic Development and Innovation/Ministry of Research and Innovation of Ontario (Canada; B.H.-K.).

AUTHOR CONTRIBUTIONS
Conception and design: Vandana Sandhu, Julie Wilson, Benjamin Haibe-Kains
Financial support: Steven Gallinger, Benjamin Haibe-Kains
Administrative support: Julie Wilson, Benjamin Haibe-Kains
Provision of study material or patients: Vandana Sandhu, Sara Hafezi-Bakhtiari, Julie Wilson, Elin H. Kure, Benjamin Haibe-Kains
Collection and assembly of data: Vandana Sandhu, Knut Jorgen Labori, Ayelet Borgida, Ilincu Lungu, John Bartlett, Sara Hafezi-Bakhtiari, Rob

AUTHORS’ DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST
The following represents disclosure information provided by authors of this manuscript. For more information about ASCO’s conflicts of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

John Bartlett
Travel, Accommodations, Expenses:.bioTheranostics

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT
The authors thank Syed Haider, MD, for courteously providing the prediction scores from the classifier for comparison with PCOSP. The authors thank all patients who participated in this study.
REFERENCES


APPENDIX

MATERIALS AND METHODS

Data Sets

We surveyed the literature and curated 17 data sets that included 1,236 patients with pancreatic ductal adenocarcinoma (PDAC) from the public domain for which transcriptome data of PDAC were available. We filtered samples on the basis of the availability of overall survival (OS) and sample size (\(n > 10\)) after dichotomization into high- and low-survival groups on the basis of an OS cutoff of 1 year. This resulted in a total of four sequencing studies and seven array-based studies providing transcriptomic and clinical data for 1,001 patients with PDAC. A total of 12,430 protein-coding genes commonly assessed across all cohorts were used for additional analysis.

Prognostic Model

Decision rules are based on the relative ordering of gene expression values within the same sample in which the k-Top scoring gene pairs are used to build the classifier. Samples were resampled 1,000 times, where 40 samples from each group were selected in each run to build a k-Top scoring disjoint pair (k-TSP) model, and the model was further tested on the 49 out-of-bag samples. Models were selected if the balanced accuracy was greater than 0.6 or the model was rejected. We then froze the parameters of the predictive model and validated it in the remaining compendium of independent data sets. The class probability of the sample was calculated as the frequency of the sample predicted as one class divided by the total number of models. Advantages of considering pairs of genes with a binary value—"1" if expression of gene \(i >\) gene \(j\), "0" otherwise—are as follows: it transforms the feature space in a way that mitigates platform biases and potential batch effects and it makes the model robust to any data processing that preserves gene order (Patil P, et al: Bioinformatics 31: 2318-2323, 2015; Eddy JA, et al: Technol Cancer Res Treat 9:149-159, 2010).

Random Classifier

To assess whether gene expression profiles were associated with survival, we shuffled the actual class labels while maintaining the expression values. To test whether the gene pairs selected in the Pancreatic Cancer Overall Survival Predictor (PCOSP) model were robustly associated with survival, we randomly assigned genes to the k-TSP model and assessed its prognostic value. Both procedures were performed 1,000 times. As a prevalidation set, we compared the balanced accuracy of all 1,000 random models generated using both approaches to PCOSP using the Wilcoxon rank sum test. Furthermore, we trained the k-TSP classifier models from both approaches in the same way as we built our consensus PCOSP model. We then froze the parameters of the prognostic model and validated it in the compendium of independent data sets and compared meta-estimates for both models against the PCOSP model.

Subtyping of PDAC Cohorts

PDAC cohorts were classified into basal and classic transcriptomic subtypes using the Moffitt classifier.13 We calculated the meta-estimates of C-index and hazard ratio for PDAC subtypes using the random effect model implemented in survcomp package in R Schroder et al: Bioinformatics 27: 3206-8, 2011.

FIG A1. Density plot showing the distribution of balanced accuracy for random models. (A and B) Distribution of meta-estimates of 1,000 models generated using (A) random reshuffling of labels and (B) random assignment of genes to k-Top scoring disjoint pair (k-TSP) models. Meta-estimates were independently calculated for all the cohorts combined, sequencing cohorts, and array-based cohorts. Pink, green, and blue dashed lines represent meta-estimate of area under the receiver operating characteristics curve from the Pancreatic Cancer Overall Survival Predictor model for overall, sequencing, and array-based cohorts, respectively.
Gene Set Enrichment Analysis


Research Reproducibility

### Table A

<table>
<thead>
<tr>
<th>Cohort</th>
<th>P value</th>
<th>C-index</th>
<th>Cohort</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCGA (n = 146)</td>
<td></td>
<td></td>
<td>TCGA (n = 146)</td>
<td></td>
</tr>
<tr>
<td>Basal subtype (n = 29)</td>
<td>7.7E-01</td>
<td></td>
<td>Basal subtype (n = 29)</td>
<td></td>
</tr>
<tr>
<td>Classical subtype (n = 117)</td>
<td>1.6E-03</td>
<td></td>
<td>Classical subtype (n = 117)</td>
<td>1.7E-02</td>
</tr>
<tr>
<td>PCSI (n = 118)</td>
<td>1.0E-03</td>
<td></td>
<td>PCSI (n = 118)</td>
<td></td>
</tr>
<tr>
<td>Basal subtype (n = 25)</td>
<td>1.6E-01</td>
<td></td>
<td>Basal subtype (n = 25)</td>
<td>5.2E-01</td>
</tr>
<tr>
<td>Classical subtype (n = 93)</td>
<td>2.2E-02</td>
<td></td>
<td>Classical subtype (n = 93)</td>
<td>1.6E-02</td>
</tr>
<tr>
<td>Kirby (n = 51)</td>
<td>1.8E-06</td>
<td></td>
<td>Kirby (n = 51)</td>
<td>7.4E-05</td>
</tr>
<tr>
<td>Basal subtype (n = 13)</td>
<td>2.3E-04</td>
<td></td>
<td>Basal subtype (n = 13)</td>
<td>1.1E-01</td>
</tr>
<tr>
<td>Classical subtype (n = 38)</td>
<td>3.8E-06</td>
<td></td>
<td>Classical subtype (n = 38)</td>
<td>6.2E-03</td>
</tr>
<tr>
<td>ICGC array (n = 178)</td>
<td>2.6E-07</td>
<td></td>
<td>ICGC array (n = 178)</td>
<td>4.2E-07</td>
</tr>
<tr>
<td>Basal subtype (n = 156)</td>
<td>1.9E-06</td>
<td></td>
<td>Basal subtype (n = 156)</td>
<td>2.4E-02</td>
</tr>
<tr>
<td>Classical subtype (n = 22)</td>
<td>2.8E-04</td>
<td></td>
<td>Classical subtype (n = 22)</td>
<td>1.2E-04</td>
</tr>
<tr>
<td>UNC (n = 125)</td>
<td>4.7E-05</td>
<td></td>
<td>UNC (n = 125)</td>
<td>2.5E-03</td>
</tr>
<tr>
<td>Basal subtype (n = 27)</td>
<td>9.3E-01</td>
<td></td>
<td>Basal subtype (n = 27)</td>
<td>9.7E-01</td>
</tr>
<tr>
<td>Classical subtype (n = 98)</td>
<td>2.4E-05</td>
<td></td>
<td>Classical subtype (n = 98)</td>
<td>4.6E-03</td>
</tr>
<tr>
<td>Chen (n = 63)</td>
<td>3.8E-02</td>
<td></td>
<td>Chen (n = 63)</td>
<td>2.2E-02</td>
</tr>
<tr>
<td>Basal subtype (n = 20)</td>
<td>1.7E-03</td>
<td></td>
<td>Basal subtype (n = 20)</td>
<td>9.1E-02</td>
</tr>
<tr>
<td>Classical subtype (n = 43)</td>
<td>2.5E-01</td>
<td></td>
<td>Classical subtype (n = 43)</td>
<td>4.4E-02</td>
</tr>
<tr>
<td>OUH (n = 48)</td>
<td>8.9E-01</td>
<td></td>
<td>OUH (n = 48)</td>
<td>2.9E-01</td>
</tr>
<tr>
<td>Basal subtype (n = 08)</td>
<td>NA</td>
<td></td>
<td>Basal subtype (n = 08)</td>
<td></td>
</tr>
<tr>
<td>Classical subtype (n = 40)</td>
<td>9.8E-01</td>
<td></td>
<td>Classical subtype (n = 40)</td>
<td>3.5E-01</td>
</tr>
<tr>
<td>Zhang (n = 42)</td>
<td>2.1E-03</td>
<td></td>
<td>Zhang (n = 42)</td>
<td>2.6E-02</td>
</tr>
<tr>
<td>Basal subtype (n = 13)</td>
<td>4.8E-02</td>
<td></td>
<td>Basal subtype (n = 13)</td>
<td>2.6E-01</td>
</tr>
<tr>
<td>Classical subtype (n = 29)</td>
<td>1.3E-01</td>
<td></td>
<td>Classical subtype (n = 29)</td>
<td>4.6E-01</td>
</tr>
<tr>
<td>Winter (n = 30)</td>
<td>8.9E-01</td>
<td></td>
<td>Winter (n = 30)</td>
<td>6.4E-01</td>
</tr>
<tr>
<td>Basal subtype (n = 8)</td>
<td>7.4E-01</td>
<td></td>
<td>Basal subtype (n = 8)</td>
<td>6.0E-01</td>
</tr>
<tr>
<td>Classical subtype (n = 22)</td>
<td>7.2E-01</td>
<td></td>
<td>Classical subtype (n = 22)</td>
<td>7.5E-01</td>
</tr>
<tr>
<td>Collisson (n = 27)</td>
<td>1.3E-02</td>
<td></td>
<td>Collisson (n = 27)</td>
<td>2.8E-02</td>
</tr>
<tr>
<td>Basal subtype (n = 1)</td>
<td></td>
<td></td>
<td>Basal subtype (n = 1)</td>
<td></td>
</tr>
<tr>
<td>Classical subtype (n = 26)</td>
<td>4.2E-02</td>
<td></td>
<td>Classical subtype (n = 26)</td>
<td>6.3E-02</td>
</tr>
</tbody>
</table>

### Table B

<table>
<thead>
<tr>
<th>Cohort</th>
<th>P value</th>
<th>Log2 D-index</th>
<th>Cohort</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCGA (n = 146)</td>
<td></td>
<td>1.7E-04</td>
<td>TCGA (n = 146)</td>
<td></td>
</tr>
<tr>
<td>Basal subtype (n = 29)</td>
<td>4.3E-01</td>
<td></td>
<td>Basal subtype (n = 29)</td>
<td></td>
</tr>
<tr>
<td>Classical subtype (n = 117)</td>
<td>1.7E-02</td>
<td></td>
<td>Classical subtype (n = 117)</td>
<td></td>
</tr>
<tr>
<td>PCSI (n = 118)</td>
<td>1.0E-02</td>
<td></td>
<td>PCSI (n = 118)</td>
<td></td>
</tr>
<tr>
<td>Basal subtype (n = 25)</td>
<td>5.2E-01</td>
<td></td>
<td>Basal subtype (n = 25)</td>
<td></td>
</tr>
<tr>
<td>Classical subtype (n = 93)</td>
<td>1.6E-02</td>
<td></td>
<td>Classical subtype (n = 93)</td>
<td></td>
</tr>
<tr>
<td>Kirby (n = 51)</td>
<td>7.4E-05</td>
<td></td>
<td>Kirby (n = 51)</td>
<td></td>
</tr>
<tr>
<td>Basal subtype (n = 13)</td>
<td>1.1E-01</td>
<td></td>
<td>Basal subtype (n = 13)</td>
<td></td>
</tr>
<tr>
<td>Classical subtype (n = 38)</td>
<td>6.2E-03</td>
<td></td>
<td>Classical subtype (n = 38)</td>
<td></td>
</tr>
<tr>
<td>ICGC array (n = 178)</td>
<td>4.2E-07</td>
<td></td>
<td>ICGC array (n = 178)</td>
<td></td>
</tr>
<tr>
<td>Basal subtype (n = 156)</td>
<td>2.4E-02</td>
<td></td>
<td>Basal subtype (n = 156)</td>
<td></td>
</tr>
<tr>
<td>Classical subtype (n = 22)</td>
<td>1.2E-04</td>
<td></td>
<td>Classical subtype (n = 22)</td>
<td></td>
</tr>
<tr>
<td>UNC (n = 125)</td>
<td>2.5E-03</td>
<td></td>
<td>UNC (n = 125)</td>
<td></td>
</tr>
<tr>
<td>Basal subtype (n = 27)</td>
<td>9.7E-01</td>
<td></td>
<td>Basal subtype (n = 27)</td>
<td></td>
</tr>
<tr>
<td>Classical subtype (n = 98)</td>
<td>4.6E-03</td>
<td></td>
<td>Classical subtype (n = 98)</td>
<td></td>
</tr>
<tr>
<td>Chen (n = 63)</td>
<td>2.2E-02</td>
<td></td>
<td>Chen (n = 63)</td>
<td></td>
</tr>
<tr>
<td>Basal subtype (n = 20)</td>
<td>9.1E-02</td>
<td></td>
<td>Basal subtype (n = 20)</td>
<td></td>
</tr>
<tr>
<td>Classical subtype (n = 43)</td>
<td>4.4E-02</td>
<td></td>
<td>Classical subtype (n = 43)</td>
<td></td>
</tr>
<tr>
<td>OUH (n = 48)</td>
<td>2.9E-01</td>
<td></td>
<td>OUH (n = 48)</td>
<td></td>
</tr>
<tr>
<td>Basal subtype (n = 08)</td>
<td></td>
<td></td>
<td>Basal subtype (n = 08)</td>
<td></td>
</tr>
<tr>
<td>Classical subtype (n = 40)</td>
<td>3.5E-01</td>
<td></td>
<td>Classical subtype (n = 40)</td>
<td></td>
</tr>
<tr>
<td>Zhang (n = 42)</td>
<td>2.6E-02</td>
<td></td>
<td>Zhang (n = 42)</td>
<td></td>
</tr>
<tr>
<td>Basal subtype (n = 13)</td>
<td>2.6E-01</td>
<td></td>
<td>Basal subtype (n = 13)</td>
<td></td>
</tr>
<tr>
<td>Classical subtype (n = 29)</td>
<td>4.6E-01</td>
<td></td>
<td>Classical subtype (n = 29)</td>
<td></td>
</tr>
<tr>
<td>Winter (n = 30)</td>
<td>6.4E-01</td>
<td></td>
<td>Winter (n = 30)</td>
<td></td>
</tr>
<tr>
<td>Basal subtype (n = 8)</td>
<td>6.0E-01</td>
<td></td>
<td>Basal subtype (n = 8)</td>
<td></td>
</tr>
<tr>
<td>Classical subtype (n = 22)</td>
<td>7.5E-01</td>
<td></td>
<td>Classical subtype (n = 22)</td>
<td></td>
</tr>
<tr>
<td>Collisson (n = 27)</td>
<td>2.8E-02</td>
<td></td>
<td>Collisson (n = 27)</td>
<td></td>
</tr>
<tr>
<td>Basal subtype (n = 1)</td>
<td></td>
<td></td>
<td>Basal subtype (n = 1)</td>
<td></td>
</tr>
<tr>
<td>Classical subtype (n = 26)</td>
<td>6.3E-02</td>
<td></td>
<td>Classical subtype (n = 26)</td>
<td></td>
</tr>
</tbody>
</table>

### Diagram

**FIG A2.** Forest plot of (A) concordance index (C-index) and (B) D-index (robust hazard ratio) for all cohorts divided on the basis of the molecular subtypes. Gray, green, and pink colors in the forest plot depict meta-estimate of C-index for overall cohort, the basal subtype, and the classic subtype of the cohorts, respectively. Squares in the forest plot represent the point estimates, horizontal bars represent CIs, and the diamond is the meta-estimate. ICGC, International Cancer Genome Consortium; OUH, Oslo University Hospital; PCSI, Pancreatic Cancer Sequencing Initiative; TCGA, The Cancer Genome Atlas; UNC University of North Carolina.
FIG A3. Scatterplot showing the meta-estimate of area under the receiver operating characteristics curve (AUROC; orange) and the total number of unique genes (blue) in the Pancreatic Cancer Overall Survival Predictor (PCOSP) model at different balanced accuracy thresholds. The threshold used in the PCOSP is marked as a dashed line at 0.6.