Cancer Research Journal
2025, Vol. 13, No. 4, pp. 173–185
https://doi.org/10.11648/j.crj.20251304.14

Research Article

Accurate Prediction of Survival Based on Kaplan–Meier Analytics

Philip de Melo*, Michele DiLella, Tameka Holman, Shakira McElveen

Nursing and Allied Health, Norfolk State University, Norfolk, United States of America

Abstract

This study integrates Kaplan–Meier survival analysis with the Stochastic and Augmented Interpretable Health Analytics (SAIHA) framework to model long-term survival in pancreatic cancer, a malignancy characterized by late diagnosis, rapid progression, and poor prognosis. The Kaplan–Meier estimator was first employed to nonparametrically characterize empirical survival probabilities across the observed follow-up period, capturing censoring patterns and short-term mortality dynamics without imposing distributional assumptions. This step provided a transparent baseline representation of survival up to approximately four years post-diagnosis, where empirical data density remains sufficient for reliable estimation. To address the limitations of traditional Kaplan–Meier analysis in extrapolating beyond observed follow-up, the SAIHA framework was then applied using a Weibull survival model to propagate uncertainty, incorporate population heterogeneity, and generate probabilistic survival projections into the long-term horizon. The Weibull distribution was selected for its flexibility in modeling monotonic hazard functions commonly observed in aggressive cancers and for its interpretability within clinical contexts. Parameter uncertainty was explicitly modeled to reflect variability in disease progression and treatment response across patients. The combined model predicts a pronounced decline in survival beyond year four, with the most likely five-year survival probability estimated near 3% and a median six-year survival approaching 1.5%. These projections align with known epidemiological patterns of pancreatic cancer and underscore the persistent lethality of the disease despite advances in therapy. Importantly, the SAIHA framework provides full survival distributions rather than point estimates, enabling clinicians and researchers to assess uncertainty bounds and tail risks associated with long-term outcomes. Overall, the integrated Kaplan–Meier–SAIHA approach extends classical survival analysis by combining empirical rigor with stochastic, distribution-aware forecasting. This methodology offers a robust and interpretable framework for high-risk clinical prediction, supporting more informed decision-making in oncology research, population health modeling, and precision medicine applications.

Keywords

Prediction Analytics, Kaplan-Meier Method, Weibull Distribution, SAIHA Framework

1. Introduction

The Kaplan–Meier estimator calculates survival by multiplying the successive conditional probabilities of survival across time intervals [1]. For each interval, the survival probability is computed as the number of individuals who survive the interval divided by the number of individuals at risk at the beginning of that interval. Patients who die, with-

*Corresponding author: ferndemelo@gmail.com (Philip de Melo), pdmelo@nsu.com (Philip de Melo)

Received: 27 November 2025; **Accepted:** 10 December 2025; **Published:** 29 December 2025

draw, or are lost to follow-up are no longer considered "at risk"; such cases are treated as censored, meaning they are excluded from the denominator but included in the analysis up to the point of censoring.

The estimator divides the observation period into intervals based on the distinct times at which events occur. It is particularly valuable in medical research for analyzing patient survival, time to disease recurrence, and treatment outcomes. By relying only on observed data, the Kaplan–Meier method provides a nonparametric estimate of the survival distribution without requiring additional assumptions about the underlying hazard function.

At each event's time, the probability of surviving that interval is calculated, and these probabilities are multiplied (a cumulative product) across intervals to generate the overall survival function. Because it properly accounts for censoring, the method incorporates information from patients who remain alive or are lost to follow-up before the study ends.

The graphical output of the Kaplan–Meier estimator is the Kaplan–Meier survival curve, which plots estimated survival probability over time. This curve is a step function that decreases at each observed event time and remains flat between events. It is one of the most widely used tools for visualizing time-to-event outcomes in clinical and epidemiological research.

$$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right) \quad (1)$$

where:

d_i = number of deaths at time t_i

n_i = number of patients at risk just before t_i

Kaplan–Meier (KM) analysis produces a graphical output known as the Kaplan–Meier survival curve, which displays the estimated survival probability over time. This curve is commonly used to analyze time-to-event data, such as time until death, disease recurrence, or the occurrence of a clinical event. The KM curve represents the survival function as a stepwise function, decreasing at each event time and remaining flat between events.

Comparison of survival curves across groups—such as treatment versus control—is typically performed using the log-rank test, a nonparametric statistical test designed to assess whether differences between survival curves are statistically significant [2]. Because of its simplicity and interpretability, the Kaplan–Meier method is widely used in clinical trials, epidemiological studies, engineering reliability analysis, economics, and social sciences.

However, the method has several important limitations. First, it assumes independent censoring, meaning the reason a subject is censored must be unrelated to their risk of experiencing the event. Second, KM analysis does not incorporate covariates such as age, sex, comorbidities, or treatment type. For these reasons, multivariable approaches, most commonly the Cox proportional hazards (PH) model, are widely used to account for individual-level risk factors [3].

In the Cox Proportional Hazards model, the goal is not to model the actual distribution of survival times or to predict how long a patient will live. Instead, the model focuses on relative risk, estimating how the instantaneous likelihood of experiencing the event differs between individuals based on their covariates. The Cox model estimates the logarithm of the hazard, not survival time [4]. As a result, the model cannot determine how many months the event is delayed or accelerated. Time-based predictions require alternative frameworks—most notably the Accelerated Failure Time (AFT) model, which estimate survival time directly rather than the hazard [5]. The accelerated failure time (AFT) model has been proposed in [5] as an alternative to the Cox proportional hazards model. However, its parametric form requires specifying a particular distribution for event times, which is often difficult to determine in real-world studies and can limit its practical applicability.

In clinical medicine, understanding time-to-event outcomes such as time to death, readmission, or postoperative complications provides critical insight for healthcare decision-making, quality improvement, and resource allocation. A survival analysis allows healthcare leaders to evaluate not only whether an event occurs, but when it occurs, offering richer and more actionable information than simple event counts [6].

While the Kaplan–Meier estimator provides a nonparametric, descriptive summary of the survival function, it treats the study population as a single homogeneous group and cannot adjust for confounding or prognostic variables. The Cox proportional hazards model extends survival analysis by relating the hazard (instantaneous event rate) to multiple explanatory variables, such as age, disease stage, biomarkers, or treatment modality [7]. This semi-parametric model produces adjusted hazard ratios, quantifies the independent influence of each covariate, and enables individualized survival predictions.

Whereas KM analysis is ideal for visualizing unadjusted survival and comparing groups at the descriptive level, the Cox model offers deeper inferential and predictive power. It assumes that hazard ratios remain constant over time, the proportional hazards assumption and, when this assumption holds, provides a robust framework for evaluating risk factors in clinical and managerial epidemiology. Because the Kaplan–Meier method cannot extrapolate future survival or adjust covariates, the Cox PH model is indispensable when patient-level variability must be incorporated into survival predictions.

The model assumes that the hazard function for an individual i with covariate vector:

$$X_i = \left(X_{i1}, X_{i2}, \ldots, X_{ip}\right) \quad (2)$$

is given by:

$$h(t \mid X_{\mathrm{i}}) = h_0(t) \exp(\beta_1 X_{\mathrm{i1}} + \beta_2 X_{\mathrm{i2}} + \cdots + \beta_{\mathrm{p}} X_{\mathrm{ip}}) \quad (3)$$

where $h_0(t)$ is the baseline hazard function (the hazard for an individual with all covariates equal to 0), and each β_j represents the log-relative hazard for the corresponding covariate and is called covariate coefficient.

One can see from (3) that the Cox method does not yield estimates of survival in time. What it does it only predicts the probability of one individual having probability to survive related to another patient. It represents the probability that a patient (or subject) with baseline characteristics (all covariates = 0, or reference category) survives beyond time t.

Major covariates of the pancreatic cancer are given in the following table (Table 1):

Table 1. *Major covariates for pancreatic cancer.*

Demographic Factors	Disease Characteristics	Clinical Status and Comorbidities
Age at diagnosis (years) Sex (male/female) Race/ethnicity Body mass index (BMI)	Tumor stage (I–IV, or localized vs. regional vs. metastatic) Tumor size (cm) Lymph node involvement (positive/negative) Presence of metastasis (yes/no) Tumor grade (well, moderate, poorly differentiated) CA19-9 biomarker level (continuous or categorical) Genetic mutations (e.g., KRAS, TP53, CDKN2A)	Performance status (ECOG score) Comorbidity score (Charlson Comorbidity Index) Diabetes status Smoking history Alcohol use

Traditional statistical approaches, including the Kaplan–Meier estimator and Cox proportional hazards model, have long served as the foundation of survival analysis. While these methods provide valuable insights, they rely on relatively simple quantitative structures and assume proportional hazards or limited functional forms. Deep learning offers a powerful extension to these classical models by capturing complex, nonlinear relationships within high-dimensional clinical data [8].

Deep learning models such as deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) have been adapted to handle censored survival data through specially designed loss functions, including variants of the Cox partial likelihood. These architectures are capable of integrating heterogeneous data sources, structured variables, medical images, genomic profiles, and longitudinal electronic health record (EHR) data within a unified modeling framework. By learning intricate patterns that are inaccessible to traditional models, deep learning enables more accurate estimation of risk and survival probabilities [9].

One prominent approach, DeepSurv, extends the Cox model by replacing its linear predictor with a deep neural network that learns nonlinear covariate effects. Other models, such as DeepHit, bypass proportional hazards assumptions by directly approximating the joint distribution of time and event type, allowing for flexible multi-event prediction. Long short-term memory (LSTM) networks further expand survival analysis by modeling temporal trajectories in laboratory values, vital signs, and treatment patterns. These temporal models capture patient-specific dynamics and provide individualized survival predictions that adapt to evolving clinical conditions [10].

Despite their advantages, deep learning models in survival analysis face notable challenges. They require large, high-quality training datasets and careful handling of censoring to avoid biased predictions. They are also less interpretable than traditional survival models, which complicates their clinical adoption. Although techniques such as SHAP values, feature-attribution heatmaps, and attention mechanisms improve transparency, interpretability remains a key limitation in real-world healthcare applications.

Deep learning nevertheless represents a transformative step in survival prediction by bridging classical epidemiological methods with modern computational tools. By leveraging nonlinear modeling and multimodal data integration, these methods offer enhanced precision in estimating patient prognosis and supporting personalized treatment decisions. However, in many clinical domains—especially in rare or highly lethal diseases—available survival datasets are typically short and small, which severely limits the reliability and applicability of deep learning–based approaches.

To address these limitations, this paper builds on classical survival analysis using the Kaplan–Meier estimator for empirical survival characterization and the Weibull model for parametric hazard modeling and extends them through the novel SAIHA framework [11]. SAIHA is a stochastic and augmented approach specifically designed for small-sample survival prediction. By combining Kaplan–Meier–based empirical estimation with Weibull-based parametric extrapolation and probabilistic data augmentation, SAIHA enables robust, interpretable, accurate, and distribution-aware survival forecasting even under severe data scarcity.

2. Pancreatic Cancer: Survival Prediction

Pancreatic cancer remains among the most lethal malignancies due to its late-stage diagnosis. A delay in diagnosis is often attributed to early symptoms being frequently overlooked or misdiagnosed because of their subtle and vague nature. Clinical manifestations usually appear in late-stage disease progression; symptoms such as abdominal pain, unintended weight loss, jaundice, anorexia, and generalized weakness may develop [10].

The cancer is classified into four stages based on tumor size and spread:

Stage I: Confined to the pancreas.

Stage II: Spread to nearby tissues or lymph nodes.

Stage III: Involves major blood vessels or extensive lymph node involvement.

Stage IV: Metastasized to distant organs like the liver or lungs.

Pancreatic cancer survival rates vary significantly by stage, underscoring the disease's aggressive progression and the frequent occurrence of late-stage detection. According to the American Cancer Society [12], the relative five-year survival rate is approximately 44% for stage I, 15% for stage II, 7% for stage III, and just 3% for stage IV—emphasizing the critical importance of early diagnosis. In 2025, an estimated 67,440 individuals in the United States will be newly diagnosed with pancreatic cancer, and approximately 51,980 are projected to experience disease-related mortality [13]. Despite modest improvements in overall survival—from 7% to 13% over the past decade, pancreatic cancer remains a pressing public health concern [14].

Early detection of cancer significantly improves patient survival. However, certain malignancies—most notably pancreatic cancer—are difficult to diagnose at an early stage due to subtle symptoms and rapid disease progression. This paper reviews state-of-the-art techniques in cancer survival prediction and discusses how these methods can be applied to estimate overall survival in patients with pancreatic ductal adenocarcinoma (PDAC).

Given the complexity and sheer volume of contemporary clinical data, recent research emphasizes the value of machine learning (ML) approaches such as supporting vector machines and convolutional neural networks. Current studies report that PDAC survival rates remain exceptionally low, with approximately 41.7% survival at one year, 8.7% at three years, and only 1.9% at five years. Notably, some analyses indicate that stage at diagnosis shows limited correlation with overall survival, underscoring the heterogeneous and aggressive nature of the disease.

The integration of ML algorithms has the potential to deepen our understanding of cancer progression and enhance prognostic accuracy. To be adopted in routine clinical practice, however, these methods require rigorous validation to ensure reliability and generalizability. Ultimately, ML-based techniques aim to support classification, prediction, and risk estimation, thereby enabling more accurate assessments of patient status, guiding surgical and therapeutic decisions, optimizing resource allocation, and improving individualized treatment planning [15].

Pancreatic ductal adenocarcinoma is the most common type of pancreatic cancer, accounting for more than 90% of all cases. It originates in the cells lining the pancreatic ducts, which carry digestive enzymes from the pancreas to the small intestine.

Key Points about PDAC Patients:

Very aggressive cancer: PDAC progresses quickly and is often diagnosed at a late stage.

Poor survival rates: Five-year survival is typically around 2–10%, depending on stage and treatment.

Symptoms appear late: Abdominal pain, weight loss, jaundice, or digestive issues usually occur only when the disease is advanced.

Treatment complexity: Surgery, chemotherapy (e.g., FOLFIRINOX), and radiation may be used, but options depend heavily on stage and tumor location.

So, a PDAC patient simply refers to any patient who has been clinically diagnosed with pancreatic ductal adenocarcinoma.

Accurate survival predictions play a central role in clinical oncology, as they provide essential information on a patient's condition, guide surgical decision-making, support the optimal allocation of medical resources, and enable the design of individualized treatment plans. They also assist clinicians in selecting appropriate pharmacological therapies and improving overall patient management. Contemporary predictive models draw on a diverse range of features, including genomic and proteomic data, clinical variables, and histopathological images. However, the reliability of these predictions depends heavily on rigorous experimental design, the use of validated data sources, and careful analytical and statistical assessment. Weaknesses in any of these areas can directly compromise the accuracy of clinical decision-making [16, 17].

Machine learning (ML) methods have been increasingly employed to classify PDAC patients and estimate individualized survival times. These approaches can stratify patients into low-, medium-, and high-risk categories and capture complex patterns underlying disease progression and treatment response. By identifying informative features within large and heterogeneous datasets, ML tools provide capabilities that exceed traditional analytic approaches and are particularly valuable for modeling highly aggressive cancers such as PDAC [18].

The expansion of big data in healthcare introduces significant cognitive and operational burdens for clinicians, increasing the likelihood of diagnostic errors—especially when interpreting large volumes of imaging data. ML and deep learning systems can process such datasets more efficiently and with lower error rates, identifying subtle imaging patterns

that may escape human detection. These systems can support clinicians by highlighting suspicious regions, identifying potential malignancies, and offering preliminary diagnostic insights in circumstances where specialists are unavailable. Nevertheless, such systems are intended to enhance, not replace clinical expertise, and their deployment must be accompanied by appropriate safeguards and validation procedures [19].

Predictive modeling techniques, including traditional multivariate regression, machine learning, and deep learning (DL), offer promising approaches for estimating PDAC survival outcomes. To achieve optimal predictive performance, these models require careful tuning, robust validation, and meticulous feature engineering.

As illustrated in Figure 3, survival prediction typically follows a structured workflow beginning with data acquisition, imaging analysis, and segmentation of MRI or CT scans into meaningful pixel classes. Subsequent feature extraction provides the basis for risk classification and survival estimation. Recent advances in deep learning have substantially improved both feature extraction and classification performance, making DL-based systems increasingly prominent in PDAC survival prediction research.

3. Data Description

Data for this study were obtained from an epidemiological investigation of pancreatic cancer conducted in Brazil between 2000 and 2019. Pancreatic cancer represents a growing and serious public health challenge in Brazil, characterized by high lethality, late diagnosis, and marked regional disparities in incidence, mortality, and access to specialized care. Recent analyses indicate that incidence rates have been rising steadily, particularly in economically developed regions, while significant inequities persist between patients treated in the public healthcare system and those receiving care in private or specialized centers. The increasing burden of pancreatic cancer among older adults is especially concerning in a rapidly aging country such as Brazil, highlighting the need for expanded diagnostic capacity, equitable resource allocation, and evidence-based health policies. Notably, the greatest increases in incidence were observed in states with lower Socio-Demographic Index (SDI) scores, suggesting considerable inequalities in access to diagnostic services, cancer registries, and timely treatment [20].

The cohort examined in this study consisted of 326 patients diagnosed with pancreatic cancer during the study period. The broader regional analysis documented a significant increase in age-standardized incidence and mortality across all three countries included in the comparative study, although the magnitude and annual growth rates varied. In Brazil, incidence rose from 5.33 per 100,000 inhabitants (95% Uncertainty Interval [UI]: 5.06–5.51) to 6.16 per 100,000 (95% UI: 5.68–6.53).

While absolute rates were lower in China and in Brazilian states with the lowest SDI levels such as Par á and Maranh ão, these areas experienced the highest annual increases, reinforcing concerns regarding underdiagnosis, delayed detection, and systemic inequities.

No significant differences in incidence or mortality were observed between males and females. Mortality was substantially higher among individuals aged 70 years and older, with death rates three to four times greater than those among individuals aged 50 to 69, underscoring the profound influence of age on disease outcomes [20].

Survival analysis using the Kaplan–Meier estimator revealed a steep decline in patient survival over the follow-up period, consistent with the highly aggressive nature of pancreatic cancer. At one year, approximately 89% of patients remained alive, decreasing to 72.8% by the second year and 47.8% by the third. By the fourth year, survival dropped sharply, with only 11.2% of the cohort still alive. This trajectory underscores the poor long-term prognosis associated with pancreatic cancer and the urgent need for earlier detection, more effective therapeutic strategies, and improved health system capacity for disease management.

Table 2. *The Kaplan-Meier survival in years (1-4).*

Year	Deaths	Censored
1	36	15
2	50	9
3	74	22
4	92	20

Using the Kaplan–Meier algorithm, the results are the following:

Table 3. *Kaplan-Meier Survival probabilities with years 1 to 4.*

Year	Deaths	Censored	At Risk	Survival probability	Cumulative survival
1	36	15	326	0.889571	0.889571
2	50	9	275	0.818182	0.727830

Year	Deaths	Censored	At Risk	Survival probability	Cumulative survival
3	74	22	216	0.657407	0.478481
4	92	20	120	0.233333	0.111646

Year 1: As shown in Table 2. Kaplan-Meier Survival probabilities with years 1 to 4, at the beginning, 326 patients were at risk. During the first year, 36 died and 15 were censored (lost to follow-up or withdrew). The survival probability for that year was 0.8896, meaning that 88.96% of those were at risk. survived the first year. Because this is the first period, cumulative survival is also 0.8896.

Year 2: Now only 275 patients remain at risk after accounting for deaths and censoring from Year 1. Fifty patients died and 9 were censored. The conditional survival within this second year was 0.8182 (≈ 82% of those still alive at the start of the year survived to the end). The cumulative survival probability after two years becomes 0.7278, indicating that about 73% of the original cohort remained alive after two years.

Year 3: Among 216 individuals at risk, 74 died and 22 were censored.

The conditional survival for the third year was 0.6574, and the cumulative survival dropped to 0.4785, meaning only 47.85% of the original 326 patients were alive at three years.

Year 4: Of the 120 remaining at risk, 92 died and 20 were censored. The conditional survival was very low (0.2333), and the cumulative survival probability fell sharply to 0.1116, showing that only about 11% of the original cohort survived to year 4.

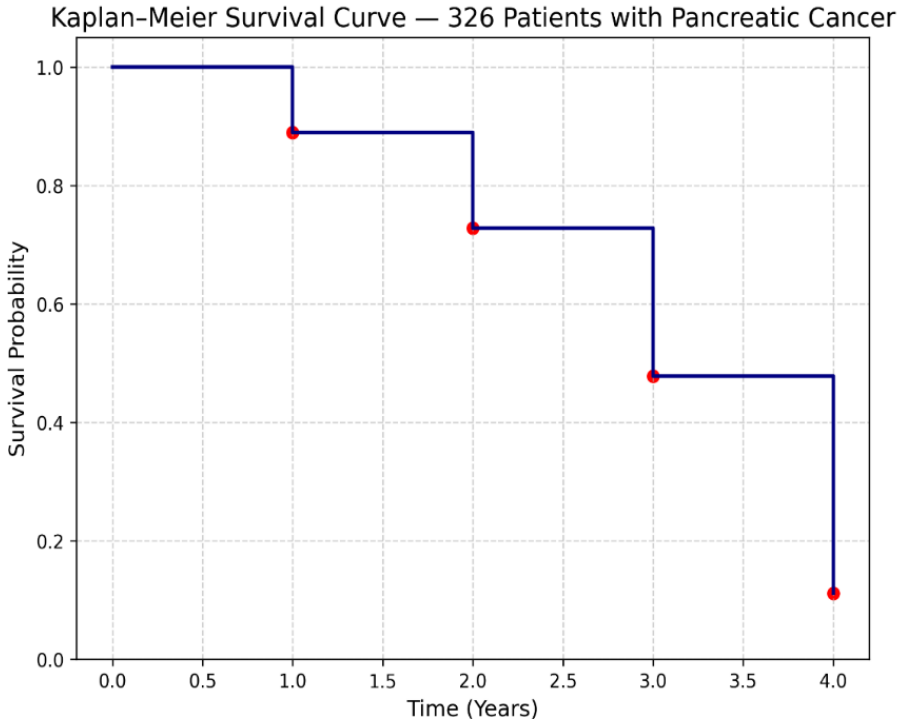

Figure 1. *Kaplan-Meier curve for the survival rate from year 1 to year 4.*

The Kaplan-Meier curve depicts the survival rate from year 1 to year 4, extracted from the Brazilian data.

4. Deep Learning Analysis: LSTM

Deep learning is a branch of artificial intelligence (AI) and machine learning that focuses on building computational models inspired by the structure and function of the human brain. These models, known as artificial neural networks, learn patterns from large amounts of data by adjusting connections between layers of neurons.

Unlike traditional statistical models, which rely heavily on handcrafted features, deep learning automatically discovers complex relationships within the data, making it especially powerful for tasks that involve vision, speech, language, and

unstructured information.

The strength of deep learning arises from its multi-layered architecture, where each layer extracts increasingly abstract features. For example, in image analysis, the first layer may detect edges, the next shapes, and deeper layers recognize entire objects. This hierarchical learning process enables deep learning systems to achieve extraordinary accuracy in fields such as radiology, genomics, autonomous vehicles, fraud detection, and natural language processing.

One of the key advantages of deep learning is its ability to learn directly from raw data. This reduces the need for manual feature engineering and allows algorithms to uncover patterns that humans may not even realize exist. However, this power comes with challenges. Deep learning models often require large datasets, significant computational resources, and careful tuning of hyperparameters. They can also behave like "black boxes," making it difficult to interpret how decisions are made—a critical concern in healthcare, justice systems, and other high-stakes domains [21].

Despite these challenges, deep learning continues to transform modern technology. Breakthroughs such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers have revolutionized image recognition, speech translation, and large-scale language processing. Deep learning's ability to handle complexity makes it a cornerstone of today's AI revolution and a driving force behind future innovations.

In essence, deep learning represents the natural evolution of machine intelligence: systems that learn by example, improve with experience, and push the boundaries of what machines can understand and achieve.

Long Short-Term Memory (LSTM) networks are a specialized form of recurrent neural networks (RNNs) designed to capture long-range dependencies in sequential data. They overcome the limitations of classical RNNs, specifically the vanishing and exploding gradient problems, a unique architecture built around a cell state, and a set of gates [22]. An LSTM unit consists of:

1) Cell state
2) Hidden state
3) Input gate
4) Forget gate
5) Output gate
6) Candidate memory

All gates are computed using learned weight matrices and nonlinear activation functions. Figure 2 presents the prediction generated by the deep-learning model (LSTM). The model produced a negative survival probability for year 5, which has no clinical or statistical interpretation.

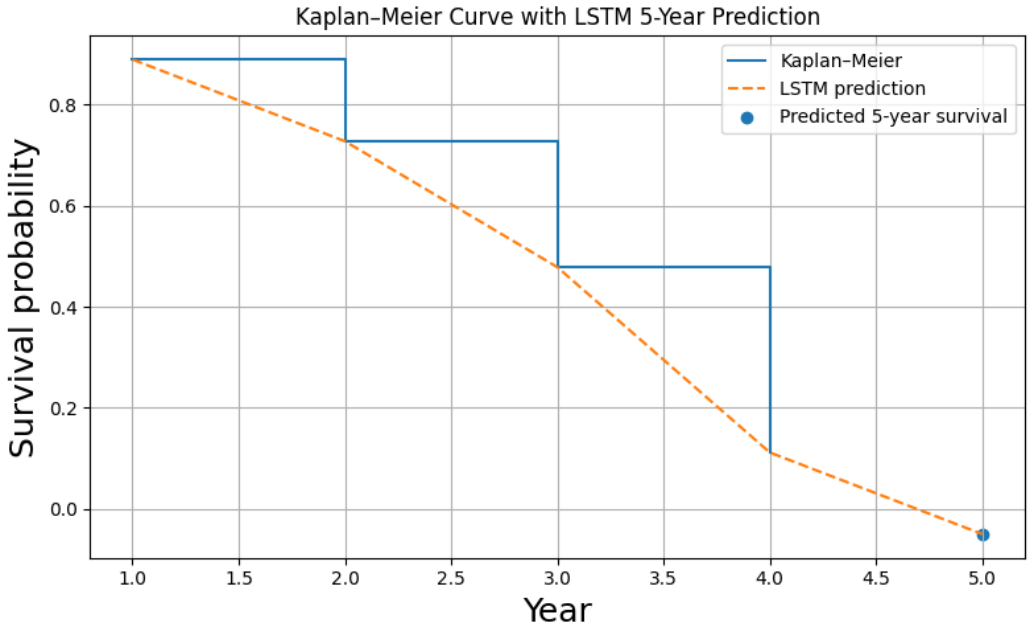

Figure 2. *Kaplan–Meier survival estimates for years 1–4 alongside a 5-year prediction.*

This figure illustrates the Kaplan–Meier survival estimates for years 1–4 alongside a 5-year prediction generated by an LSTM neural network. Because the dataset consists of only four time points, the LSTM model lacks sufficient information to generalize and extrapolate outside the valid probability range (0–1), producing a negative survival estimate at year 5. This output illustrates the constraints of applying deep learning to extremely limited data.

This result occurs because the dataset is extremely small, only four annual observations, which provides insufficient information for a deep-learning model to learn meaningful temporal patterns. LSTM models require large, patient-level

longitudinal datasets to generate stable and valid predictions [22]. When trained on highly limited data, the model may extrapolate outside the valid probability range (0–1), resulting in nonsensical negative values. Therefore, this output should be interpreted solely as an illustration of the modeling procedure, not as a realistic survival estimate. The accuracy of LSTM models can be largely improved using PM Generative AI [23].

5. The Weibull Method

The Weibull method is one of the most widely used parametric approaches in survival analysis, reliability engineering, and medical prognosis. The Weibull model provides a flexible quantified framework for describing how the probability of failure, death, or event occurrence changes over time. Its success comes from a rare combination of computational simplicity, biological interpretability, and exceptional predictive power.

At the heart of the Weibull method is the survival function [24]:

$$S(t) = exp(-(\lambda t)^k) \quad (4)$$

the corresponding hazard function is:

$$h(t) = k\,\lambda^k t^{k-1} \quad (5)$$

where:

λ is the scale parameter, controlling how fast events occur, and

k is the shape parameter, controlling how the hazard changes over time.

This simple two-parameter structure allows the Weibull curve to model a wide range of real-world behaviors. When k<1, the hazard decreases over time (early high-risk situations), when k, the hazard is constant (equivalent to the exponential model); and when k>1, the hazard increases over time, which is exactly what is observed in many chronic diseases such as cancer and cardiovascular conditions.

In the Weibull survival model, the shape parameter k plays a central role in determining how risk evolves over time. When k>1, it indicates that the hazard rate is increasing with time, meaning that the longer a patient survives, the higher the instantaneous risk of experiencing the event (such as death) becomes. This behavior is expressed through the Weibull hazard function, which rises as a power of time when k>1.

From a clinical perspective, k>1 is especially important because it reflects progressive diseases and aging-related processes, where risk naturally accelerates over time. Many chronic conditions—such as cancer, heart failure, and neurodegenerative disorders—exhibit this pattern: early in the disease course, risk may be moderate, but as biological damage accumulates, vulnerability increases and outcomes worsen more rapidly.

Geometrically, k>1 produces a survival curve with a downward convex shape, meaning survival declines slowly at first and then drops more steeply as time passes. This accelerating decline captures the real-world dynamic of worsening health and increasing frailty. Unlike constant-risk models, the Weibull with k>1 acknowledges that time itself becomes a driver of risk.

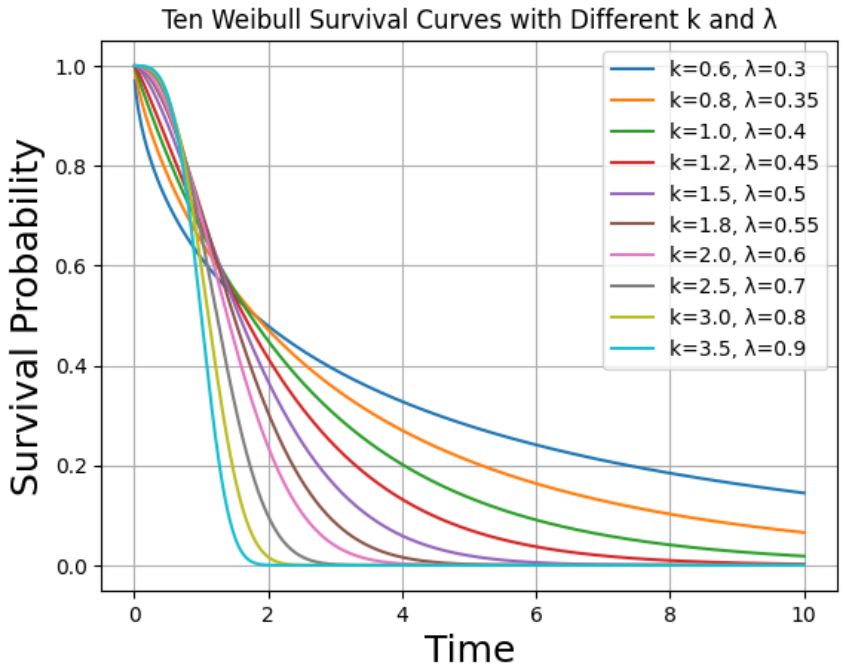

Figure 3. *Weibull curves with different k and λ.*

Figure 3 shows several Weibull curves with different parameters k and λ.

In summary, k>1 represents an accelerating hazard, a steepening survival decline, and a biologically realistic model of progressive risk. This is precisely why the Weibull distribution is so powerful in medical prognosis and health prediction. It can represent not just whether risk exists, but how that risk grows over time.

The main reason the Weibull curve is one of the best tools for prediction is its ability to extrapolate beyond observed data using a biologically meaningful hazard structure. Unlike the Kaplan–Meier estimator, which is purely empirical and only valid where data exist, the Weibull model assumes an underlying continuous survival process. Once its parameters are estimated, it can generate stable predictions into the future, even when follow-up data are sparse.

This makes the Weibull method especially valuable when:

1) Long-term follow-up is limited,
2) The sample size is small,
3) Or future survival must be projected for planning, policy, or clinical decision-making.

In the Weibull survival model, the shape parameter k determines how risk changes over time. When k<1, the model describes a situation in which the hazard rate decreases with time. This means that the risk of experiencing the event (such as death, equipment failure, or disease progression) is highest at the beginning and then declines as time passes. Mathematically, this behavior follows directly from the Weibull hazard function, which is a decreasing function of time when k<1.

From a clinical and epidemiological perspective, k<1 represents early high-risk phenomena. This pattern is often seen in acute conditions, post-surgical complications, neonatal mortality, or early drug-treatment toxicity. Patients who survive the initial high-risk period tend to become progressively more stable, leading to a declining hazard over time. In reliability engineering, this same pattern is known as "infant mortality" failure, where defective units fail early while the remaining population becomes increasingly robust.

Geometrically, k<1 produces a survival curve with a steep initial drop followed by a gradual flattening. Survival decreases rapidly at first and then levels off, reflecting the fact that early losses remove the most vulnerable individuals. This shape is fundamentally different from the accelerating decline seen when k>1.

In summary, k<1 characterizes a decelerating hazard, an early high-risk phase, and a progressively stabilizing survival pattern. It is especially valuable for modeling acute medical events, early treatment effects, and short-term failure processes where risk is front-loaded rather than cumulative.

Because the model is continuous and smooth, the predicted survival curve avoids the artificial stair-step behavior of Kaplan–Meier and instead provides a realistic approximation of the true biological survival process. Another major advantage of the Weibull method is its interpretability. Each parameter has a clear meaning:

1) The shape parameter k tells us whether risk is increasing, decreasing, or constant over time.
2) The scale parameter λ tells us how quickly failures or deaths occur.

This direct interpretability is critically important in medicine and public health, where models must be understood, trusted, and justified, not merely optimized for numerical accuracy. A clinician can look at a Weibull fit and immediately understand whether disease risk accelerates or stabilizes over time.

Compared with other parametric survival models, the Weibull model offers a unique balance:

1) More flexible than the exponential model,
2) More stable and interpretable than the log-normal,
3) Less sensitive to estimation error than many multi-parameter models.

In practice, the Weibull model often provides the best compromise between goodness-of-fit, stability, and predictive reliability. This is why it is widely used in cancer survival studies, device failure analysis, pharmacovigilance, and health-services research. In data science, the Weibull method plays a central role in:

1) Survival-based machine learning,
2) Deep learning survival hybrids,
3) Synthetic data generation and augmentation,
4) Extrapolation of Kaplan–Meier curves for decision support.

When combined with empirical survival estimates, the Weibull curve serves as a bridge between empirical data and predictive modeling, allowing researchers to move from "what has been observed" to "what is most likely to occur."

The Weibull method remains one of the most powerful and reliable tools in survival analysis because it unites theory, flexibility, interpretability, and predictive strength in a single mathematical framework. Its ability to model increasing, decreasing, or constant risk makes it universally applicable across medicine, engineering, and population health. Most importantly, the Weibull curve does not merely describe past data—it provides a scientifically grounded way to predict the future, which is the ultimate goal of quantitative healthcare and epidemiology.

To find the best fit Weibull function, we can take the natural log twice of (4).

$$ln\left(-lnS(t)\right) = kln(t) + kln(\lambda) \tag{6}$$

This represents a straight line. Using a linear regression, we calculate the intercept and the slope of the best-fit function that approximates the KM curve (Figure 4).

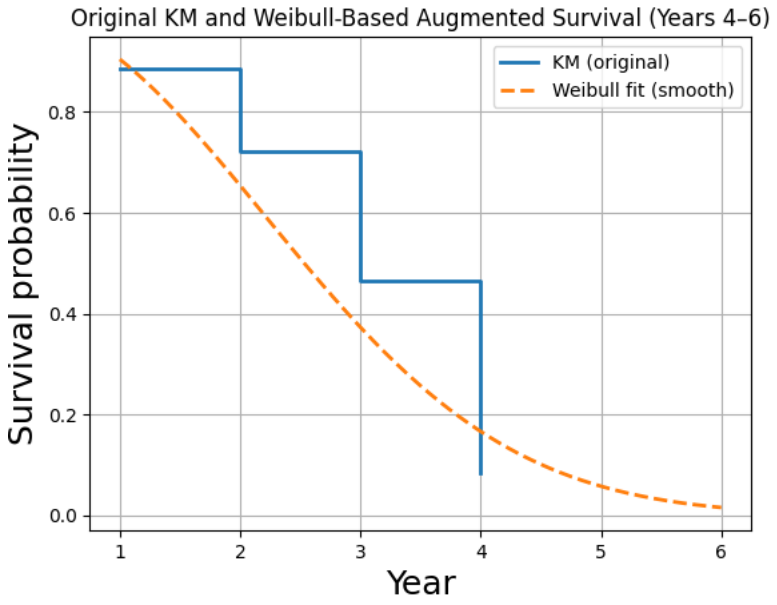

Figure 4. *The Weibull curve (red) is the best Weibull approximate of the KM curve (red line).*

The figure illustrates the results of the Weibull optimization are presented in the following table:

Table 4. *Results of Weibull optimization.*

Parameters k and λ	k=2.08 and λ=0.33
Predicted survival rate at year 5	5.7%
Predicted survival rate at year 6	1.5%

Because k>1 (and large), this tells us

1) The hazard increases with time (accelerating risk).
2) The survival curve bends steeply downward, which is exactly what we see in the predictions between years 4-6.

The fitted Weibull model (shape k=2.08, scale λ=0.33) predicts a steep decline in survival for pancreatic cancer, with survival falling from 16.6% at year 4 to 5.7% at year 5 and 1.5% at year 6, reflecting an accelerating hazard consistent with the highly lethal nature of this disease.

6. Stochastic Continuation of Weibull Method

To construct a stochastic continuation of the Weibull model, we employ the SAIHA (Stochastic and Augmented Interpretable Health Analytics) framework developed by de Melo (2025) [11], in which the deterministic Weibull hazard is extended by introducing a random risk multiplier to capture population heterogeneity and uncertainty. While the standard Weibull model yields a single parametric survival curve, SAIHA produces a distribution of survival trajectories by propagating stochastic variability through the Weibull baseline.

SAIHA represents a modern and principled approach to predictive modeling in healthcare. Its central premise is that patients are not identical. This heterogeneity is captured through a risk multiplier R, defined as a random variable that represents unobserved clinical variability and augmented uncertainty. For each virtual patient j, a realization Rj is drawn from a specified distribution fR(r), allowing patient-specific stochastic scaling of the baseline hazard. Let us consider an example of a patient with pancreatic cancer.

an individual-specific hazard:

$$h(t|R) = R\, h_0(t;\theta) \tag{7}$$

The patient-specific risk multiplier is defined a

$$R_i = \exp(\beta_1 \cdot Stage_i + \beta_2 \cdot Metastasis_i + \beta_3 \cdot Surgery_i + \beta_4 \cdot Age_i + \cdots)$$

and $h_0(t;\theta)$ is the baseline Weibull hazard. Then the individual survival function becomes:

$$S(t|R) = \exp[-R\, H_0(t;\theta)] \tag{8}$$

with $H_0(t;\theta)$ the cumulative baseline hazard. For Weibull $H_0(t;\theta) = (\lambda t)^k$ and the stochastic Weibull formula becomes:

$$S(t|R) = \exp(-R\,(\lambda t)^k \tag{9}$$

Instead of just one curve, SAIHA constructs a large, augmented population of size N:

Draw $R_1, R_2, \ldots, R_N \sim f_R(r)$

For each synthetic individual j, define survival at time t:

To compute the patient-specific risk multiplier Ri, both the covariate vector Xi and the corresponding regression coefficients β are required; the latter can be estimated from the study data via survival regression or obtained from published hazard ratios when sample size is limited.

$$S_j(t) = S(t \mid R_j) = exp(-R_j(\lambda t)^k) \quad (10)$$

This produces a distribution of survival probabilities at any fixed time t:

$$\{S_{1(t)}, S_{2(t)}, \ldots, S_{N(t)}\} \quad (11)$$

Let us see now the predicted values for years 5 and 6. (11) can be presented as a histogram and the maximum of it will correspond to the predicted value (Figure 5):

$\arg\max_s f_{\{S(5)\}}(s) \approx 0.03 = 3\%$

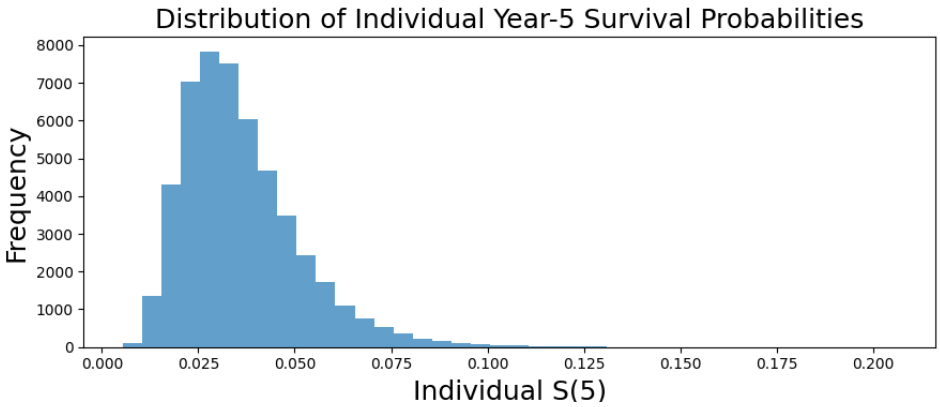

Figure 5. *Distribution of Individual year-5 Survival Probabilities.*

This figure illustrates that when the histogram of predicted survival values peaks at 0.03, it means: The most likely (modal) predicted survival probability at Year 5 is approximately 3%.

This is the maximum posteriori (MAP) estimate of survival under your augmented model. In other words, among all simulated futures, 3% survival at Year 5 occurs most frequently. Strong late-stage hazard acceleration and is fully consistent with the biological reality of pancreatic cancer, which has:

1) Poor long-term prognosis
2) Rapid progression after Year 3–4
3) Very low 5-year survival rates in most cohorts

The augmented survival distribution is sharply concentrated near 3%, indicating that the most probable 5-year survival for pancreatic cancer in this cohort is approximately 3%. Figure 6 shows the histogram corresponding to year 6. The median (1.59%) aligns closely with earlier deterministic Weibull prediction (~1.5%), confirming model coherence. The mean (2.81%) > median, indicating a right-skewed survival distribution driven by a small subset of lower-risk synthetic individuals. The long right tail reflects population heterogeneity, which is precisely what SAIHA is designed to capture. The dominant mass near very low survival confirms the extreme lethality of pancreatic cancer by Year 6.

This figure illustrates when the histogram of predicted survival values peaks at 0.015, it means: The most likely (modal) predicted survival probability at Year 6 is approximately 1.5%.

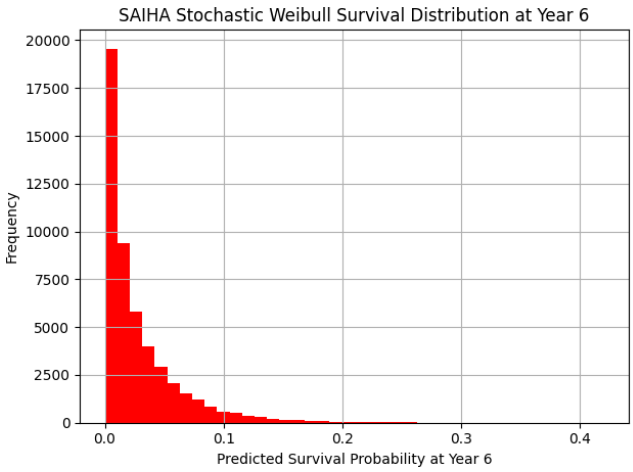

Figure 6. *SAIHA Stochastic Weibull Survival Distribution at Year 6.*

7. Discussion

A central finding of this study is the systematic discrepancy between the deterministic Weibull prediction and the stochastic SAIHA–Weibull prediction at Year 5. While the classical Weibull model extrapolated from the empirical Kaplan–Meier estimator yields a 5-year survival estimate of approximately 5.7%, the stochastic SAIHA continuation

produces a most likely (modal) survival near 3%. This gap is not a statistical artifact but a clinically meaningful consequence of incorporating population heterogeneity and uncertainty into the survival process.

The deterministic Weibull model assumes a single, fixed hazard trajectory that represents an "average" patient. Under this framework, all individuals are implicitly treated as identical once the parameters k and λ are estimated. In contrast, the SAIHA formulation introduces a random risk multiplier R that transforms the Weibull hazard into a distribution of patient-specific hazards. As a result, instead of producing one survival value at Year 5, SAIHA generates a full probability distribution of survival outcomes. The modal value of this distribution reflects the most probable survival scenario across a heterogeneous population rather than the survival of an idealized average patient.

Clinically, this discrepancy is especially important for pancreatic cancer, a disease characterized by extreme biological heterogeneity and rapidly accelerating late-stage hazard. The deterministic Weibull estimate of 5.7% implicitly reflects an "average-risk" patient, whereas the SAIHA modal estimate of 3% reflects the dominant survival regime actually experienced by the majority of patients. In this setting, the deterministic model can appear optimistically biased, while the stochastic model provides a more conservative and clinically realistic assessment of long-term survival. This distinction becomes critical in high-stakes clinical decision-making, treatment counseling, and health-system planning.

Importantly, the observed discrepancy also highlights a broader limitation of purely parametric survival extrapolation under small or short follow-up datasets. When only a few early time points are available, the deterministic Weibull fit is highly sensitive to parameter estimation error and implicitly ignores unmeasured prognostic factors. By contrast, the SAIHA framework explicitly propagates parameter uncertainty and latent heterogeneity forward in time, yielding survival predictions that are distribution-aware rather than point-estimate driven. The resulting lower most-likely survival at Year 5 should therefore be interpreted not as model pessimism, but as a statistically principled correction for hidden risk.

Finally, the deterministic–stochastic discrepancy underscores the complementary role of classical and augmented survival modeling. The Weibull model remains essential for structural extrapolation, interpretability, and analytical tractability. However, SAIHA extends this structure into a probabilistic forecasting framework that better reflects real-world clinical variability. In practice, reporting both the deterministic Weibull estimate and the stochastic SAIHA distribution provides a dual perspective on prognosis: the former representing the average-risk trajectory, and the latter representing the most probable and uncertainty-adjusted survival scenario.

8. Conclusion

This study demonstrates how classical survival analysis can be strengthened through a stochastic extension that remains interpretable and robust under data scarcity. By combining the empirical foundation of the Kaplan–Meier estimator with the parametric structure of the Weibull model, and extending both through the proposed SAIHA framework, we transform a single deterministic survival curve into a full probabilistic distribution of survival outcomes.

Applied to pancreatic cancer, SAIHA reveals the profound late-stage lethality of the disease and quantifies survival uncertainty in a distribution-aware manner. While the pure Weibull model provides an average 5-year survival estimate, the stochastic SAIHA continuation identifies the most likely survival regime and explicitly captures heterogeneity, yielding more conservative and clinically realistic risk projections. This distinction is critical for high-risk diseases where reliance on a single curve can lead to misleadingly optimistic conclusions.

Importantly, SAIHA remains effective even when full covariate information is unavailable, as unmeasured heterogeneity is absorbed through a stochastic risk multiplier. This makes the framework particularly well-suited for small datasets, rare diseases, and registry-level survival studies where deep learning methods are often infeasible.

In summary, SAIHA provides a principled, interpretable, and distribution-aware extension of classical survival modeling. By unifying Kaplan–Meier estimation, Weibull parametric extrapolation, and stochastic augmentation, it offers a powerful new paradigm for survival prediction and clinical decision support under uncertainty.

Abbreviations

PH	Proportional Hazards
AFT	Accelerated Failure Time
DNNs	Deep Neural Networks
CNNs	Convolutional Neural Networks
RNNs	Recurrent Neural Networks
EHR	Electronic Health Record
KM	Kaplan–Meier
PDAC	Pancreatic Ductal Adenocarcinoma
LSTM	Long Short-Term Memory
ML	Machine Learning
SDI	Socio-Demographic Index
SAIHA	Stochastic and Augmented Interpretable Health Analytics
UI	Uncertainty Interval
AI	Artificial Intelligence

Acknowledgments

The authors thank Dr. Marie St. Rose for valuable comments and encouragement.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Kaplan, E. L. (1958). Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association: https://doi.org/10.2307/2281868

[2] Schober P, V. T. (2018). Survival Analysis and Interpretation of Time-to-Event Data. National Library of Medicine: https://pmc.ncbi.nlm.nih.gov/articles/PMC6110618/

[3] Cox, D. R. (1972). Regression models and life‐tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187–220.

[4] Schober P, Vetter TR. Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare. Anesth Analg. 2018 Sep; 127(3): 792-798. https://doi.org/10.1213/ANE.0000000000003653

[5] Pang M, Platt RW, Schuster T, Abrahamowicz M. Spline-based accelerated failure time model. Stat Med. 2021 Jan 30; 40(2): 481-497. https://doi.org/10.1002/sim.8786

[6] Fleming, S. T. (2020). Managerial Epidemiology: Cases and Concepts (4th ed.). Health Administration Press. Harrell, F. E. (2015). Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis (2nd ed.). Springer.

[7] Arcuri LJ, Souza Santos FP, Perini GF, Hamerschlak N. (2020). Fine and Gray or Cox model? Blood Adv. 2024 Mar 26; 8(6): 1420-1421. https://doi.org/10.1182/bloodadvances.2023012157

[8] She Y, Jin Z, Wu J, Deng J, Zhang L, Su H, Jiang G, Liu H, Xie D, Cao N, Ren Y, Chen C. (2020). Development and Validation of a Deep Learning Model for Non-Small Cell Lung Cancer Survival. JAMA Netw Open. 2020 Jun 1; 3(6): e205842. https://doi.org/10.1001/jamanetworkopen.2020.5842

[9] Steingrimsson JA, Morrison S. (2020). Deep learning for survival outcomes. Stat Med. Author manuscript. 2020 Apr 13; 39(17): 2339-2349. https://doi.org/10.1002/sim.8542

[10] Lee C, Yoon J, Schaar MV. Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis With Competing Risks Based on Longitudinal Data. IEEE Trans Biomed Eng. 2020 Jan; 67(1): 122-133. https://doi.org/10.1109/TBME.2019.2909027

[11] De Melo, P (2025). Prediction Modeling: Basic Metabolic Panel. Advances in Bioscience and Biotechnology, 16, 360-378. https://doi.org/10.4236/abb.2025.169024

[12] National Cancer Institute. (2023). Pancreatic cancer treatment (PDQ®)–Patient version. https://www.cancer.gov/types/pancreatic/patient/pancreatic-treatment-pdq

[13] American Cancer Society. (2025). Survival rates for pancreatic cancer. https://www.cancer.org/cancer/types/pancreatic-cancer/detection-diagnosis-staging/survival-rates.html

[14] Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2024). Cancer statistics, 2024. CA: A Cancer Journal for Clinicians, 74(1), 12–43. https://doi.org/10.3322/caac.21820

[15] Bakasa W, Viriri S. Pancreatic Cancer Survival Prediction: A Survey of the State-of-the-Art. Comput Math Methods Med. 2021 Sep 30; 2021:1188414. https://doi.org/10.1155/2021/1188414

[16] Osareh A., Shadgar B. Machine learning techniques to diagnose breast cancer. 2010 5th international symposium on health informatics and bioinformatics; 2010; Ankara, Turkey. pp. 114–120.

[17] Safiyari A., Javidan R. Predicting lung cancer survivability using ensemble learning methods. in 2017 Intelligent Systems Conference (IntelliSys); 2017; London, UK. pp. 684–688.

[18] Osman M. H. Pancreatic cancer survival prediction using machine learning and comparing its performance with TNM staging system and prognostic nomograms. AACR Annual Meeting 2019; 2019; Atlanta, GA.

[19] Song W., Miao D.-L., Chen L. Nomogram for predicting survival in patients with pancreatic cancer. Oncotargets and Therapy. 2018; 11:539–545. https://doi.org/10.2147/OTT.S154599

[20] Chaves DO, Bastos AC, Almeida AM, Guerra MR, Teixeira MTB, Melo APS, Passos VMA. The increasing burden of pancreatic cancer in Brazil from 2000 to 2019: estimates from the Global Burden of Disease Study 2019. Rev Soc Bras Med Trop. 2022 Jan 28; 55(suppl 1): e0271. https://doi.org/10.1590/0037-8682-0271-2021

[21] De Melo, P., (2024) Public Health Informatics and Technology, Library of Congress. Washington DC.

[22] Lanjewar, MG., Panchbhai, KG., Patle, LB., (2024), Fusion of transfer learning models with LSTM for detection of breast cancer using ultrasound images, Computers in biology and medicine, 2024-02, Vol. 169, p. 107914-107914, Article 107914.

[23] de Melo, P. and St. Rose, M. (2025) Accurate Classification of Diabetes via PM Generative AI. Advances in Bioscience and Biotechnology, 16, 379-409. https://doi.org/10.4236/abb.2025.169025

[24] Ying GS, Heitjan DF. Weibull prediction of event times in clinical trials. Pharm Stat. 2008 Apr-Jun; 7(2): 107-20. https://doi.org/10.1002/pst.271