Addressing Incomplete Data in Survival and Quality of Life Prediction: A Pancreatic Cancer Case Study

Aneta Lisowska[1,2✉], Floris den Hengst[1], Syed Ihtesham Hussain Shah[1], Annette ten Teije[1], Omar Bohoudi[3], Pauline Vissers[4,5], Mahsoem Ali[2], Laura Leeuwenburgh[2,4], Martijn Stommel[5], Amber Lamoré[2,6], Inez Verpalen[2,6], Marjolein Homs[7], Bas Groot Koerkamp[7], Judith De Vos-Geelen[8], Sven Mieog[9], Jeanin E. Van Hooft[9], Lois Daamen[10], Vincent E. De Meijer[11], Hanneke Wilmink[2,12], Hanneke van Laarhoven[2,12], Marc Besselink[2,13], and Geert Kazemier[2,14] for the Dutch Pancreatic Cancer Group

[1] Vrije Universiteit, Dept. Computer Science, Amsterdam, the Netherlands
a.j.lisowska@vu.nl
[2] Cancer Center Amsterdam, the Netherlands
[3] Amsterdam UMC, loc. VU, Dept. Radiotherapy, Amsterdam, the Netherlands
[4] IKNL, Dept. Research and Development, Utrecht, the Netherlands
[5] Radboud UMC, Dept. Surgery, Nijmegen, Netherlands;
[6] Amsterdam UMC, loc. UvA, Dept. Radiology, Amsterdam, the Netherlands
[7] Erasmus MC, Rotterdam, the Netherlands
[8] Maastricht UMC, Dept. Internal Medicine, GROW – Research Institute for Oncology & Reproduction, Maastricht, the Netherlands;
[9] Leiden UMC, Dept. Gastroenterology & Hepatology, Leiden, the Netherlands
[10] Utrecht UMC, Division of Imaging & Oncology, Utrecht, the Netherlands
[11] University of Groningen and UMC, Dept. Surgery, Groningen, the Netherlands
[12] Amsterdam UMC, loc. UvA, Dept.Medical Oncology, Amsterdam, the Netherlands
[13] Amsterdam UMC, loc. UvA, Dept. Surgery, Amsterdam, the Netherlands
[14] Amsterdam UMC, loc. VU, Dept. Surgery, Amsterdam, the Netherlands

Abstract. Models that predict the survival and quality of life (QoL) in patients with malignancies can support prognostic counselling and patient-centered evaluation of disease trajectories. Challenges in the development of such models include incomplete data in real-world registries, including missing baseline measurements and irregular, sparse availability of longitudinal QoL outcomes during treatment. These limitations complicate the incorporation of QoL into survival prediction and the reliable modeling of QoL over time. We address these limitations with a novel framework to predict survival and quality of life in the face of sparse data. The framework consists of a regularized missingness-avoiding random survival forest (MA-RSF) for survival prediction, and a conditional diffusion-based approach to predict QoL at various follow-up intervals. We apply the approach to a pancreatic cancer use-case, and evaluate it on a nationwide Dutch cohort. We find that our approach maintains high predictive accuracy (C-index 0.77) for survival prediction and minimizes reliance on sparse pre-treatment data, including QoL PROMs. In addition, we demonstrate that the QoL prediction

expressed as the Static-Dynamic diffusion architecture reduces trajectory prediction error (RMSE) at key follow-up intervals compared to standard conditional generation (e.g. 12.9 vs 14.0 RMSE at 7 months). The proposed framework facilitates the prediction of survival and QoL outcomes in settings characterized by incomplete data and could enhance prognostic counseling in patients with malignancies, including pancreatic cancer.

Keywords: Quality of Life · Survival Analysis · Generative Diffusion Models · Missing Data / Missingness-Avoidance · Pancreatic Cancer

1 Introduction

Shared decision-making for patients with poor prognosis involves a consideration of both therapeutic goals such as extending survival and palliative goals such as preserving quality of life (QoL). For example, the median overall survival duration for patients diagnosed with pancreatic ductal adenocarcinoma (PDAC) is 4 months and 5-year survival is as low as 4%, and QoL is an important factor in shared decision-making [5]. Additionally, QoL has been established as an independent predictor of overall survival [6]. QoL and its relation with survival are therefore becoming increasingly important in shared decision-making.

Treatment planning in settings with a poor survival outlook is characterized by weighing potential survival benefits of interventions against their associated impact on QoL [9]. Prognostic survival models have been developed to assist with patient counselling with time-to-event analysis at Concordance Index (C-index) values in the range of 0.70–0.73 for overall survival [2, 4], and have demonstrated improved accuracy over traditional staging systems. However, these models often do not incorporate QoL in their predictions, limiting their usefulness for holistic shared decision-making.

QoL is typically quantified using patient-reported outcome measures (PROMs) that are obtained directly from patients via questionnaires. Gold standards for QoL PROMs have long been established for a multitude of conditions [1]. However, as these instruments rely on active patient participation, they suffer from limited longitudinal availability. Specifically, patients can be enrolled at various phases of their care journey which results in missing pre-treatment QoL measures, and measures at irregular intervals during treatment.

These characteristics are illustrated by the Dutch Pancreatic Cancer Project (PACAP), which collects clinical data and patient-reported outcome measures (PROMs) from a multicenter cohort of patients with PDAC [7]. The resulting sparsity at both the onset of and during treatment provides technical challenges for standard prognostic modeling techniques.

In this paper, we propose a joint modeling framework to predict survival and QoL in the face of incomplete data. The framework integrates two components:

Missingness-Aware Random Survival Forest (MA-RSF) which employs RSF-specific regularization during node splitting to minimize reliance on features with high missingness. [15]

Static-Dynamic Architecture a ConDOR [3] adaptation that explicitly disentangles conditioning factors, restricting the Regional Diffusion Model (RDM) to time-invariant patient covariates and the Temporal Diffusion Model (TDM) to dynamic treatment impulses, to better capture treatment-associated temporal patterns in the QoL progression.

We validate our framework through a clinical case study. Results demonstrate that MA-RSF maintains high survival prediction accuracy with lower reliance on sparse pre-treatment data. Furthermore, the Static-Dynamic diffusion architecture significantly reduces QoL trajectory prediction error (RMSE) at key follow-up intervals compared to static conditional generation.

2 Methodology

This section details our contributions to the challenge of predicting survival and QoL outcomes in the face of data incompleteness. First, we address missing explanatory variables prior to treatment with a random forest robust to missing data, including QoL, clinical, and demographics data. Second, we address missing QoL data during the treatment with a generative conditional diffusion model.

2.1 Missingness Avoidance Random Survival Forest

We address missing pretreatment explanatory variables for survival prediction by extending random survival forests (RSF) with techniques from the missingness-avoidance (MA) paradigm, motivated by Stempfle et al. who argued that existing approaches, such as mean imputation or missingness indicators, often introduce bias or obscure model interpretability [8]. We first provide a high-level overview of RSF and then describe the extension to the MA framework.

The RSF algorithm constructs an ensemble of de-correlated decision trees, where each tree is grown through a recursive process of node splitting. At every node, the algorithm evaluates a subset of candidate features to find the optimal data partition that maximizes the survival difference between the resulting child nodes. In traditional RSF, this is achieved via a log-rank split criterion $L_{\mathrm{LR}}(j, \tau)$ for feature j and split thresholds τ, which calculates a test statistic to identify the feature and threshold that best separate distinct survival distributions for the data subset associated with the current node.

We adopt the MA paradigm to survival analysis by extending its log-rank split criterion with an additional term to minimize reliance on sparse pretreatment variables. Our approach penalizes the use of missing data in its splitting criterion by using what we refer to as the missingness-aware log-rank criterion:

$$L_{\mathrm{MALR}}(j, \tau) := L_{\mathrm{LR}}(j, \tau) - \alpha N_{\mathrm{missing},j} \tag{1}$$

[15] We release the open-source Python package https://pypi.org/project/masurvival/

where j and τ are as before, and $N_{\text{missing},j}$ the number of samples in the current node with a missing value for feature j, and $\alpha \geq 0$ a regularization hyperparameter to control the penalty strength. The algorithm identifies a model within the Rashomon set that optimizes predictive accuracy while minimizing reliance on sparse variables. While this missingness-avoiding regularization mitigates instability from Missing At Random patterns, and α ensures that informative Missing Not At Random predictors are retained if their utility outweighs the risk to model reliability.

2.2 Longitudinal QoL Prediction with Conditional Diffusion

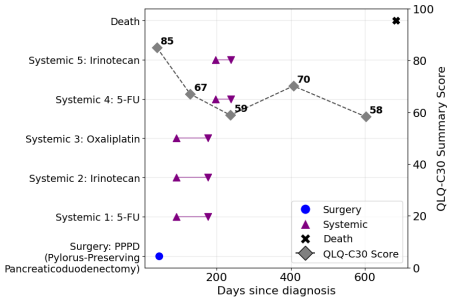
Longitudinal QoL prediction is complicated by non-linear dynamics arising from the stochastic interplay between disease progression and interventions. Conventional imputation, such as linear interpolation, fails to characterize these fluctuations, while standard forecasting architectures like RNNs and Transformers are typically optimized for discrete, equidistant sequences. In clinical settings, assessments often occur at irregular intervals. To address the resulting temporal sparsity, we utilize ConDOR (Conditional Diffusion with Ordinal Regression)[3].

Our framework instantiates ConDOR as a two-stage conditional diffusion model and introduces a static–dynamic decomposition. A first denoising network (RDM) estimates the pretreatment QoL state x_0 at $t = 0$:, $x_0 = g_\phi(s)$, based on static covariates s (e.g., demographics and pretreatment laboratory measurements), providing a patient-specific initialization. A second denoising network (TDM) models longitudinal state updates from t to $t + 1$: $x_{t+1} = x_t + f_\theta(x_t, t, s, u_t)$, where u_t represents time-varying treatment factors at time t, including intervention type (e.g., surgery/systemic therapy), treatment initiation indicators, and elapsed time since treatment start. Therefore, Stage 2 learns how treatment exposure modifies the QoL trajectory over time, conditional on the patient baseline context s. The two-stage design separates (i) baseline heterogeneity across patients (captured by $g_\phi(s)$) from (ii) treatment-driven temporal evolution (captured by $f_\theta(x_t, t, s, u_t)$), enabling generation of individualized QoL trajectories under different intervention schedules.

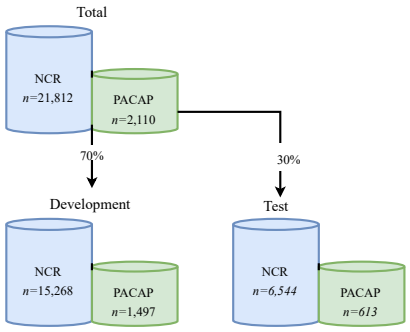
Unlike the original ConDOR model, we drop the cohort-level priors and rely solely on individual trajectories to prevent sparse-data averages from smoothing over acute changes stemming from interventions.

3 Experimental Setup

A retrospective cohort study is used with data from two merged national Dutch sources. The study period includes patients diagnosed between January 1, 2015, and December 31, 2024. Clinical, demographic, and tumor-related data are extracted from the Netherlands Cancer Registry (NCR), a comprehensive, nationwide population-based registry. Longitudinal QoL PROMs are obtained from PACAP [10] using the EORTC QLQ-C30 questionnaire [1].



(a) **Example patient records.** The gray squares represent time points at which the QoL PROM was completed by the patient (from PACAP). Treatment time points, here surgery and systemic therapy, are taken from the NCR registry.



(b) **Development and test splits.** PACAP is a subset of NCR, n the number of patients.

Fig. 1: Visualisation of NCR registry joined with PACAP self-reported QoL data.

The PACAP cohort represents a 9.8% subset of the full NCR registry (See Figure 1b). Merging these datasets enables the integration of static pretreatment clinical variables with dynamic longitudinal QoL trajectories as depicted in Figure 1a. The 2,110 patients in the PACAP development set provide a total of 5,441 PROMs throughout their care trajectory, with a mean of 3.6 responses per patient (median 3, max 15). Both selection bias (due to informed consent for PACAP-PROMS) and response bias make it unlikely that the PACAP-PROMS cohort is representative of the entire NCR cohort. The study uses two outcomes defined as follows:

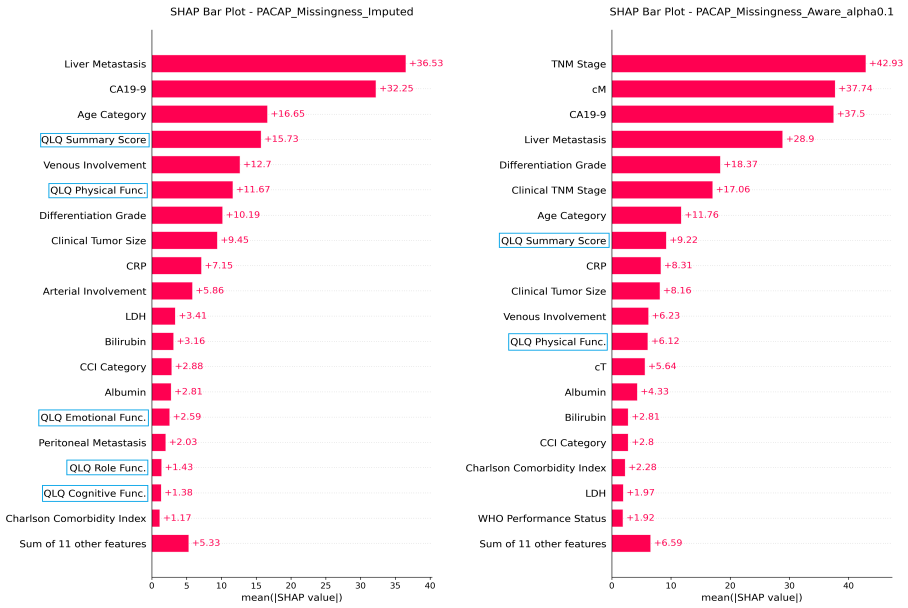
Overall Survival (OS) : the time from diagnosis to death from any cause,

Quality of Life (QoL) : the EORTC QLQ-C30 Summary Score [1], calculated following EORTC guidelines as the mean of 13 component scales, with symptom scales reversed so that higher scores indicate better QoL consistently.

Models were trained to predict these outcomes using 5-fold cross-validation on the development set. For MA-RSF, we additionally performed a 10-fold parameter search, identifying $\alpha = 0.1$ as the optimal regularization value. Performance was evaluated via the C-index and RMSE on held-out test sets stratified by cohort (PACAP vs. Full NCR) to assess generalization to broader populations.

3.1 Data Pre-processing and Feature Representation

For the MA-RSF survival prediction model, static pretreatment variables captured at diagnosis were included, including demographics (age, sex), WHO performance status, comorbidities (Charlson Comorbidity Index), laboratory values (e.g., CA19-9, Albumin, CRP), and tumor characteristics (e.g. metastasis status, vasculature involvement). The functioning and summary scores of the QLQ-C30



(a) RSF (Missing Imputed), $\alpha = 0.0$ (b) MA-RSF, $\alpha = 0.1$

Fig. 2: Feature importance (SHAP) comparison between standard imputation and missingness-avoiding regularization ($\alpha = 0.1$) on the PACAP test set.

were included where available. In the PACAP development cohort, 9% lacked valid PROMs. Baseline PROMs were defined as the first valid PROM prior to treatment initiation, or the first valid PROM overall for untreated patients. Baseline PROMs were available for 47% of patients and occurred on average 20 days after diagnosis. Features were organized for the static-dynamic ConDOR model for QoL trajectory prediction as follows. Both models were conditioned on time since diagnosis (continuous) and clinical tumor stage (ordinal). For the *static* diffusion model that predicts pretreatment baselines, time-invariant features including age, sex, season of diagnosis, CCI, and pretreatment laboratory values were used. For the *dynamic* model a time-dependent vector capturing treatment events was used. Both surgery and systemic therapy were encoded using: 1) a binary feature indicating treatment occurrence; 2) a *cumulative status* (post-event binary flag); and 3) a *time-since-event* counter to model non-linear recovery.

3.2 Survival Experiment Design

MA-RSF was evaluated against three common strategies: 1) **Missing Vars Dropped** (exclusion of features with missing values); 2) **Missing Imputed** (mean/mode imputation); and 3) **Missing Indicated** (sentinel constant -999).

Table 1: Survival Prediction Performance on the test set (C-index ↑ [std]). The best and second-best performance within each training–test set block are highlighted.

Training set	Missingness handling strategy	PACAP	NCR
PACAP	Missing Vars Dropped	0.691 [0.002]	0.664 [0.003]
	Missing Imputed	0.705 [0.002]	0.696 [0.005]
	Missing Indicated	0.704 [0.002]	0.693 [0.005]
	Missing Avoiding	0.706 [0.002]	0.701 [0.003]
	Missing Vars Dropped + Treatment	0.760 [0.002]	0.731 [0.002]
	Missing Imputed + Treatment	0.760 [0.002]	0.749 [0.006]
	Missing Indicated + Treatment	0.760 [0.002]	0.740 [0.009]
	Missing Avoiding + Treatment	0.761 [0.002]	0.748 [0.005]
NCR	Missing Vars Dropped	0.701 [0.003]	0.683 [0.001]
	Missing Imputed	0.729 [0.002]	0.744 [0.000]
	Missing Indicated	0.727 [0.002]	0.743 [0.001]
	Missing Avoiding	0.731 [0.002]	0.744 [0.001]
	Missing Vars Dropped + Treatment	0.769 [0.001]	0.730 [0.001]
	Missing Imputed + Treatment	0.775 [0.002]	0.772 [0.001]
	Missing Indicated + Treatment	0.774 [0.001]	0.770 [0.001]
	Missing Avoiding + Treatment	0.775 [0.001]	0.770 [0.001]

3.3 QoL Synthesis Design

The diffusion networks were implemented in PyTorch and optimized jointly using Adam with a decoupled learning rate strategy (1×10^{-3} for RDM; 5×10^{-4} for TDM) to stabilize training on sparse treatment signals. The objective function minimized Mean Squared Error (MSE) for noise prediction (RDM) and state transitions (TDM). Early stopping was applied based on validation loss within a 5-fold cross-validation scheme. [16]

To quantify architectural contributions, we compared four configurations in an *ablation study*: (i) **Baseline**, conditioned solely on tumor stage and time since diagnosis; (ii) **RDM (Clinical)**, adding static pretreatment features to the RDM; (iii) **TDM (Treatment)**, adding dynamic treatment vectors to the TDM; and (iv) **Static-Dynamic (Full)**, the integrated architecture utilizing demographic/clinical static features and dynamic treatment impulses.

4 Results

4.1 Overall Survival.

We validated the effectiveness of the MA penalty by analyzing feature importance (SHAP) distributions (Figure 2). In the unregularized model ($\alpha = 0.0$), the QLQ Summary Score ranked as the fourth most important predictor. Conversely, the MA-RSF ($\alpha = 0.1$) successfully deprioritized this frequently missing variable, dropping it to eighth place while maintaining predictive performance.

[16] https://github.com/anli66/panc-surv-qol

Table 2: QoL Synthesis Ablation (RMSE).

Config.	T_0	Full
Baseline	25.60	12.85
RDM (Clin.)	25.10	13.50
TDM (Treat.)	26.37	12.58
Full Model (S-D)	26.26	12.20

Table 3: Monthly RMSE: Baseline vs. S-D

Month	Base.	S-D	p-val
1	25.99	24.53	0.149
3	14.07	13.97	0.711
4	13.41	12.90	0.005^{**}
6	14.18	13.80	0.014^{*}
7	14.01	12.92	0.006^{**}
10	13.15	11.89	0.052
13	11.93	10.83	0.031^{*}

Notably, the importance of variables with low missingness (e.g., arterial involvement, 3.5% missing) was also reduced, indicating the regularization successfully isolates the most robust core predictors.

Table 1 reports predictive performance. As anticipated, dropping incomplete variables significantly degraded performance (paired Student's t-test on aligned fold-specific C-index $p < 0.01$). While standard imputation performed comparably to the missingness-aware (MA) approach within the training distribution (PACAP), the MA-RSF showed more robust generalization to the broader NCR cohort, achieving competitive accuracy while reducing reliance on sparsely observed inputs. However, this advantage diminishes when treatment variables are introduced or when the model is trained on the full NCR registry. The differences between MA-RSF and Imputation are not significant.

4.2 Longitudinal QoL.

Ablation results (Table 2) confirm the efficacy of the Static-Dynamic architecture. While paired Student's t-tests showed no significant differences in baseline (T_0) prediction errors, the S-D configuration (RDM Clinical + TDM Treatment) achieved the lowest overall trajectory error. This full model significantly outperformed the RDM (Clin.) ($p = 0.03$) and demonstrated a borderline-significant improvement over the baseline ($p = 0.05$). For the PACAP study, questionnaires are sent every three months; our analysis focused on months in the test set where at least 100 patients responded. Monthly RMSEs were obtained from 5-fold CV on the Black Envelope set, and Baseline vs. Static–Dynamic (S-D) differences were assessed with paired two-sided t-tests on fold-matched RMSEs (n=5 folds), since errors are not independent across models. Table 3 shows significant improvements ($p < 0.05$) at 4, 6, 7, and 13 months. Across-month p-values are correlated and should be interpreted as exploratory.

5 Discussion and Conclusion

This study addresses the challenges of baseline feature missingness and longitudinal sparsity in oncological survival and QoL modeling. Our results demonstrate that incorporating missingness-avoidance regularization into survival forests (MA-RSF) effectively shifts model reliance away from sparse baseline variables toward robust clinical variables. Unlike standard imputation, which risks introducing

bias when data are structurally missing (e.g., patients referred late in the care pathway), the MA approach learns a valid decision policy within the Rashomon set that prioritizes observable features. Consequently, the model maintains generalization performance across the broader NCR population where pretreatment QoL is frequently unavailable.

For longitudinal prediction, the proposed Static-Dynamic adaptation of CondOR offers a methodological advancement over static conditional diffusion. By architecturally disentangling static patient phenotypes from dynamic treatment impulses, the model is constrained to learn state transitions as responses to interventions rather than demographic correlations. This yielded improvements in RMSE at 4, 7, and 10 months—critical windows for evaluating treatment efficacy. However, the absolute RMSE values remain relatively high, reflecting the inherent stochasticity of self-reported well-being and the complexity of unmeasured confounders in registry data.

Limitations: Our validation is restricted to a Dutch cohort, which may limit generalizability to healthcare systems with different care pathways. In addition, the PACAP-PROMS cohort is subject to both selection bias and response bias, which together reduce the representativeness of the QoL trajectory relative to the full population of patients. Our current framework also relies on pretreatment clinical variables that are updated only by treatment events for the primary tumor and does not incorporate time-varying clinical biomarkers (e.g., longitudinal CA19-9 or albumin), which could provide important physiological context to further refine trajectory predictions.

Conclusion: We present a robust pipeline for PDAC prognosis that navigates the irregularities of real-world clinical data. By harmonizing missingness-avoidant survival analysis with conditional QoL trajectory generation, this framework moves closer to enabling personalized, holistic risk assessment in oncology.

Declaration of AI assistance technologies during manuscript preparation. The authors used Google Gemini for AI Activity Class 1 and Cursor for AI Activity Classes 6 and 7. The authors reviewed the results and take full responsibility for the contents of the manuscript.

Acknowledgments This work has been partly supported by the PersOn project (P21-03) and the Hybrid Intelligence center (grant no. 024.004.022), both funded by Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).

References

1. Aaronson, N.K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N.J., Filiberti, A., Flechtner, H., Fleishman, S.B., Haes, J.C.d., et al.: The european organization for research and treatment of cancer qlq-c30: a quality-of-life instrument for use in international clinical trials in oncology. JNCI: Journal of the National Cancer Institute **85**(5), 365–376 (1993)
2. van den Boorn, H.G., Dijksterhuis, W.P., van der Geest, L.G., de Vos-Geelen, J., Besselink, M.G., Wilmink, J.W., van Oijen, M.G., van Laarhoven, H.W.: Sourcepanc: a prediction model for patients with metastatic pancreatic ductal adeno-

carcinoma based on nationwide population-based data. Journal of the National Comprehensive Cancer Network **19**(9), 1045–1053 (2021)
3. Cho, H., Wei, Z., Lee, S., Dan, T., Wu, G., Kim, W.H.: Conditional diffusion with ordinal regression: Longitudinal data generation for neurodegenerative disease studies. In: The Thirteenth International Conference on Learning Representations (2025)
4. Deng, Q.L., Dong, S., Wang, L., Zhang, C.Y., Ying, H.F., Li, Z.S., Shen, X.H., Guo, Y.B., Meng, Z.Q., Yu, J.M., et al.: Development and validation of a nomogram for predicting survival in patients with advanced pancreatic ductal adenocarcinoma. Scientific reports **7**(1), 11524 (2017)
5. Latenstein, A.E., van der Geest, L.G., Bonsing, B.A., Koerkamp, B.G., Mohammad, N.H., de Hingh, I.H., de Meijer, V.E., Molenaar, I.Q., van Santvoort, H.C., van Tienhoven, G., et al.: Nationwide trends in incidence, treatment and survival of pancreatic ductal adenocarcinoma. European Journal of Cancer **125**, 83–93 (2020)
6. Mackay, T.M., Latenstein, A.E., Sprangers, M.A., Van Der Geest, L.G., Creemers, G.J., Van Dieren, S., De Groot, J.w.B., Koerkamp, B.G., De Hingh, I.H., Homs, M.Y., et al.: Relationship between quality of life and survival in patients with pancreatic and periampullary cancer: a multicenter cohort analysis. Journal of the National Comprehensive Cancer Network **18**(10), 1354–1363 (2020)
7. Seelen, L.W., Augustinus, S., Stoop, T.F., Bouwense, S.A., Busch, O.R., Cirkel, G.A., van Eijck, C.H., de Vos-Geelen, J., Koerkamp, B.G., Mohammad, N.H., et al.: Quality of life among patients with locally advanced pancreatic cancer: A prospective nationwide multicenter study. Journal of the National Comprehensive Cancer Network **23**(3), 97–104 (2025)
8. Stempfle, L., Matsson, A., Mwai, N., Johansson, F.D.: Prediction models that learn to avoid missing values. In: Forty-second International Conference on Machine Learning (2025)
9. Valverde, C.P.T., Ebrahimi, G., Sprangers, M.A., Pateras, K., Bruynzeel, A.M., Jacobs, M., Wilmink, J.W., Besselink, M.G., Crezee, H., van Tienhoven, G., et al.: Impact of short-course palliative radiation therapy on pancreatic cancer-related pain: prospective phase 2 nonrandomized painpanc trial. International Journal of Radiation Oncology* Biology* Physics **118**(2), 352–361 (2024)
10. Van Den Braak, R.C., Van Rijssen, L., Van Kleef, J., Vink, G., Berbee, M., Henegouwen, M.V.B., Bloemendal, H., Bruno, M., Burgmans, M., Busch, O., et al.: Nationwide comprehensive gastro-intestinal cancer cohorts: the 3p initiative. Acta Oncologica **57**(2), 195–202 (2018)