Neuro-Oncology

Neuro-Oncology 18(5), 609–623, 2016
doi:10.1093/neuonc/nov255
Advance Access date 8 December 2015

Statistical considerations on prognostic models for glioma

Annette M. Molinaro, Margaret R. Wrensch, Robert B. Jenkins, and Jeanette E. Eckel-Passow

Department of Neurological Surgery, University of California San Francisco (UCSF), San Francisco, California (A.M.M., M.R.W.); Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California (A.M.M., M.R.W.); Institute of Human Genetics, University of California San Francisco, San Francisco, California (M.R.W.); Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota (R.B.J.); Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota (J.E.E.-P.)

Corresponding Author: Annette M. Molinaro, PhD, UCSF Department of Neurosurgery, 400 Parnassus Ave A850b, Room A 808, San Francisco, CA 94143-0372 (annette.molinaro@ucsf.edu).

Given the lack of beneficial treatments in glioma, there is a need for prognostic models for therapeutic decision making and life planning. Recently several studies defining subtypes of glioma have been published. Here, we review the statistical considerations of how to build and validate prognostic models, explain the models presented in the current glioma literature, and discuss advantages and disadvantages of each model. The 3 statistical considerations to establishing clinically useful prognostic models are: study design, model building, and validation. Careful study design helps to ensure that the model is unbiased and generalizable to the population of interest. During model building, a discovery cohort of patients can be used to choose variables, construct models, and estimate prediction performance via internal validation. Via external validation, an independent dataset can assess how well the model performs. It is imperative that published models properly detail the study design and methods for both model building and validation. This provides readers the information necessary to assess the bias in a study, compare other published models, and determine the model's clinical usefulness. As editors, reviewers, and readers of the relevant literature, we should be cognizant of the needed statistical considerations and insist on their use.

Keywords: glioma, model building, prognostic models, statistics, validation.

In clinical practice, neuro-oncologists need tools to help patients understand their prognoses based on clinical and biological measurements. Prognostic models estimate the probability that a specific event (eg, progression, death) will occur in the future. Such models are important to patients, their families, and physicians, as they can inform therapeutic decision making and life planning as well as be used for stratifying patients in clinical trials.

Recently several studies have defined subtypes of glioma using combinations of molecular markers and have shown that patient's with different molecular subtypes have different survival.[1–6] Most used markers that were defined *a priori*, while one used an agnostic approach to identify molecular markers that defined groups of gliomas with different outcomes. Here, we review how to build and validate prognostic models, explain the models presented in the current glioma literature, and discuss advantages and disadvantages of each model. We then summarize the results of current glioma models and provide considerations for the development of future prognostic models for clinical use.

Study Design

The 2 fundamental steps for establishing prognostic models are model building and model validation. Prior to beginning either step, which patients are to be studied must be carefully determined. If the entry criteria for the study are too broad, important prognostic variables could be missed due to the heterogeneous selection of patients, resulting in a model with poor prognostic ability. However, if the entry criteria are too specific, then the results may not be generalizable to a wider population and thus unusable. Additionally, if the goal is to develop a prognostic model using genomic data, the importance and availability of other relevant clinical and pathologic data must be considered; for example, age, gender, disease stage/grade, and treatment all affect survival. Ultimately, the most important aspect of study design is to first consider the primary objective of the prognostic model and, likewise, identify the population for whom the prognostic model will be applied.

The preferable study design for building prognostic models is a cohort of patients collected prospectively, which allows the

Received 6 April 2015; accepted 14 September 2015

collection of all pertinent clinical and biological measurements in addition to a planned adequate sample size.[7] However, a cohort design usually is not practical for diseases with long dormancy or for a rare disease such as brain cancer. As a substitute for a cohort design, data collected as part of a clinical trial can be used to develop prognostic models. While clinical trial data are convenient, it is imperative to consider the primary objective of the corresponding clinical trial and, importantly, the inclusion and exclusion criteria that were used. That is, the primary objective of a clinical trial is typically to compare treatment therapeutics, and thus inclusion and exclusion are defined with respect to this primary objective. Thus, using clinical trial data to develop a prognostic model requires careful consideration to ensure that the results will be generalizable to the patient population of interest. As an alternative, a retrospective design is commonly used for rare diseases. While retrospective data collection allows for longer follow-up times, the accuracy of the data is dependent on recall of risk factors, and missing data are common.

In addition to deciding whether to use prospective or retrospective data collection, the number of patients to analyze is an extremely important consideration. While formal sample size calculators exist, a general rule of thumb is that a minimum of 10 events (eg, deaths in a survival analysis, the number of observations in the least frequent group in a logistic regression analysis) are necessary for each variable that will be considered in the prognostic model.[8–10] Estimating necessary sample sizes for validation datasets is similarly important and discussed below.

Model Building and Validation

After deciding on the cohort of patients to be studied and collecting the data (referred to as the discovery dataset), the first step of model building is to combine multiple predictors in a statistical model. The second step is to assess how well this model predicts future patients' outcomes (ie, prediction performance). Subsequently, the model should be further validated with an independent (external) dataset.

Statistical Models

The most common statistical approaches used to develop prognostic models include logistic regression for binary outcomes (eg, progression vs progression-free survival at 6 months) to assess the probability of progression at 6 months and Cox proportional hazards models for survival outcomes to assess risk of death/progression as a function of time.[11] Alternatively, more algorithmic methods such as recursive partitioning analysis can be employed for either type of outcome. Recursive partitioning (RP) analysis methods, such as Classification and Regression Trees (CART),[12] are useful for separating patients into groups with similar outcomes. For example, RP performs this task by starting with all patients in a dataset (the "root node") and splitting this group into 2 subgroups ("daughter nodes"), with the aim of maximizing the homogeneity with respect to the outcome within each subgroup. Further binary splitting occurs, and this process is continued until there are a prespecified minimum number of observations in a subgroup (node). Subsequently, a properly sized tree, as determined by cross-validation (described below), is chosen. The final subgroups are the "terminal nodes." An advantage of RP over logistic or Cox regression is that RP is less subjective in selecting and combining variables, as the algorithm, rather than the user, selects which and how the variables are included. RP is useful for accommodating and modeling nonlinear relationships, interactions, and variables with high correlation. It should be noted that the terminal nodes are not guaranteed to be statistically significantly different; RP does not provide statistical tests for the variables or corresponding nodes.

An important consideration, especially when using retrospective data, is how to accommodate missing data. Deleting observations with missing data can lead to bias and reduces the power of the model. One popular option for missing data in logistic and Cox regression is to use a multiple imputation approach before beginning variable selection and model building.[13] In multiple imputation, missing values are filled in with plausible values based on the other variables in the dataset. In RP, for each variable that the algorithm splits on, an ordered list of "next best variables" is automatically generated. The next best variable for making the current split is used in case the variable for the current split is missing. Sensitivity analyses can be performed to verify that the imputation did not skew the results.

Another important consideration is that new prognostic models should improve on previously established clinical models. For example, when looking at molecular markers for prognosis, the models should be compared with models with commonly available clinical variables known to be associated with the outcome. Such variables include age, extent of resection, and performance status. Inclusion of these variables should increase prediction accuracy as well as ease the introduction of the model into the clinic.

Once a statistical model is selected, the analyst must choose how many and which variables to include. This step is referred to as variable/model selection, the goal of which is to select the best subset of predictors. In logistic and Cox regression, typical approaches are forward and backward selection. Backward selection begins with all variables in a model, then one variable is removed at a time based on a prespecified test and a nominal *P*-value or a criterion like the Akaike information criterion (AIC). Forward selection begins with no variables in the model and adds one variable at a time based on a test and *P*-value or the AIC. There are noted complications with both forward and backward approaches[14]; therefore, a mix of the 2 is preferred. Alternatively, statistical methods that include automatic variable selection can be used, such as branch-and-bound algorithm, RP, Lasso, and least angle regression (LARS).[15,16] Lasso and LARS are both model selection algorithms that are an advanced form of forward selection. Both algorithms are easy to implement via statistical software such as R. As model building is a vast area of statistical research and much of it is beyond the scope of this manuscript, the reader is directed to Hastie et al[15] for further guidance.

Regardless of the model chosen, any assumptions of the statistical model should be verified. For example, in Cox regression, proportional hazards must be verified. In all regression models, the assumption of linearity of a continuous variable (eg, the risk associated with age increases as age increases) should be verified.

Prediction performance assessment. Subsequent to model and variable selection, the prediction performance of the chosen model is assessed, that is, how well the chosen model predicts the outcome. Prediction performance is estimated via calibration, discrimination, and misclassification.[17] **Calibration** is the agreement between the predicted risk probabilities and the observed frequencies of the outcome. Calibration is typically assessed by plotting the observed versus predicted outcomes; a 45-degree line denotes a perfectly calibrated model. **Discrimination** is the ability of the model to differentiate between those individuals that experienced the outcome and those that did not. Typical measures of discrimination include R^2 for regression models and the concordance index (c-index) for logistic and survival models. **Misclassification** is the correct/incorrect assignment into a risk group.

A difficulty with developing prognostic models is that due to either the analysis and/or study design, the results may not be reproducible. For example, a model may be "overfit." **Overfitting** means that a model is overly specific to the dataset from which it was developed and therefore is not replicable in an independent dataset or, more importantly, is not generalizable to the broader population.[18–20] There are 2 ways to check a model for overfitting and generalizability: internal validation and external validation. The best way to show that results are reproducible is to assess a model in a completely independent dataset (known as external validation, see below). However, due to small numbers or difficulty with obtaining similar cohorts, this is not always feasible.

Internal validation. **Internal validation** is a statistical technique to quantify overfitting and to get an unbiased estimate of prediction performance without an external dataset. **Resubstitution** refers to when a model is built using all of the data and subsequently prediction performance is evaluated on the same data.[17–19] Because resubstitution estimates model performance on the exact same data that were used to construct the model, estimates of prediction performance are **biased**, and thus the model may not generalize to the broader population.[19,21]

Other more favorable forms of internal validation include cross-validation and bootstrapping.[22] Bootstrapping and cross-validation allow the model to be built on a subset of the data and assessed on an entirely different subset.[19,22] Bootstrapping uses sampling with replacement such that the training set is the same size as the discovery cohort but has repeated patients (ie, the data from the same patient can appear multiple times). The test set is all patients not included in the training set and does not include any repeats. Alternatively, in 10-fold cross-validation, each patient is randomly assigned to one of 10 partitions, or folds (Fig. 1). As a result, the 10 partitions are of approximately the same size. For example, by default, RP uses 10-fold cross-validation for selecting the best size tree and estimating the prediction performance. For each of the 10 folds, a tree is built with the training set that contains all but one of the partitions, labeled the test set. Therefore, each partition acts as the test set exactly once. The prediction performance for the tree built on the training set is estimated with the corresponding test set and subsequently the estimate from the 10 test sets are averaged. This allows a prediction performance estimate for trees with various numbers of terminal nodes, that is, 1 (the root node) to 20 nodes representing a sequence of nested subtrees. The tree with the best prediction performance is chosen as the best size tree. The goal of maximizing the prediction performance is to choose the tree that will perform the best with an independent dataset by not overfitting the discovery data. For this chosen tree, the terminal nodes represent the stratification of the observations into similar risk groups.

The same resampling methods (such as cross-validation and bootstrapping) can be used for further partitioning the training set into learning and evaluation sets for variable/model selection (Fig. 2). As shown in Fig. 2, the discovery data can be split twice. The first split is for the purpose of estimating prediction performance; the data are split into a training set and a test set. A subsequent split of only the training set, into learning and evaluation sets, can be used for variable/model selection. For example, two-thirds of the discovery dataset could be used for the training set, leaving one-third for the test set. Subsequently, 10-fold cross-validation could be employed on the training set for variable/model selection. Once the variables are chosen and a model built, the test set would be used for evaluation.

The 2 splits into training/test sets and learning/evaluation sets can occur multiple times via resampling tools such as bootstrapping and cross-validation.[22] By repeating the splits, bias is reduced. Additionally, it allows the chance that every patient is included at least once in each of the 4 sets.[21,22]

There are several limitations to estimating internal validity. Most importantly, it is frequently performed incorrectly. For example, variable selection performed prior to using resampling methods for variable/model selection induces biased estimates of prediction performance. Thus, variable selection should be done within the resampling process (eg, within each fold of the cross-validation procedure) and not before implementation of the resampling process (eg, cross-validation).[22–24] Another frequent error is that often discovery data are only split into a training and test set once, which is referred to as the split sample method. It has been shown that the split sample method results in biased estimates of performance due to the reduced training set size that limits its ability to effectively select variables and fit a model.[22] Additionally, internal validation does not protect against biases due to problems associated with selection and handling of the relevant biological specimens (eg, different assays were used for a molecular marker for the tumors from all alive patients vs those that were dead). Furthermore, if the patient population is not representative of other patient populations, no amount of statistical maneuvering will guarantee good performance of the model in other settings. Thus, it is important to start with as representative a patient group as possible.

External validation. **External validation** means validating the model via an independent dataset (eg, a different cohort from another institution or study).[22,25] If need be, a new cohort of patients from the same institution, but diagnosed at a later date, can be collected; however, this does not avoid the aforementioned study design limitations for internal validity. That is, there may still be inherent biases as to which patients come to a particular institution or in the way tumor tissues are handled. For sample size, a rule of thumb is to have an independent test

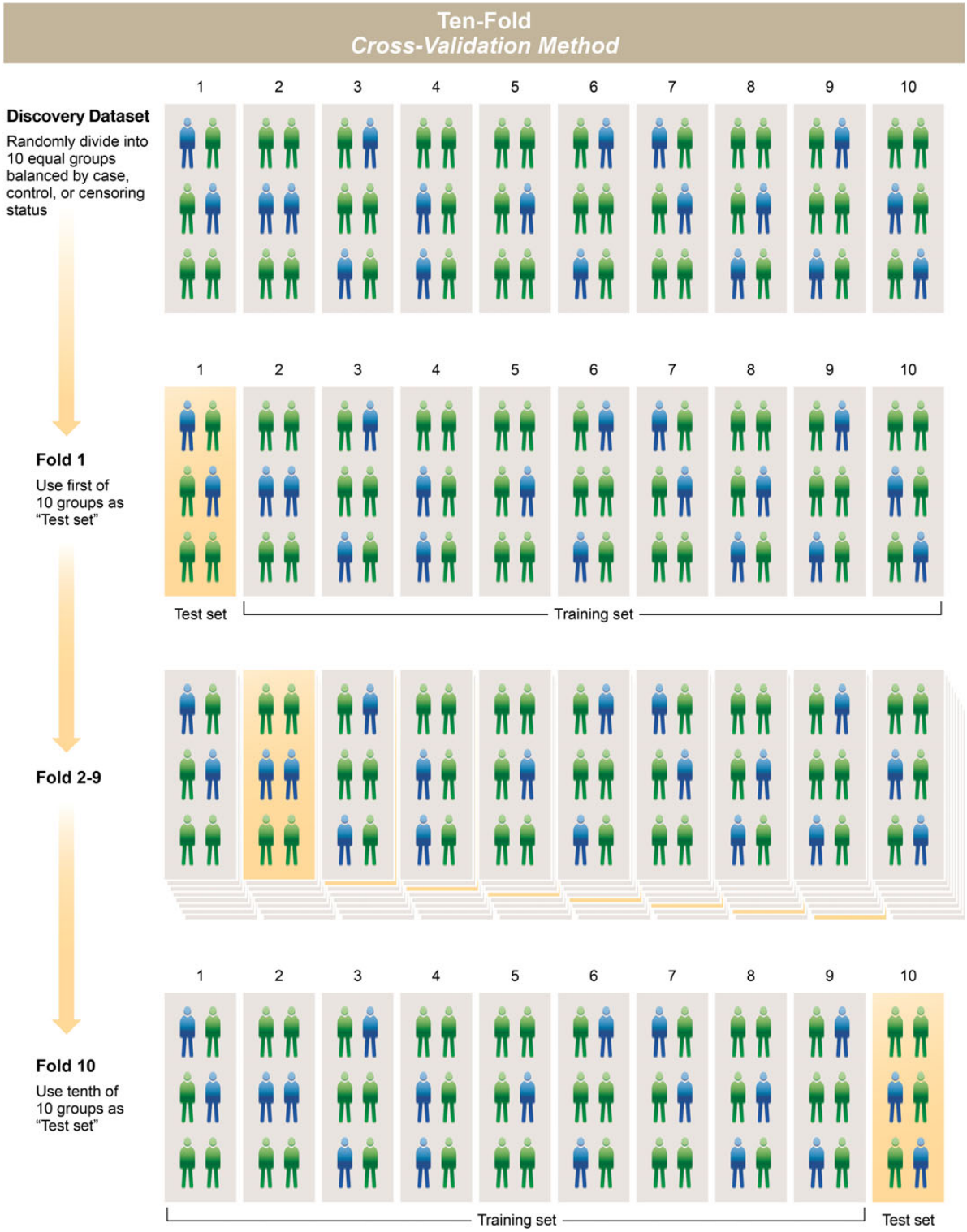

Fig. 1. Depiction of training and test sets for the iterations of 10-fold cross-validation.

set with 100 events and 100 non-events for binary outcomes or 100 events for survival outcomes.[26] Smaller test sets can lack the power to test differences in model performance.

To evaluate a model in an independent dataset, the same model that was built from the discovery data is applied to the external dataset and prediction performance is evaluated.

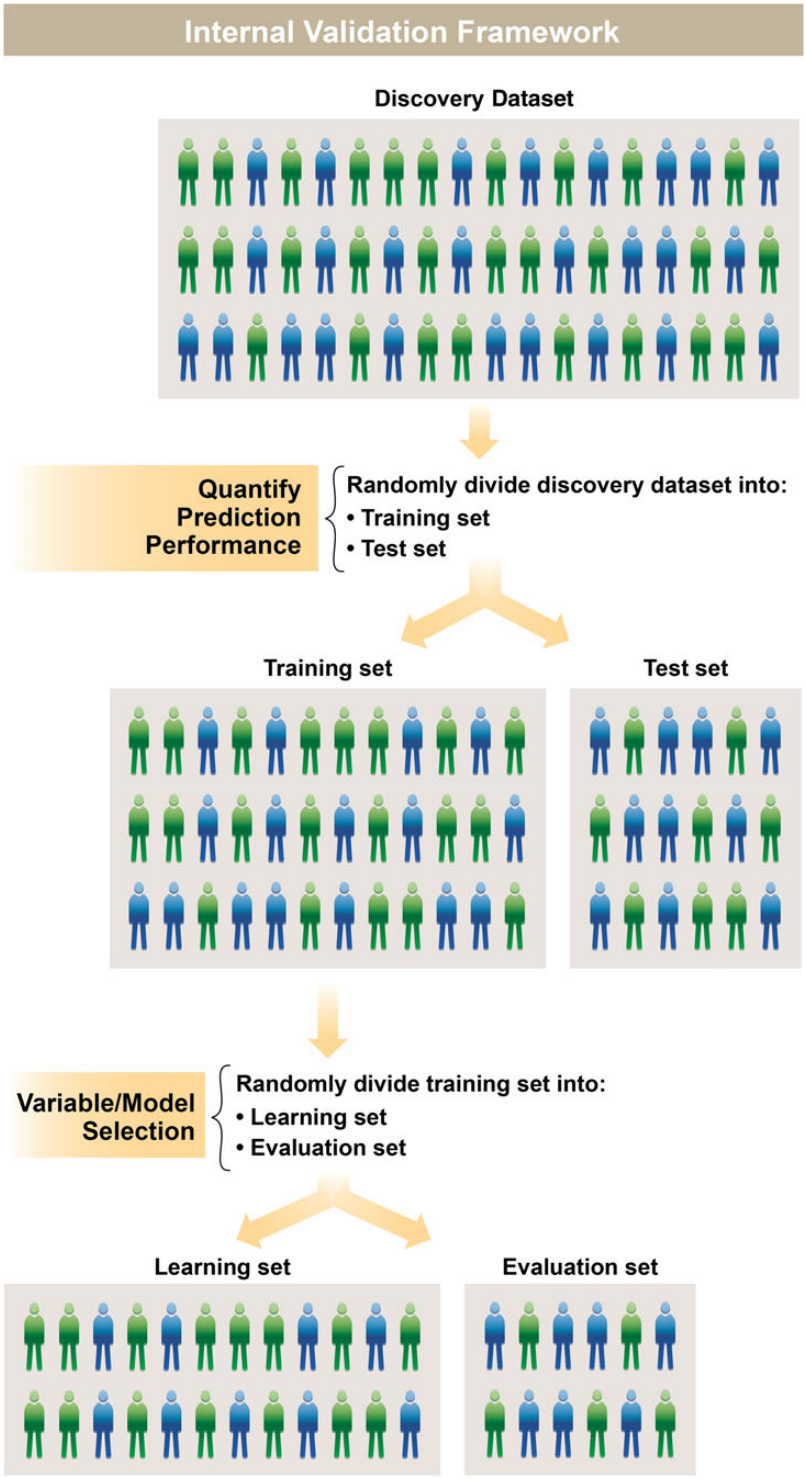

Fig. 2. Framework on internal validation for allocating data into training, learning, evaluation, and test sets for the purposes of quantifying prediction performance and variable/model selection.

Typically, the prediction performance will not be as high as that from internal validation. If the prediction performance is poor, then instead of starting again or building a new model, the original model can be updated.[27,28] For example, if changes in clinical practice have happened over time, specific adjustments to a subset of predictors can be made. However, the updated model should similarly be subsequently validated in an external dataset.

Recent Examples in Neuro-oncology

To illustrate the aforementioned statistical considerations, we present several recently published models in the neuro-oncology literature. In this section we detail the advantages and disadvantages of each model and compare the prediction performance of the models in an independent test set of glioma patients (low-grade gliomas and glioblastoma multiforme [GBM]) from The Cancer Genome Atlas (TCGA) and Mayo Clinic. TCGA data are publicly available and include 127 grade II, 138 grade III, and 153 grade IV glioma patients with clinical and biomarker data. Mayo Clinic data include 88 grade II, 112 grade III, and 117 grade IV glioma patients with clinical and biomarker data. Unless noted, all studies defined overall survival (OS) as the time from diagnosis to death or last follow-up. The patient and other features of these studies are summarized in Table 1.

Panageas et al, 2014

Panageas and colleagues[1] studied 587 patients with anaplastic oligodendrogliomas and oligoastrocytomas with known 1p/19q codeletion status from a retrospective database. Their goal was to assess the prognostic value of 1p/19q codeletion in the presence of other known important clinical variables for the purpose of clinical management/prognostication. The variables investigated included: age (continuous), history of prior low-grade glioma (yes vs no), 1p/19q status (codeleted vs not codeleted), histology (presence or absence of an astrocytic component), tumor lobe (frontal vs other), tumor hemisphere (right vs left vs bilateral), gender, extent of resection (biopsy vs debulking), postoperative treatment (radiation only vs chemotherapy), and KPS status at diagnosis ($\geq$70 vs $<$70).

To classify patients into risk groups, the authors used RP, where the final tree had 7 nodes, chosen by cross-validation. As several of the terminal nodes had similar hazards, the authors subsequently condensed the 7 nodes into 5 (Table 1, Fig. 3A). The median OS ranged from 9.3 years for patients younger than 60 with 1p/19q codeletion, to 0.6 years in those patients 70 and older without 1p/19q codeletion. The authors discussed that they did not perform external validation, as potential datasets had too much heterogeneity in treatment and exclusion criteria. Thus, we performed external validation using the combined TCGA and Mayo Clinic data, restricted to 147 patients with anaplastic oligodendrogliomas or oligoastrocytomas (Table 1). In the TCGA + Mayo data, only the patients without 1p/19q codeletion who were 60–69 or 43–59 years old with non–frontal lobe tumor locations were statistically different from those under 60 with 1p/19q codeletion (hazard ratio [HR] = 3.3, $P = .04$; Fig. 3A, Table 1). In summary, while the median OS estimates differed between what was reported by Panageas et al[1] and what was observed in the TCGA + Mayo data, there is an overlap in the 95% confidence intervals. Additionally, the survival curves (Fig. 3) demonstrate a similar trend. The observed differences are most likely due to the fact that the TCGA + Mayo validation set does not have enough events and non-events overall as well as in each subtype to adequately validate the model. Additional reasons could be differences in treatment across the datasets or inclusion criteria of the 3 different studies.

The advantages of the Panageas et al[1] prognostic model were (i) the relatively large sample size for model building; (ii) the objective nature of building a prognostic model, where the authors did not choose how they defined the risk groups but instead let the RP algorithm select the variables (from those that the authors included, which can add subjectivity) and combinations thereof; (iii) the inclusion of known prognostic clinical variables (eg, age at diagnosis and tumor location); (iv) internal validation via 10-fold cross-validation; and (v) a restriction to 2 types of low-grade glioma as opposed to all low-grade gliomas, which are known to be heterogeneous. The disadvantages were (i) the ad hoc combination of terminal nodes based on similar hazards – this step reduces the objectivity of the method and was performed after cross-validation precluding an accurate estimate of prediction performance[29]; (ii) the lack of external validation; and (iii) the *P*-value for the log-rank test was included but the hazard ratios and accompanying *P*-values were not. One way to address the first limitation is to include a multilevel categorical variable representing membership in the terminal nodes and then let RP group the terminal nodes, in effect running RP twice. The last limitation is important, as the *P*-value from a log-rank test can be $<.05$ if only one of the 5 groups is different from the other 4. That is, a log-rank *P*-value $<.05$ does not imply that all pairwise comparisons are significantly different from each other. Additionally, a *P*-value, unlike a hazard ratio, does not provide an estimate of the magnitude of difference. An alternative is to choose one of the groups as the reference (this group should have a large number of observations) and report the hazard ratios and confidence intervals for each of the other groups in comparison to the reference.

Labussière et al, 2014, Neurology

Labussière et al[2] studied 395 newly diagnosed cases of GBM with no history of seizure or previous low-grade glioma. Their goal was to assess the prognostic value of telomerase reverse transcriptase promoter (TERTp) mutations as well as the association with common molecular alterations, including isocitrate dehydrogenase 1 (IDH1) mutation, epidermal growth factor receptor (EGFR) amplification, cyclin-dependent kinase inhibitor 2A (CDKN2A) homozygous deletion, loss of chromosome 10, O^6-DNA methylguanine-methyltransferase (MGMT) promoter methylation, and tumor protein 53 (TP53) mutation.

The authors tested numerous possible combinations of TERTp mutation and associated markers and concluded that TERTp, IDH mutation, and EGFR amplification resulted in prognostic stratification, but the others did not. One finding they highlight is an interaction between EGFR amplification and TERTp mutation, that is, among patients with TERTp mutation, those whose tumor had EGFR amplification did better than those without. Using the 3 resulting markers (TERTp, IDH, and EGFR), the authors made 4 combinations (Table 1, Fig. 3B). The median OS ranged from 1.1 years for the TERTp–wild type (WT)/EGFR-amp group to 3.13 years for the TERTp-WT/EGFR-WT/IDH-mutation (MT) group. Although the authors did not include the number of patients in each of the 4 groups, it appears that there were approximately 8 deaths in the TERTp-WT/EGFR-amp group and 10 in the TERTp-WT/EGFR-WT/IDH-MT

Table 1. For recent published glioma model studies, the original definition of the model and reported median OS and results of the validation in TCGA + Mayo data

Author	Journal, Year	Grade	IDH	TERTp	1p/19q, Codeleted	EGFR	Age, y	Lobe	*N*	Median OS, y (95% CI)	*n*	Median Mayo and TCGA OS, y (95% CI)	Cox Proportional Hazards HR	*P*
Panageas	Neuro Oncol, 2014	III			Yes		<60		256	9.3 (8.4–16.0)	62	11.17 (8.4–NA)	1 (baseline)	
					No		<43		174	8.9 (5.5–10.8)	49	11.13 (6.3–NA)	1.05	.88
					No Yes		43–59 ≥60	Frontal	64	4.3 (2.9–6.0)	19	7.1 (4.3–NA)	1.96	.16
					No No		43–59 60–69	Not frontal	75	2.0 (1.6–2.3)	15	NA (0.97– NA)	3.3	.04
					No		≥70		18	0.6 (0.5–0.9)	2	9.53 (NA–NA)	1.64	.63
Labussière et al	Neurology, 2014	IV		MT						1.15	213	1.12 (1.03–1.23)	1 (baseline)	
				WT		Amplified				1.11	11	2.11 (1.50–NA)	0.52	.07
			MT	WT		WT				3.13	16	2.80 (2.02–NA)	0.42	.005
			WT	WT		WT				2.21	14	3.37 (0.33–NA)	0.89	.76
Labussière et al	BJC, 2014	II–IV	MT	MT					122	>17	155	11.95 (9.2–NA)	1 (baseline)	
			MT	WT					180	8.12	260	9.49 (7.09–13.27)	1.4	.1
			WT	WT					114	2.63	69	1.81 (1.55–2.89)	5.79	<.001
			WT	MT					340	1.28	250	1.15 (1.08–1.27)	13.02	<.001
Killela et al	Oncotarget, 2014	II	MT	WT					57	10.89 (7.93–12.08)	121	10.9 (8.86– NA)	1 (baseline)	
			MT	MT					46	17.13 (7.15–21.49)	78	20.3 (11.3–NA)	0.54	.08
		III–IV	WT	MT					187	0.96	245	1.15 (1.04–1.240)	1 (baseline)	
			WT	WT					50	1.43	58	1.77 (1.49–2.42)	0.52	<.001
			MT	WT					82	4.75	139	6.74 (5.61–13.11)	0.14	<.001
			MT	MT					40	10.42	77	8.54 (5.16–NA)	0.13	<.001
Brat et al[a]	NEJM, 2015	II–III	MT		Yes				84	8	74	20.08 (11.95–NA)	1 (baseline)	
			MT		No				139	6.3	105	13.27 (9.42–NA)	1.58	.11
			WT						55	1.7	21	2.49 (1.85–8.24)	6.38	<.001
Eckel-Passow et al[b]	NEJM, 2015	II–III	MT	MT	Yes				69	20.3 (11.96–NA)	64	7.94 (6.5–NA)	0.68	.33
			MT	MT	Mo				9	7.6 (5–NA)	5	2.15 (2.15 NA)	3.39	.24
			MT	WT	No				96	14.25 (9.5–NA)	144	6.25 (5.28–9.75)	1 (baseline)	
			WT	WT	Mo				8	9.0 (2.2–NA)	20	1.66 (0.87–NA)	3.02	.02
			WT	MT	No				13	2.2 (1.7–NA)	32	1.47 (0.73–NA)	14.03	<.001
		IV	MT	MT	Yes				1		0			
			MT	MT	No				6	1.06 (0.5–NA)	1			
			MT	WT	No				8	3.18 (13–NA)	8	2.8 (2.02–NA)	0.26	.01
			WT	WT	No				19	1.8 (1.4–4.5)	22	1.49 (0.43–NA)	0.55	.06
			WT	MT	No				83	1.2 (1.11–4)	122	1.04 (0.90–1.17)	1 (baseline)	

[a]Validated with Mayo data only.
[b]Validated with TCGA data only.

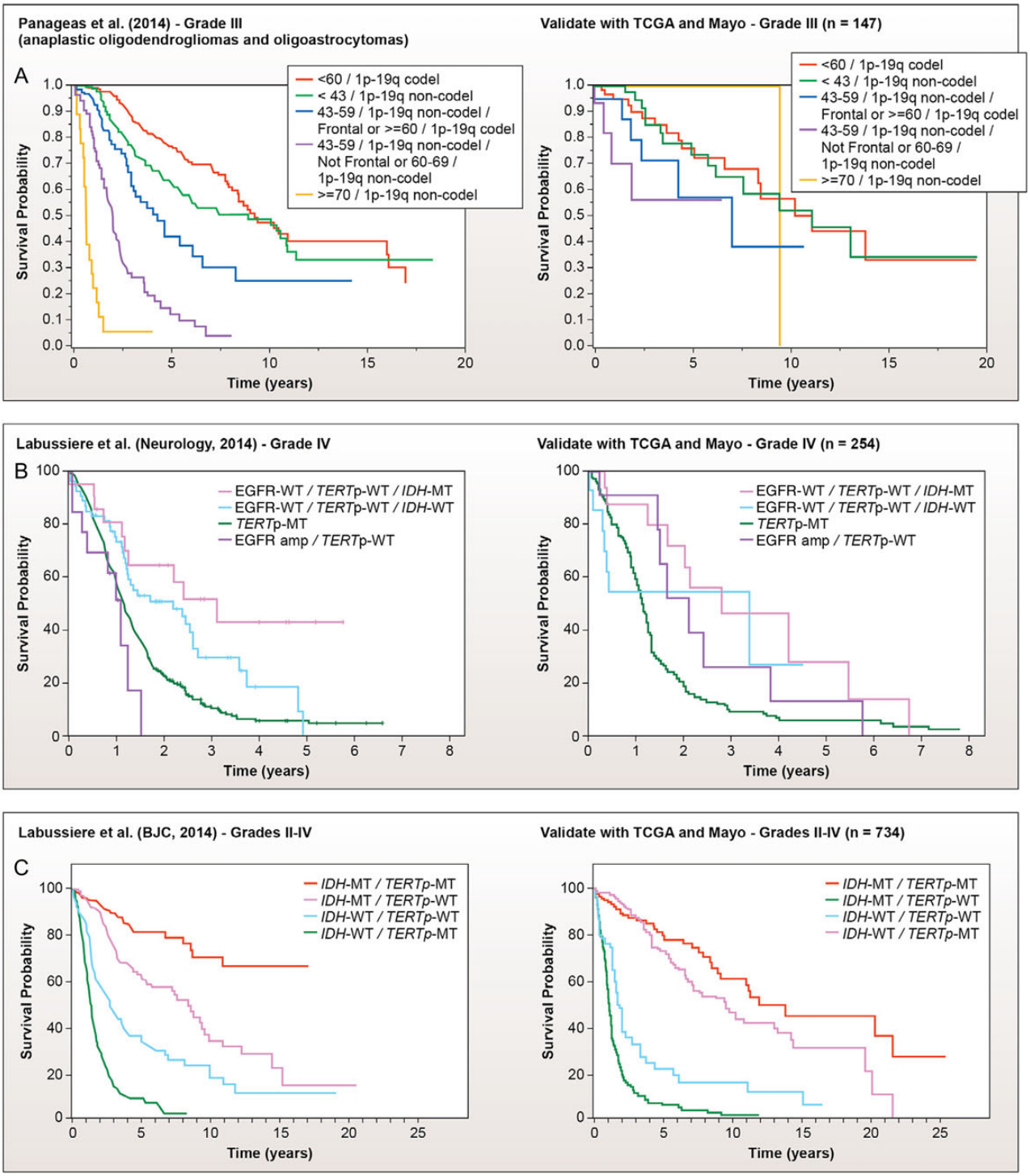

Fig. 3. For each of recent published glioma model studies, the original Kaplan–Meier curves on the left-hand side and the validation in TCGA data on the right-hand side.

group with multiple censorings, while the majority of patients fell into the other 2 groups. Since external validation was not performed, we performed external validation using the TCGA + Mayo data, restricted to 254 GBM patients (Table 1, Fig. 3B). It should be noted that we did not have access to seizure history and thus we did not restrict to GBM patients with a history of seizures. The TERTp-MT group had similar median OS between the published paper and the TCGA + Mayo validation, while the all-WT group differed by more than a year.

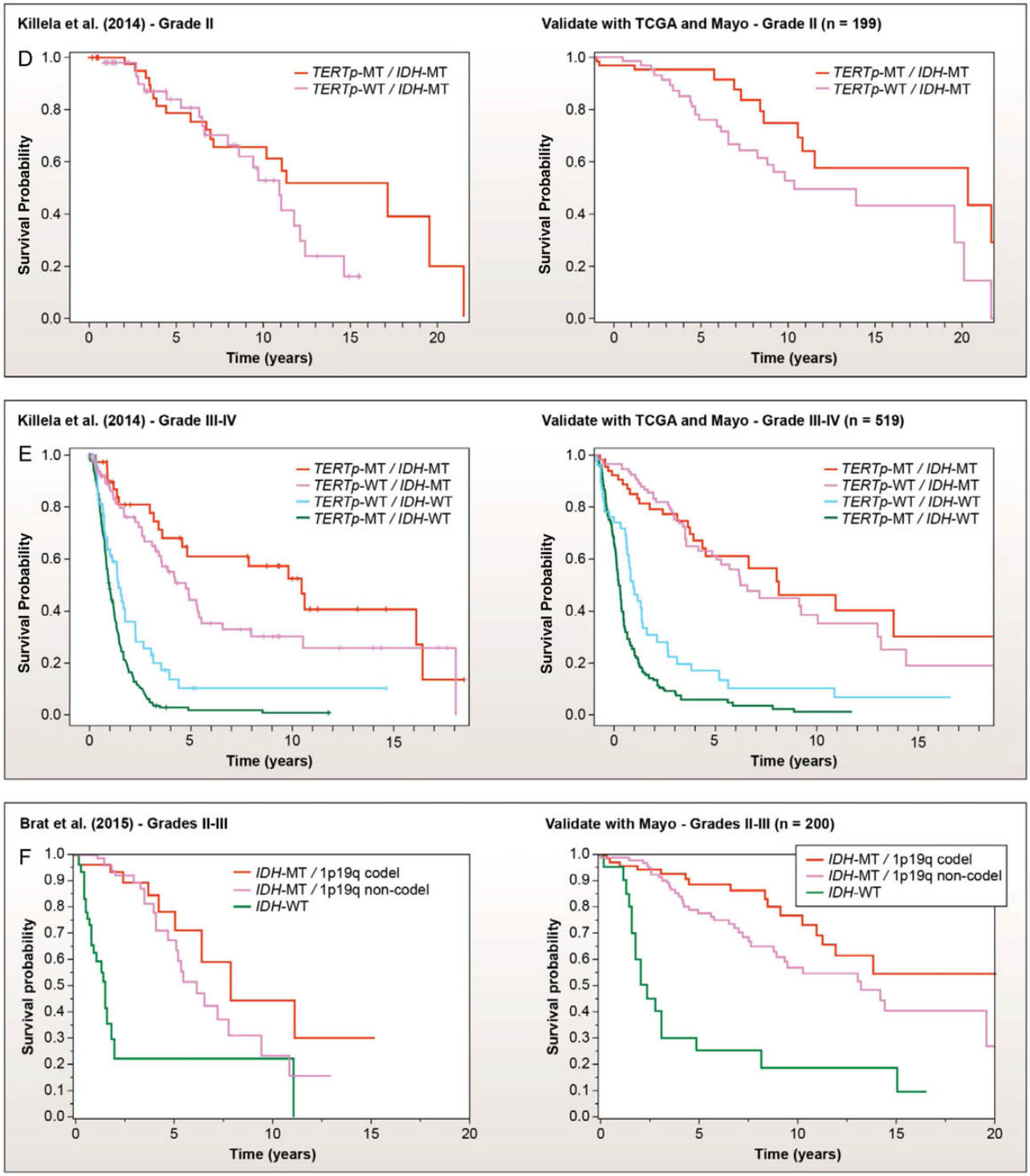

Fig. 3. *Continued.*

In the TCGA + Mayo data, only the TERTp-WT/EGFR-WT/IDH-MT group was statistically different from the TERTp-MT group (HR = 0.42, $P = .005$). In Labussière et al,[2] the TERTp-WT/EGFR-amp group did worse than the TERTp-MT patients and all died within 20 months. Conversely, in the TCGA + Mayo data, the TERTp-WT/EGFR-amp group had more favorable but not significantly different survival than the TERTp-MT group (HR = 0.52, $P = .07$). In addition, in the published manuscript, the all-WT group experienced intermediate survival, but in the TCGA + Mayo data these patients began

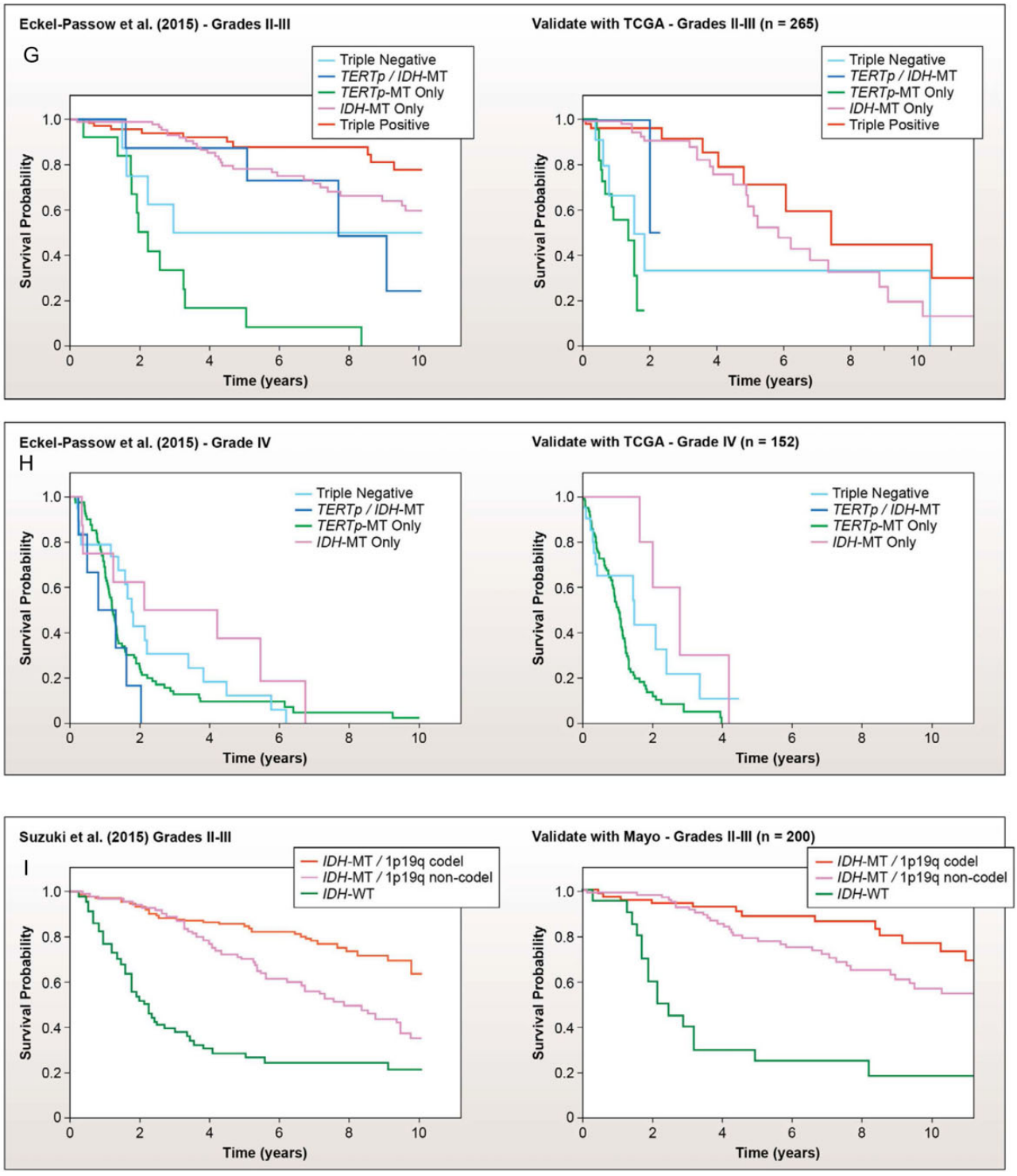

Fig. 3. *Continued.*

with a much worse survival, with 6 of the 14 patients dying by 5.2 months. In summary, the all-WT group and TERTp-WT/EGFR-amp groups differed between what was reported by Labussière et al[2] and what was observed in the TCGA + Mayo data for the median OS estimates. Additionally, the survival curves (Fig. 3B) demonstrate a similar trend for 2 of the groups: EGFR-WT/TERTp-WT/IDH-MT and TERTp-MT. The observed differences may be due to not excluding patient history of seizures in the TCGA + Mayo data and/or there were not enough events in the subgroups.

The advantage of this prognostic model was the large dataset. The disadvantages were (i) internal and external validation were lacking; (ii) there was no adjustment for known prognostic clinical variables; and (iii) no group sample sizes, HRs and corresponding CIs, or significance levels were provided. Interestingly, the noted interaction between TERTp and EGFR was not incorporated into the final 4 groups defined by Labussière et al.[2]

Labussière et al, 2014, British Journal of Cancer

Labussière et al[3] studied 807 primary grades II–IV glioma tumor patients. The authors did not state whether the GBM patients in this cohort overlapped with those in the *Neurology* manuscript.[2] Here, their goal was to assess the prognostic value of TERTp mutations and to determine whether TERTp mutation status provides additional prognostic value beyond IDH mutation status.

Using the 2 markers, the authors made 4 combinations (Table 1, Fig. 3C). The median OS ranged from 1.28 years for the worst group (TERTp-MT/IDH-WT) to >17 years in the best group (TERTp-MT/IDH-MT). Since external validation was not performed, we performed external validation using the TCGA + Mayo data, with all 734 gliomas (Table 1, Fig. 3C). The groups had similar median OS between the published results and the TCGA + Mayo validation set. In the TCGA + Mayo data, the TERTp-WT/IDH-WT (HR = 5.79, $P < .001$) and TERTp-MT/IDH-WT (HR = 13.02, $P < .001$) groups had significantly worse survival than the TERTp-MT/IDH-MT group, but the TERTp-WT/IDH-MT group did not (HR = 1.4, $P = .1$). In summary, the median OS estimates were not substantially different between the Labussière et al[3] data and the TCGA + Mayo data. Additionally, the survival curves (Fig. 3C) demonstrated very similar relationships across the 4 groups. The ability to validate the Labussière model is in part due to the fact that the TCGA + Mayo data had numerous events both overall and within each of the 4 subgroups.

The advantages of this prognostic model were the large dataset and the simple classification into 4 groups using only 2 markers that were defined *a priori*. The disadvantages were (i) the lack of internal and external validation; (ii) no confidence intervals or significance provided among the 4 groups (via HRs adjusted and unadjusted for known prognostic clinical characteristics); and (iii) lack of consideration of known prognostic clinical variables and other biomarkers.

Killela et al, 2014

Killela and colleagues[4] studied 473 grades II–IV glioma patients from a repository at Duke University. Their goal was to assess the prognostic value of IDH1/2 and TERTp mutations. The outcome was OS defined as time from surgery (instead of diagnosis) to death or last follow-up; time from surgery was most likely used, since secondary GBMs were included.

The authors formed 4 groups based on the presence or absence of the 2 markers and evaluated survival of each of the groups separately by grade. They found that of the 112 grade II patients, 103 (92%) could be identified by mutations in TERT and IDH or IDH alone (forming 2 groups); whereas in the 121 grade III and 240 primary/secondary GBM patients, all possible combinations of IDH and TERT were represented. The median OS for grade II ranged from 10.9 years for the TERTp-WT/IDH-MT group to 17.13 years for the TERTp-MT/IDH-MT group. Since external validation was not performed, we used the TCGA + Mayo data, with 199 grade II patients (Table 1, Fig. 3D) for validation. The median OS values for the 2 groups were similar between TCGA + Mayo and Killela et al.[4]

For the grades III–IV patients, the median OS ranged from 0.96 years for the TERTp-MT/IDH-WT group to 10.42 years for the TERTp-MT/IDH-MT group. For external validation, we used 519 grades III–IV patients available in the TCGA + Mayo data (Table 1, Fig. 3E). The median OS was similar among all groups. The other 3 groups were all significantly different from the TERTp-MT/IDH-WT group. However, it should be noted that the TERTp-MT/IDH-MT group was not statistically different from the TERTp-WT/IDH-MT group (HR = 1.1, $P = .72$). In summary, the median OS estimates were similar across the datasets, and the survival curves showed similar overall trends.

The advantage of the prognostic model was that it was simple: 2 markers were defined *a priori*. The disadvantages were (i) internal and external validation was not performed; (ii) known prognostic clinical variables were not considered; and (iii) HR for each risk group was not provided. It would be important to verify that the patients who were included as secondary GBM cases were not also included as primary grade II or grade III cases.

Brat et al, 2015

Brat et al[5] studied 293 grades II–III gliomas from TCGA. Their goal was to integrate multiple genomic data types (DNA copy number, DNA methylation, and mRNA and microRNA expression) to discover molecular groups with distinct outcomes in grades II–III. First, clustering was applied to each of the data types separately. Subsequently, clustering of clustering was performed, which yielded 3 molecular groups. Ultimately, the authors decided that 2 known markers (IDH mutation and 1p/19q codeletion) could largely describe these 3 groups.

Using IDH mutation and 1p/19q codeletion, the authors made 3 combinations (Table 1, Fig. 3F). The median OS ranged from 1.7 years for the IDH-WT group to 8 years for the IDH-MT/1p/19q codeleted group. We performed external validation using only Mayo Clinic data, which included 200 grades II–III gliomas (Table 1, Fig. 3F). If we had also used TCGA data we would have performed resubstitution, which, as discussed above, is not appropriate. In the validation Mayo Clinic data, the median survival for the IDH-MT groups was much greater than in the Brat et al[5] training data. The IDH-WT group was significantly different than the IDH-MT/1p/19q codeleted group, while the IDH-MT and 1p/19q non-codeleted group was not (HR = 1.58, $P = .11$). In summary, the median OS estimates were greater in the Mayo Clinic data, and the survival curves show similar overall trends.

The advantages of this prognostic model were the large dataset and the final simple classification that utilized only 2 markers. The disadvantages were (i) internal and external validation were lacking; (ii) the classifier was derived subjectively after doing many genomic analyses, including clustering and then combining those results into 3 groups; (iii) given the subjectivity, this approach would be difficult to reproduce.

Of interest, concurrently, Suzuki et al[30] published the same 3 groups as Brat et al[5] (Fig. 3I) using a combination of TCGA and Japanese patient data.[30] They did not include an external validation or estimates of median OS. In Fig. 3I, we show the same curves as in Fig. 3F using the Mayo Clinic data only for validation, as the Suzuki et al[30] training set included TCGA data. The median OS appears similar for the IDH-WT group but different for the other 2 groups.

Eckel-Passow et al, 2015

Eckel-Passow et al[6] used 317 glioma patients from a Mayo Clinic case-control series as the discovery set and then validated the results in 2 datasets: 351 glioma patients from the UCSF Adult Glioma Study (AGS) and 419 gliomas from TCGA. Their goal was to assess the value of 5 glioma molecular groups defined by 3 markers: TERTp mutation, IDH mutation, and 1p/19q codeletion. The authors were interested in assessing the value of these groups in terms of similarity of clinical variables, acquired somatic alterations, and germline variants. They also showed the age-adjusted survival experience of the 5 groups stratified by grade (II/III vs IV).

Using the 3 markers, the 5 combinations accounted for >95% of the Mayo Clinic glioma cases. For grades II–III, the median OS ranged from 2.2 years for the TERTp-MT only group to 20.3 years for the triple-positive group. In the published manuscript, 2 external test sets were utilized: UCSF AGS and TCGA. Here, we used only TCGA data, which included 265 grades II–III glioma (Table 1, Fig. 3G). Note, due to the small number of TCGA cases in the triple-positive group, the baseline for calculating HRs was the IDH-MT only group, as it was the largest. All groups had substantially lower median OS in TCGA validation versus the Mayo study, which is what we also observed when validating the results of Brat et al[5] with Mayo data. In the validation TCGA data, the triple-positive and TERTp-MT/IDH-MT groups were not significantly different than the IDH-MT only group, while the triple negative and TERTp-MT only groups were (HR = 3.02, $P = .02$ and HR = 14.03, $P < .001$, respectively).

For validation in the GBM subset, we used TCGA data with all 152 GBM cases (Table 1, Fig. 3H). Note, due to the small number of TCGA cases in the triple-positive group, the baseline for the HR was the TERTp-MT only group, as it was the largest. In the validation TCGA data, the TERTp-MT/IDH-MT group had only one observation, thus the estimate for the HR was not reliable. The IDH-MT only and triple-negative groups were significantly different than the TERTp-MT only (HR = 0.26, $P = .01$ and HR = 0.55, $P = .0583$, respectively). In summary, the validation set of TCGA did not have enough events in some of the subtypes to adequately validate the model. However, the survival curves show the same patterns, and the median OS is close in the GBM data. The differences in median OS are the same as those reflected in the validation of the Brat et al[5] model, indicating a difference in the 2 cohorts.

The advantages of this classification scheme were the use of a discovery dataset (Mayo Clinic) and 2 independent validation datasets (UCSF AGS and TCGA), and Cox models with adjustment for known prognostic clinical covariates were provided. The disadvantages were that the groups were derived based on 3 markers adding subjectivity, while additional important prognostic markers (molecular and clinical) may have been overlooked.

Discussion

Given the dearth of beneficial treatments in glioma (except for some oligodendrogliomas), there is a need for prognostic models to assist neuro-oncologists and their patients in therapeutic decision making and life planning. These models could also advance research in future treatments by appropriately stratifying patients in clinical trials.

The 3 statistical considerations to establishing clinically useful prognostic models are: study design, model building, and validation. The most preferable study design is a prospective cohort; nonetheless, retrospective cohorts are frequently more convenient and allow for longer follow-up times. During model building, a discovery cohort of patients should be used to choose variables, construct models, and estimate prediction performance via internal validation. Subsequently, external validation should be utilized to assess how well the model performs on patients from another institution or study.

Several models have recently been published in the neuro-oncology literature. Herein, we reviewed 7 models, and since most did not perform external validation, we performed external validation using glioma patients from TCGA and the Mayo Clinic. In this issue of *Neuro-Oncology*, Aldape et al review the results of 2 of the papers that defined molecular groups of glioma based on a combination of 3 well-recognized glioma markers: TERTp mutation, IDH mutation, and 1p/19q codeletion.[5,6,30] Eckel-Passow et al[6] defined 5 molecular groups based on TERTp mutation, IDH mutation, and 1p/19q codeletion, while Brat et al[5] (and Suzuki et al[30]) defined 3 molecular groups based on IDH mutation and 1p/19q codeletion. Three others have similarly defined molecular groups using IDH and TERTp mutation[3,4] in addition to considering EGFR.[2] In a different approach, Panageas et al[1] combined clinical measures (age and tumor location) with 1p/19q codeletion to define prognostic groups specifically in anaplastic oligodendrogliomas and oligoastrocytomas.

Table 2 summarizes the steps each study took for model construction and assessment. It should be noted that because this manuscript focuses on prognostic stratification at the cohort level, our conclusions do not apply to prediction at the individual level. Panageas et al[1] correctly performed model construction and prediction performance using internal validation by applying 10-fold cross-validation for variable selection within RP. In fact, the study by Panageas et al was the only one that performed internal validation. However, due to reasons they enumerate, they did not perform external validation. Eckel-Passow et al[6] were the only authors who performed external validation; they tested their model in 2 independent datasets. While definitely preferred for the reasons stated above, external validation may not always be feasible. Thus, at a minimum, internal validation should be applied in order to quantify overfitting and to get an unbiased estimate of prediction performance. However, as we demonstrated above, there are publicly available resources (eg, TCGA) that allow external validation to be financially feasible. These free resources should be taken advantage of when appropriate. As previously noted, the proportional hazards assumption should be verified for Cox models. However, most of these studies did not specifically state that they verified this assumption.

Table 2. A summary table illustrating which models were included, the applicable histological grades, discovery sample size, and number of risk groups in model

Model	Grades	Discovery Sample Size	Groups[a]	Model Building			Model Validation
				Variable Selection[b]	Internal Validation[c]	Hazard Ratio for Difference in Survival across Groups	Performed External Validation[c]
Panageas et al (Neuro Oncol 2014)	III	587	5	Yes	Yes	No	No
Labussière et al (Neurology 2014)	IV	395	4	Yes	No	No	No
Labussière et al (BJC 2014)	II–IV	807	4	No	No	No	No
Killela et al (Oncotarget 2014)	II–IV	473	4	No	No	No	No
Brat et al (NEJM 2015)	II–III	293	3	Yes	No	Yes	No
Suzuki et al (Nat Genet 2015)	II–III	665	3	No	No	No	No
Eckel-Passow et al (NEJM 2015)	II–IV	317	5	No	No	Yes	Yes

Additional information is included on whether a study used variable/model selection, internal and external validation, and reported HRs for the difference between risk groups.
[a]External validation: assess the model via an independent dataset.
[b]Variable selection: from a large set of variables include only those that explain signal in the data.
[c]Internal validation: employ resampling methods to quantify overfitting and get an unbiased estimate of prediction performance.

Using the combined TCGA + Mayo data, we attempted to validate all 7 models (Table 1, Fig. 3). While TCGA data were specifically chosen because they are free and publicly available, Mayo Clinic data were added in order to have an adequate validation sample size. In doing so, we acknowledge that TCGA tumors are typically large tumors and tend to have poorer outcomes, as was observed herein. Another important limitation of our validation is that some subgroups in the combined TCGA + Mayo data had small sample sizes. As was seen in the attempted validation of the Panageas et al[1] and Labussière et al[2] models, the combined TCGA + Mayo validation set did not have enough events. However, we included these validations to illustrate external validation and commented on the limited sample sizes.

Overall agreements can be seen across the models. With respect to predicting OS, based on studies published thus far, IDH mutation and TERTp mutation are the most relevant markers, as initially demonstrated by Labussière et al[3] and Killela et al.[4] Examining the Labussière et al[2] groupings defined by IDH, TERT, and EGFR, it is apparent in the TCGA + Mayo validation set that only the IDH-MT/TERTp-WT/EGFR-WT group had a significantly different outcome than the TERTp-MT group (HR = 0.42, $P = .005$). While Brat et al[5] used IDH and 1p/19q codeletion to define groups, there was no significant difference in survival between the 1p/19q codeletion and non-codeleted groups within the IDH-MT (HR = 1.58, $P = .11$). However, there was a significant difference in survival between the IDH-MT/1p/19q codeleletion group and the IDH-WT group (HR = 6.83, $P < .001$). Similarly, while Eckel-Passow et al[6] defined groups based on TERTp, IDH, and 1p/19q codeletion, ultimately only IDH and TERTp mutation defined groups that were significantly associated with survival outcome. It is important to note that since some of the molecularly defined groups have very few subjects (eg, those with IDH-MT/TERTp-MT but lacking 1p/19q codeletion), there is limited statistical power to determine whether survival for these groups differs from that of other groups.

While to date studies have reproducibly defined distinct OS with tumor IDH and TERTp mutation status, there are likely additional molecular markers as well as clinical variables that are important in understanding the biology and development of gliomas, as well as predicting response to treatment. For example, Eckel-Passow et al[6] demonstrated that 1p/19q codeletion is important for defining groups that have distinct age at diagnosis and associations with germline risk variants. Additionally, Cairncross et al[31] demonstrated that 1p/19q codeletion and IDH mutation significantly predicted better response to treatment in grade III oligodendrogliomas and mixed oligoastrocytomas than IDH alone. Future model building efforts should include additional markers and seek to improve on the predictive accuracy of models constructed using only commonly available clinical variables, such as age, grade, performance status (which is an intermediate endpoint), and extent of resection.

In conclusion, as the neuro-oncology community moves toward the goal of establishing prognostic models in glioma, it is imperative that the published models properly detail methods of both model building and validation. Guidelines for reporting multivariate models for prognosis and diagnosis have recently been enumerated.[32] This will provide readers the information necessary to assess bias in a study, compare other models in the literature, and determine clinical usefulness. Without such information, prognostic models should not be integrated into clinical use. As editors, reviewers, and readers of the relevant literature, we should be cognizant of the needed statistical considerations and insist on their use.

Funding

This study was supported by R01 CA163687 (Annette M. Molinaro, Principal Investigator), the UCSF and Mayo Brain Spores (P50CA097257 and P50CA108961, respectively), and R01CA52689 (Margaret R. Wrensch, Co-Principal Investigator).

Glossary of terms

Overfitting: a model is overly specific to the dataset from which it was developed and therefore does not replicate in an independent dataset
Resubstitution: when a model is built and has prediction performance assessed with the same data
Internal validation: employ resampling methods to quantify overfitting and get an unbiased estimate of prediction performance
External validation: assess the model via an independent dataset
Bias: the difference between the estimated value and the true value
Resampling methods: approaches for assessing internal validity including cross-validation and bootstrapping
Cross-validation: a form of resampling where model is built on a subset of the data and assessed on an entirely different subset
Bootstrapping: a form of resampling where a training set is defined by sampling with replacement such that the training set is the same size as the discovery set but has repeated patients. The test set is all patients in the discovery set not included in the training set.
Prediction performance: the accuracy of a given statistical model
Variable Selection: from a large set of variables choose only those that explain signal in the data

Acknowledgments

We would like to thank Paul Decker, Matt Kosel, and Hugues Sicotte for assistance with accessing and assembling TCGA and Mayo Clinic data. We would also like to thank Jennifer Clarke and Dan Lachance for their clinical input.

Conflict of interest statement. There are no conflicts to report.

References

1. Panageas KS, Reiner AS, Iwamoto FM, et al. Recursive partitioning analysis of prognostic variables in newly diagnosed anaplastic oligodendroglial tumors. *Neuro Oncol*. 2014;16(11):1541–1546.
2. Labussière M, Boisselier B, Mokhtari K, et al. Combined analysis of TERT, EGFR, and IDH status defines distinct prognostic glioblastoma classes. *Neurology*. 2014;83(13):1200–1206.
3. Labussière M, Di Stefano AL, Gleize V, et al. TERT promoter mutations in gliomas, genetic associations and clinico-pathological correlations. *Br J Cancer*. 2014;111(10):2024–2032.
4. Killela PJ, Pirozzi CJ, Healy P, et al. Mutations in IDH1, IDH2, and in the TERT promoter define clinically distinct subgroups of adult malignant gliomas. *Oncotarget*. 2014;5(6):1515–1525.
5. Brat D, Verhaak R, Aldape KD, et al. Comprehensive, integrative genomic analysis of diffuse lower grade gliomas. *N Engl J Med*. 2015;372(26):2484–2498.
6. Eckel-Passow JE, Lachance DH, Molinaro AM, et al. TERT, IDH and 1p/19q alterations in five principal glioma groups. *N Engl J Med*. 2015;372(26):2499–2508.
7. Moons KGM, Royston P, Vergouwe Y, et al. Prognosis and prognostic research: what, why, and how? *BMJ*. 2009;338:b375.
8. Harrell FE, Lee KL, Califf RM, et al. Regression modelling strategies for improved prognostic prediction. *Stat Med*. 1984;3(2):143–152.
9. Peduzzi P, Concato J, Feinstein AR, et al. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol*. 1995;48(12):1503–1510.
10. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49(12):1373–1379.
11. Vittinghoff E, Glidden DV, Shiboski SC, et al. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. New York: Springer Science & Business Media; 2012.
12. Breiman L, Friedman J, Olshen R, et al. *Classification and Regression Trees*. Boca Raton, FL: Wadsworth Books; 1984.
13. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc*. 1996;91(434):473–489.
14. Jewell NP. *Statistics for Epidemiology*. Boca Raton, FL: Chapman and Hall/CRC; 2003.
15. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. New York: Springer New York Inc.; 2001.
16. Efron B, Hastie T, Johnstone I, et al. Least angle regression. *Ann Stat*. 2004;32(2):407–451.
17. Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart*. 2012;98(9):683–690.
18. Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer*. 2004;4(4):309–314.
19. Molinaro AM, Lostritto K. Statistical resampling for large screening data analysis such as classical resampling bootstrapping, Markov chain Monte Carlo, and statistical simulation and validation strategies. In: Lee JK, ed. *Statistical Bioinformatics: A Guide for Life and Biomedical Science Researchers*. Hoboken, NJ: John Wiley & Sons, Inc.; 2010.
20. Taylor JMG, Ankerst DP, Andridge RR. Validation of biomarker-based risk prediction models. *Clin Cancer Res*. 2008;14(19):5977–5983.
21. Steyerberg EW, Harrell FE Jr., Borsboom GJJM, et al. Internal validation of predictive models. *J Clin Epidemiol*. 2001;54(8):774–781.
22. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*. 2005;21(15):3301–3307.
23. Simon R, Radmacher MD, Dobbin K, et al. Pitfalls in the Use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst*. 2003;95(1):14–18.
24. Quackenbush J. Meeting the challenges of functional genomics: from the laboratory to the clinic. *Preclinica*. 2004;2:313–316.
25. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012;98(9):691–698.
26. Vergouwe Y, Steyerberg EW, Eijkemans MJC, et al. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol*. 2005;58(5):475–483.

27. Janssen KJM, Moons KGM, Kalkman CJ, et al. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008;61(1):76–86.
28. Steyerberg EW, Borsboom GJJM, van Houwelingen HC, et al. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med*. 2004;23(16): 2567–2586.
29. Lostritto K, Strawderman RL, Molinaro AM. A partitioning deletion/substitution/addition algorithm for creating survival risk groups. *Biometrics*. 2012;68(4):1146–1156.
30. Suzuki H, Aoki K, Chiba K, et al. Mutational landscape and clonal architecture in grade II and III gliomas. *Nat Genet*. 2015;47(5): 458–468.
31. Cairncross JG, Wang M, Jenkins RB, et al. Benefit from procarbazine, lomustine, and vincristine in oligodendroglial tumors is associated with mutation of IDH. *J Clin Oncol*. 2014;32(8):783–790.
32. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55–63.