

Exploring Clinical Phenotypes of Pancreatic Cancer in 2-year Survival and Treatment Patterns: A Retrospective Registry-Based Cluster Analysis

Louisa Schwarz

`louisa.schwarz@uni-mainz.de`

Johannes Gutenberg University Mainz

Christina Justenhoven

Institut für digitale Gesundheitsdaten Rheinland-Pfalz

Research Article

Keywords: Pancreatic Cancer, Ward's Linkage, Clustering, Machine Learning, Cox-Regression, Survival Analysis, Data Analysis

Posted Date: August 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10418248/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Exploring Clinical Phenotypes of Pancreatic Cancer in 2-year Survival and Treatment Patterns: A Retrospective Registry-Based Cluster Analysis

Louisa Schwarz^{1*} and Christina Justenhoven²

^{1*}Johannes Gutenberg University, Mainz, Germany.

²Cancer Registry of Rhineland-Palatinate in the Institute for Digital
Health Data RLP, Mainz, Germany.

*Corresponding author(s). E-mail(s): louisa.schwarz@uni-mainz.de;
Contributing authors: justenhoven@idg-rlp.de;

Abstract

Purpose: Pancreatic cancer (PC) is a highly lethal type of cancer and a highly heterogeneous disease. The current treatment of PC is primarily determined by tumor severity and influences the subsequent course of the disease. Our goal is to address this heterogeneity by identifying distinct PC phenotypes and assessing their relationship to survival and received therapies.

Methods: We used a registry-based dataset of 1,158 pancreatic cancer patients provided by the Cancer Registry Rhineland-Palatinate and performed an unsupervised hierarchical Ward's linkage analysis using basic patient, tumor, and treatment characteristics. We then explored the relationship between the identified phenotypes and the two-year event-free survival probability, as well as treatment associations within phenotypes using Cox regressions.

Results: We identified six phenotypes consisting of homogeneous patient, tumor, and treatment characteristics. The young, low-grade, node-negative phenotype 2 defines the favorable prognosis baseline. Male patients in phenotype 5 and elderly female patients in phenotype 1, representing the most typical clinical profile, have twice the risk of death or an event. An even higher risk is observed in advanced profiles characterized by a high proportion of high-grade phenotypes that received surgery (phenotype 4 with partial resection and phenotype 3 with pancreatectomy). Patients with pancreatic cancer (PC) who are metastatic define the high-risk end (phenotype 6).

Conclusion: The six identified phenotypes capture existing clinical profiles and treatment guidelines. Additionally, they vary significantly in terms of two-year event-free survival, suggesting that they represent clinically meaningful PC profiles that could facilitate the development of novel therapeutic strategies and promote more informed treatment decisions, which are crucial for PC heterogeneity.

Keywords: Pancreatic Cancer, Ward’s Linkage, Clustering, Machine Learning, Cox-Regression, Survival Analysis, Data Analysis

1 Introduction

Pancreatic cancer (PC) is a highly lethal and aggressive type of cancer and a highly heterogeneous disease. Around 80% of all PC patients are diagnosed at a locally advanced or distant metastatic stage, due to its asymptomatic course which makes early detection so difficult [1]. In 2020, PC is the seventh leading cause of cancer death in both sexes with approximately 466,000 deaths and almost as many new diagnosis (496,000) worldwide [2]. With increasing incidence yearly, PC is projected to become the second leading cause of cancer death by 2030, underscoring the need for optimal or novel therapeutic strategies to improve event-free survival [3, 4].

Current treatment of PC is primarily determined by the severity of the tumor and influences the subsequent course of the disease [5–7]. As of yet, surgical resection of the pancreas with adjuvant chemotherapy is the only known curative treatment. Effective therapies for non-resectable tumors remain limited, whereby systemic therapies yield modest survival benefits but high inter-patient heterogeneity in treatment response [6, 8]. This heterogeneity may also be reflected in different biological phenotypes, which are based on clinically observable patient and tumor characteristics as well as the therapies received [9, 10]. Thus, there is a need to explore and understand the heterogeneity of PC patients in addition to the development of improved or novel therapeutic strategies [3, 9].

Therefore, the main contribution of this study is to assess the variety of PC by identifying distinct phenotypes and evaluating their relation to therapies received and 2-year event-free survival. The results may contribute to a better understanding of PC subgroups and might support improved personalized clinical management of this disease in the future [1, 9].

The goal of this study is to apply hierarchical clustering with Ward’s linkage to identify distinct and homogeneous clusters of PC patients, using all available information regarding patient, tumor, and treatment characteristics. Subsequently, we evaluate the relationship between the identified phenotypes and 2-year event-free survival.

2 Methods

2.1 Study Participants

The registry-based dataset used in this study was provided by the Cancer Registry Rhineland-Palatinate in the Institute for Digital Health Data RLP (CR-RLP) and consists of adult patients diagnosed with pancreatic cancer (ICD-10-GM C25) between 2016/01/01 and 2023/12/31, including information on follow-up until 2025/12/31. Only primary tumors and PC diagnosis that are histologically confirmed are included. PC patients who died within the first three months after diagnosis or who died from causes unrelated to PC two years after diagnosis are excluded. Further data cleansing and feature selection follow clinical practice and are performed in agreement with CR-RLP specialists and medical experts [6, 8, 11, 12].

The final dataset includes 1,158 patients, of whom 459 had an event-free of at least two years. For the analysis, the following events were included: death, tumor progression, recurrence, newly developed distant metastasis or metastatic recurrence. The median age of this cohort is 69.52 years, with 51.55 % (n=597) male patients. The majority of 65.8% (n=762) of PC is diagnosed in the head of the pancreas (C25.0),

using the International Statistical Classification of Diseases and Related Health Problems - 10. Revision - German Modification (IDC-10-GM). According to the TNM classification system, the most common diagnoses of PC are T2 (39.46% n=457) or T3 (28.32%, n=328), and M0 (82.73%, n=958). Regarding therapies, 57.6% (n=667) received surgical resection, 8.72% (n=101) underwent pancreatectomy, and 40.41% (n=468) received systemic therapy.

2.2 Current Staging and Treatment

Depending on the degree of vascular involvement, the National Comprehensive Cancer Network (NCCN) guideline [13] categorizes the anatomical stage of PC into the following resectability criteria: resectable (RPC), borderline resectable (BRPC), locally advanced (LAPC), or metastatic tumors [3]. For the latter, around 57% of PC patients are diagnosed with metastatic, non-resectable disease, for which the standard of care is palliative chemotherapy [3].

The standard treatment for patients with resectable PC is surgery with chemotherapy, providing the best chance for long-term survival [14].

For patients with RPC, the current standard of care involves surgery followed by adjuvant chemotherapy [3]. However, about 33% do not receive adjuvant chemotherapy due to surgical complications that impact functional recovery [15]. For BRPC patients, neoadjuvant chemotherapy followed by surgery and sometimes adjuvant chemotherapy demonstrated superior survival. For LAPC patients, primary treatment involves induction chemotherapy for 4-6 months, sometimes followed by surgical resection [3].

Additionally, the surgical procedure is also determined by the location of the tumor within the pancreas itself (ICD-10-GM classification), in either the pancreatic head (C25.0), body (C25.1), or tail (C25.2) [16]. Surgery includes the pancreatic-head resection with so-called partial pancreatoduodenectomy, pancreatic-tail resection

with so called distal pancreatectomy, or the total removal of the pancreas, so called pancreatoduodenectomy [3, 16].

2.3 Statistical Analysis: Clustering, Distances, and Survival

In this work an unsupervised machine learning approach to identify clinically meaningful phenotypes of pancreatic cancer patients was used [17, 18]. We include all available patient, tumor, and treatment characteristics in the clustering analysis and focused purely on phenotypic stratification [17, 18].

For this purpose, we use hierarchical clustering with Ward’s linkage, that has already shown to be a promising approach for exploring phenotypes of heterogeneous diseases [19–21]. Ward’s linkage is a hierarchical, agglomerative clustering method based on a bottom-up approach, and focuses on minimizing the total within-cluster variance [22]. Initially, each data point is in its own cluster; iteratively, the algorithm merges pairs of clusters that result in the smallest possible increase in total within-cluster variance (sum of squared errors) [23, 24]. The optimal number of clusters K is determined by the Dunn index, that measures the minimum distance between two clusters divided by maximum distances of data points within a given cluster. The higher the index, the better the cluster quality [24, 25].

As we face a real-world dataset with mixed data types and varying scales and units, we use the Gower distance as a suitable distance metric [26]. For numeric variables, the distance between two data points is calculated as normalization of the absolute difference. For categorical and binary features, the distance is 0 if both feature levels are equal, and 1 if they are unequal. Finally, the Gower distance between two points is calculated as the average of the distances of all variables [26, 27].

We use Kaplan-Meier curves to visualize and evaluate survival probabilities between identified phenotypes. The curves show the probability that an event has not yet occurred [28]. In this study, we investigate the 2-year event-free survival,

including death, tumor progression, recurrence, newly developed distant metastasis or metastatic recurrence as events. Additionally, we apply Cox Proportional Hazards models, also known as Cox regressions. The baseline hazard function is defined as the reference when all input features $X = 0$ at time t , given the estimated coefficients β of the predictors X . The hazard ratios $HR = \exp(X, \beta)$ indicate whether the risk of an event increases or decreases with a unit increase or decrease in X . The statistic significance of HR is indicated by p -values [29]. In this study, we first associate the identified phenotypes with the two-year event-free survival probability. Second, we explore treatment associations within phenotypes.

All analyses are done in R (4.5.2) using the packages `hclust` (2.1.8.2), `NbClust` (3.0.1), `survival` (3.8-6), and `survminer` (0.5.2).

3 Results

The goal of this study is to identify distinct and homogeneous clusters of PC patients, using all available information regarding patient, tumor, and treatment characteristics. Subsequently, we evaluate the relationship between the identified phenotypes and 2-year event-free survival using survival analysis, including Kaplan-Meier curves and Cox regressions.

3.1 Identified phenotypes

Based on the cohort of 1,158 PC patients, we identified six distinct phenotypes, Figure 3.1 shows the dendrogram of the hierarchical clustering using Ward’s Linkage, illustrating the bottom-up approach and the grouping of the identified phenotypes. Table 1 provides an overview of the patient characteristics for each cluster. The identified phenotypes are characterized by distinct patient, tumor, and treatment profiles.

Table 1 Patient characteristics of each cluster identified using all available patient, tumor, and therapeutic characteristics.

Parameter	Phenotype					
	1	2	3	4	5	6
	elderly females with low grading N=307	young, low grade and node-negative profile N=92	elderly patients receiving pancreatectomy N=191	high grade and node-positive profile N=162	male patients with low grading N=234	young metastatic profile N=172
Demographics						
Age, median (sd)	70.93 (10.27)	66.33 (11.59)	70.52 (9.39)	69.76 (10.39)	69.55 (9.3)	65.27 (10.87)
Males, % (n)	3.91% (12)	54.35% (50)	62.83% (120)	50.62% (82)	99.57% (233)	58.14% (100)
Females, % (n)	96.09% (295)	45.65% (42)	37.17% (71)	49.38% (80)	0.43% (1)	41.86% (72)
ICD-10,% (n)						
C25.0	76.22% (234)	30.43% (28)	69.63% (133)	66.05% (107)	76.92% (180)	46.51% (80)
C25.1	6.84% (21)	16.3% (15)	10.99% (21)	8.64% (14)	5.56% (13)	14.53% (25)
C25.2	12.05% (37)	38.04% (35)	6.28% (12)	12.96% (21)	15.38% (36)	20.35% (35)
other	4.89% (15)	15.22% (14)	13.09% (25)	12.35% (20)	2.14% (5)	18.6% (32)
Histology,% (n)						
8500/3	47.23% (145)	4.35% (4)	50.26% (96)	75.93% (123)	51.71% (121)	8.14% (14)
8140/3	47.88% (147)	1.09% (1)	41.88% (80)	7.41% (12)	44.44% (104)	65.7% (113)
H_other	4.89% (15)	94.57% (87)	7.85% (15)	16.67% (27)	3.85% (9)	26.16% (45)
Grading,% (n)						
Low	75.24% (231)	94.57% (87)	32.98% (63)	1.23% (2)	76.92% (180)	30.81% (53)
High	20.2 % (62)	0 % (0)	63.35 % (121)	98.77 % (160)	19.66 % (46)	20.93 % (36)
unknown	4.56% (14)	5.43% (5)	3.66% (7)	0% (0)	3.42% (8)	48.26% (83)
T status,% (n)						
TIS	0.33% (1)	0% (0)	0% (0)	1.23% (2)	0.43% (1)	0% (0)
T1	9.45% (29)	22.83% (21)	5.76% (11)	5.56% (9)	9.4% (22)	2.33% (4)
T2	40.39% (124)	36.96% (34)	45.55% (87)	48.15% (78)	49.57% (116)	10.47% (18)
T3	30.94% (95)	23.91% (22)	35.08% (67)	33.95% (55)	28.63% (67)	12.79% (22)
T4	4.56% (14)	5.43% (5)	5.76% (11)	3.09% (5)	2.56% (6)	23.26% (40)
TX	14.33% (44)	10.87% (10)	7.85% (15)	8.02% (13)	9.4% (22)	51.16% (88)
N status,% (n)						
N0	28.34% (87)	61.96% (57)	37.17% (71)	29.63% (48)	28.63% (67)	11.05% (19)
N1	40.07% (123)	25% (23)	38.74% (74)	38.27% (62)	39.74% (93)	28.49% (49)
N2	14.66% (45)	3.26% (3)	15.71% (30)	22.22% (36)	20.51% (48)	4.07% (7)
NX	16.94% (52)	9.78% (9)	8.38% (16)	9.88% (16)	11.11% (26)	56.4% (97)
M status,% (n)						
M0	93.49% (287)	90.22% (83)	95.29% (182)	90.12% (146)	95.3% (223)	21.51% (37)
M1	6.51% (20)	9.78% (9)	4.71% (9)	9.88% (16)	4.7% (11)	78.49% (135)
Therapies,% (n)						
Pancreatectomy	1.95% (6)	3.26% (3)	46.07% (88)	1.23% (2)	0.43% (1)	0.58% (1)
Partial resection	78.18% (240)	67.39% (62)	0% (0)	97.53% (158)	86.75% (203)	2.33% (4)
Systemic therapy	41.69% (128)	2.17% (2)	38.22% (73)	41.98% (68)	45.3% (106)	52.91% (91)

sd = standard deviation; ICD-10: ICD-10-GM with location of the tumor in the pancreatic head (C25.0), body (C25.1), or tail (C25.2); Histology: 8500/3 = invasive ductal adenocarcinoma, 8140/3 = adenocarcinoma; T-, N-, M-status follows the current TNM-classification system; Pancreatectomy = pancreatoduodenectomy; Partial resection = pancreatic- head or pancreatic-tail resection.

The largest cluster identified is **Phenotype 1 (elderly females with low grading, n=307)**. This profile is characterized by elderly women (96.1% female; median age 70.9 years), mostly low grade tumors (75.2%) that are non-metastatic (93.5%), and predominantly located in the pancreatic head (C25.0, 76.2%). Most patients received surgical resection (78.2%) and systemic therapy (41.7%).

A second largest profile is **Phenotype 5 (male patients with low grading, n=234)**. These patients have similar characteristics with Phenotype 1: low grade (76.9%), non-metastatic tumors (95.3%), and location in the pancreatic head (76.9%), but mostly male and slightly younger (99.6% male; median age 69.55 years). The therapeutic profile resembles Phenotype 1, with high rates of surgical resection (86.8%) and systemic therapy (45.3%).

Phenotype 2 (young, low grade and node-negative profile, n = 92), is the smallest cluster with young patients (median age 66.3 years), uniformly low grade (94.6%), and the highest proportion of node-negative patients (62.0%) across all clusters. This profile is also characterized by uncommon histology (94.6% are other than the most common adenocarcinoma subtypes). Some patients underwent surgical resection (67.39%), systemic therapy is almost missing (2.2%).

In contrast to Phenotype 2, **Phenotype 4 (high grade and node-positive profile, n=162)** shows a more advanced profile. Nearly all tumors are high grade (98.8%), and the most lymph node involvement (N1=38.27% and N2=22.22%). There is a predominance of invasive ductal adenocarcinoma histology (8500/3, 75.9%). Almost all patients received surgical resection (97.5%), the highest rate among all herein defined phenotypes.

Phenotype 3 (elderly patients receiving pancreatectomy, n=191) is characterized by elderly patients (median age 70.5 years) and as the only cluster with a significant share of patients who underwent a pancreatectomy (46.07%), while none

underwent surgical resection. They have mostly high grad tumors (63.35%) that are most frequently located in the pancreatic head (69.63%).

The cluster with the youngest patients identified is **Phenotype 6 (young, metastatic profile, n=172)** that is characterized by an advanced profile: 78.5% of patients are diagnosed with distant metastases (M1), the highest proportion across all phenotypes. Tumor and nodal status are frequently unknown, given that for M1 patients, tumor and nodal status are commonly not determined anymore (TX=51.16%, NX=56.4%). Accordingly, systemic therapy is the most common treatment (52.91%).

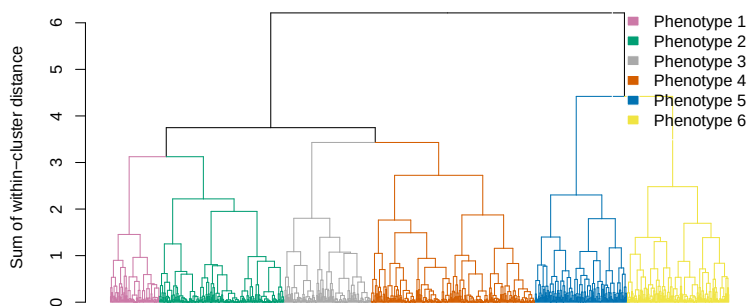


Fig. 1 Dendrogram of hierarchical clustering using Ward's Linkage, clustered on patient, tumor, and therapeutic characteristics.

3.2 Association of phenotypes with 2-year survival

699 of the 1,158 patients (60.4%) experiences no 2-year event-free survival. To further analyze the 2-year event-free survival between each phenotype identified, we apply univariate Cox regression with phenotype membership as the single co-variate. Table 3.2 provides an overview of the findings from the Cox regression analysis. Figure 2 illustrates the survival curves for each phenotype using Kaplan-Meier curves.

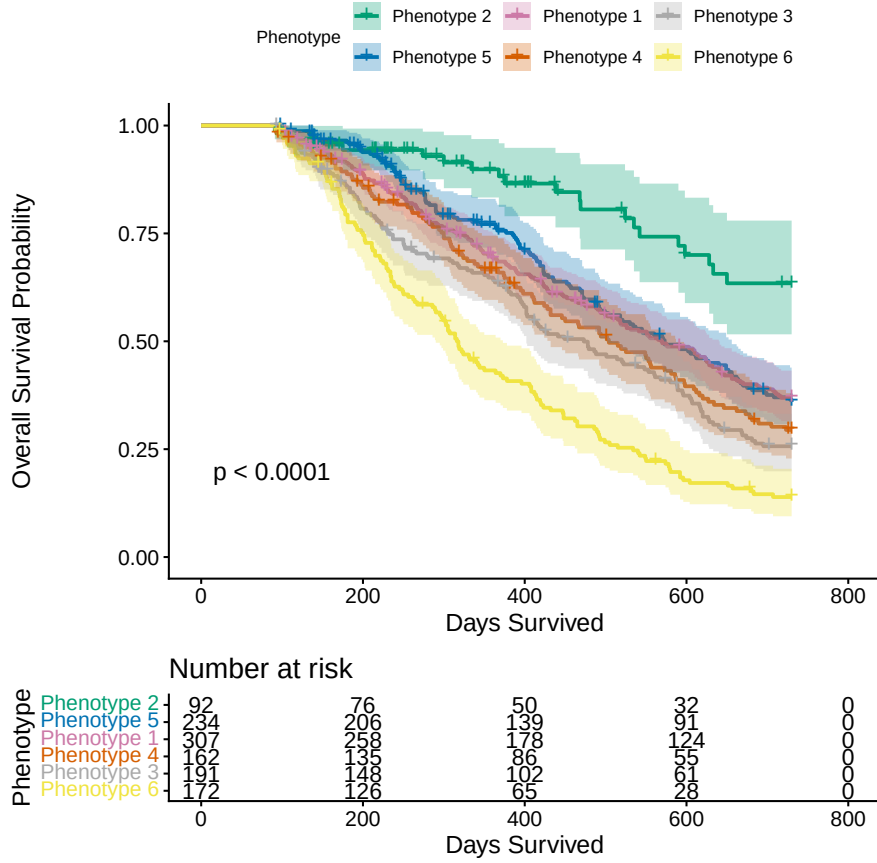


Fig. 2 Incidence rates of 2-year event-free survival illustrated using Kaplan-Meier curves for each cluster identified. Low event-free survival rate in phenotype 6 can be explained by its high-proportion of metastatic patients. The high event-free survival rate in phenotype 2 can be explained by predominantly younger patients with low grading without lymph node involvement that had undergone partial surgical resection.

Phenotype 2 had the highest event-free survival rate and is therefore defined as reference group, which can be explained by predominantly younger patients with low grading and without lymph node involvement, undergoing surgical resection. The results of the Cox regression show a substantially higher event-free survival probability for all other phenotypes with high significant p-values ≤ 0.001 . Additionally, the risk increases substantially across all six phenotypes, ranging from low-grade profiles to advanced diseases, as described in Section 3.1. Phenotypes 1 and 5 are defined by

similar characteristics of low grading, similar treatment including partial resection of the pancreas in conjunction with systemic therapies. These phenotypes demonstrate a relatively similar hazard of event within two years, but statistically significant worse and twice as high as the reference (Phenotype 1: HR 2.32, CI 1.48-3.65; Phenotype 5: HR 2.23, CI 1.41-3.54). An even higher level of risk is evident in Phenotypes 4 and 3, both characterized by an advanced stage with high-grade tumors and a relatively high rate with lymph node involvement. Notably, the treatment primarily differs between those two Phenotypes, with Phenotype 4 receiving partial resection having an even lower risk than Phenotype 3, where the patients showed the highest rate of total pancreatectomy (Phenotype 4: HR 2.81, CI 1.76-4.49; Phenotype 3: HR 3.19, CI 2.01-5.05). The most significant hazard is observable for Phenotype 6, which is consistent with its predominantly metastatic profile and an HR five times higher than the baseline (HR 5.06, CI 3.20-8.01).

To conclude, the ordering of survival rates are consistent with the patient, tumor, and treatment characteristics of the identified phenotypes: The young, low-grade, and node-negative phenotype 2 defines the favorable prognosis baseline. Male (phenotype 5) and elderly female (phenotype 1) representing the most typical clinical profile confront twice the risk of an event. An even higher risk is observable in the advanced profiles, characterized by a high proportion of high-grade phenotypes receiving surgery (phenotype 4 with partial resection, phenotype 3 with pancreatectomy). PC patients who are metastatic define the high-risk end (phenotype 6).

3.3 Treatment patterns

To further analyze the association between the treatment received and 2-years event-free survival, we apply separate multivariate Cox regression analysis for each identified phenotype. The results of these analyses are presented in Table 3. The analysis of each

Table 2 Summary of the result of Cox proportional hazards regression model, comparing each phenotype identified with respect to the hazard of event within 2 years and Phenotype 2 as reference group.

Phenotype	HR	95% CI	p-value
2 young, low grade and node-negative profile	Ref.	Ref.	Ref.
5 male patients with low grading	2.2343	1.4085 - 3.5441	≤ 0.001
1 elderly females with low grading	2.3206	1.4753 - 3.6501	≤ 0.001
4 high grade and node-positive profile	2.8072	1.7566 - 4.4863	≤ 0.001
3 elderly patients receiving pancreatectomy	3.1856	2.0093 - 5.0500	≤ 0.001
6 young metastatic profile	5.0634	3.2007 - 8.0096	≤ 0.001

HR = hazard ratios, CI = confidence interval, Ref. = reference.

phenotype is conducted independently. The findings are primarily exploratory due to the reduced sample size and must be interpreted with caution.

Within the identified phenotypes, treatment shows few significant associations with 2-year event-free survival. The only significant effect at $\alpha \leq 0.1$ is observed for systemic therapy in Phenotype 1 and 5, where receiving systemic therapy is associated with a reduced hazard of event (Phenotype 1: HR 0.67, CI 0.49-0.92; Phenotype 5: HR 0.72, CI 0.51-1.04). In the remaining phenotypes, systemic therapy is not associated with 2-year event-free survival.

Regarding surgery, there is no association with 2-year event-free survival in any identified phenotype in this study. The impact of partial resection is not estimable in Phenotype 3, nor the impact of pancreatectomy to Phenotypes 2, 5, and 6, as no or few patients within these phenotypes underwent these treatments (see Section 3.1).

4 Discussion

The aim of this study is to identify homogeneous and distinct phenotypes within the heterogeneous population of PC patients using standard characteristics that describe patient, tumor, and treatment. Therefore, we applied unsupervised hierarchical clustering using Ward's Linkage. We identified six distinct phenotypes of PC patients and

Table 3 Results of six Cox regression models for each identified phenotype to the 2-year event-free survival, including all treatment characteristics, namely systematic therapy, partial resection, and pancreatectomy. * indicates p -value below $\alpha \leq 0.1$.

Phenotype	term	HR	95% CI	p -value
2 young, low grade and node-negative profile	Systemic therapy	1.2025	0.1594-9.0726	0.8581
	Partial resection	1.1153	0.4471-2.7822	0.8150
	Pancreatectomy	n.a.	n.a.	n.a.
5 male patients with low grading	Systemic therapy	0.7242	0.5063-1.0359	0.0772*
	Partial resection	1.3037	0.7705-2.2059	0.3229
	Pancreatectomy	n.a.	n.a.	n.a.
1 elderly females with low grading	Systemic therapy	0.6727	0.4915-0.9206	0.0133*
	Partial resection	0.9273	0.6470-1.3290	0.6810
	Pancreatectomy	0.5604	0.1766-1.7789	0.3258
4 high grade and node-positive profile	Systemic therapy	1.0402	0.7018-1.5419	0.8443
	Partial resection	1.7641	0.3712-8.3839	0.4754
	Pancreatectomy	1.8584	0.3897-8.8623	0.4368
3 elderly patients receiving pancreatectomy	Systemic therapy	0.9925	0.6911-1.4254	0.9676
	Partial resection	n.a.	n.a.	n.a.
	Pancreatectomy	1.0635	0.7481-1.5119	0.7315
6 young metastatic profile	Systemic therapy	0.7677	0.5517-1.0684	0.1169
	Partial resection	0.2204	0.0307-1.5799	0.1324
	Pancreatectomy	n.a.	n.a.	n.a.

subsequently analyzed their association with 2-years event-free survival and treatment patterns.

Phenotype 2 represents the most favorable prognostic group and serves as the reference for survival analysis. It is characterized by young age, low grade, and node negativity, indicating anatomically resectable disease. As already described, combining surgery and systemic therapy provides the best chance of long-term survival [13, 14]. However, most patients received only surgical resection and no systemic therapy (only 2 PC patients), which may be reflected by the predominance of uncommon histological subtypes in this phenotype with over 94%. These findings are consistent with the review of Collisson et al., that underscores the importance of histopathological subtyping and addresses the current status, that uncommon histologies have not yet been defined for making informed treatment decisions for PC [9].

Phenotypes 1 and 5 are the two largest profiles, representing around 46% of the entire cohort. Both phenotypes represent a resectable and the most typical clinical profile, and differ primarily by sex. PC is located in the pancreatic head and treated with high rates of surgical resection. This maps them to resectable disease, for which the standard of care is surgery combined with systemic therapy [3, 13]. Regarding therapy, only 42% and 45% of these patients received systemic therapy, indicating that most patients cannot be treated with systemic therapy due to surgical complications or impaired functional recovery [3, 15]. However, there is clinical relevance when patients are not treated with systemic therapy, as systemic therapy is the only treatment associated with a reduced risk of events in both of these phenotypes (Phenotype 1: HR 0.67, p -value 0.0133; Phenotype 5: HR 0.72, p -value 0.0772; see Section 3.3). Compared to phenotype 2 (see Figure 2), the increased risk can be attributed to the high prevalence of adenocarcinomas, as well as to the high percentage of patients that did not receive systemic therapy. These findings suggest that systemic therapy may serve as a potentially modifiable treatment option for these otherwise resectable patients, again aligning with oncological knowledge that combined treatment improves event-free survival rate [16].

Phenotypes 4 and 3 represent anatomically resectable but more advanced diseases. Both profiles are characterized by predominantly high-grade tumors and a high incidence of lymph node involvement. Phenotype 4 is characterized by a high rate of invasive ductal adenocarcinomas (8500/3, 75%), which are treated almost exclusively by partial resection (97.5%). Although this profile is anatomically resectable, it has a poor prognosis because lymph node involvement further worsens the resectability criteria [13]. This profile likely reflects BRPC or LAPC patients for whom neoadjuvant or induction chemotherapy is indicated [3, 14]. Phenotype 3 is characterized by 46% undergoing total pancreatectomy, a more extensive complex surgery associated with higher risks of complications that is a typical risk of more advanced disease [16].

Both phenotypes are associated with a significantly higher HR than the low-risk group (Phenotype 4: HR 2.81; Phenotype 3: HR 3.19). Although the HR of phenotype 3 is higher among these two, the within-phenotype analysis revealed no effect of the surgical method on 2-year event-free survival (section 3.3); suggesting that the higher risk is likely due to the even more advanced stage of the disease rather than the type of surgery itself.

Lastly, **Phenotype 6** defines the high-risk end. Despite being the youngest profile, most of its patients have distant metastases (78.5%), an unknown or advanced tumor, and an unknown nodal status, which is frequently not determined for M1 patients anymore. This phenotype represents non-resectable PC for which palliative systemic chemotherapy is the standard of care, which is reflected by the highest rate of systemic therapy across all identified phenotypes (52.91%) [3, 8]. This profile has the lowest observed 2-year event-free survival probability (HR 5.06). Moreover, receiving systemic therapy has no effect on 2-year event-free survival (HR 0.77; Section 3.3). These findings are consistent with the palliative rather than curative intention of chemotherapy as well as its limited survival benefit for metastatic PC patients [6, 8].

Taken together, the identified phenotypes cover the full clinical staging of PC, from early detected resectable disease (Phenotype 2), through the most typical clinical profile with resectable low-grade disease that appears undertreated regarding systemic therapy (Phenotypes 1 and 5), to high-grade advanced but still resectable disease (Phenotypes 4 and 3) and finally, metastatic disease (Phenotype 6). This data-driven ordering is consistent with established staging and treatment guidelines, supporting the clinical plausibility of the identified phenotypes [1, 3, 9]. The hint of beneficial systemic therapy highlights a potentially modifiable option for improving outcomes in patients who are otherwise resectable (see phenotype 1 and 5). Additionally, our findings support the development of a histopathological taxonomy that could improve

current treatment options for common and uncommon histological subtypes separately (see phenotype 2) [9].

This work has some limitations. First, our dataset only includes the most basic characteristics. In ongoing studies, variables could be expanded to include additional information about non-modifiable risk factors, such as family history or comorbidities, as well as modifiable risk factors related to lifestyle [1]. Second, our investigation of the association with 2-year event-free survival includes death, tumor progression, recurrence, newly developed distant metastasis or metastatic recurrence as events. This analysis could be more detailed to investigate the HR of each event individually. Lastly, an external validation is needed to determine the robustness and generalizability of our findings.

To conclude, we applied an unsupervised data-driven approach using baseline characteristics describing patient, tumor, and treatment information of PC patients. It is noteworthy that our analysis is conducted on a real-world dataset from the population-based CR-RLP that additionally includes clinical information including diagnosis, treatment and disease progression from the inpatient and outpatient sector [12], resulting in a extensive cohort of 1,158 PC patients. We identified six homogeneous phenotypes that capture existing clinical profiles and treatment guidelines. Most importantly, these phenotypes vary significantly in terms of two-year event-free survival, suggesting that they represent clinically meaningful PC profiles that could facilitate the development of novel therapeutic strategies and promote more informed treatment decisions, which are crucial for PC heterogeneity [1, 3, 9]. Future work should strengthen this analysis by including more detailed treatment information, particularly the specific surgical procedures performed and the sequencing of chemotherapy relative to surgery [3].

Declarations

Acknowledgments. We would like to thank the experts of the Cancer Registry of Rhineland-Palatinate in the Institute for Digital Health Data RLP, for the preparation and extraction of the registry data.

Funding. Not applicable.

Competing interests. The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval and consent to participate. Neither informed consent to participate nor Institutional Review Board statement is required for our investigation, because population-based reporting of cancer cases to state cancer registries is mandated by law in Germany by the German Cancer Early Detection and Registry Act (Krebsfrüherkennungs- und -registergesetz, §65c Sozialgesetzbuch V) and its regional implementation (Landeskrebsregistergesetz). In addition, analyses were done in a retrospective design using anonymized data. Administrative approval for access to anonymized patient data was granted by the Cancer Registry Rhineland-Palatinate. All procedures were conducted in accordance with the ethical principles of the Declaration of Helsinki.

Consent for publication. Not applicable.

Data availability. The dataset supporting the results of this study are available from the Cancer Registry Rhineland-Palatinate, but are not readily available due to legal restrictions under the [KFRG](#) and [LKRG](#). They are available from the authors upon reasonable request and with the permission of the Cancer Registry Rhineland-Palatinate, or can be requested directly at reporting@idg-rlp.de.

Author contribution. LS was responsible for the methodology, software, validation, formal analysis, data curation, visualization, and writing of the original draft.

CJ provided resources, especially the data management within the CR-RLP and was responsible for supervision. LS and CJ conceptualized, investigated, contributed to the review and editing of the manuscript and the project administration. Both read and approved the final manuscript.

References

- [1] Hu, J.-X. *et al.* Pancreatic cancer: A review of epidemiology, trend, and risk factors. *World journal of gastroenterology* **27**, 4298 (2021).
- [2] Sung, H. *et al.* Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **71**, 209–249 (2021).
- [3] Stoop, T. F. *et al.* Pancreatic cancer. *The Lancet* **405**, 1182–1202 (2025).
- [4] Ferlay, J., Partensky, C. & Bray, F. More deaths from pancreatic cancer than breast cancer in the eu by 2017. *Acta oncologica* **55**, 1158–1160 (2016).
- [5] Zhao, Z. & Liu, W. Pancreatic cancer: a review of risk factors, diagnosis, and treatment. *Technology in cancer research & treatment* **19**, 1533033820962117 (2020).
- [6] Torphy, R. J., Fujiwara, Y. & Schulick, R. D. Pancreatic cancer treatment: better, but a long way to go. *Surgery today* **50**, 1117–1125 (2020).
- [7] Kamisawa, T., Wood, L. D., Itoi, T. & Takaori, K. Pancreatic cancer. *The Lancet* **388**, 73–85 (2016).
- [8] Mizrahi, J. D., Surana, R., Valle, J. W. & Shroff, R. T. Pancreatic cancer. *The Lancet* **395**, 2008–2020 (2020).

- [9] Collisson, E. A. *et al.* Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nature medicine* **17**, 500–503 (2011).
- [10] Sinkala, M., Mulder, N. & Martin, D. Machine learning and network analyses reveal disease subtypes of pancreatic cancer and their molecular characteristics. *Scientific reports* **10**, 1212 (2020).
- [11] Stegmaier, C. *et al.* *Das Manual der Krebsregistrierung* (W. Zuckschwerdt Verlag München, 2019).
- [12] Katalinic, A. *et al.* Population-based clinical cancer registration in germany. *Cancers* **15**, 3934 (2023).
- [13] Tempero, M. A. *et al.* Pancreatic adenocarcinoma, version 2.2021, nccn clinical practice guidelines in oncology. *Journal of the National Comprehensive Cancer Network* **19**, 439–457 (2021).
- [14] Liao, W.-C. *et al.* Adjuvant treatments for resected pancreatic adenocarcinoma: a systematic review and network meta-analysis. *The Lancet Oncology* **14**, 1095–1103 (2013).
- [15] Mackay, T. M. *et al.* The risk of not receiving adjuvant chemotherapy after resection of pancreatic ductal adenocarcinoma: a nationwide analysis. *Hpb* **22**, 233–240 (2020).
- [16] Strobel, O., Neoptolemos, J., Jaeger, D. & Buechler, M. W. Optimizing the outcomes of pancreatic cancer surgery. *Nature reviews Clinical oncology* **16**, 11–26 (2019).
- [17] Agrawal, U. *et al.* Combining clustering and classification ensembles: A novel pipeline to identify breast cancer profiles. *Artificial intelligence in medicine* **97**,

27–37 (2019).

- [18] Li, L. & Wang, X. Identification of gastric cancer subtypes based on pathway clustering. *npj Precision Oncology* **5**, 46 (2021).
- [19] Burgel, P. R. *et al.* Clinical copd phenotypes: a novel approach using principal component and cluster analyses. *European Respiratory Journal* **36**, 531–539 (2010).
- [20] Iding, A. F. *et al.* Exploring phenotypes of deep vein thrombosis in relation to clinical outcomes beyond recurrence. *Journal of Thrombosis and Haemostasis* **21**, 1238–1247 (2023).
- [21] Robles, A. P. *et al.* Unsupervised clustering of venous thromboembolism patients by clinical features at presentation identifies novel endotypes that improve prognostic stratification. *Thrombosis Research* **227**, 71–81 (2023).
- [22] Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **58**, 236–244 (1963).
- [23] Murtagh, F. & Legendre, P. Ward’s hierarchical agglomerative clustering method: which algorithms implement ward’s criterion? *Journal of classification* **31**, 274–295 (2014).
- [24] Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. Nbclust: an r package for determining the relevant number of clusters in a data set. *Journal of statistical software* **61**, 1–36 (2014).
- [25] Dunn, J. C. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics* **4**, 95–104 (1974).

- [26] Gower, J. C. A general coefficient of similarity and some of its properties. *Biometrics* 857–871 (1971).
- [27] Kaufmann, L. & Rousseeuw, P. J. Finding groups in data: an introduction to cluster analysis. *New York: John Wiley* 274–295 (1990).
- [28] Rich, J. T. *et al.* A practical guide to understanding kaplan-meier curves. *Otolaryngology—Head and Neck Surgery* **143**, 331–336 (2010).
- [29] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 187–202 (1972).