SPECIAL SERIES: STATISTICAL METHODS IN PRECISION ONCOLOGY

review articles

Developing and Validating Risk Assessment Models of Clinical Outcomes in Modern Oncology

Susan Halabi, PhD[1]; Cai Li, PhD[1]; and Sheng Luo, PhD[1]

abstract

The identification of prognostic factors and building of risk assessment prognostic models will continue to play a major role in 21st century medicine in patient management and decision making. Investigators often are interested in examining the relationship among host, tumor-related, and environmental variables in predicting clinical outcomes. We distinguish between static and dynamic prediction models. In static prediction modeling, variables collected at baseline typically are used in building models. On the other hand, dynamic predictive models leverage the longitudinal data of covariates collected during treatment or follow-up and hence provide accurate predictions of patients' prognoses. To date, most risk assessment models in oncology have been based on static models. In this article, we cover topics related to the analysis of prognostic factors, centering on factors that are both relevant at the time of diagnosis or initial treatment and during treatment. We describe the types of risk prediction and then provide a brief description of the penalized regression methods. We then review the state-of-the art methods for dynamic prediction and compare the strengths and limitations of these methods. Although static models will continue to play an important role in oncology, developing and validating dynamic models of clinical outcomes need to take a higher priority. A framework for developing and validating dynamic tools in oncology seems to still be needed. One of the limitations in oncology that may constrain modelers is the lack of access to longitudinal biomarker data. It is highly recommended that the next generation of risk assessments consider longitudinal biomarker data and outcomes so that prediction can be continually updated.

JCO Precis Oncol. © 2019 by American Society of Clinical Oncology

INTRODUCTION

The identification of prognostic factors and building of risk assessment prognostic models will continue to play a major role in 21st century medicine in patient management and decision making.[1] Prognostic factors in oncology associate host and tumor variables with clinical outcomes independent of treatment.[2] Gospodarowicz et al[3] classified factors as either tumor related, host, or environmental. Tumor-related factors are variables related to the tumor and reveal the tumor biology and pathology, such as size of tumor, lymph node involvement, presence of metastasis, and molecular markers (overexpression of *PTEN* gene, presence of androgen receptor variant AR-V7). Host factors are associated with patient characteristics, such as age and comorbidities. Finally, environmental factors are external to the patient, such as access to health care.[3] Prognostic models are increasingly used in the design, conduct, and analysis of clinical trials. For example, in several trials of prostate cancer, randomization was stratified by the predicted survival probability determined by the prognostic model of overall survival.[4-7] In the TAILORx trial, Oncotype DX, a 21-gene score that predicts the likelihood of recurrence, is used to classify women with breast cancer by their risk group of recurrence.[8] Prognostic factors also have been used for enriching the patient population in trials with targeted therapies. For example, in the ToGA trial, patients with human epidermal growth factor receptor 2–positive gastric cancer were randomly assigned to either trastuzumab plus chemotherapy or chemotherapy alone.[9]

In this article, we cover topics that are related to the analysis of prognostic factors, centering on factors that are relevant both at the time of diagnosis or initial treatment and during treatment or follow-up. We use the terms prognostic models, risk models, and risk assessments interchangeably. This article is organized in the following way. First, we describe the type of risk prediction and then provide a brief description of the penalized regression methods. Second, we review the state-of-the-art methods for dynamic prediction and compare the strengths and limitations of these methods. Third, we offer a concise discussion of validation and metrics for assessing models. Finally, we present recommendations for the next generation of risk assessment methods to be built in modern oncology.

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on September 12, 2019 and published at ascopubs.org/journal/po on October 24, 2019: DOI https://doi.org/10.1200/PO.19.00068

CONTEXT SUMMARY

Key Objective

Building prognostic models will continue to play a role in 21st century medicine in patient management. We make a distinction between static and dynamic predictive models and provide a review of the state-of-the-art methods for dynamic predictive models to promote them for future use.

Knowledge Generated

To date, most risk assessment models in oncology have been based on static prognostic models. An understanding of the longitudinal relationship between host and tumor-related factors and their impact on clinical outcomes is critical. Regardless of whether static or dynamic modeling is the primary objective, we envision that this review will encourage investigators to take risk assessment as a discipline by itself.

Relevance

We expect to see an upsurge in dynamic risk assessments, and as such, it is highly recommended that the next generation of models consider the longitudinal data and outcomes so that predictions are updated.

TYPES OF RISK PREDICTIONS

Investigators are interested in examining the relationship between host and tumor-related factors in predicting clinical outcomes (Fig 1A). We distinguish between static and dynamic prediction. In static prediction modeling, variables collected at baseline typically are used in building models. For example, prostate-specific antigen (PSA) measurements at baseline are used for prediction of recurrence. On the other hand, dynamic predictive models explicitly leverage the longitudinal data of covariates that are collected during treatment or follow-up. In patients with advanced cancer, the disease has substantially evolved

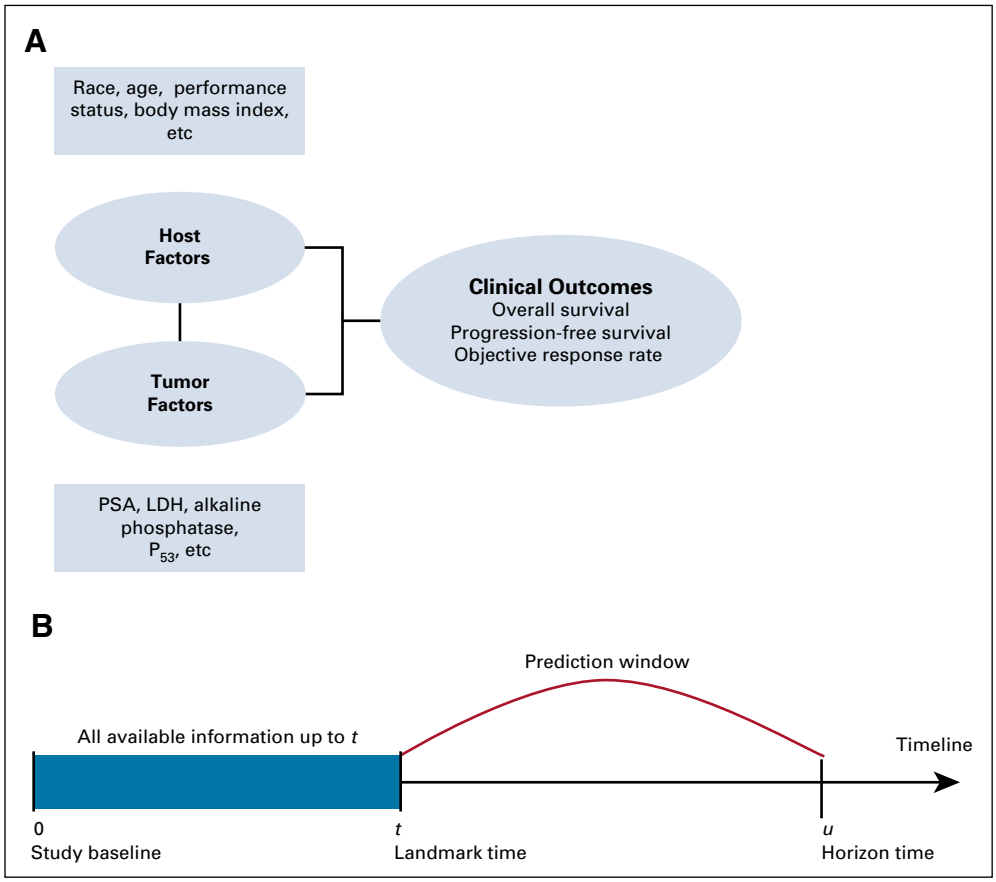

FIG 1. (A) Relationship between host and tumor-related factors and clinical outcomes. Modified from Halabi and Owzar.[2] (B) Dynamic risk prediction framework. LDH, lactate dehydrogenase; PSA, prostate-specific antigen.

 © 2019 by American Society of Clinical Oncology

and has a heterogeneous presentation within the patient.[10] The inter- and intrapatient variability should be taken into account in statistical modeling.[11,12] Dynamic prediction incorporates time-dependent covariates so that risk prediction would be continually updated with new observations to reflect the patient's prognosis.

We define terminology that is typically used in dynamic risk prediction. The term landmark time is defined as a current time point at which we have data (host, tumor-related variables, and outcomes). The term horizon time is defined as a future time point at which we want to predict a time-to-event outcome, such as overall survival. Dynamic predictive models essentially capture the historical information of the longitudinal measurements from the study baseline to the landmark time (t) such that the risk at horizon time (u) can be accurately predicted (Fig 1B).

IDENTIFICATION OF PROGNOSTIC FACTORS

Several popular strategies exist for identifying prognostic factors in static risk assessment. Standard variable selection approaches, such as forward selection, backward selection, and so forth, with logistic regression for binary end points,[13] and proportional hazards regression for time-to-event end points,[14] have been applied. Criticism for the stepwise variable selection has been well documented.[15,16] Of note, classification trees, such as recursive partitioning for both binary and time-to-event end points,[17-20] frequently have been used.[2,21-26]

We concentrate on penalized methods that fit and shrink p predictors and in doing so, reduce the variance of the coefficient estimates.[27] Thus, these methods would improve the accuracy of the model.[28] The least absolute shrinkage and selection operator (LASSO) and adaptive LASSO (ALASSO) have been widely used to develop prognostic models of clinical outcomes.[29,30] We will briefly describe ridge regression and penalized methods. Ridge regression minimizes the residual sum of squares function, but a caveat is that it does not reduce all the coefficients exactly to 0.[28] Let y_i be the response, x_{ij} the jth covariate value ($j = 1, 2, \ldots, p$) corresponding to the ith individual, β_j the regression coefficient jth covariate, and λ a tuning parameter. Similar to ridge regression, the first term in LASSO is the residual sum of squares, and LASSO minimizes this function subject to the l_1 penalty (Eq 1):

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|. \qquad (1)$$

A large tuning parameter causes the coefficient estimates to be equal to 0, thus the LASSO will have the sparsity property.[28] LASSO is an improvement over ridge regression, although it has the main limitation of tending to select too many unimportant variables, and it performs poorly in situations when $p > n$.[20,31-33] ALASSO has been proposed as an improvement over LASSO to overcome the limitation of LASSO.[34] ALASSO minimizes this function (Eq 2):

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}w_j|\beta_j|. \qquad (2)$$

ALASSO uses a weighted penalty term in the l_1 penalty where $w = (w_1, w_2, \ldots, w_p)$ is the weight vector. If $\hat{\beta}$ is a $\sqrt{n}$-consistent estimator [eg, $\hat{\beta}$(OLS)] of $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$, then an appropriate choice of the weight w is $1 / |\hat{\beta}|$. ALASSO is considered to be an improvement over LASSO because it has consistent variable selection as well as lower prediction error. Consequently, ALASSO tends to select fewer nonzero coefficients than the LASSO despite having a smaller prediction error. The ALASSO enjoys the oracle property.[30,34]

Elastic net regression uses a combination of l_1 penalty and ridge l_2 penalty and is a compromise between LASSO and ridge regression. Furthermore, one of its main advantages when $p > n$ is that it retains more than n variables in the model.[35] Hastie et al[28] provided a thorough comparison of these shrinkage techniques.

We applied LASSO and ALASSO from CALGB 90401, a phase III clinical trial in advanced prostate cancer, with the overall goal of building a model of overall survival.[5] We had 22 variables that were common between CALGB 90401 and the Enthuse trial (external data set).[36] Because of missing data in the covariates, we used methods to impute them as described by White and Royston.[37] The regression's estimates from the Cox proportional hazards model, LASSO, and ALASSO, are listed in Table 1. We considered LASSO and ALASSO and applied both the Akaike information criterion and the Bayesian information criterion to choose the optimal model of overall survival. LASSO and ALASSO selected eight and nine variables, respectively (Table 1). We determined the ALASSO model to be the final optimal model since it included the site of metastases for bone disease. Figure 2 shows the solution path for ALASSO, and we observe that lactate dehydrogenase (LDH) greater than the upper limit of normal and Eastern Cooperative Oncology Group (ECOG) performance status were selected early in the l_1 path compared with the other variables. This is followed by visceral disease, alkaline phosphatase, albumin, hemoglobin, pain, bone metastases, and then PSA (the Bayesian information criterion stopped at PSA). The final model selected the following prognostic factors: LDH greater than the upper limit of normal, ECOG performance status, metastatic site (presence of visceral disease, presence of bone metastases), PSA, alkaline phosphatase, albumin, hemoglobin, and analgesic opioid use.

We have focused on variable selection when the number of predictors is small relative to the sample size. There are two main challenges in identifying potential prognostic factors

TABLE 1. Identified Prognostic Factors by LASSO and ALASSO in the Training Set

Description of Variable (variable name)	Cox Model, $\hat{\beta}$	LASSO, $\hat{\beta}$	ALASSO, $\hat{\beta}$
Site of metastases			
Bone (DS2)	0.234	—	0.058
Visceral (DS3)	0.400	0.056	0.293
Liver (LIVER)	0.013	—	—
Lung (LUNG)	0.199	—	—
Opioid analgesic use (PAIN)	0.136	0.077	0.088
Age in years (AGE)	–0.003	—	—
Body mass index (BMI)	–0.021	—	—
White race (Caucasian)	0.034	—	—
ECOG performance status (ECOG)	0.278	0.190	0.305
Comorbidity (Comorb)	0.070	—	—
Gleason score (GLEAST)	0.052	—	—
Prior treatment with radiotherapy (Radio)	0.105	—	—
LDH ≥ ULN (LDH.High)	0.325	0.203	0.335
Albumin (ALB)	–0.133	–0.080	–0.122
Bilirubin (BILI)	–0.017	—	—
Hemoglobin (HGB)	–0.094	–0.085	–0.065
Platelets (PLT)	0.000	—	—
WBCs (WBC)	0.055	—	—
Alkaline phosphatase (ALKPHOS)	0.145	0.138	0.145
Aspartate aminotransferase (AST)	0.025	—	—
Prostate-specific antigen (PSA)	0.063	0.026	0.015
Testosterone (TESTO)	–0.088	—	—
Training C-index		0.662	0.660
Integrated time AUC		0.742	0.740

Abbreviations: ALASSO, adaptive least absolute shrinkage and selection operator; AUC, area under the curve; $\hat{\beta}$, estimated regression coefficient; LASSO, least absolute shrinkage and selection operator; LDH, lactate dehydrogenase; ULN, upper limit of normal.

in high-dimensional space: computational intensity and a high false discovery rate.[38,39] Of note, several prescreening methods are useful in identifying prognostic features both in the large *p*, small *n* problem and in ultra-high-dimensional space.[31,32,39-41]

The concept of variable selection is more challenging in building dynamic models because the main goal is to identify important factors that are related to the longitudinal process and the outcomes. In recent years, a few statistical studies extended the penalized method for the joint modeling of longitudinal data and survival outcomes.[42,43] The general idea is to postulate the joint likelihood linking the two submodels through latent random variables and to add shrinkage operators to select fixed and random effects. He et al[42] proposed a variable selection method for joint modeling with a univariable longitudinal outcome, and Chen and Wang[43] extended the framework to incorporate multiple longitudinal outcomes. While these methods have not been implemented in oncology, the statistical development paves the way for dynamic risk prediction.

Heterogeneity of treatment effect (HTE) is another important area to consider when building prognostic models. HTE is the nonrandom, explainable variability in the direction and magnitude of treatment effects for individuals within a population.[44] There are different sources for HTE, and HTE may arise from an underlying causal mechanism, artifacts, measurements, or methods.[45,46] One main goal of HTE analyses is to predict whether a patient might benefit from a treatment. Traditionally, the Cox regression method has been used to identify subgroups of patients who may benefit from treatment.[14] Recursive partitioning also has been used to identify a subgroup of patients.[47] While these classification tree methods have several advantages, they can create complicated structures and produce overfitting.[19,41,47] Other methods, such as permutation methods, SIDES (subgroup identification based on differential effect search), doubly robust augmented inverse probability–weighted estimator, and virtual twins, have been developed to take HTE into consideration.[45,48-52] The personalized prediction can be adequately addressed by the dynamic prediction framework, although it is an area for future research.

ESTIMATING PATIENT-SPECIFIC OUTCOME PREDICTION AND CONSTRUCTING RISK GROUPS

Once the final model is chosen, the next step is estimating patient-specific outcome prediction. The estimated survival function at time t is (Eq 3)

$$\hat{S}(t) = [\widehat{S}_0(t)]^{expR}, \quad (3)$$

where R is the estimated linear predictor or risk score for the ith individual ($= x_i^T\hat{\boldsymbol{\beta}} = \sum_{j=1}^{p}\hat{\beta}_j x_{ij}$), the baseline survival function is $\widehat{S}_0(t) = \exp(-\hat{H}_0(t))$, and the baseline cumulative hazard function is $\hat{H}_0(t)$. When we turned to our prognostic model of overall survival in prostate cancer, we computed a risk score from the estimated regression coefficients and the predicted survival at 24 months using the baseline cumulative hazard. We present the profiles of two patients with different baseline prognostic factors and their predicted overall survival at 24 months[5] (Table 2). We observe that patient 2 had a worse predicted survival probability at 24 months than patient 1.

Another important task in static predictive modeling is to construct prognostic risk groups, which can be formed on the basis of their quantiles from the estimated linear predictor. In our overall survival model, we constructed two and three prognostic risk groups and determined the cut points from the training set on the basis of quantiles (33rd, 50th, and 67th percentiles).[5] While demonstrating that the overall survival curves differ across the three risk groups is appropriate, this approach is not sufficient.[53] The optimal

 © 2019 by American Society of Clinical Oncology

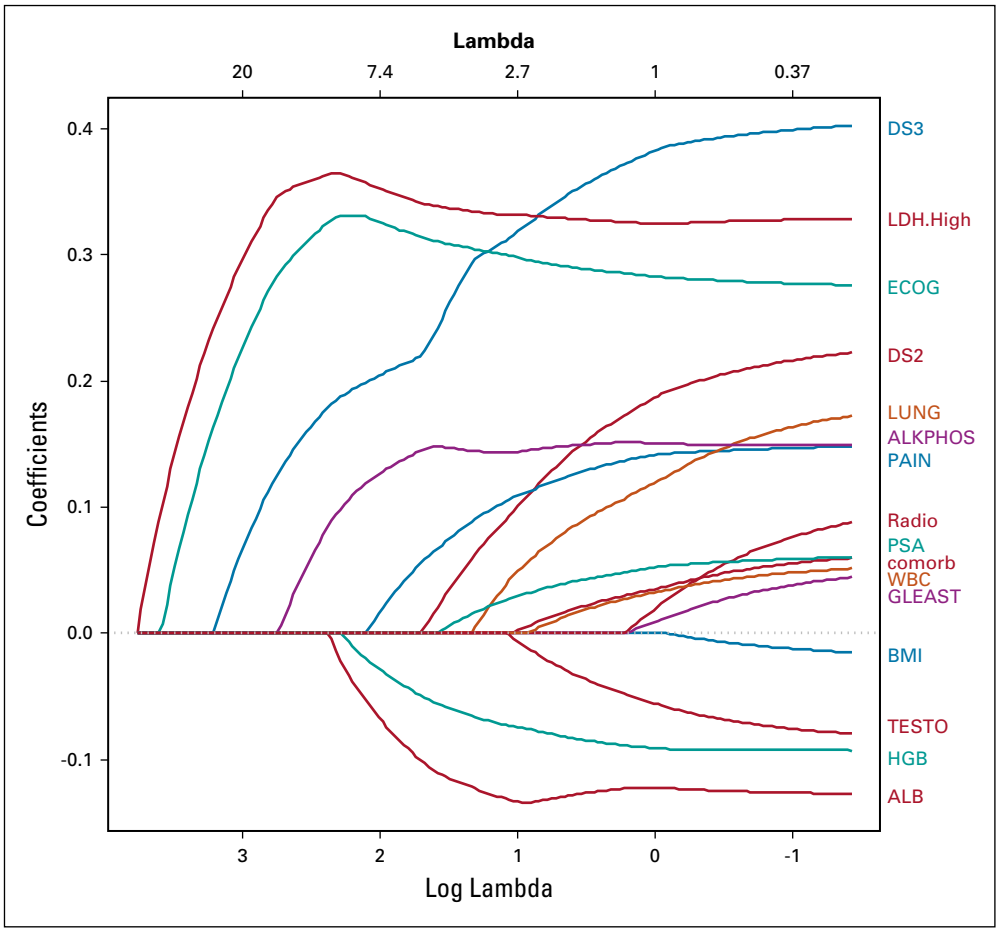

FIG 2. Solution path for adaptive least absolute shrinkage and selection operator. ALB, albumin; ALKPHOS, alkaline phosphatase; BMI, body mass index; comorb, comorbidity; DS2, bone metastases; DS3, visceral metastases; ECOG, Eastern Cooperative Oncology Group performance status; GLEAST, Gleason score; HGB, hemoglobin; LDH.High, lactate dehydrogenase greater than or equal to the upper limit of normal; PAIN, opioid analgesic use; PSA, prostate-specific antigen; Radio, prior treatment with radiotherapy; TESTO, testosterone. Reproduced with permission.[96]

strategy would be to compute a measure of discriminative ability of the model.

METHODS FOR DYNAMIC MODELING

Let T_i denote the true failure time and $f_i(t)$ a set of longitudinal measurements at some time points up to landmark time t. We are interested in predicting the probability that a new patient i^* is event free at least up to horizon time $u > t$ given survival up to t. The conditional probability is defined as (Eq 4)

$$\pi_{i^*}(u|t) = \Pr\left(T_{i^*} \geq u | T_{i^*} > t, f_{i^*}(t), D_n\right), \quad (4)$$

where D_n denotes a sample from the target population and on which the prediction is based. This formulation enables a dynamic updating scheme. Indeed, if a new

TABLE 2. Profiles of Patient-Specific Predicted Probabilities

Patient No.	Disease Site	Opiate Use	ECOG	LDH > ULN	ALB (g/dL)	HGB (g/dL)	ALK (U/L)	PSA (ng/mL)	Predicted Probability at 24 Months (95% CI)*
1	Bone	Yes	1	No	4	14	130	90	0.47 (0.42 to 0.52)
2	Visceral	Yes	1	Yes	3	10	90	75	0.28 (0.19 to 0.38)

NOTE. From Halabi et al.[5]

Abbreviations: ALB, albumin; ALK, alkaline phosphatase; ECOG, Eastern Cooperative Oncology Group performance status; HGB, hemoglobin; LDH, lactate dehydrogenase; PSA, prostate-specific antigen; ULN, upper limit of normal.

*Computed using the First-Line Metastatic Castrate-Resistant Prostate Cancer Patients risk classification tool.[97]

measurement for patient i^* is observed at time $t' > t$, we can update the risk prediction by calculating $\pi_{i^*}(u|t')$.

There are two general dynamic risk prediction frameworks: joint modeling and landmark analysis. Joint modeling comprises two linked submodels, one for the longitudinal process and one for the time-to-event data, and both depend on a common set of latent random variables.[54,55] In particular, the longitudinal data usually are modeled by a linear mixed-effects model. The time-to-event data are modeled by the proportional hazards model, with true longitudinal processes as time-varying covariates. The Cox regression coefficient quantifies the association between the latent longitudinal process and the hazard rate at time t. The longitudinal process and the event time process are assumed to be independent given the latent random effects, and the joint likelihood can be derived. The model can be estimated either using a frequentist approach that attains maximum likelihood through an expectation-maximization algorithm[54,56-58] or a Bayesian approach that uses Markov chain Monte Carlo to obtain posterior means.[59-61] While assuming that the parameters are readily estimated from the observed data, the conditional probability $\pi_{i^*}(u|t)$ can be computed. A Monte Carlo estimate of $\pi_{i^*}(u|t)$ can be obtained by sampling the random effects and the parameters from the corresponding distributions.[62]

On the other hand, landmarking[63-66] consists of a series of related Cox regression models, where each one is defined at a distinct landmark time t.[63-66] For each pair of $\{u,t\}$, a separate model is fitted to the individuals who remain in the study and have not yet experienced the event of interest. The baseline hazard can be estimated using Breslow's estimator.[67] Then $\pi_{i^*}(u|t)$ is computed as the survival probability that treats the longitudinal observation at time t as a baseline covariate.

COMPARISON BETWEEN JOINT MODELING AND LANDMARKING

Joint modeling and landmarking approaches differ in three aspects: model assumptions, information used, and computational complexity. Joint modeling models the dual distribution of the longitudinal process and the failure times and hence satisfies the consistency conditions for dynamic prediction.[68] Moreover, it exploits the full information of collected data and takes into account the measurement error of the longitudinal data. This latter point is critical because it implies that joint modeling is more efficient than landmarking. However, joint modeling needs to specify a correct model for the longitudinal process and requires stronger assumptions than landmarking. Joint modeling also takes a considerable amount of effort to estimate the parameters, and the computational cost is high because it involves complicated joint distribution and numerical integration. In contrast, landmarking circumvents the aforementioned model assumptions and computational burden, but it is not a comprehensive probability model of the longitudinal process and failure times and, as such, does not satisfy the consistency conditions. Another major shortcoming is that landmarking only considers the patients at risk at the landmark time and does not fully explore the information.

Several articles have focused on the comparison of joint modeling and landmarking. Rizopoulos et al[69] compared the two prediction frameworks and proposed a compromise between joint modeling and landmarking. Suresh et al[70] contrasted joint modeling and landmarking for dynamic risk prediction in the context of a binary longitudinal marker and applied these approaches to a prostate cancer study. Ferrer et al[71] compared the two approaches in the case of model misspecification, and they aimed for predicting competing risks of prostate cancer from PSA history.

Dynamic risk prediction (joint modeling) relies on model assumptions, and hence, its performance suffers from model misspecification. Functional data analysis, which is a nonparametric framework, has received considerable attention in medical studies because it is a flexible tool for modeling nonlinear longitudinal processes.[72-74] In particular, these methods have been incorporated with joint modeling[75] and are exploited to construct dynamic prediction models.[61,76-78]

EXAMPLES

A few examples exist in cancer where longitudinal data were modeled with clinical outcomes. We have previously explored whether PSA decline at different landmark times is prognostic for overall survival in patients with advanced prostate cancer.[79] Fontein et al[80] developed a dynamic model for predicting overall survival in patients with breast cancer. The authors validated the overall survival model using a dynamic cross-validated C-index and reported C-indices of 0.72, 0.76, and 0.79 at 1, 2, and 3 years, respectively. Suresh et al[70] and Ferrer et al[71] extended the landmarking approach and used prostate cancer studies as the testing bed. In addition, Proust-Lima and Taylor[81] developed a dynamic prognostic tool that is based on joint modeling using PSA as a biomarker for prostate cancer recurrence. Other applications of dynamic models have been implemented to prostate cancer[82-85] and colorectal cancer.[86]

We demonstrated the application of dynamic risk prediction using the DATATOP study,[87] a clinical trial that examined the benefits of deprenyl and α-tocopherol in slowing the progression of Parkinson disease (PD).[88] Multiple longitudinal biomarkers were collected in the DATATOP study, including Unified PD Rating Scale total score, modified Hoehn and Yahr scale, and Schwab and England activities of daily living. The biomarkers measured at baseline, 1 month, and every 3 months showed strong correlation between PD symptoms and the terminal event. We applied a joint modeling framework to account for the informative event times. We developed a Bayesian approach for parameter estimation and predicted patients' future outcome

 © 2019 by American Society of Clinical Oncology

trajectories (Fig 3A) and risk of functional disability (Fig 3B). Patient 169, who had more severe disease with earlier development of functional disability, and patient 718, who had less severe disease, were selected to illustrate the patient-specific predictions at clinically relevant future time points conditional on their available longitudinal measurements. The predicted Unified PD Rating Scale trajectories were biased, with wide uncertainty bands when only baseline measurements were used. Although dynamic prediction for longitudinal trajectories is an advantage of the

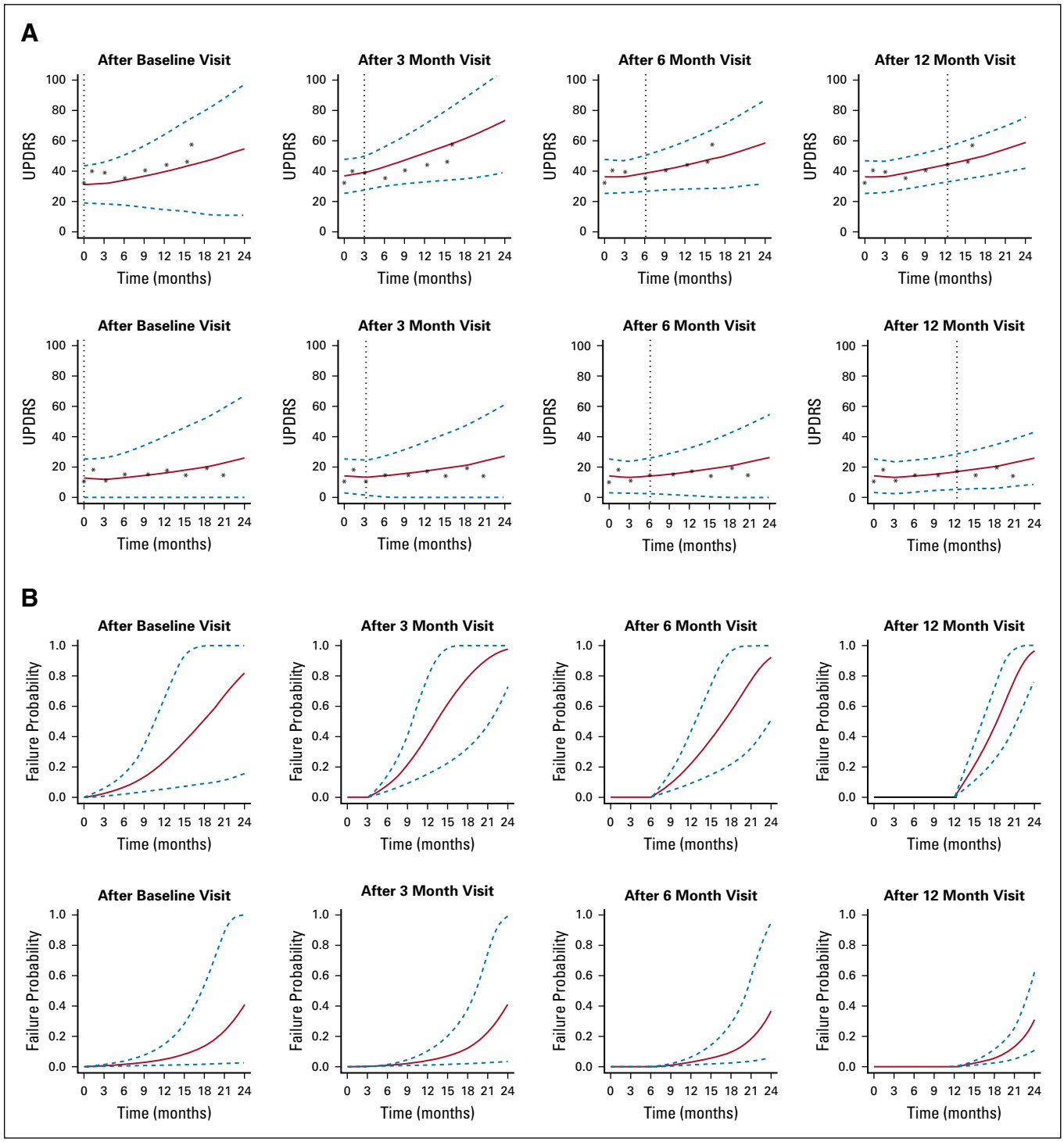

FIG 3. (A) Predicted Unified Parkinson Disease Rating Scale (UPDRS) trajectories and (B) predicted conditional failure probabilities for patient 169 (top rows) and patient 718 (bottom rows). Solid lines are the means of 2,000 Markov chain Monte Carlo samples. Dashed lines are the 2.5 and 97.5 percentiles of the 2,000 Markov chain Monte Carlo samples. The vertical lines represent the landmark time. Reproduced with permission.[87]

joint modeling, our major interest was to predict the probability of functional disability after visits at time *t* given the patients' longitudinal profiles and event-free status up to time *t*. We found a similar pattern in that the risk predictions with higher accuracy were achieved on the basis of more longitudinal observations. For example, according to the longitudinal profiles of the first 12 months, the predicted probabilities in the next 3, 6, 9, and 12 months were 0.21, 0.46, 0.78, and 0.97, respectively, for patient 169 (Fig 3B, last plot, top row) and 0.02, 0.06, 0.13, and 0.30, respectively, for patient 718 (Fig 3B, last plot, bottom row). Therefore, patient 169, who had a higher risk of functional disability in the next few months, might need attentive medical intervention to control disease progression. Meanwhile, the Brier's scores were 0.216 and 0.108, respectively, which implied a better prediction in terms of calibration given more follow-up data.

VALIDATION AND ASSESSMENT OF PROGNOSTIC MODELS

The primary goal of a risk assessment is to provide accurate outcome prediction in new patients.[16,89] Overfitting remains one of the main challenges in model building. Overfitting occurs when a high predictive accuracy is estimated from a model that has been applied to the training set but has low accuracy when assessed in an independent data set.[90] A good example of overfitting is provided in Halabi and Owzar.[2]

The validation of a prognostic model is considered a critical step after a risk assessment model has been built. There are two types of validation: external and internal.[16,89,90] External validation, where the frozen model from the training data is applied to an independent data set, is the most rigorous approach. However, investigators often may not have access to the external data set. Of note, other types of resampling methods, such as cross-validation, bootstrapping, and bootstrapping using 0.632+, are considered appropriate approaches to model validation.[32,53,91,92]

Assessment of the model's performance usually is conducted by examining the calibration and discriminative ability of the model. Calibration signifies the extent of the match between the predicted and observed outcome.[16] Often, investigators plot the predicted versus the observed outcome. The model would be calibrated if the data fall on a 45° line. Using data from two phase III clinical trials, we evaluated the overall survival model for calibration at different landmarks.[36,93] Figure 4A shows that the predicted survival probabilities at 18 months in the Enthuse 33 trial were close to the proportion of patients who survived 18 months. On the other hand, Figure 4B demonstrates that the model was not well gauged because the observed-predicted data points did not fall on the 45° line. The first two points (circles) show that the model overpredicted the proportion of patients who survived 12 months, whereas the third and fifth data points show that the model underpredicted the proportion of patients who survived 12 months.

Discrimination describes the ability of a prognostic model to distinguish between patients with and without the outcome of interest.[16] Several metrics are used to report the performance of a model. A widely used measure is concordance (C-index), which is the agreement between observed outcomes and prediction. Another widely used measure of predictive accuracy is the time-dependent area under the receiver operating characteristic curve (tAUROC),[94] which can be combined to form an integrated AUROC for the whole range of the study.[95] Circling back to our prognostic model in prostate cancer, we evaluated the performance of the model by implementing the integrated measure for the tAUROC by Uno et al,[95] which was 0.73 (95% CI, 0.70 to 0.73) and 0.76 (95% CI, 0.72 to 0.76) in the testing and validation sets, respectively.[5]

Dynamic models also need to be assessed for their discriminative ability and calibration. These measures can evaluate the performance of the model at various time points of the prediction. The tAUROC and Brier's score are widely used for dynamic prediction validation.[69,70,82,87] In the DATATOP study,[87] we applied fivefold cross-validation to evaluate the predictive performance of our framework. Conditional on the longitudinal history up to month 3 and month 12, our model yielded tAUROCs of 0.744 and 0.766, respectively, for correctly assigning a higher risk of functional disability by month 15 to patients with more severe disease.

Criteria for evaluating risk assessments have been published by the Precision Medicine Core of the American Joint Commission on Cancer,[53] the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis,[92] and the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies.[91] Investigators are encouraged to follow these guidelines as more rigorous tools of clinical outcomes would be developed in oncology. These models are anticipated to be implemented in both the design and the conduct of future trials.

Although static models will continue to play an important role in oncology, developing and validating dynamic models of clinical outcomes need to take a higher priority. A framework for developing and validating dynamic tools in oncology seems to be needed. One of the limitations is that modelers may be constrained by the lack of access to the longitudinal biomarker data; therefore, the next generation of risk assessments are highly recommended to take into consideration the longitudinal biomarker data and outcomes so that predictions are updated.

In summary, risk assessment will remain an important research task in precision oncology. We advocate for good clinical practice in risk assessment studies and recommend that investigators design these studies prospectively to obtain accurate individual outcome prediction and prognostic risk group classification. Prognostic studies should begin by asking fundamental questions that are

 © 2019 by American Society of Clinical Oncology

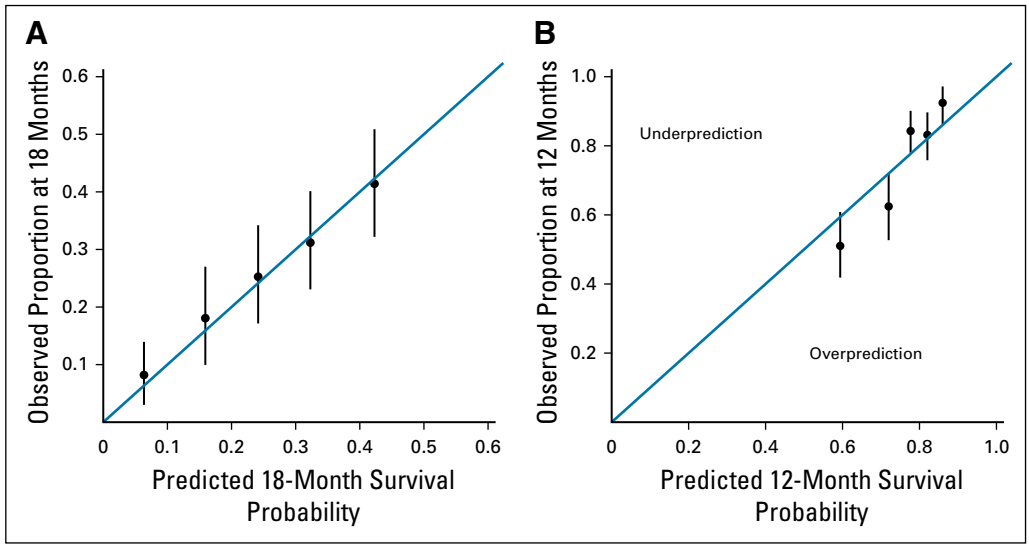

FIG 4. Calibration of the overall survival model for observed and predicted survival probability at (A) 18 months and (B) 12 months.

pertinent to patient outcomes, define the primary end point a priori, justify the sample size, and describe the appropriate methods for variable selection and model assessment. Lastly, they should be validated using external data sets if available.

An understanding of the longitudinal relationship between host and tumor-related factors and their impact on clinical outcomes is critical. We expect to see an upsurge in dynamic risk assessments in oncology, and as such, the American Joint Committee on Cancer and the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis guidelines should be extended to dynamic predictive modeling. Regardless of whether static or dynamic modeling is the primary objective, we envision that this review will bridge gaps in knowledge and motivate investigators to take risk assessment as a discipline by itself. Funding opportunities with the primary goal of building and validating high-quality prognostic models will be critical for personalized predictions.

AFFILIATION

[1]Duke University Medical Center, Durham, NC

CORRESPONDING AUTHOR

Susan Halabi, PhD, Department of Biostatistics and Bioinformatics, Duke University, 2424 Erwin Rd, Ste 11088, Durham, NC 27705; Twitter: @DukeCancer; e-mail: susan.halabi@duke.edu.

SUPPORT

Supported in part by US Army Medical Research Awards W81XWH-15-1-0467 and W81XWH-18-1-0278 and the Prostate Cancer Foundation. S.L.'s and C.L.'s research was partly supported by the National Institutes of Health (R01NS091307).

AUTHOR CONTRIBUTIONS

Conception and design: All authors
Collection and assembly of data: All authors
Data analysis and interpretation: All authors
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/po/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Susan Halabi
Employment: ASCO TAPUR
Consulting or Advisory Role: Eisai, Ferring Pharmaceuticals, Bayer AG

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

This article draws from and elaborates on the work of Halabi and Owzar,[2] Kelly et al,[41] and Halabi and Michiels.[96]

REFERENCES

1. O'Sullivan B, Brierley J, Gospodarowicz M: Prognosis and classification of cancer, in O'Sullivan B, Brierley JD, D'Cruz AK, et al: UICC Manual of Clinical Oncology (ed 9). West Sussex, UK, Wiley, 2015, pp 23-33
2. Halabi S, Owzar K: The importance of identifying and validating prognostic factors in oncology. Semin Oncol 37:e9-e18, 2010
3. Gospodarowicz MK, O'Sullivan B, Sobin LH: Prognostic Factors in Cancer (ed 3). Hoboken, NJ, Wiley-Liss, 2006
4. Halabi S, Small EJ, Kantoff PW, et al: Prognostic model for predicting survival in men with hormone-refractory metastatic prostate cancer. J Clin Oncol 21:1232-1237, 2003
5. Halabi S, Lin CY, Kelly WK, et al: Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. J Clin Oncol 32:671-677, 2014
6. Kelly WK, Halabi S, Carducci M, et al: Randomized, double-blind, placebo-controlled phase III trial comparing docetaxel and prednisone with or without bevacizumab in men with metastatic castration-resistant prostate cancer: CALGB 90401. J Clin Oncol 30:1534-1540, 2012
7. Morris MJ, Heller G, Bryce AH, et al: Alliance A031201: A phase III trial of enzalutamide (ENZ) versus enzalutamide, abiraterone, and prednisone (ENZ/AAP) for metastatic castration resistant prostate cancer (mCRPC). J Clin Oncol 37, 2019 (suppl; abstr 5008)
8. Sparano JA, Gray RJ, Makower DF, et al: Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. N Engl J Med 379:111-121, 2018
9. Bang YJ, Van Cutsem E, Feyereislova A, et al: Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): A phase 3, open-label, randomised controlled trial. Lancet 376:687-697, 2010
10. Nowell PC: The clonal evolution of tumor cell populations. Science 194:23-28, 1976
11. Bedard PL, Hansen AR, Ratain MJ, et al: Tumour heterogeneity in the clinic. Nature 501:355-364, 2013
12. Junttila MR, de Sauvage FJ: Influence of tumour micro-environment heterogeneity on therapeutic response. Nature 501:346-354, 2013
13. Hosmer DW, Lemeshow S, Sturdivant RX: Applied Logistic Regression. Oxford, UK, Wiley-Blackwell, 2013
14. Cox DR: Regression models and life-tables. J R Stat Soc B 34:187, 1972
15. Altman DG: Studies investigating prognostic factors: Conduct and evaluation, in Gospodarowicz MK, O'Sullivan B, Sobin LH: Prognostic Factors in Cancer (ed 3). Hoboken, NJ, Wiley-Liss, 2006, pp 39-54
16. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis Introduction (Springer Series in Statistics) (ed 2). New York, NY, Springer, 2015
17. Breiman L, Friedman JH, Olshen RA, et al: Classification and Regression Trees. Boca Raton, FL, Chapman & Hall/CRC, 1984
18. Leblanc M, Crowley J: Survival trees by goodness of split. J Am Stat Assoc 88:457-467, 1993
19. Hothorn T, Hornik K, Zeileis A: Unbiased recursive partitioning: A conditional inference framework. J Comput Graph Stat 15:651-674, 2006
20. Banerjee M, George J, Song EY, et al: Tree-based model for breast cancer prognostication. J Clin Oncol 22:2567-2575, 2004
21. Zhou X, Liu K-Y, Wong STC: Cancer classification and prediction using logistic regression with Bayesian gene selection. J Biomed Inform 37:249-259, 2004
22. Sparano JA: Prognostic gene expression assays in breast cancer: Are two better than one? NPJ Breast Cancer 4:11, 2018
23. Miller KD, O'Neill A, Gradishar W, et al: Double-blind phase III trial of adjuvant chemotherapy with and without bevacizumab in patients with lymph node-positive and high-risk lymph node-negative breast cancer (E5103). J Clin Oncol 36:2621-2629, 2018
24. Wishart GC, Azzato EM, Greenberg DC, et al: PREDICT: A new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Res 12:R1, 2010 [Erratum: Breast Cancer Res 12:401, 2010]
25. Rudloff U, Jacks LM, Goldberg JI, et al: Nomogram for predicting the risk of local recurrence after breast-conserving surgery for ductal carcinoma in situ. J Clin Oncol 28:3762-3769, 2010
26. Halbesma N, Jansen DF, Heymans MW, et al: Development and validation of a general population renal risk score. Clin J Am Soc Nephrol 6:1731-1738, 2011
27. Hastie T, Tibshirani R, Friedman JH: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY, Springer, 2017
28. Hastie T, Friedman J, Tibshirani R: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY, Springer, 2001
29. Tibshirani R: The lasso method for variable selection in the Cox model. Stat Med 16:385-395, 1997
30. Zhang HH, Lu WB: Adaptive lasso for Cox's proportional hazards model. Biometrika 94:691-703, 2007
31. Fan J, Feng Y, Wu Y: High-dimensional variable selection for Cox's proportional hazards model. Inst Med Stat Collect 6:70-86, 2010
32. Fan J, Lv J: Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Series B Stat Methodol 70:849-911, 2008
33. Meinshausen N, Bühlmann P: High-dimensional graphs and variable selection with the Lasso. Ann Stat 34:1436-1462, 2006
34. Zou H: Theory and methods—the adaptive lasso and its oracle properties. J Am Stat Assoc 101:1418-1429, 2006
35. Zou H, Hastie T: Regularization and variable selection via the Elastic Net. J R Stat Soc Series B Stat Methodol 67:301-320, 2005
36. Fizazi K, Higano CS, Nelson JB, et al: Phase III, randomized, placebo-controlled study of docetaxel in combination with zibotentan in patients with metastatic castration-resistant prostate cancer. J Clin Oncol 31:1740-1747, 2013 [Erratum: J Clin Oncol 32:3461, 2014]
37. White IR, Royston P: Imputing missing covariate values for the Cox model. Stat Med 28:1982-1998, 2009
38. Kim S, Halabi S: High dimensional variable selection with error control. BioMed Res Int 2016:8209453, 2016
39. Pi L, Halabi S. Combined performance of screening and variable selection methods in ultra-high dimensional data in predicting time-to-event outcomes. Diagn Progn Res 2:2, 2018
40. Zhao SD, Li Y: Principled sure independence screening for Cox models with ultra-high-dimensional covariates. J Multivariate Anal 105:397-411, 2012
41. Halabi SP, Pi L: Statistical considerations for developing and validating prognostic models of clinical outcomes, in Kelly DO, Kevin W, Halabi S (eds): Oncology Clinical Trials: Successful Design, Conduct, and Analysis (ed 2). New York, NY, Springer, 2018, pp 313-322
42. He Z, Tu W, Wang S, et al: Simultaneous variable selection for joint models of longitudinal and survival outcomes. Biometrics 71:178-187, 2015
43. Chen Y, Wang Y: Variable selection for joint models of multivariate longitudinal measurements and event time data. Stat Med 36:3820-3829, 2017
44. Varadhan R, Segal JB, Boyd CM, et al: A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. J Clin Epidemiol 66:818-825, 2013
45. Gabler NB, Duan N, Liao D, et al: Dealing with heterogeneity of treatment effects: Is the literature up to the challenge? Trials 10:43, 2009
46. Kravitz RL, Duan N, Braslow J: Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 82:661-687, 2004
47. Ruberg SJ, Chen L, Wang Y: The mean does not mean as much anymore: Finding sub-groups for tailored therapeutics. Clin Trials 7:574-583, 2010
48. Wang R, Schoenfeld DA, Hoeppner B, et al: Detecting treatment-covariate interactions using permutation methods. Stat Med 34:2035-2047, 2015

 © 2019 by American Society of Clinical Oncology

49. Foster JC, Taylor JM, Ruberg SJ: Subgroup identification from randomized clinical trial data. Stat Med 30:2867-2880, 2011
50. Dane A, Spencer A, Rosenkranz G, et al: Subgroup analysis and interpretation for phase 3 confirmatory trials: White paper of the EFSPI/PSI working group on subgroup analysis. Pharm Stat 18:126-139, 2019
51. Lipkovich I, Dmitrienko A: Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. J Biopharm Stat 24:130-153, 2014
52. Lipkovich I, Dmitrienko A, B R: Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials. Stat Med 36:136-196, 2017
53. Kattan MW, Hess KR, Amin MB, et al: American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. CA Cancer J Clin 66:370-374, 2016
54. Wulfsohn MS, Tsiatis AA: A joint model for survival and longitudinal data measured with error. Biometrics 53:330-339, 1997
55. Tsiatis AA, Davidian M: Joint modeling of longitudinal and time-to-event data: An overview. Stat Sin 14:809-834, 2004
56. Henderson R, Diggle P, Dobson A: Joint modelling of longitudinal measurements and event time data. Biostatistics 1:465-480, 2000
57. Rizopoulos D: JM: An R package for the joint modelling of longitudinal and time-to-event data. J Stat Softw 35:1-33, 2010
58. Rizopoulos D: Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. Boca Raton, FL, Chapman & Hall/CRC, 2012
59. Rizopoulos D, Hatfield LA, Carlin BP, et al: Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging. J Am Stat Assoc 109:1385-1397, 2014
60. Rizopoulos D: The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. J Stat Softw 72:1-46, 2016
61. Li K, Luo S: Dynamic predictions in Bayesian functional joint models for longitudinal and time-to-event data: An application to Alzheimer's disease. Stat Methods Med Res 28:327-342, 2019
62. Rizopoulos D: Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. Biometrics 67:819-829, 2011
63. Anderson JR, Cain KC, Gelber RD: Analysis of survival by tumor response. J Clin Oncol 1:710-719, 1983
64. Van Houwelingen H, Putter H. Dynamic prediction in clinical survival analysis. Boca Raton, FL, CRC Press, 2011
65. Van Houwelingen HC: Dynamic prediction by landmarking in event history analysis. Scand J Stat 34:70-85, 2007
66. Zheng Y, Heagerty PJ: Partly conditional survival models for longitudinal data. Biometrics 61:379-391, 2005
67. Breslow NE: Discussion of Professor Cox's paper. J R Stat Soc B 34:216-217, 1972
68. Jewell NP, Nielsen JP: A framework for consistent prediction rules based on markers. Biometrika 80:153-164, 1993
69. Rizopoulos D, Molenberghs G, Lesaffre EMEH: Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. Biom J 59:1261-1276, 2017
70. Suresh K, Taylor JMG, Spratt DE, et al: Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model. Biom J 59:1277-1300, 2017
71. Ferrer L, Putter H, Proust-Lima C: Individual dynamic predictions using landmarking and joint modelling: Validation of estimators and robustness assessment. Stat Methods Med Res 28:3649-3666, 2019
72. Yao F, Müller H-G, Wang J-L: Functional data analysis for sparse longitudinal data. J Am Stat Assoc 100:577-590, 2005
73. Yao F, Müller HG, Clifford AJ, et al: Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. Biometrics 59:676-685, 2003
74. Yao F, Müller H-G, Wang J-L: Functional linear regression analysis for longitudinal data. Ann Stat 33:2873-2903, 2005
75. Yao F: Functional principal component analysis for longitudinal and survival data. Stat Sin 17:965-983, 2007
76. Yan F, Lin X, Huang X: Dynamic prediction of disease progression for leukemia patients by functional principal component analysis of longitudinal expression levels of an oncogene. Ann Appl Stat 11:1649-1670, 2017
77. Yan F, Lin X, Li R, et al: Functional principal components analysis on moving time windows of longitudinal data: Dynamic prediction of times to event. J R Stat Soc Ser C 67:961-978, 2018
78. Li K, Luo S: Bayesian functional joint models for multivariate longitudinal and time-to-event data. Comput Stat Data Anal 129:14-29, 2019
79. Halabi S, Conaway MR, Small EJ, et al: Serum prostate specific antigen as a predictor of survival in prostate cancer patients treated with second-line hormonal therapy (CALGB 9181). TPR Prostate J 3:18-25, 2001
80. Fontein DB, Klinten Grand M, Nortier JW, et al: Dynamic prediction in breast cancer: Proving feasibility in clinical practice using the TEAM trial. Ann Oncol 26:1254-1262, 2015
81. Proust-Lima C, Taylor JM: Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: A joint modeling approach. Biostatistics 10:535-549, 2009
82. Sène M, Taylor JM, Dignam JJ, et al: Individualized dynamic prediction of prostate cancer recurrence with and without the initiation of a second treatment: Development and validation. Stat Methods Med Res 25:2972-2991, 2016
83. Taylor JM, Park Y, Ankerst DP, et al: Real-time individual predictions of prostate cancer recurrence using joint models. Biometrics 69:206-213, 2013
84. Yu M, Law NJ, Taylor JM, et al: Joint longitudinal-survival-cure models and their application to prostate cancer. Stat Sin 14:835-862, 2004
85. Yu M, Taylor JMG, Sandler HM: Individual prediction in prostate cancer studies using a joint longitudinal survival–cure model. J Am Stat Assoc 103:178-187, 2008
86. Król A, Ferrer L, Pignon JP, et al: Joint model for left-censored longitudinal data, recurrent events and terminal event: Predictive abilities of tumor burden for cancer evolution with application to the FFCD 2000-05 trial. Biometrics 72:907-916, 2016
87. Wang J, Luo S, Li L: Dynamic prediction for multiple repeated measures and event time data: An application to Parkinson's disease. Ann Appl Stat 11:1787-1809, 2017
88. Shoulson I, Group PS: DATATOP: A decade of neuroprotective inquiry. Ann Neurol 44:S160-S166, 1998
89. Steyerberg EW: Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York, NY, Springer, 2010
90. Steyerberg EW, Harrell FE Jr: Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 69:245-247, 2016
91. Moons KG, de Groot JA, Bouwmeester W, et al: Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS checklist. PLoS Med 11:e1001744, 2014
92. Moons KG, Altman DG, Reitsma JB, et al: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Ann Intern Med 162:W1-73, 2015

93. Petrylak DP, Vogelzang NJ, Budnik N, et al: Docetaxel and prednisone with or without lenalidomide in chemotherapy-naive patients with metastatic castration-resistant prostate cancer (MAINSAIL): A randomised, double-blind, placebo-controlled phase 3 trial. Lancet Oncol 16:417-425, 2015
94. Heagerty PJ, Zheng Y: Survival model predictive accuracy and ROC curves. Biometrics 61:92-105, 2005
95. Uno H, Cai TX, Tian L, et al: Evaluating prediction rules for t-year survivors with censored regression models. J Am Stat Assoc 102:527-537, 2007
96. Halabi S, Pi L, Lin C-Y: Developing and validating prognostic models of clinical outcomes, in Halabi S, Michiels S (eds): Textbook of Clinical Oncology: A Statistical Perspective. New York, NY, Chapman & Hall/CRC, 2019, pp 347-374
97. Duke University: First-Line Metastatic Castrate-Resistant Prostate Cancer Patients. https://www.cancer.duke.edu/Nomogram/firstlinechemotherapy.html

 © 2019 by American Society of Clinical Oncology