Research Article ISSN 2639-8478

Cancer Science & Research

From Population Health Data to Personalized Medicine: Breast Cancer Survival Prediction

Philip De Melo* and Audrey Douglas-Cooke

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

ORCID: 0009-0003-3607-137x

***Correspondence:**
Philip De Melo, Department of Nursing and Allied Health, Norfolk State University, United States of America.

Received: 01 Oct 2026; **Accepted:** 02 Nov 2026; **Published:** 14 Nov 2026

Citation: Philip De Melo, Audrey Douglas-Cooke. From Population Health Data to Personalized Medicine: Breast Cancer Survival Prediction. Cancer Sci Res. 2026; 9(4): 1-11.

ABSTRACT

This study presents a new framework that translates population health data into individualized breast cancer survival and mortality predictions. Using a dataset containing more than 4,000 patient records, the framework defines each virtual patient by an existing record's demographic and clinical characteristics. Each selected patient is excluded from training, and a Cox proportional hazards model is fitted to the remaining records using survival duration and vital status while accounting for right censoring. Population-level observations support estimation of the cumulative baseline hazard and baseline survival function. These estimates are combined with the patient's covariates to obtain individualized survival. The corresponding hazard structure describes time-dependent mortality risk under the proportional hazards' assumption. In the illustrated case, a patient had an estimated survival probability of approximately 94% at 60 months and 89% near 107 months; the patient was documented as alive at 60 months, with subsequent survival unknown. Additional examples contrasted Patient 7, who remained alive at 60 months and had approximately 3% predicted cumulative mortality at that time, while Patient 9, who died at month 14 and had approximately 6% predicted mortality by month 14 and 45% by month 60 having a steep mortality curve. These cases illustrate differences in estimated risk while highlighting limitations in predicting individual death timing. Changes in cumulative mortality over defined intervals, together with hazard estimates, may provide useful prognostic information, subject to validation. The framework demonstrates how population health data can support personalized prognosis using existing patient profiles rather than synthetic individuals.

Keywords

Anticipatory Oncology, Breast Cancer, Latent state, Hidden progression of cancer. Personalized medicine, Virtual Patient.

Introduction

Breast cancer survival prediction requires an understanding of how patient characteristics relate to outcomes over time. Patients differ in age, tumor characteristics, disease stage, hormone receptor status, and lymph node involvement. These differences motivate approaches that translate patterns observed in a patient population into individualized estimates of survival.

Survival analysis provides a statistical framework for examining time-to-event outcomes while accounting for patients whose eventual outcomes remain unknown at the end of follow-up. In this study, a Cox proportional hazards model combines demographic and clinical characteristics to estimate a patient-specific survival function. The resulting curve describes the predicted probability of surviving beyond successive time points measured from the dataset's defined survival time origin.

We introduce a new virtual patient framework using a breast cancer dataset containing more than 4,000 patient records. A virtual patient is a computational representation of an existing patient's recorded characteristics. Each selected record is withheld from model training, and the model is fitted to the remaining records. The selected patient's characteristics are then entered into the fitted model to generate an individualized survival curve. Repeating this procedure produces a series of virtual patients, allowing examination of how predicted survival varies across clinical profiles.

The framework separates the patient's observed follow-up from the model's predictions. For example, a patient documented as alive at 60 months has a right-censored outcome: survival is known through that follow-up point, but the eventual time of death is unknown. A predicted curve extending beyond 60 months provides a model-based estimate and does not establish what subsequently happened to that patient. Similarly, a curve generated from baseline characteristics describes survival probabilities over time without representing repeated measurements of disease progression.

The objective of this study is to demonstrate how population-level data can be used to construct individualized survival curves for a series of held-out patient profiles. The initial example illustrates the process for VP-001, while the broader framework supports its application to additional records. Excluding each selected patient from training prevents direct use of that patient's outcome in fitting their prediction, although this alone does not establish predictive accuracy. Evaluation of discrimination, calibration, model assumptions, subgroup performance, and external validity will be necessary to determine the framework's reliability and potential value for personalized prognosis.

Literature Review

Individualized breast cancer survival prediction draws on established survival analysis, clinical prognostic models, and computational representations of patients. These areas provide the foundation for a framework in which existing patient records are represented as virtual patients and assigned individualized survival curves.

Cox [1] introduced a regression framework for investigating relationships between patient characteristics and time-to-event outcomes. The model links explanatory variables to the hazard of an event while leaving the baseline hazard unspecified. It accommodates right-censored observations, including patients who remain alive at their last recorded follow-up. When combined with an estimated baseline survival function, the model can produce survival probabilities for individual patient profiles. This provides the statistical foundation for the present framework, although its use requires assessment of proportional hazards and the specification of predictor effects.

Individualized breast cancer prognosis is already an established application of clinical prediction modeling. Wishart et al. developed PREDICT to estimate overall and breast cancer-specific survival following surgery for early invasive breast cancer. Their study included external validation using a separate dataset of 5,468 patients. Subsequent work incorporated HER2 status into PREDICT Plus and revised the model with additional independent validation. These studies demonstrate that routinely collected clinical characteristics can support individualized prognosis, while emphasizing the importance of evaluating predictions in populations beyond the development cohort.

For the present study, PREDICT provides an important precedent: translating population-level relationships into patient-specific survival estimates is not itself a new capability. The proposed contribution should therefore be framed around the systematic representation of held-out patient profiles, generation of their survival curves, and evaluation of these predictions. Any comparison with established tools must also consider differences in patient eligibility, available predictors, survival time origin, and outcome definition.

Survival modeling has expanded beyond conventional regression. Ishwaran et al. introduced random survival forests for right-censored outcomes, using ensembles of survival trees to accommodate complex predictor relationships. Katzman et al. developed DeepSurv, which uses a neural network to model the risk component of a Cox proportional hazards model. These approaches offer alternatives when relationships between clinical characteristics and outcomes are inadequately captured by a conventional linear predictor. Their published results, however, do not establish that they will outperform Cox regression in every breast cancer dataset.

Accordingly, Cox regression offers a reasonable initial model for the proposed framework. More flexible methods could subsequently serve as comparators, provided that all models are evaluated using the same held-out observations and consistent preprocessing. Model selection should depend on demonstrated predictive performance and interpretability rather than algorithmic complexity alone.

The term "virtual patient" has several meanings in computational oncology. Wang et al. describe virtual patients in quantitative systems pharmacology as model parameterizations that produce physiologically plausible outputs. Such representations can support simulations of treatment response and variation across patient populations. Their discussion distinguishes these virtual populations from digital twins, which emphasize correspondence with an individual patient and periodic updating using new measurements.

De Melo et al. [2] integrated Kaplan–Meier estimation with Stochastic and Augmented Interpretable Health Analytics (SAIHA) to investigate pancreatic cancer survival. Kaplan–Meier analysis characterized survival during the observed follow-up period, while a Weibull-based stochastic model generated longer-term projections incorporating parameter uncertainty and population heterogeneity. The study emphasized distributions of predicted survival probabilities rather than single estimates, distinguishing empirical survival estimation from extrapolation dependent on model assumptions.

This work provides a methodological foundation for examining survival prediction together with uncertainty. The present breast cancer framework addresses a complementary objective: generating individualized survival curves for patient profiles

excluded from model training. Whereas the earlier Kaplan–Meier–SAIHA approach focused on population survival and stochastic extrapolation, the current Cox-based approach incorporates demographic and clinical predictors to estimate patient-specific survival. Incorporating SAIHA-inspired uncertainty analysis could further strengthen the virtual patient framework, but this would require additional implementation and validation.

The present study uses a narrower, explicitly defined interpretation: a virtual patient is a computational representation of an existing patient's baseline demographic and clinical profile, accompanied by a predicted survival function. It does not generate additional synthetic individuals or simulate detailed biological mechanisms. Because the proposed curves are derived from baseline records without ongoing patient-specific updates, the framework is best described as individualized survival prediction rather than a dynamic digital twin. This distinction helps align the terminology with the actual capabilities of the model.

A patient-specific curve must be evaluated for both discrimination and calibration. Uno et al. examined concordance measures for censored survival data and showed why censoring must be considered when assessing a model's ability to distinguish patients with different event times. Van Calster et al. [3] emphasized calibration—the agreement between predicted probabilities and observed outcomes—as a central requirement for reliable risk prediction. A model may rank patients appropriately while still producing inaccurate absolute probabilities.

These findings support evaluating the proposed framework across multiple held-out patients rather than interpreting one illustrated curve as evidence of accuracy. Withholding each selected record prevents its direct inclusion in model fitting, but the entire preprocessing and modeling procedure must also respect that separation. Predictions should be assessed at clinically relevant time horizons, with attention to censoring, uncertainty, and the amount of follow-up supporting longer-term estimates.

Transparent reporting is equally necessary. The TRIPOD+AI statement provides reporting guidance for prediction models developed using either regression or machine learning. PROBAST+AI provides a complementary framework for assessing model quality, risk of bias, and applicability. Together, these publications support clear documentation of participant selection, predictor availability, outcome definitions, missing data, model development, and evaluation.

The literature supports the feasibility of generating individualized survival estimates from population data. The present study applies this principle to more than 4,000 breast cancer records by representing selected patients as held-out virtual profiles and constructing their predicted survival curves. Its contribution is an application and evaluation framework built on established survival methods. Demonstrating its value will require evidence that the resulting probabilities are accurate, calibrated, and transferable to independent patient populations, together with a clear distinction between observed follow-up and model-based predictions.

Clinical Significance

The virtual patient framework offers a way to translate population-level breast cancer data into individualized survival estimates. By incorporating demographic and clinical characteristics, it generates curves that describe predicted survival probabilities over time for distinct patient profiles. If adequately validated, these estimates could support prognostic discussions and help clinicians communicate how outcomes vary among patients with different clinical characteristics.

A series of virtual patients could also support research and education by illustrating the relationships between disease stage, tumor characteristics, lymph node involvement, and predicted survival. Evaluating predictions across patient subgroups could identify differences in model reliability and guide improvements in its applicability to diverse populations.

Our earlier Kaplan–Meier–SAIHA work emphasized uncertainty in survival projections. Incorporating uncertainty intervals into the present framework would similarly help communicate the range of plausible estimates and reduce reliance on a single predicted curve.

Clinical significance depends on demonstrated accuracy, calibration, and external validity. The illustrated curve is a model-based prognosis, and one patient's observed follow-up does not establish its predictive accuracy. Furthermore, survival differences between patient profiles do not establish the benefits of changing treatments. Following rigorous validation, the framework could complement clinical judgment and established prognostic tools in personalized breast cancer care.

A Step to Personalized Medicine

Personalized medicine considers the characteristics of each patient when assessing prognosis and planning care. The proposed virtual patient framework takes a step in this direction by using demographic and clinical information to generate individualized breast cancer survival curves. Although the model learns from population data, its predictions reflect the particular combination of characteristics recorded for each patient.

Applying this approach across more than 4,000 patient records can produce a series of virtual patient profiles with distinct survival estimates. These curves provide a means of examining variation in prognosis and, following validation, could support discussions between clinicians and patients. Incorporating uncertainty estimates, informed by our earlier SAIHA work, would further help communicate the range of plausible outcomes.

The framework represents an initial contribution to personalized prognosis. Its potential role in care depends on reliable calibration, independent validation, and consistent performance across patient

groups. Future incorporation of molecular information, treatment history, and repeated clinical measurements could enable more detailed predictions that update as new patient information becomes available.

Formulation of a Virtual Patient

A virtual patient is formulated by representing an individual's recorded clinical characteristics as a computational profile. In this study, the first patient record is selected from a breast cancer dataset containing more than 4,000 patients. The profile incorporates available predictors, including age, tumor size, disease stage, tumor grade, hormone receptor status, and lymph node involvement. The selected patient's survival time and outcome are excluded from the predictors, and the record is withheld from model training to prevent direct information leakage.

A Cox proportional hazards model is fitted using the remaining patients' clinical characteristics, observed survival times, and censoring indicators. The fitted model is then applied to the withheld patient's profile to estimate an individualized survival function.

The algorithm consists of the following steps 1) Calculate relative risks, 2) Calculate baseline hazard increments, 3) Calculate baseline survival, 4) predict the chosen patient's survival, 5) calculate probability of death by time.

Breast Cancer

Disease Characteristics and Biological Heterogeneity

Breast cancer is a heterogeneous group of malignant neoplasms arising predominantly from the epithelial structures of the breast. Ductal carcinoma in situ is confined to the ductal system, whereas invasive carcinomas penetrate the basement membrane and may disseminate through lymphatic or hematogenous pathways. Invasive carcinoma of no special type and invasive lobular carcinoma are among the principal histologic forms [4]. Although advances in screening and systemic treatment have substantially reduced breast cancer mortality, the disease remains a major public health challenge, with outcomes differing according to tumor biology, stage at diagnosis, receipt of effective treatment, and broader conditions affecting access to care [5,6].

Breast cancer is not a single biological entity. Clinically actionable phenotypes are primarily defined by estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) expression. Hormone receptor-positive tumors are generally candidates for endocrine therapy; HER2-positive tumors may respond to HER2-directed agents; and triple-negative breast cancers lack ER and PR expression and are HER2-negative by conventional criteria. These categories encompass considerable variation in proliferative activity, molecular alterations, treatment responsiveness, and patterns of recurrence. Tumor size, histologic grade, axillary lymph node involvement, and extent of distant spread remain fundamental indicators of disease burden and prognosis [4,6].

The distinction between HER2-positive and HER2-negative disease has become more nuanced because selected HER2-negative tumors with low HER2 expression may benefit from antibody-drug conjugates in advanced disease. Nevertheless, current HER2 testing guidance does not establish "HER2-low" as a separate formal interpretive category; accurate reporting of immunohistochemical scores and confirmatory testing remains essential [7,8].

Major Clinical and Diagnostic Features

Breast cancer may be detected in an asymptomatic individual through screening or present clinically as a palpable mass, focal thickening, nipple retraction or discharge, skin changes, or axillary lymphadenopathy. Mammography can identify suspicious masses, architectural distortion, and clustered microcalcifications, including lesions that are not palpable. Diagnostic evaluation integrates clinical history and examination with breast imaging and pathologic confirmation. Bilateral mammography, including digital breast tomosynthesis when appropriate, and targeted ultrasonography are central imaging modalities; breast magnetic resonance imaging is reserved for selected circumstances, such as inconclusive conventional imaging, assessment of disease extent, or certain high-risk clinical settings [4,6].

Image-guided core needle biopsy establishes the histologic diagnosis and permits characterization of tumor type and grade. Pathology assessment ordinarily includes ER and PR immunohistochemistry, HER2 immunohistochemistry with in-situ hybridization when indicated for equivocal findings, and consideration of proliferative markers such as Ki-67. Axillary evaluation and selective staging investigations identify regional or distant disease when clinically warranted. These findings inform the clinical and pathologic stage and help distinguish localized, regional, and metastatic disease; extensive metastatic imaging is not routinely required for every patient with asymptomatic early-stage cancer [6,8].

Additional molecular information refines risk stratification and, for eligible patients, therapeutic decision-making. Multigene expression assays may inform adjuvant chemotherapy decisions in selected early-stage, hormone receptor-positive, HER2-negative cancers, but their interpretation depends on clinical context, including menopausal status and nodal involvement. Germline testing for pathogenic variants such as BRCA1 and BRCA2 is also clinically important for eligible patients because it may influence systemic treatment, surgical discussions, and familial risk assessment. These genomic evaluations supplement rather than replace established imaging, histopathology, receptor testing, and staging [9,10].

Contemporary Treatment Approaches

Contemporary breast cancer management combines locoregional treatment with systemic therapies selected according to stage, receptor phenotype, genomic characteristics, recurrence risk, patient health status, and treatment preferences. For many

patients with localized disease, surgery consists of breast-conserving excision followed by radiotherapy or mastectomy when clinically indicated. Sentinel lymph node biopsy and, in selected circumstances, less extensive axillary treatment reduce unnecessary morbidity without abandoning oncologic assessment. Radiotherapy contributes to local control and may be adapted to surgical findings and recurrence risk. Chemotherapy and biologically targeted treatments may be administered before surgery (neoadjuvant therapy) or afterward (adjuvant therapy), depending on disease characteristics and therapeutic goals [6].

For hormone receptor-positive disease, endocrine therapy with tamoxifen, an aromatase inhibitor, and/or ovarian function suppression constitutes a major component of treatment [6]. Chemotherapy is reserved according to the balance of clinical and genomic risk and expected benefit; the RxPONDER trial demonstrated that the value of chemotherapy in selected node-positive, hormone receptor-positive, HER2-negative disease differs by menopausal status [10]. For appropriately selected patients with higher-risk early-stage disease, adjuvant cyclin-dependent kinase 4/6 inhibition can improve invasive disease–free survival [11]. In metastatic hormone receptor-positive, HER2-negative disease, endocrine therapy combined with a CDK4/6 inhibitor is a major treatment strategy, with subsequent biomarker-guided options for tumors developing resistance, including therapies directed at ESR1 mutations or PI3K-AKT pathway alterations [12].

HER2-positive breast cancer is treated with HER2-directed therapy, commonly incorporated into chemotherapy-based neoadjuvant or adjuvant regimens. Subsequent treatment is adapted to response and residual disease [12]. The introduction of antibody-drug conjugates has also transformed the management of certain advanced tumors: trastuzumab deruxtecan improved outcomes in previously treated metastatic HER2-low breast cancer in the DESTINY-Breast04 trial, illustrating how careful biomarker assessment can expand therapeutic eligibility [7,12].

Chemotherapy is a central component of treatment for many patients with triple-negative breast cancer because standard endocrine and HER2-targeted therapies are generally ineffective against its defining receptor profile [12]. For eligible patients with stage II-III disease, perioperative pembrolizumab combined with neoadjuvant chemotherapy, followed by adjuvant pembrolizumab, has produced an overall survival benefit [13]. For selected patients with high-risk, HER2-negative early breast cancer and pathogenic germline BRCA1 or BRCA2 variants, adjuvant olaparib offers an additional targeted strategy after standard local and systemic treatment [14]. Across subtypes, metastatic treatment is generally directed toward prolonging survival, controlling disease, preserving function, and minimizing treatment burden rather than guaranteeing cure [12].

Survival Challenges and Sources of Prognostic Variation

Survival following breast cancer diagnosis varies substantially among patients because prognosis reflects the combined influence of disease extent, biological subtype, treatment exposure and response, competing illnesses, and time elapsed since diagnosis. Regional nodal involvement and distant metastases generally signify greater disease burden, while tumor grade and receptor status help differentiate biological aggressiveness and treatment sensitivity [5,4]. Importantly, the timing of recurrence is not uniform: hormone receptor–positive cancer may recur after prolonged disease-free intervals, making longer-term follow-up relevant even when early outcomes are favorable. A systematic review of late-recurrence prediction in hormone receptor–positive disease highlights both the usefulness and limitations of clinicopathologic risk tools [15].

Survival disparities further demonstrate that tumor characteristics alone do not explain clinical outcomes. In the United States, Black women experience higher breast cancer mortality than White women despite a lower overall incidence, and uneven access to screening, timely diagnostic resolution, specialist services, and guideline-concordant treatment contributes to persistent inequities [5]. In a cohort of Black women with breast cancer, neighborhood disadvantage was associated with increased breast cancer specific mortality, even after adjustment for several individual and clinical factors [16]. Accordingly, demographic variables should not be treated as direct biological causes of survival differences, and clinical prognostic models should be evaluated for differential performance across populations and care settings [17].

Another challenge is that survival estimates depend on the outcome being predicted. Overall survival concerns death from any cause, whereas breast cancer–specific survival considers death attributable to breast cancer; recurrence-free and progression-free outcomes measure different events [18]. The distinctions are clinically consequential, particularly among older patients and those with significant comorbidity. Treatment advances can also change survival patterns over calendar time, and a baseline clinical profile may not capture subsequent treatment modifications, toxicity, adherence, acquired resistance, or metastatic progression [6,12].

Implications for Individualized Survival Prediction

These clinical considerations provide the rationale for the present study's application of population-derived survival models to individual patient profiles. The available predictors—including age, tumor size, disease stage, grade, hormone receptor status, and lymph node involvement—represent clinically meaningful correlates of prognosis [5,6]. A Cox proportional hazards model can combine those baseline characteristics with observed survival times and censoring indicators to generate an individualized survival function [1]. However, a predicted curve expresses an estimated probability conditional on model assumptions and the measured predictors; it does not identify the date of an individual patient's death or establish that an unobserved treatment would improve survival [18].

When systemic treatment histories, longitudinal disease

assessments, selected molecular alterations, social-context measures, or comorbidity information are unrecorded, the clinical completeness of such predictions may be limited [6,12]. Furthermore, possible differences in hazard patterns across molecular subtypes or treatment groups warrant examination of the proportional hazards assumption. Model performance must therefore be assessed through discrimination, calibration, appropriate handling of censored observations, assessment of uncertainty, subgroup analyses, and validation in independent patient populations [17,18]. In this context, individualized survival prediction is best understood as a potentially useful adjunct to and not a replacement for clinical assessment, shared decision-making, or evidence-based treatment selection.

Algorithm Validation and Data Description

Data Description

The breast cancer dataset contains 4,024 patient records and 16 variables describing demographic characteristics, tumor features, lymph node involvement, hormone receptor status, and survival outcomes. Fourteen variables are potential predictors, while Survival Months and Status provide the survival outcome information. Each row represents a patient record with a single set of clinical characteristics and an associated follow-up duration. The dataset therefore supports time-to-event analysis but does not contain repeated measurements of disease progression, treatment response, or changing clinical characteristics.

Patient age ranges from 30 to 69 years, with a mean of 54.0 years, a standard deviation of 9.0 years, and a median of 54 years. The recorded racial categories are White, Black, and Other, comprising 3,413 patients (84.8%), 291 patients (7.2%), and 320 patients (8.0%), respectively. Marital status includes Married, Single, Divorced, Widowed, and Separated; married patients account for 65.7% of the records. The predominance of White patients and the restricted age range should be considered when evaluating the model's applicability to other populations.

Tumor characteristics include T-stage, N-stage, sixth-edition stage, differentiation, grade, A-stage, and tumor size. T-stage ranges from T1 to T4, with T2 (44.4%) and T1 (39.8%) being the most frequent categories. N-stage includes N1, N2, and N3, with N1 representing 67.9% of patients. Recorded overall stage categories are IIA, IIB, IIIA, IIIB, and IIIC. A-stage distinguishes Regional disease in 3,932 records from Distant disease in 92 records. Tumor size ranges from 1 to 140, with a median of 25; its measurement unit requires confirmation from the source documentation. Moderately differentiated tumors are the most common. Differentiation and grade correspond exactly in this file and therefore represent redundant information that should be addressed before model fitting.

Estrogen receptor status is positive in 3,755 patients (93.3%), while progesterone receptor status is positive in 3,326 patients (82.7%). The number of regional lymph nodes examined ranges from 1 to 61, with a median of 14. Positive regional lymph nodes range from 1 to 46, with a median of 2. All records contain at least one positive node, indicating that node-negative patients are not represented.

Recorded survival or follow-up duration ranges from 1 to 107 months, with a median of 73 months. There are 616 patients classified as Dead (15.3%) and 3,408 classified as Alive (84.7%). For survival modeling, Dead is coded as an observed event and Alive as right-censored at the recorded follow-up time. The median recorded duration is not an estimate of median overall survival. The file does not specify the survival time origin or cause of death, so the endpoint should not be described as breast cancer-specific mortality without supporting documentation.

No missing values were detected, although one duplicate row, inconsistent spacing, and an irregular grade label require review. In the virtual patient analysis, withholding one record leaves 4,023 records for training before any additional exclusions. The selected patient's clinical characteristics are then used to estimate an individualized survival curve, while their observed survival time and status remain excluded from training.

Algorithm Validation

Validation of the new virtual patient survival prediction algorithm must establish whether the model produces accurate, reliable, and interpretable predictions for patients whose outcomes were not used during model development. The algorithm combines a Cox proportional hazards model with Breslow estimation of the baseline cumulative hazard to generate individualized survival curves. Although withholding a selected patient prevents that record from directly contributing to model fitting, a survival curve for one patient does not establish predictive validity. Validation therefore requires repeated evaluation across held-out patients, assessment of the model's assumptions, and examination of performance across clinically relevant groups. The following protocol describes the analyses needed to evaluate the algorithm; it does not report completed validation results.

The first stage of validation involves confirming the accuracy and consistency of the input data. The available dataset contains 4,024 records, including 616 recorded deaths and 3,408 patients classified as alive at their recorded follow-up time. Before analysis, categorical labels should be standardized, numerical values checked for plausibility, and the identified duplicate row reviewed. Because the file does not include a patient identifier, identical rows cannot automatically be assumed to represent the same individual. If the duplicate is confirmed to be an accidental repetition, it should be removed. If its origin remains uncertain, a sensitivity analysis should compare results with and without the repeated row.

The survival endpoint must also be clearly defined. For the proposed analysis, Status is coded as an event indicator, with Dead indicating an observed death and Alive indicating right censoring. Survival Months represents the elapsed time to the recorded event

or the end of observation. Source documentation should confirm the starting point for this duration, such as diagnosis, and establish whether Status refers to death from any cause or a specific cause. Without this information, predictions should be described as probabilities associated with the recorded death endpoint rather than breast cancer-specific mortality.

Predictor availability must be checked against the intended prediction time. Every predictor should represent information available when the survival prediction would be made. Measurements obtained later in the patient's course could introduce information leakage if used to predict survival from an earlier starting point. The correspondence between differentiation and grade also requires attention because these variables encode the same classification in the supplied file. Including both without adjustment can create redundant model terms. Similarly, the relationship among T-stage, N-stage, and combined stage should be examined when specifying the predictor set.

Separation of Model Development and Evaluation

The central requirement of predictive validation is to separate the information used to develop the model from the information used to evaluate it. In the illustrative virtual patient procedure, one record is withheld and the model is trained on the remaining eligible records. The withheld patient's clinical characteristics are then entered into the fitted model to obtain an individualized survival curve. Their observed survival time and event indicator are used only for evaluation.

This procedure can be repeated using leave-one-out validation, in which each patient is withheld in turn. However, evaluating several thousand separately fitted models is computationally demanding, and withholding one record at a time does not create an independent external population. A practical primary approach is repeated, event-stratified five-fold cross-validation. The dataset is divided into five groups with approximately similar proportions of recorded deaths. Four groups are used for training and the remaining group for evaluation. This process continues until every patient has received a prediction from a model trained without their record. Repeating the partition with different random seeds helps assess sensitivity to the allocation of patients.

All preprocessing must occur within the training portion of each split. This includes selecting predictors, defining category encodings, estimating scaling parameters, and handling missing values if they arise in later datasets. The evaluation portion must be transformed using the rules learned from the corresponding training portion. If regularization strength, predictor combinations, or other modeling choices are selected by comparing performance, that selection should take place within an inner validation loop. The outer held-out groups then provide an assessment of the complete model development procedure.

Known records from the same patient must remain in the same group. Where identifiers are unavailable, exact duplicate records should be kept together during splitting as a precaution against placing a potentially repeated individual in both training and evaluation groups.

Assessment of Discrimination

Discrimination measures whether the model can distinguish patients with different survival experiences. A principal measure is the concordance index, which evaluates whether higher predicted risk corresponds to earlier observed events among comparable patients. A value near 0.5 indicates performance comparable to random ordering, while larger values indicate better ranking. Because censoring affects which pairs can be compared, the analysis should report a censoring-adjusted concordance measure and specify the evaluation period.

Time-dependent area under the receiver operating characteristic curve provides complementary information at selected horizons. For example, performance may be examined at 12, 36, and 60 months, provided that each horizon has adequate follow-up and sufficient observed events. These measures evaluate the model's ability to distinguish patients who experience the event by a given time from those who survive beyond that time, while accounting for censoring.

Discrimination should be evaluated using held-out predictions. Performance calculated on the same patients used to fit the model is likely to be optimistic. In addition, good discrimination does not establish that predicted survival probabilities are numerically accurate. A model may correctly rank patients while systematically overestimating or underestimating their probability of survival.

Assessment of Calibration

Calibration examines agreement between predicted probabilities and observed outcomes. At each selected horizon, held-out patients can be grouped according to their predicted survival probabilities. The average prediction within each group is compared with a censoring-adjusted estimate of observed survival, such as a Kaplan–Meier estimate. A calibration plot places predicted survival against estimated observed survival, with the diagonal representing agreement.

These comparisons should include uncertainty intervals and information about the number of patients contributing to each estimate. The number of groups should be chosen to avoid unstable estimates, especially when events are uncommon. A smooth calibration curve may provide additional detail when supported by the sample size.

For an individual patient, a predicted survival probability cannot be directly verified from a single outcome. A patient either survives beyond a particular time or experiences the event before that time, whereas the model produces a probability. Calibration therefore requires comparison across multiple patients with similar predicted probabilities. This distinction is essential when interpreting the virtual patient paradigm: individualized predictions are evaluated

through their performance across patients, rather than by expecting each prediction to match a single observed outcome exactly.

Overall Prediction Error

The Brier score evaluates the difference between predicted survival probabilities and observed survival status at a specified time. Because survival status is unknown beyond censoring for some patients, the score should be calculated using an appropriate censoring adjustment, such as inverse probability of censoring weighting. Lower values indicate smaller prediction error.

An integrated Brier score summarizes prediction error over a prespecified time interval. That interval should be restricted to a period with adequate follow-up support. Although the largest recorded duration is 107 months, predictions near the end of follow-up may depend on a relatively small number of patients. Reporting error over a well-supported interval is more informative than extending evaluation to the maximum observed duration without considering the declining risk set.

The individualized model should be compared with a reference approach that assigns the training cohort's Kaplan–Meier survival curve to every held-out patient. A simpler Cox model using a limited set of clinically relevant predictors can provide a second comparator. These comparisons help determine whether the additional patient characteristics improve prediction beyond an overall cohort estimate or a simpler model.

Model Assumptions and Robustness

The Cox model assumes proportional hazards: the relative hazard associated with a predictor remains constant over time. This assumption should be assessed using Schoenfeld residuals and graphical diagnostics. Substantial violations may indicate a need for time-varying effects, stratification, or an alternative survival model. Any modification should be selected within the model development process and evaluated on held-out patients.

The functional form of numerical predictors also requires examination. Age, tumor size, and lymph node counts may not have linear relationships with the log hazard. Flexible terms, such as restricted cubic splines, can be considered when clinically and statistically justified. Correlated predictors, sparse categories, and influential observations should be examined because they can produce unstable coefficient estimates.

Right-censored analysis also relies on assumptions about the censoring process. Patients whose observation ends may differ from those who remain under follow-up. The available data should be examined for differences in follow-up across clinical and demographic groups, while recognizing that the file may not contain enough information to explain why observation ended.

Robustness analyses should examine whether conclusions change after addressing duplicate records, removing redundant predictors, altering reasonable model specifications, or changing validation partitions. Confidence intervals for performance measures can be estimated through resampling procedures that reflect the model development process. Bootstrap refitting can also help characterize uncertainty in selected survival predictions, although it does not capture every source of uncertainty, such as population differences or unmeasured predictors.

Subgroup and External Validation

Performance should be examined across age groups, recorded racial categories, disease stages, and hormone receptor groups where sufficient patients and events are available. The dataset is predominantly White and contains patients aged 30–69 years with positive regional lymph nodes. These characteristics limit conclusions about older patients, node-negative disease, and populations poorly represented in the file. Subgroup estimates should include sample sizes, event counts, and uncertainty intervals; small numerical differences should not be interpreted as established disparities when estimates are imprecise.

External validation requires applying the finalized model and preprocessing rules to an independent dataset without refitting them before the initial assessment. The external population should have compatible predictors, an aligned survival time origin, and the same event definition. Differences in discrimination and calibration would reveal how well the model transfers across settings. If recalibration is subsequently performed, results before and after recalibration should be reported separately.

The final validation report should document the patient flow, preprocessing decisions, predictor specification, validation design, evaluation horizons, performance estimates, and uncertainty. Together, these analyses would determine whether the virtual patient algorithm provides useful individualized survival estimates. Clinical usefulness would require further evaluation in the intended care setting, particularly before predictions are used to guide treatment or patient counseling.

Figure 1 shows predicted mortality at 14, 36, and 60 months using five-fold internal validation. Each point represents one of five groups defined by quantiles of held-out predicted risk, comparing mean predicted mortality with Kaplan–Meier-estimated observed mortality. Vertical bars indicate pointwise 95% confidence intervals for observed mortality. The dashed diagonal denotes perfect agreement; points above and below it indicate underestimation and overestimation of mortality, respectively. Each patient's prediction was generated by a model trained without that patient.

Table 1 shows the predicted cumulative mortality for the first 10 patients at 14, 36, and 60 months. Values represent the model-estimated probability of death by each time point, expressed as a percentage and calculated as $100 \times [1 - \hat{S}_i(t)]$, where $\hat{S}_i(t)$ is the predicted survival probability for patient i at time t. These estimates describe cumulative risk from the model's time origin, rather than the probability of death during an individual month or between successive time points.

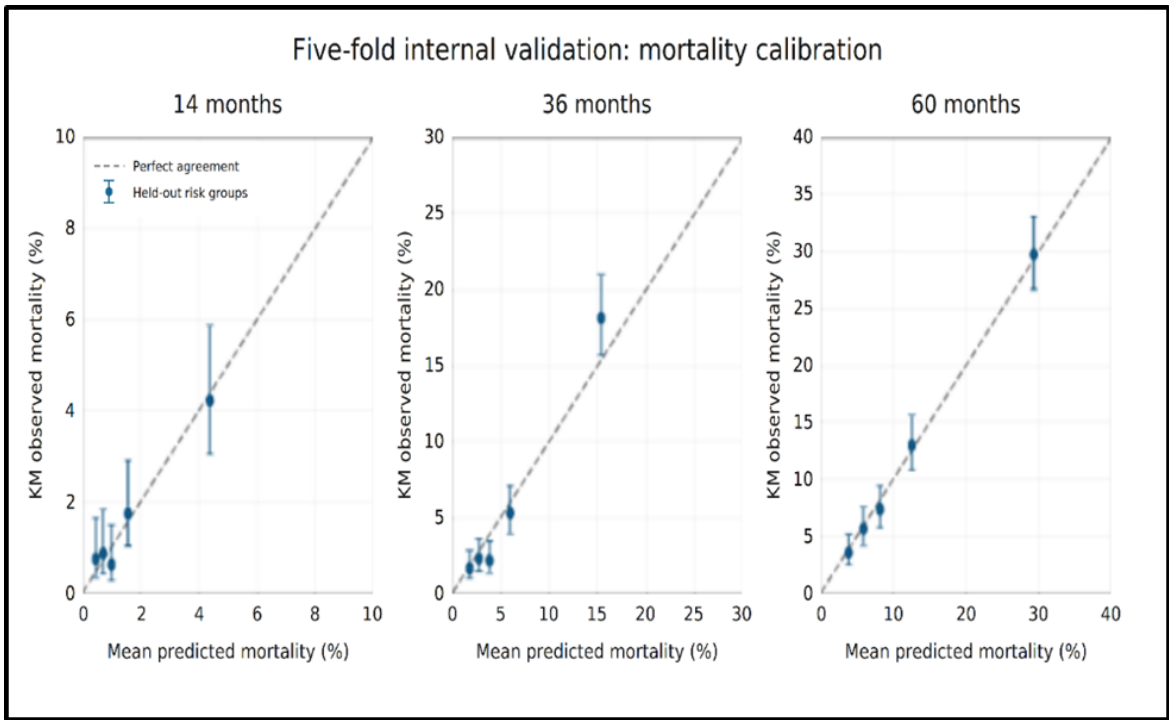


Figure 1: Predicted mortality at 14, 36, and 60 months using five-fold internal validation.

Predicted mortality increases over time for all patients, with substantial differences in the magnitude of risk. Patient 9 has the highest predicted mortality at each horizon, increasing from 6.87% at 14 months to 24.05% at 36 months and 45.13% at 60 months. Patients 10 and 3 have the next highest estimates, reaching 26.27% and 19.96%, respectively, at 60 months. Patient 7 has the lowest predicted mortality throughout follow-up, with estimates of 0.39%, 1.51%, and 3.26%. Across the 10 patients, predicted mortality ranges from 0.39% to 6.87% at 14 months, 1.51% to 24.05% at 36 months, and 3.26% to 45.13% at 60 months.

PREDICTED MORTALITY FOR THE FIRST 10 PATIENTS

Patient	Mortality at 14 months (%)	Mortality at 36 months (%)	Mortality at 60 months (%)
1	0.68	2.59	5.57
2	1.28	4.84	10.25
3	2.61	9.70	19.96
4	1.08	4.13	8.78
5	1.24	4.69	9.94
6	0.55	2.12	4.57
7	0.39	1.51	3.26
8	0.61	2.38	5.12
9	6.87	24.05	45.13
10	3.55	13.03	26.27

Table 1: Predicted cumulative mortality for the first 10 patients at 14, 36, and 60 months.

The table illustrates how the model assigns different mortality trajectories to individual patients while maintaining the same risk ranking across the three prediction horizons. These probabilities are estimates, not observed outcomes or predictions of an exact time of death. Differences between patients demonstrate variation in model-assigned risk but do not, by themselves, establish predictive accuracy or clinical usefulness. Confidence intervals are not shown; assessment of reliability requires comparison with observed survival outcomes using methods that account for censoring, together with evaluation of calibration and discrimination in independent data.

Let us now plot a survival curve for patient 7 with the lowest risk of mortality and patient 9 with the highest rate of mortality.

Figure 2 shows the predicted cumulative mortality for two patients with contrasting observed outcomes. Left: Patient 7, who remained alive at 60 months, had a low predicted probability of death, approximately 3% by month 60. Right: Patient 9, who died at month 14, had a substantially higher predicted mortality trajectory, reaching approximately 45% by month 60. The curves represent cumulative mortality probabilities, (1-S(t)), measured from the dataset time origin, rather than the probability of death within each individual month. Although the model assigned higher risk to the patient who died, the predicted probability of death by month 14 remained relatively low, approximately 6%. These examples illustrate differences in estimated individual risk but do not establish accurate prediction of death timing or overall model performance. The trajectory beyond month 14 for Patient 9

represents model estimates, not observed survival.

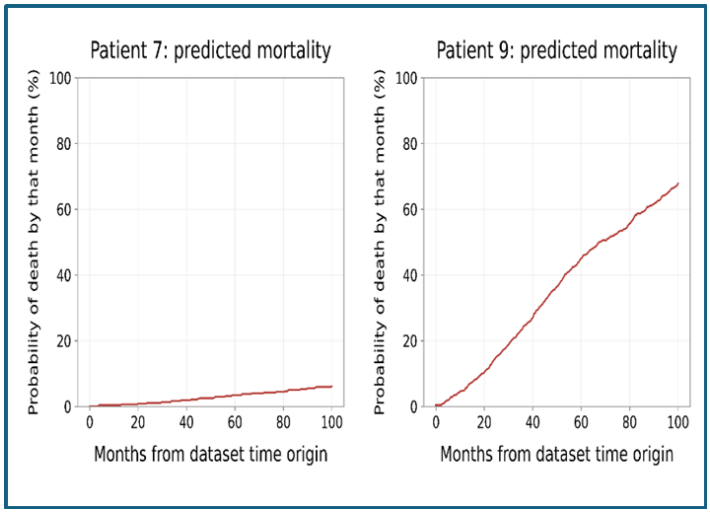


Figure 2: The predicted cumulative mortality for two patients with contrasting observed outcomes. Left: Patient 7 who remained alive at 60 months and Patient 9, who died at month 14.

Discussion

The present analysis highlights an opportunity to make greater use of population health data for individualized breast cancer prognosis. Cancer registries and clinical datasets contain information on patient characteristics, tumor features, and survival outcomes that can support analyses beyond population averages. When suitable patient-level data are available, survival models can translate these observations into individualized mortality curves, estimating the cumulative probability of death overtime for a particular patient profile. This approach offers a practical connection between population health research and personalized medicine, although the reliability of each estimate depends on data quality, representativeness, and model validation.

The illustrated cases demonstrate both the potential and the limitations of this approach. The patient who remained alive at 60 months had a relatively flat predicted mortality curve, whereas the patient who died at month 14 had a substantially steeper trajectory. This difference suggests that the model distinguished their relative risk profiles. However, the low predicted mortality at month 14 for the patient who died also shows that identifying a higher-risk patient does not necessarily mean accurately predicting when death will occur. These examples support further investigation rather than establishing clinical accuracy.

Beyond the cumulative probability at a selected time point, the slope of an individualized mortality curve may provide useful prognostic information. A steeper segment indicates that predicted cumulative mortality is increasing more rapidly during that interval. Mathematically, if cumulative mortality is expressed as ($F(t)=1-S(t)$), its slope is ($f(t)=dF(t)/dt$). The corresponding hazard, ($h(t)=f(t)/S(t)$), describes the instantaneous mortality rate among patients who have survived to that time. Thus, the slope and the hazard provide related but distinct descriptions of how mortality risk is distributed over time. For the stepwise curves shown here, changes over defined intervals are more interpretable than derivatives at individual steps.

With appropriate validation, these temporal measures could become useful tools for clinical prognostic assessment and for investigating periods that may warrant closer clinical attention. They should not be interpreted as direct evidence of tumor progression or as diagnostic tests for breast cancer. Moreover, if the dataset records death without identifying its cause, the resulting curves estimate all-cause mortality among patients with breast cancer, rather than breast cancer–specific mortality. Future work should evaluate calibration, discrimination, uncertainty, and performance across patient subgroups in independent cohorts. The central opportunity is to use existing population data more effectively to generate individualized, time-dependent risk estimates that can be evaluated for clinical usefulness.

This study presents a virtual patient framework that connects population health data with personalized breast cancer prognosis. By estimating baseline survival and cumulative hazard from population-level observations and combining these estimates with individual demographic and clinical characteristics, the framework generates patient-specific survival and mortality curves. These curves demonstrate how existing patient records can support individualized estimates of risk over time, extending the value of population data beyond aggregate survival statistics.

The temporal pattern of predicted mortality, including changes in the curve's slope and the corresponding hazard, may provide additional prognostic information by identifying intervals during which estimated risk accumulates more rapidly. However, the illustrated cases show that distinguishing higher-risk patients does not necessarily establish accurate prediction of individual death timing. Independent validation, calibration assessment, and evaluation across patient subgroups are therefore essential. With these requirements addressed, this framework could support more informative risk communication and contribute to the translation of population health evidence into personalized medicine.

Conclusion

This study presents a virtual patient framework that connects population health data with personalized breast cancer prognosis. By estimating baseline survival and cumulative hazard from population-level observations and combining these estimates with individual demographic and clinical characteristics, the framework generates patient-specific survival and mortality curves. These curves demonstrate how existing patient records can support individualized estimates of risk over time, extending the value of population data beyond aggregate survival statistics.

The temporal pattern of predicted mortality, including changes in the curve's slope and the corresponding hazard, may provide additional prognostic information by identifying intervals during

which estimated risk accumulates more rapidly. However, the illustrated cases show that distinguishing higher-risk patients does not necessarily establish accurate prediction of individual death timing. Independent validation, calibration assessment, and evaluation across patient subgroups are therefore essential. With these requirements addressed, this framework could support more informative risk communication and contribute to the translation of population health evidence into personalized medicine.

References

1. Cox DR. Regression models and lifetables. *Journal of the Royal Statistical Society Series B Methodological*. 1972; 34: 187-202.
2. de Melo P, DiLella M, Holman T, et al. Accurate prediction of survival based on Kaplan Meier analytics. *Cancer Research Journal*. 2025; 13: 173-185.
3. Van Calster B, McLernon DJ, van Smeden M, et al. Topic Group Evaluating Diagnostic Tests and Prediction Models of the STRATOS Initiative. Calibration The Achilles Heel of Predictive Analytics. *BMC Medicine*. 2019; 17: 230.
4. Loibl S, Poortmans P, Morrow M, et al. Breast cancer. The Lancet. 2021; 397: 1750-1769.
5. Giaquinto AN, Sung H, Newman LA, et al. Breast Cancer Statistics 2024. CA Cancer J Clin. 2024; 74: 477-495.
6. Loibl S, André F, Bachelot T, et al. Early Breast Cancer ESMO Clinical Practice Guideline for Diagnosis Treatment and Follow-up. Ann Oncol. 2024; 35: 159-182.
7. Modi S, Jacot W, Yamashita T, et al. Trastuzumab Deruxtecan in Previously Treated HER2-Low Advanced Breast Cancer. N Engl J Med. 2022; 387: 9-20.
8. Wolff AC, Somerfield MR, Dowsett M, et al. Human epidermal growth factor receptor 2 testing in breast cancer ASCO–College of American Pathologists guideline update. Journal of Clinical Oncology. 2023; 41: 3867-3872.
9. Bedrosian I, Somerfield MR, Achatz MI, et al. Germline Testing in Patients with Breast Cancer ASCO–Society of Surgical Oncology guideline. J Clin Oncol. 2024; 42: 584-604.
10. Kalinsky K, Barlow WE, Gralow JR, et al. 21-gene assay to inform chemotherapy benefit in node-positive breast cancer. N Engl J Med. 2021; 385: 2336-2347.
11. Slamon D, Lipatov O, Nowecki Z, et al. Ribociclib Plus Endocrine Therapy in Early Breast Cancer. N Engl J Med. 2024; 390: 1080-1091.
12. Lloyd MR, Jhaveri K, Kalinsky K, et al. Precision Therapeutics and Emerging Strategies for HR-Positive Metastatic Breast Cancer. Nat Rev Clin Oncol. 2024; 21: 743-761.
13. Schmid P, Cortes J, Dent R, et al. Overall Survival With Pembrolizumab in Early-Stage Triple-Negative Breast Cancer. N Engl J Med. 2024; 391: 1981-1991.
14. Tutt ANJ, Garber JE, Kaufman B, et al. Adjuvant Olaparib for Patients with BRCA1- or BRCA2-Mutated Breast Cancer. N Engl J Med. 2021; 384: 2394-2405.
15. Shrestha A, Cullinane C, Evoy D, et al. Clinical Treatment Score post-5 years as a predictor of late distant recurrence in hormone receptor-positive breast cancer Systematic review and meta-analysis. Br J Surg. 2022; 109: 411-417.
16. Holder EX, Barnard ME, Xu NN, et al. Neighborhood Disadvantage Individual Experiences of Racism and Breast Cancer Survival. JAMA Network Open. 2025; 8: e253807.
17. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI An updated quality risk of bias and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025; 388: e082505.
18. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024; 385: e078378.

© 2026 Philip De Melo, et al. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License